The London School of Economics and Political Science

Reproducibility of Empirical Findings: Experiments in Philosophy and Beyond

Hamid Seyedsayamdost

A thesis submitted to the Department of Philosophy, Logic, and Scientific Method of the London School of Economics for the degree of Doctor of Philosophy, London, September 2014

Declaration

I certify that the thesis I have presented for examination for the MPhil/PhD degree of the London School of Economics and Political Science is solely my own work other than where I have clearly indicated that it is the work of others (in which case the extent of any work carried out jointly by me and any other person is clearly identified in it).

I warrant that this authorisation does not, to the best of my belief, infringe the rights of any third party.

I declare that my thesis consists of 53,782 words.

Statement of conjoint work (if applicable)

Not applicable

Statement of inclusion of previous work (if applicable)

Not applicable

Statement of use of third party for editorial help (if applicable)

Not applicable

Abstract

The field of experimental philosophy has received considerable attention, essentially for producing results that seem highly counter-intuitive and at the same time question some of the fundamental methods used in philosophy.

A substantial part of this attention has focused on the role of intuitions in philosophical methodology. One of the major contributions of experimental philosophy on this topic has been concrete evidence in support of intuitional diversity; the idea that intuitions vary systematically depending on variables such as ethnicity, socioeconomic background, or gender.

Because of the important implications, these findings have been the subject of extensive debate. Despite the seeming significance of the findings and despite all the debates that the experimental philosophy movement has prompted, what has not been examined systematically is the reproducibility of the results. Instead, the reported findings have been simply accepted as established facts.

We set out to replicate a wide range of experiments and surprisingly failed to reproduce many of the reported findings, some of which are from the most cited and attention grabbing papers of the field.

We draw two conclusions from our findings. The first is that the instability of intuitions has been exaggerated by experimental philosophers. Intuitions appear to be more uniform across different demographic groups. The argument that intuitions need to be discarded

because they depend on arbitrary factors such as ethnicity, socioeconomic background, or gender does not seem tenable anymore. The second conclusion is that experimental philosophy needs a better system to ensure the reproducibility of published findings. The current research-publication system of various empirical fields, especially those employing statistical methods, leads to an overproduction of false-positive findings in the published literature. Unless changes are made to the current research-publication system, this overproduction is likely to continue, in experimental philosophy as well as other disciplines.

Table of Contents

List of Tables and Figures

Page Intentionally Left Blank

Introduction

Recently, philosophers have started using the tools of experimental psychology to study philosophical questions; this movement has been termed experimental philosophy. Experimental philosophy has attracted great attention, essentially for producing results that seem highly counter-intuitive and at the same time question some of the fundamental methods used in philosophy. For example, one of the earlier papers reported that individuals from different ethnic as well as socioeconomic backgrounds displayed different epistemic intuitions (Weinberg, Nichols, & Stich, 2001). More recently, differences in women's and men's responses have been reported for a host of scenarios, including Gettier-type questions, compatibilism cases, as well as some classical scenarios such as Putnam's Twin-Earth and Searle's Chinese Room thought experiment (Buckwalter & Stich, 2013). Researchers have also claimed that simple manipulations can influence moral judgments on decisions as grave as whether to sacrifice an innocent bystander in order to save the lives of a greater number of individuals (Valdesolo & DeSteno, 2006).

One of the main branches of experimental philosophy is concerned with the role of intuitions in philosophy and in particular examines intuitional diversity (Nadelhoffer & Nahmias, 2007), i.e. the idea that intuitions vary systematically depending on variables such as ethnicity, socioeconomic status, gender, or age (Buckwalter & Stich, 2013; Colaço, Buckwalter, Stich, & Machery, 2014; Machery, Mallon, Nichols, & Stich, 2004; Weinberg et al., 2001; Zamzow & Nichols, 2009). These are variables that most philosophers agree, should not have any bearing on how philosophical questions are evaluated. Experimental philosophers have also taken an interest in manipulations of moral judgments (Tobia,

Chapman, & Stich, 2013; Zhong & Liljenquist, 2006; Zhong, Strejcek, & Sivanathan, 2010) as evidence of the instability of intuitions. One of the major contributions of experimental philosophy in the discussion on the role of intuitions has been (supposedly) concrete evidence in support of intuitional diversity and intuitional instability. Some prominent philosophers have taken these results to mean that philosophical practice as conducted over millennia needs to be changed drastically and that in fact one of the main methods (reliance on intuitions) used over the last 2400 years has "been a terrible mistake" (Stich, 2001).

The role of intuitions in philosophy had been the subject of extensive debate before the emergence of experimental philosophy (Bealer, 1996, 2000; Cummins, 1998; Goldman & Pust, 1998; Gopnik & Schwitzgebel, 1998; Gutting, 1998; Kornblith, 1998; Osbeck, 1999; Ramsey, 1992; Shafir, 1998; E. Sosa, 1998; Stich, 1988; Wild, 1938; Wisniewski, 1998); however, the findings of experimental philosophers have given the debate a new urgency (Alexander, 2012; Alexander, Mallon, & Weinberg, 2010; Alexander & Weinberg, 2007; Buckwalter & Stich, 2013; Cappelen, 2012; Chudnoff, 2011; Cullen, 2010; Deutsch, 2009, 2010; Dowell, 2008; Feltz, 2008, 2009a; Gendler, 2007; Goldman, 2007; Ichikawa, 2014; Levin, 2005; Liao, 2008; Mallon, Machery, Nichols, & Stich, 2009; Nagel, 2012; Nichols, 2004; D. Sosa, 2006; E. Sosa, 2007, 2009; Stich, 2001; Symons, 2008; Weinberg et al., 2001; Williamson, 2004; Wright, 2010).

Intuitions play a unique role in philosophy that is very different from other disciplines. We provide a closer account of the role of intuitions in Paper 1. Briefly, intuitions are used as data points in philosophical theorizing. There are numerous descriptions of this method and although these vary somewhat, an account of intuitions as data or evidence lies at the

core. We will refer to this approach to philosophy as the Intuition as Evidence approach (IAE), following Liao (2008).

A degree of uniformity of intuitions is of special importance for IAE because without it the practice would be on extremely unstable grounds. The findings on intuitional diversity, many philosophers believe, would make it problematic for philosophers to rely on intuitions as a source of evidence. If intuitions were to vary in this way, they would merely be reflective of cultural or socioeconomic background or gender rather than be informative on details of philosophical scenarios; this is the suggestion that a branch of experimental philosophy advances, following Weinberg et al. (2001).

Given the important implications of the results and given all the debate that the experimental philosophy movement has prompted, we believed, when we started this project, that a careful examination of the findings was appropriate. It is not that there had been a lack of engagement by philosophers with the findings of the experimental philosophy literature (Bengson, 2013; Cullen, 2010; Deutsch, 2009, 2010; Grundmann, 2010; Liao, 2008; Nagel, 2012, 2013; Shieber, 2010; Weatherson, 2003; Williamson, 2004, 2011).

However, with some exceptions (Cullen, 2010; Nagel, 2013; Nagel, Juan, & Mar, 2013), a great majority of these responses had taken the findings as fact and begun their analysis and criticism from there. When we first became interested in this topic in 2007, what had not been attempted were systematic replications of the experiments that these philosophers engaged with so intensely, to see if the effects actually existed. We believed that a first

step in a close examination of the experimental philosophy literature entailed a test of the reproducibility of the reported findings.

We wanted to test the reproducibility of a diverse set of results and so selected the following topics for examination: differences in epistemic intuitions cross-culturally and socioeconomically (Paper 1); gender differences on a range of intuitions (Paper 2); and finally, the instability of moral intuitions (Paper 3). We were unable to reproduce most of these findings. Concerns regarding the reproducibility of these findings should have been a first point of departure. Replications should have been integral to a careful examination that should have taken place many years ago, before all the discussion that the findings prompted, before all the back and forth between proponents and opponents of the views expressed by some experimental philosophers and before all the lengthy discussions on the merits of specific articles as well as the merits of the movement as a whole (Alexander, 2010; Feltz, 2009b; Ichikawa, 2012; Kauppinen, 2007; Nadelhoffer & Nahmias, 2007; Stich, 2013).

When we started this project on the reproducibility of experimental findings, we could not find any discussions on replications in the published experimental philosophy literature or otherwise in more informal outlets such as blogs. Replications were of little (or no) prominence in psychology and completely absent in experimental philosophy. When we started our endeavor, it was before the current 'crisis of confidence' in psychology and somewhat ironically, conducting replications constituted a novel pursuit. Replications in themselves are not necessarily original undertakings; however, when we started the work, the originality consisted in doing something that others neglected entirely. As it turned out, some of the most cited and attention-grabbing papers in the field turned out to be non-

reproducible; see Paper 1 (Seyedsayamdost, forthcoming) and Paper 2 (Seyedsayamdost, 2014).

In the early stages of our work, colleagues advised us against pursuing this project. This advice stemmed from various considerations. One, just like other researchers who engaged with experimental philosophers, these colleagues took the findings as given. From this perspective, replications would have simply reaffirmed the findings and that would have been the end of the story without any fruitful outcomes. Second, replications are notoriously difficult to publish. Many of the most prestigious journals in psychology have policies against publishing replications (see Paper 4). Had our studies successfully reproduced the findings of the original articles, it would have been impossible to publish this work and we would have simply 'wasted' an enormous amount of time and resources. When we started this project, it entailed a great amount of risk in terms of investing resources in something that was likely to be a 'non-result'.

Aside from advice from colleagues, we also had our own doubts. For Paper 1 (Seyedsayamdost, forthcoming) we examined Weinberg et al. (2001) on differences of epistemic intuitions based on cross-cultural and socioeconomic backgrounds. This article had been extremely influential and the foundation on which a whole new subfield of philosophy had been built. When we started data collection, the paper had been public for close to ten years and our best guess was that other researchers had already tested the robustness of the findings. To our surprise, we could not find any replication attempts; however, it was very likely that any prior replication had been successful and hence had had little chance of being published or gaining any further attention in informal outlets such as blogs.

For Paper 2 (Seyedsayamdost, 2014) we examined Buckwalter & Stich (2013) on gender differences of intuitions. The data in Buckwalter & Stich (2013) gave no indication that the findings may not have been reproducible. Almost all of their experiments reported relatively large samples and the procedures were clearly outlined and seemed to be without flaws. For all we knew, the results were likely to replicate successfully and that would have been the end of this pursuit.

For Paper 3 we examined the instability of moral intuitions by attempting replication of (Valdesolo & DeSteno, 2006) and (Zhong et al., 2010). These articles were published in reputable psychology journals and except for the surprise factor, little else initially hinted at the non-reproducibility of the findings.

When we first attained the results of Papers 1-3 (failures of replication), we were very surprised. In fact, our first reaction was that there may have been flaws, either in our experimental procedures or our data analysis. One of the reasons why all our reports, with the exception of the first part of Paper 3, are multi-study attempts is to a great extent because we did not trust our own initial findings. After collecting new data and after re-examining our data analyses multiple times, we made our findings public on the Social Science Research Network (SSRN) in 2012 (Seyedsayamdost, 2012a, 2012b, 2012c), always including a word of caution that our studies should not be taken as definitive and that our hope was that other researchers would attempt to independently replicate the findings in order to attain further verification.[1]

[1] Not all of our replication attempts failed; see results sections of Papers 1-3 and the thesis conclusion.

Since we uploaded our papers to SSRN in 2012, presented the work at conferences, and journals published our papers (Seyedsayamdost, 2014, forthcoming), a trend in replications has started in experimental philosophy; three examples that emerged in 2014 are (Adleberg, Thompson, & Nahmias, 2014; Kvanvig, 2014; Minsun & Yuan, ms). We discuss the significance of this trend in some more detail in the thesis conclusion. Furthermore, some of these groups attempted to independently replicate the papers we examined and have reproduced our findings (Adleberg et al., 2014; Minsun & Yuan, ms). As further evidence of the reliability of our findings, Nagel and colleagues also report no effect of ethnicity and gender on epistemic intuitions (Nagel et al., 2013).[2] There is now growing evidence that our findings are robust and that indeed some of the most cited and attention-grabbing papers in the field of experimental philosophy are not reproducible.

To some extent the failures of replication should not have been completely unexpected and the fact that we were greatly surprised is at least partly because we were not familiar with the historical literature on the reproducibility of findings in psychology and certain other empirical fields. It is not that we were carelessly neglectful of this literature. This literature is simply not part of the basic training psychologists (or perhaps more generally, empirical scientists) receive. The author of this thesis completed a two year fellowship in a psychology department and issues surrounding reproducibility were (and perhaps still are) just not a major concern, or a concern at all. The current literature on reproducibility (some of which is addressed in the following papers) emerged in response to the current crisis, that is, after we embarked on our project and was not available to us when we conducted most of the work for Papers 1-3. Any references in Papers 1-3 to the literature on

[2] In contrast to the other studies, this work was conducted independently of our efforts; Nagel and colleagues had not seen our findings before starting their work.

reproducibility (historical and current) in empirical fields was added in the later stages of manuscript submissions, close to two years after we made the findings public.

Experimental philosophy is often described as the study of philosophical questions, using the tools of experimental psychology (Alexander, 2012; Knobe & Nichols, 2008; Nadelhoffer & Nahmias, 2007). By importing the methods of experimental psychology, philosophers will likely import the problems of that field and one of the problems that has afflicted experimental psychology for some time is the high rate of false-positive results in the published literature (Bakker, van Dijk, & Wicherts, 2012; Ioannidis, 2005, 2012; Pashler & Harris, 2012). As one researcher contends, in "several fields of investigation, including many areas of psychological science, perpetuated and unchallenged fallacies may comprise the majority of the circulating evidence" (Ioannidis, 2012, p. 645).

In Paper 4 we review the literature on the problems of reproducibility in empirical fields with an emphasis on psychology, as it is the methods of this field that experimental philosophers have adopted. In light of the shortcomings discussed in Paper 4, many researchers have suggested solutions to alleviate these problems. Paper 5 provides a review of this literature and concludes with what we believe to be important components of any sustainable solution.

We are presenting our papers in chronological order of time of composition and hence the literature review (typically presented first) is presented last in Papers 4 and 5 after the original findings are presented in Papers 1-3. For one, this captures the development of our project better. Our approach may have been somewhat different had we been familiar with the historical literature on reproducibility before conducting the work presented in Papers

1-3. We have also decided to present the papers in chronological order because we believe that presenting the concrete findings of Papers 1-3 first, makes the materials of Papers 4 and 5 more accessible.

As our research progressed, we came to hold two objectives. One, the initial aim, was to test the effects reported in the experimental philosophy literature: do people from different ethnic backgrounds really have different epistemic intuitions; do women and men really have different intuitions on common scenarios such as the Brain in the Vat and the Twin-Earth cases? The other objective was to demonstrate that replications are an important part of the scientific process and that without replications false positives are likely to persist and flourish in the published literature. We hope that the trend of replication studies in experimental philosophy can persist and that this trend can prevent experimental philosophy from going down a similar path to some areas of psychology where false-positive results likely make up the majority of published findings.

We draw two conclusions from our work. The first is that the instability of intuitions has been exaggerated by experimental philosophers. Intuitions appear to be more uniform across different demographic groups and are generally more stable than the experimental philosophy literature indicates. Whether intuitions should be considered legitimate data points in philosophical theorizing is a different question (which we will not engage with in this thesis); however, the argument that intuitions need to be discarded because they depend on 'arbitrary' factors such as ethnicity, socioeconomic background, or gender does not seem tenable anymore.

The second conclusion is that experimental philosophy, like some other empirical fields, needs a better system to test for the reproducibility of published findings. As it stands, current research practices lead to an overproduction of false positives; be it simply as a result of standard statistical procedures (Ioannidis, 2005; Pashler & Harris, 2012) or questionable research practices (Fanelli, 2009; John, Loewenstein, & Prelec, 2012; Martinson, Anderson, & De Vries, 2005). Unless changes are made to the current research-publication system, this overproduction is likely to continue, in experimental philosophy as well as other disciplines.

Page Intentionally Left Blank

Paper 1: On Normativity and Epistemic Intuitions[3]

Failure of Replication

Abstract

The field of experimental philosophy has received considerable attention, essentially for producing results that seem highly counter-intuitive and at the same time question some of the fundamental methods used in philosophy. One of the earlier influential papers that gave rise to the experimental philosophy movement titled "Normativity and Epistemic Intuitions" by Jonathan M. Weinberg, Shaun Nichols and Stephen Stich (2001), reported that respondents displayed different epistemic intuitions depending on their ethnic background as well as socioeconomic status. These findings, if robust, would have important implications for philosophical methodology in general and epistemology in particular. Because of the important implication of its findings, Weinberg et al. (2001) has been very influential – currently with more than four hundred citations – and the subject of extensive debate. Despite the paper's significance and despite all the debates this paper has generated, there has not been a replication attempt of its results. We collected data from four different sources (two on-line and two in-person) to replicate the experiments. Despite several different data sets and in various cases larger sample sizes, we failed to detect significant differences between the above-mentioned groups. Our results suggest that epistemic intuitions are more uniform across ethnic and socioeconomic groups than Weinberg et al. (2001) indicates. Given our data, we believe that the notion of differences in epistemic intuitions among different ethnic and socioeconomic groups advanced by Weinberg et al. (2001) and accepted by many researchers needs to be corrected.

[3] The author would like to thank Susan Carey, Donal Cahill, all of the class teachers at the LSE who allocated class time for data collection, and all of the students who participated. This paper is forthcoming in *Episteme* (Cambridge University Press) with some minor changes.

The field of experimental philosophy has received considerable attention, essentially for producing results that seem highly counter-intuitive and at the same time question some of the fundamental methods used in philosophy. Much of the debate that this field has generated focuses on the role of intuitions in philosophy. This debate predates the emergence of experimental philosophy; however, the findings of experimental philosophers have given the debate a new urgency. One of the major contributions of experimental philosophy in the discussion on the role of intuitions in philosophy has been concrete evidence in support of intuitional diversity, i.e. the idea that intuitions vary systematically depending on variables such as ethnicity, socioeconomic status, gender, or age (Buckwalter & Stich, 2013; Colaço et al., 2014; Weinberg et al., 2001).

The paper by Weinberg and colleagues titled "Normativity and Epistemic Intuitions" published in 2001, in large part gave rise to the movement now known as experimental philosophy. The authors of the paper presented data showing that the epistemic intuitions of East and South Asian individuals differed significantly from that of their 'Western' counterparts on a host of scenarios. Since publication of this paper, "one of the most widely discussed kinds of intuitional diversity has been cultural diversity," namely the hypothesis that "our philosophical intuitions seem to be sensitive to our own cultural background" (Alexander, 2012, p. 72). In addition to data on cultural diversity, Weinberg et al. also provided data indicating that individuals from different socioeconomic backgrounds have differing epistemic intuitions.

Weinberg et al. (2001) has been very influential and the topic of extensive discussion. However, despite its influence and reach, the findings presented in the paper have not been tested for their reproducibility but typically simply accepted as given. We collected data

through four different sources, two on-line and two in-person where we presented individuals with scenarios identical in wording to those asked by Weinberg et al. (2001). Our results strongly suggest that epistemic intuitions are not significantly different among individuals from different ethnic or socioeconomic backgrounds. Given our findings, we have to conclude that the results of Weinberg et al. (2001) are not reproducible and hence the notion of intuitional diversity for these cases among different ethnic and socioeconomic groups needs to be corrected.

We will advance as follows. The next section gives an introduction on the role of intuitions in philosophy. We keep this section somewhat brief. Given the central role that intuitions play in philosophy, this issue has been the subject of extensive debate and any attempt at providing a comprehensive review here would not do the topic justice. For some of the discussions that took place before the emergence of experimental philosophy, see (Bealer, 1996, 2000; Cummins, 1998; Goldman & Pust, 1998; Gopnik & Schwitzgebel, 1998; Gutting, 1998; Kornblith, 1998; Osbeck, 1999; Ramsey, 1992; Shafir, 1998; E. Sosa, 1998; Stich, 1988; Wild, 1938; Wisniewski, 1998); and for some of these debates in the context of experimental philosophy, see (Alexander, 2012; Alexander et al., 2010; Alexander & Weinberg, 2007; Buckwalter & Stich, 2013; Cappelen, 2012; Chudnoff, 2011; Cullen, 2010; Deutsch, 2009, 2010; Dowell, 2008; Feltz, 2008, 2009a; Gendler, 2007; Goldman, 2007; Ichikawa, 2014; Levin, 2005; Liao, 2008; Mallon et al., 2009; Nagel, 2012; Nichols, 2004; D. Sosa, 2006; E. Sosa, 2007, 2009; Stich, 2001; Symons, 2008; Weinberg et al., 2001; Williamson, 2004; Wright, 2010). Section 1.2 examines ethnic differences. Subsection 1 of Section 1.2 provides a description of our four data sources and the methods and materials used in data collection. In subsection 2 of Section 1.2 we present our results for East Asian and Western participants and compare these to the outcomes of Weinberg et

al. (2001). In the third subsection of Section 1.2 we present the results for South Asian and Western participants and again compare these to the relevant data from Weinberg et al. (2001). Section 1.3 examines intuitional differences based on socioeconomic status. Section 1.4 briefly discusses statistical power and Section 1.5 concludes with a discussion.

1.1: Intuitions and Philosophy

Intuitions play a unique role in philosophy that is very different from other disciplines. Levin (2005), for example, writes that the use of intuitions "has been characteristic, perhaps definitive, of philosophical argumentation throughout its history" (Levin, 2005, p. 194). Goldman (2007) explains that "one thing that distinguishes philosophical methodology from the methodology of the sciences is its extensive and avowed reliance on intuition" (Goldman, 2007, p. 1). In the philosophical literature, intuitions are (broadly understood) simply "spontaneous judgments" to philosophical questions (Stich, 2001). For example, when presented with a scenario, a reader's direct reaction as to whether something was morally permissible or not counts as an intuition. Often, individuals do not immediately have an explanation for these reactions. For example, a person may judge one action as permissible and another as impermissible, without being able to point out the relevant features (e.g. intention) that led to differences in judgments. A common description of the use of intuitions in philosophy is as follows: philosophers construct thought experiments and test their intuitive responses. These intuitions serve as data points that are used to substantiate or challenge theories. Within this framework "the role and corresponding epistemic status of intuitional evidence in philosophy is similar to the role and corresponding epistemic status of perceptual evidence in science" (Alexander, 2012, p. 11). There are numerous descriptions of this method and although these vary somewhat, an

account of intuition as data or evidence lies at the core. We will refer to this approach to philosophy as the Intuition as Evidence approach (IAE), following Liao (2008).

To fill in IAE further, consider the following two descriptions. Stich (2001), who traces this methodology to Plato, writes that philosophers proceed

> to test normative claims against people's spontaneous judgments about real and hypothetical cases. Contemporary philosophers often call these spontaneous judgments "intuitions." If the normative claim and people's intuitions agree, the claim is vindicated. But if […] a normative principle conflicts with people's intuitions, then something has to give. Sometimes we may hold on to the normative claim and ignore a recalcitrant intuition. But if a normative principle conflicts with lots of intuitions or […] if it conflicts with an intuition that we would be very reluctant to give up, then Plato's method requires that we reject the principle and try to come up with another one. (Stich, 2001, p. 36)

Goldman (2007) describes.

> To decide what is knowledge, reference, identity, or causation […], philosophers routinely consider actual and hypothetical examples and ask whether these examples provide instances of the target category or concept. People's mental responses to these examples are often called "intuitions" and these intuitions are treated as evidence for the correct answer. At a minimum, they are evidence for the examples' being instances or non-instances of knowledge, references, causation, etc. Thus, intuitions play a particularly critical role in a certain sector of philosophical activity. (Goldman, 2007, p. 1)

An exemplary instance of IAE often cited is Gettier's 1963 paper titled "Is Justified True Belief Knowledge?" (Gettier, 1963). At the time of Gettier's writing – and further back to Plato (Burnyeat, 1990) – typical accounts equated knowledge with justified true belief. Gettier provided two thought experiments wherein although an individual was described to have justified true belief, readers did not have the intuition that this individual actually had knowledge. Bealer (1996) summarizes that "at one time many people accepted the doctrine that knowledge is justified true belief. But today we have good evidence to the contrary,

namely, our intuitions that situations like those described in the Gettier literature are possible and that the relevant people in those situations would not know the things at issue" (Bealer, 1996, p. 122).

What made Gettier's examples convincing was that the intuitions they elicited were shared widely. Goldman (2007) notes that it "was the fact that almost everybody who read Gettier's examples shared the intuition that these were not instances of knowing. Had their intuitions been different, there would have been no discovery" (Goldman, 2007, p. 2). A degree of uniformity of intuitions is of special importance for IAE because without it the practice would be on extremely unstable grounds. The equivalent of this in scientific practice would be if data from experiments were to vary depending on who carried out the experiments. Systematically differing intuitions based on individuals' backgrounds such as ethnicity, socioeconomic status or gender, are often referred to as intuitional diversity. As mentioned before, one of the major contributions of experimental philosophy to general philosophy has been the (alleged) concrete evidence in support of intuitional diversity. These findings, if robust, would make it very problematic for philosophers to rely on intuitions as a source of evidence. If intuitions were to vary in this way, they would merely be reflective of cultural or socioeconomic background or gender rather than be informative on details of philosophical scenarios; this is the suggestion that Weinberg et al. (2001) advance.

Given the important implications of the findings of Weinberg et al. (2001) for philosophical methodology, we set out to examine the experiments more closely. In our paper we focus almost exclusively on Weinberg et al. (2001) and carry out an exact replication of its experiments. There are several reasons for focusing on Weinberg et al. (2001).

First, at least according to some, Weinberg et al. (2001) gave rise to the movement now known as experimental philosophy (Koppl, 2011; Williamson, 2011) and it would be of historical interest if the findings that gave rise to this (now very popular) movement were non-reproducible.

Secondly, and more importantly, Weinberg et al. (2001) has been very influential, currently with a citation count of 434.[4] In addition to references in philosophy journals, the paper has also been cited in economics, law, and mathematics journals (Koppl, 2011; Löwe, Müller, & Müller-Hill, 2009; Pardo, 2005). The paper also appeared in the edited volume *Experimental Philosophy* which is itself a popular and influential outlet. To give a sense of the status this paper enjoys, Knobe (2007) in discussing intuitional diversity writes that "perhaps the most celebrated work in this vein is Weinberg, Nichols and Stich's paper on intuitions about knowledge" (Knobe, 2007, p. 82). As another example, Doris (2005) writes that ""Experimental philosophy" pertaining to various topics is now–happily–appearing with increasing frequency […] The locus classicus is Weinberg, Nichols, and Stich (2001)" (Doris, 2005, p. 674).

Because of this wide exposure, there are numerous instances in published papers and books citing Weinberg et al. (2001) as evidence that East and South Asian individuals have different epistemic intuitions from their Western counterparts. In fact, this difference in intuitions is often treated as an established fact. For example, Bishop & Trout (2005) write that "in a fascinating study, Weinberg, Nichols, and Stich (2001) found that people in different cultural and socioeconomic groups make significantly different epistemic

[4] http://scholar.google.com/scholar?cites=2305777674912570473&as_sdt=2005&sciodt=0,5&hl=en

judgments" (Bishop & Trout, 2005, p. 705). Buckwalter (2010) notes that "famously, Weinberg et al. 2001 shows that […] divergence of intuition extends to the epistemic domain" (Buckwalter, 2010, p. 396). Mallon et al. (2009) write that "Weinberg et al. found that there are indeed systematic cross-cultural differences in epistemic intuitions" (Mallon et al., 2009, p. 340). There are numerous other examples of papers citing Weinberg et al. (2001) as evidence for intuitional diversity; for some of these, see (Alexander, 2012; Alexander & Weinberg, 2007; Beebe & Buckwalter, 2010; Feltz, 2009a; Knobe, 2007; Knobe & Nichols, 2008; Nadelhoffer & Nahmias, 2007; Zamzow & Nichols, 2009). Furthermore, in conversations with colleagues, we have also often heard this supposed difference in epistemic intuitions between different ethnic groups being treated as fact.

Aside from this widespread acceptance of the results, many philosophers have engaged with the findings of Weinberg et al. (2001) and much effort has been expended in attempting to explain why the findings do not bear heavily on IAE or why despite the results, IAE is sufficiently robust as an approach to withstand intuitional diversity. For some of these discussions, see (Deutsch, 2009, 2010; Grundmann, 2010; Liao, 2008; Nagel, 2012; Shieber, 2010; Weatherson, 2003; Williamson, 2004, 2011). However, responses (again, with some exceptions) have typically taken the findings of Weinberg et al. (2001) as a given and started their replies from there.

The final reason for carrying out an exact replication of Weinberg et al. (2001) is the following. Experimental philosophy uses the tools of experimental psychology to study questions of interest to philosophers. By adopting these tools, philosophers have inevitably adopted some of the shortcomings of psychology as well. Various fields of psychology currently face a 'crisis of confidence' which amounts to an overproduction of false-positive

results in the published literature (Bakker et al., 2012; Ioannidis, 2005, 2012; Pashler & Harris, 2012; Pashler & Wagenmakers, 2012). There are several reasons for this, one of them being a lack of interest by researchers to conduct replications and relatedly an aversion by journals to publishing this kind of work (Bozarth & Roberts, 1972; Makel, Plucker, & Hegarty, 2012; Neuliep & Crandall, 1991; Nosek, Spies, & Motyl, 2012). We hope that by providing this data from our exact replication, we can contribute in small part to preventing experimental philosophy from going down a path similar to some of the areas of psychology. We hope to show that there is an important role for replications in the scientific process. Theoretical descriptions of the scientific process typically place a high value on carrying out replications and consider reproducibility integral to science (Braude, 1979; Collins, 1992; Francis, 2012; Lamal, 1991; Nosek et al., 2012; Popper, 2002). However, this importance granted to replications in theory does not generally translate into practice (Amir & Sharon, 1991; Collins, 1992; Hendrick, 1991; Makel et al., 2012; Smith, 1970). Many psychologists agree that replications are critical in lowering the rate of false positives (Amir & Sharon, 1991; Koole & Lakens, 2012; Nosek et al., 2012; Pashler & Harris, 2012; Ritchie, Wiseman, & French, 2012b) and calls for more replications have been made frequently during various crises of the past decades. However, in practice these calls have mostly remained unanswered – with some notable exceptions (Nosek, 2012) – and in experimental philosophy systematic replications were completely lacking until we started some of our efforts.

Aside from what we believe to be the significance of Weinberg et al. (2001) and why we believed it warranted a full replication of its experiments, the authors of the paper,

themselves, underline the importance of their work by noting that their evidence, if robust,[5] shows that "a sizeable group of epistemological projects – a group which includes much of what has been done in epistemology in the analytic tradition – would be seriously undermined" (Weinberg et al., 2001, p. 429). Furthermore, one of the co-authors writes that "indeed, in light of these new findings some philosophers – I am one of them – have come to think that after 2400 years it may be time for philosophy to stop relying on Plato's method" (Stich, 2001, p. 36). These are strong claims (by the authors' own admission) that need to be examined carefully. Our paper is an attempt at a careful examination that we believe should have been carried out many years ago, before all the discussion that Weinberg et al. (2001) prompted, before all the back and forth between proponents and opponents of the views expressed in the paper and before all the effort was spent discussing the claims of the paper.

1.2: Ethnicity and Epistemic Intuitions

1.2.1: Methods and Materials

The experiments for this paper were conducted between February 2011 and March 2012. A first draft of this paper was uploaded to SSRN on October 29, 2012 (Seyedsayamdost, 2012b). First drafts of this paper were sent to conferences starting in July 2012.

Scenarios

[5] Although Weinberg et al. (2001) make some very strong and definitive claims about intuitional diversity, the authors also point out in at least three instances that the robustness of their results is not a given.

Below are the scenarios as we presented them to participants. All scenarios were taken from Weinberg et al. (2001).

Car Case

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

We used the same wording in all of our surveys except for in Data Set 4 where we replaced the names of the cars from Buick and Pontiac to Toyota and Honda, respectively and also changed the origin of the cars from 'American' to 'Japanese', accordingly.

Individualistic Truetemp Case

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

In addition to the Individualistic Truetemp case, Weinberg et al. (2001) also collected data on two variations named "Elders" and "Community Wide Truetemp" scenarios. We did

not collect data on these scenarios as Weinberg and colleagues themselves report no significant differences between different ethnic groups on these cases.

Conspiracy Case

> It's clear that smoking cigarettes increases the likelihood of getting cancer. However, there is now a great deal of evidence that just using nicotine by itself without smoking (for instance, by taking a nicotine pill) does not increase the likelihood of getting cancer. Jim knows about this evidence and as a result, he believes that using nicotine does not increase the likelihood of getting cancer. It is possible that the tobacco companies dishonestly made up and publicized this evidence that using nicotine does not increase the likelihood of cancer, and that the evidence is really false and misleading. Now, the tobacco companies did not actually make up this evidence, but Jim is not aware of this fact. Does Jim really know that using nicotine doesn't increase the likelihood of getting cancer, or does he only believe it?
>
> REALLY KNOWS ONLY BELIEVES

Zebra Case

> Pat is at the zoo with his son, and when they come to the zebra cage, Pat points to the animal and says, "that's a zebra." Pat is right–it is a zebra. However, given the distance the spectators are from the cage, Pat would not be able to tell the difference between a real zebra and a mule that is cleverly disguised to look like a zebra. And if the animal had really been a cleverly disguised mule, Pat still would have thought that it was a zebra. Does Pat really know that the animal is a zebra, or does he only believe that it is?
>
> REALLY KNOWS ONLY BELIEVES

Data Sets

Throughout the rest of this paper we will use the terms and abbreviations Western (W), East Asian (EA) and South Asian or Indian Subcontinent (SC), following the terminology in Weinberg et al. (2001) for consistency.

Data Set 1

Procedure

For this data set we visited undergraduate classes at the London School of Economics (LSE). Participation was voluntary although no one refused. After a brief introduction, we handed out a one-page questionnaire. Each student only saw one scenario. We explained that there were several different questionnaires and that therefore some would complete the questionnaire faster than others. We did hand out several different surveys but only one of them included a scenario surveying epistemic intuitions (the Car case). In all, the whole process took about five minutes.

We mainly visited philosophy classes, but, given the relatively small size of the philosophy department, we also visited some classes in the International Relations department to complement the data. About 14 percent of the data came from non-philosophy classes. We will provide a breakdown of the numbers in the results section. There was no significant difference between data collected in philosophy and non-philosophy classes with p-exact = 0.557 (N = 153).

Participants

We will provide the number of participants that fell into each one of the categories that we analyzed (EA, SC, and W) since this is the focus of our discussion. Data set 1 consisted of

41 EAs, 35 SCs and 79Ws for a total sample of 155. For the exact criteria used to categorize participants for all of the data sets in the current paper, see Appendix A.

Scenarios Presented

For this data set we only presented the Car case to participants.

Data Set 2

Procedure

For our second study we used the resources at the London School of Economics' Behavioural Research Lab (BRL). The BRL compiles a database of individuals interested in participating in studies. Participants then receive email notifications whenever studies are being conducted. Individuals received 5 pounds sterling to participate in a 30-minute study that consisted of several different tasks including answering questions from a wide variety of different fields in philosophy. Upon arrival, participants were given a brief introduction. Then they were brought to a workstation in a computer lab where they started the survey.

Participants

This data set consisted of 60 Ws, 60 EAs, and 59 SCs for a total sample of 179.

Scenarios Presented

We surveyed the Truetemp and Conspiracy cases for this data set.

Data Set 3

Procedure

For the third data set we launched questionnaires on SurveyMonkey (SM) that consisted of six questions, the four described in Section 1.2.1 and two on semantic intuitions taken from Machery et al. (2004). Participants sign up with SM and receive links to surveys from time to time. For every survey completed, SM donates $0.50 to a charity of the participant's choice. In addition, participants are entered into a draw for a chance to win a $100 gift card of an online store.[6] The first page of the survey was a brief introduction giving some background information. This included, for example, that the survey was for an academic study and the approximate time the study would take. After seeing the six questions, participants filled out a demographic questionnaire and finally there was also a text box to leave comments.

Participants

This data set consisted of 75 Ws, 36 EAs and 12 SCs for a total of 123.

Scenarios Presented

We tested all four scenarios in this data set. We did not carry out significance tests for the SC samples, as these were too small.

Data Set 4

Procedure

[6] For more information, see https://contribute.surveymonkey.com/how-it-works

The data for this study was collected through Harvard University's Moral Sense Test (MST) website.[7] Participants visited the MST website without being solicited and took part in the surveys that consisted of several different tasks and the Car scenario was included as a filler question. Some of the tasks included watching video clips or visualizing certain situations.

Participants

This sample consisted of 193 Ws and 15 SCs. Given the small sample of SCs, we mainly include this data set for completeness, as there was a statistical difference on the Car scenario.

Scenarios Presented

We only tested the Car case here. We did not have sufficient EAs to carry out a meaningful comparison.

1.2.2: Results for East Asians and Westerners

When comparing EAs and Ws, Weinberg and colleagues found statistically significant differences for the Car and Truetemp scenarios; however, failed to find differences for the other two cases. In our replication attempts, we did not attain a significant difference for any of the scenarios. A summary table of the results, including the results of Weinberg et al. (2001), is presented below.[8] All tests of the original as well as replication studies are two-sided. In all tables, * denotes $p < .05$; ** denotes $p < .01$.

[7] http://moral.wjh.harvard.edu/index2.html

[8] Given that we did not detect a difference in our samples, we carried out post-hoc power analyses to determine whether our samples provided sufficient power. For all power analyses in this paper, we took the original experiments as estimates of the population effect sizes. Calculations were conducted according to (Faul, Erdfelder, Lang, & Buchner, 2007).

Scenario	Study	*N*	Ethnicity	*n*	Response (%)		power	*p*-exact
					Really Knows	Only Believes		
Car	Weinberg et al.	89	EA	23	56.5	43.5	0.71	0.006**
			W	66	25.8	74.2		
	Data Set 1	120	EA	41	26.8	73.1	0.90	0.146
			W	79	15.2	84.8		
	Data Set 1 (Philosophy Only)[9]	102	EA	35	22.9	77.1	0.83	0.604
			W	67	17.9	82.1		
	Data Set 3	111	EA	36	22.2	77.8	0.86	1.000
			W	75	22.7	77.3		
Ind. Truetemp	Weinberg et al.	214	EA	25	12.0	88.0	0.55	0.020*
			W	189	32.3	67.7		
	Data Set 2	60	EA	31	16.1	83.9	0.42	0.527
			W	29	24.1	75.9		
	Data Set 3	111	EA	36	27.8	72.2	0.60	1.000
			W	75	29.3	70.7		
Conspiracy	Weinberg et al. (Note 1)							no sig.
	Data Set 2	66	EA	31	9.7	90.3	n/a[10]	0.713
			W	35	14.3	85.7		
	Data Set 3	111	EA	36	22.2	77.8	n/a	0.800
			W	75	18.7	81.3		
Zebra	Weinberg et al. (Note 1)							no sig.
	Data Set 3	111	EA	36	30.6	69.4	n/a	0.346
			W	75	21.3	78.7		

Table 1.1: Epistemic Intuitions – EA and W

Note 1: Weinberg et al. mention in their section on South Asians that there were no differences between EAs and Ws for the Conspiracy and Zebra cases; however, they do not provide any further details of sample sizes or *p* values.

[9] Data Set 1 was collected in philosophy as well as political science classes. This row presents data collected in philosophy classes only.

[10] We do not present power values in instances where Weinberg et al. do not give details of their outcomes because this leaves us without estimates of the population effect size.

The Car scenario produced the clearest disparity between the original and replication studies. Whereas Weinberg and colleagues report that a majority of East Asian individuals had the 'Really Knows' intuition, none of the replication studies reproduced this finding. In fact, in one of the replication studies (Data Set 3) the percentage of 'Really Knows' answers was slightly lower for East Asians than for Western participants. Although the percentage of 'Really Knows' answers for EAs was higher than for Ws in Data Set 1, the difference was nowhere as extreme as the result that Weinberg et al. (2001) report.

With regard to the other scenarios, we did not detect a difference for Truetemp, whereas Weinberg et al. did and Weinberg et al. themselves did not detect any differences for the last two scenarios (Conspiracy and Zebra) and neither did we; that is, there was no disparity between the original and replication studies.

1.2.3: Results for South Asians and Westerners

With the exception of the Truetemp case, Weinberg et al. report significant differences between SCs and Ws for all of the four scenarios. Our findings, again, paint a different picture.

Scenario	Study	*N*	Ethnicity	*n*	Response (%)		power	*p*-exact
					Really Knows	Only Believes		
Car	Weinberg et al.	89	SC	23	60.9	39.1	0.83	0.002**
			W	66	25.8	74.2		
	Data Set 1	113	SC	34	14.7	85.3	0.93	1.000
			W	79	15.2	84.8		
	Data Set 1 (Philosophy Only)	96	SC	29	10.3	89.7	0.89	0.542
			W	67	17.9	82.1		
	Data Set 4	208	SC	15	46.7	53.3	0.77	0.011*
			W	193	17.1	82.9		
Ind. Truetemp	Weinberg et al. (Note 2)							n/a
	Data Set 2	54	SC	25	24.0	76.0	n/a	1.000
			W	29	24.1	75.9		
Conspiracy	Weinberg et al.	89	SC	25	28.0	64.0	0.46	0.025*
			W	66	10.6	89.4		
	Data Set 2	69	SC	34	11.8	88.2	0.42	1.000
			W	35	14.3	85.7		
Zebra	Weinberg et al.	86	SC	24	50.0	50.0	0.34	0.05*
			W	62	30.6	69.4		
	Replication (Note 3)							no data

Table 1.2: Epistemic Intuitions – SC and W

Note 2: Weinberg and colleagues do not specify whether there was or was not a significant difference for this scenario.

Note 3: We did not have sufficient data for this scenario in any of our data sets to carry out a meaningful comparison.

The Car case, again, produced the largest difference between the original and replication studies. The outcome of the Car case Weinberg et al. present for SCs and Ws is similar to the sample of EAs and Ws. In both cases a larger number of non-Western participants respond that Bob really knows that Jill drives an American car, whereas this relationship is

reversed for Western participants. Western individuals, as opposed to non-Westerners, according to the original paper, predominantly choose the 'Only believes' answer choice. Data Set 1, where data was collected in classrooms and was closest to the original paper in procedures, yielded a very different outcome. The percentages of South Asian and Western participants were almost identical. We did attain a significant difference between these two groups for Data Set 4; however, as mentioned before, the SC sample size was small and the outcome may not be very meaningful. We mainly include this data for completeness.

1.3: Socioeconomic Status and Epistemic Intuitions

In their section on socioeconomic backgrounds, Weinberg et al. (2001) conclude that socioeconomic status (SES) has a "major impact on subjects' epistemic intuitions" (Weinberg et al., 2001, p. 453). As a reason for why individuals from different SES may have different epistemic intuitions, Weinberg and colleagues write that a "possibility is that high SES subjects accept much weaker knowledge-defeaters than low SES subjects because low SES subjects have lower minimum standards for knowledge" (Weinberg et al., 2001, p. 447). The authors continue that "whatever the explanation turns out to be, the data we've reported look to be yet another serious embarrassment for the advocates of [IAE]" (Weinberg et al., 2001, p. 447). Our replication attempts do not support this conclusion.

For this part of the paper, we setup two questionnaires on SurveyMonkey (SM) to test the epistemic intuitions of individuals from different socioeconomic backgrounds on the same scenarios for which Weinberg et al. (2001) report differences.

1.3.1: Methods and Materials

Scenarios

We used the same wording as in Weinberg et al. (2001) for all of the scenarios with the exception of the Car case where we replaced the names of the cars from Buick and Pontiac to Ford and Jeep, respectively, in order to make the scenario more current.[11] For the wording of the scenarios, see Section 1.2.1.

Procedure

The SM procedure was the same as described in Section 1.2.1. Participants sign up with SM and receive links to surveys from time to time. For every survey completed, SM donates $0.50 to a charity of the participant's choice. In addition, participants are entered into a draw for a chance to win a $100 gift card of an online retailer.

We set up two templates on SM, which we will refer to as Template 1 and Template 2 from here on. The templates were identical with the exception of the order in which the scenarios were presented. Participants first saw a brief introduction stating that we were conducting the questionnaire for an academic research project in the field of philosophy. Next, participants saw the four scenarios from Section 1.2.1. For the first template the order was Conspiracy, Zebra, Truetemp and Car. In the second template the order was Zebra, Car, Conspiracy, and Truetemp. The survey concluded with a very short

[11] This may not have been a good choice of car brands, as Jeep became the subject of the U.S. presidential campaign, which we were not aware of at the time. There were some campaign ads circulating about Jeep's purchase by Fiat, an Italian company and that production of Jeep vehicles would be outsourced to China. This topic remained an issue after the elections. For further details, see http://www.politifact.com/truth-o-meter/statements/2012/oct/30/mitt-romney/mitt-romney-obama-chrysler-sold-italians-china-ame/ and http://www.politifact.com/truth-o-meter/article/2012/dec/12/lie-year-2012-Romney-Jeeps-China/ .

demographic section where we asked about ethnic background and education.[12] SM furthermore provided us data on gender, age range, household income[13] and education. For our data analysis we used data on education that participants submitted in our surveys and not data provided by SM. There was some discrepancy between the two sources, which may be partly explained by the fact that the information is not always up to date with SM and individuals make progress in their educational attainments.

Weinberg et al. (2001) reported significant differences for the Conspiracy and the Zebra cases (from their paper, it appears that the other two scenarios did not yield a difference, although this is not mentioned explicitly). Hence, we chose the specific sequence mentioned above in order to have the Conspiracy case as the first scenario in Template 1 and the Zebra case as the first scenario in Template 2.

Participants

Weinberg and colleagues used an education proxy to categorize participants as either low or high socioeconomic status.[14] Individuals who indicated that they had never attended college were classified as low SES, whereas participants who indicated that they had taken one or more courses at the college level were classified as high SES. We used the same criteria in classifying participants.

The survey with the second template was initiated about two weeks after the first survey and we asked SM not to send out invitations to any of the individuals who participated in

[12] These SM runs are entirely different from Data Set 3 presented in Section 1.2.
[13] Data on income was missing for one of the data sets, namely for the low SES data from Template 1.
[14] In order to maintain continuity with the terminology used in Weinberg et al. (2001), we will use the terms low and high SES throughout this paper.

the first questionnaire. For the first template, we asked SM to restrict participation to individuals who were 24 years of age or older. We were concerned that given the criteria for distinguishing low and high SES by an education proxy we might get many young respondents for the low SES group. After reviewing the data for the first template, we realized that our concern was unfounded and we omitted this requirement for the second template.

For Template 1 our sample consisted of 107 participants (38 low SES, 69 high SES). For Template 2 our sample consisted of 134 individuals (47 low SES, 87 high SES).

1.3.2: Results for Socioeconomic Status

A summary table of the results for socioeconomic status is presented below.

Scenario	Study	*N*	SES	*n*	Response (%)		power	*p*-exact
					Really Knows	Only Believes		
Car	Weinberg et al. (Note 4)							n/a
	Template 1	106	low	38	44.7	55.3	n/a	0.014*
			high	68	20.6	79.4		
	Template 2	133	low	46	23.9	76.1	n/a	0.177
			high	87	35.6	64.4		
Truetemp	Weinberg et al. (Note 4)							n/a
	Template 1	106	low	38	28.9	71.1	n/a	0.211
			high	68	42.6	57.4		
	Template 2	132	low	45	33.3	66.7	n/a	0.696
			high	87	29.9	70.1		
Conspiracy	Weinberg et al.	59	low	24	50.0	50.0	0.74	0.007**
			high	35	17.1	82.9		
	Template 1	107	low	38	18.4	81.6	0.94	0.790
			high	69	15.9	84.1		
	Template 2	132	low	45	22.2	77.8	0.97	0.476
			high	87	16.1	83.9		
Zebra	Weinberg et al.	58	low	24	33.3	66.7	0.44	0.038*
			high	34	11.8	88.2		
	Template 1	106	low	38	31.6	68.4	0.72	0.824
			high	68	27.9	72.1		
	Template 2	134	low	47	27.7	72.3	0.81	0.675
			high	87	23.0	77.0		

Table 1.3: Epistemic Intuitions – SES

Note 4: Weinberg and colleagues do not state explicitly whether the Car and Truetemp scenarios yielded a significant difference or whether no data was collected. The implication seems to be that data was collected but no difference was detected.

Template 1

For Template 1 none of the scenarios yielded a significant difference with the exception of the Car case.

Template 2

In creating Template 2, we made some changes to the first template. First, we changed the order in which the scenarios were presented. Since Weinberg et al., in addition to the Conspiracy case also reported a significant difference for the Zebra scenario, we wanted this case to be placed at the beginning, so we could rule out order effects. Second, given that there was a significant difference for the Car case in our first template we wanted to place this scenario further toward the beginning of the survey in order to rule out participation fatigue as one of the reasons for the difference.

There are several things worth pointing out here. First, the Zebra case, again, did not yield a significant difference when presented as the first scenario. In fact, this time none of the scenarios yielded a significant difference. The Car case produced the closest *p* value to a significant level ($p = 0.177$); however, this time the direction of the responses was reversed when compared to the first template. This time low SES participants had a lower percentage of 'Really Knows' responses than high SES participants, which contradicts Weinberg et al.'s explanation about the role of socioeconomic status on epistemic judgments.

Further Analyses

There were various other tests we ran to examine the data. First, we ran an analysis of the combined data from the two templates. Despite the large sample ($N = 240$), none of the scenarios produced a significant outcome or a *p* value close to 0.10. We do not think that this is merely because of cancelling order effects (the Car case was significant in Template 1, however, the direction was reversed in Template 2 and these effects could be cancelling

each other out when combining the data). Rather it seems to be that despite the increased sample size, we still could not find a difference between the two groups. We tested for order effects by comparing the data of the two templates and the only scenario that produced a significant difference was the Car case.

Next, we wanted to see if there would be a significant difference between the two groups if we made the difference in educational attainment greater. So, for the high SES group we included in our next analysis only participants who had at least completed their Bachelor's degree. Low SES was coded as before. The outcomes (statistical significance) for the four scenarios did not change for either one of the templates.

We further ran an analysis excluding participants where the self-reported education level and that provided by SM did not match. None of the outcomes changed. Finally, we ran analyses excluding all participants who fell in the age range 18-29. This made the Car case for the second template significant (again, in the opposite direction of Template 1) but otherwise all other outcomes remained unchanged.

1.4 Statistical Power

With one exception (data on Conspiracy for SC/W), in all experiments on ethnicity and socioeconomic background, we had at least one sample where we attained greater power[15] than Weinberg and colleagues. If differences actually existed for these conditions, it would have been more likely that our data would have revealed it. In several cases the power we attained was above 0.90 and so clearly above the 0.80 conventional mark. In two instances

[15] Power is the probability of detecting an effect in the sample if an effect exists in the population.

(Truetemp for EA/W and Conspiracy for SC/W) we fell short of attaining this conventional mark; however, in these cases Weinberg et al. had comparable values. Overall, if effects existed in the population, our data would have been more likely to detect these, yet we still failed to find statistical differences.

1.5: Discussion and Concluding Remarks

Discussion

When discussing our results at conferences or informally with colleagues, we have been told on multiple occasions that given that differences between East Asian and Western participants have been shown for some cognitive tasks in the Nisbett literature (Nisbett, Peng, Choi, & Norenzayan, 2001), it should not come as a surprise that these ethnic groups may exhibit different epistemic intuitions. There are two responses that we can offer. First, differences in cognitive tasks in some areas do not necessarily predict differences in other areas. Secondly, and much more importantly, although the work of Nisbett and colleagues has been influential, there have been some notable failures of replication in that field (Evans, Rotello, Li, & Rayner, 2009; Lu, Daneman, & Reingold; Miellet, Zhou, He, Rodger, & Caldara, 2010; Rayner, Castelhano, & Yang, 2009; Rayner, Li, Williams, Cave, & Well, 2007; Zhou, Gotch, Zhou, & Liu, 2008). As it stands, more work is needed to attain a better picture.

Additionally, Nagel et al. (2013) report a failed conceptual replication of Weinberg et al. (2001) on the effect of ethnicity on epistemic intuitions.[16] Further, one experiment in Turri

[16] We describe this as a conceptual replication as the procedures in Nagel et al. (2013) differed from those of the reference experiment. The answer options participants could choose from were also different.

(2013) can be understood as a failed conceptual replication of Weinberg et al. (2001).[17] Finally, at the time of preparing this manuscript's final changes, we were contacted by a group who informed us that they tested the Car scenario (exact replication) in Northeast United States and this group, too, was unable to reproduce the effect reported by Weinberg and colleagues (Minsun & Yuan, ms). In all, there is now increasing evidence that the findings of Weinberg et al. (2001) are not reproducible. The main reason we see for why there is a difference between the original and replication studies and hence the failures of replication is that the sample sizes in the original study are relatively small; on average the sample size of the EA/SC and low SES samples is 24.

Weinberg and colleagues draw some strong conclusions in their article as well as elsewhere. For example, one of the co-authors (Stich) writes that "high SES Americans and low SES Americans have different epistemic intuitions! Moreover, in many cases these differences are quite dramatic." Stich continues that a "reasonable conclusion is that philosophy's 2400 year long infatuation with Plato's method has been a terrible mistake" (Stich, 2001, p. 38). The simple takeaway from our study is that much more and better evidence is needed to make such strong claims.

For the most part of this paper we attempted to remain neutral on the debate concerning the role of intuitions in philosophy. We wanted the focus to be on the findings presented; however, our findings and the results of the three groups mentioned above all weaken one of the main arguments experimental philosophy has brought against the use of intuitions in philosophy, in as far as these arguments relied on Weinberg et al. (2001).

[17] Turri (2013) is not a straightforward conceptual replication of Weinberg et al. (2001) because Turri attempts to manipulate participants' responses. However, in the context of our work, Turri (2013) suggests uniformity of epistemic intuitions among South Asian and Western individuals.

It may be the case that had we surveyed individuals born and residing in East and South Asia, we may have detected responses different from Westerners.[18] However, there are several things to note in this regard. Turri (2013) surveyed Indians living in India on a Gettier-type[19] scenario and although the form of presentation and the scenario were different from Weinberg et al. (2001), Turri found much lower rates of 'Really Knows' answer choices (15%) and also no difference when compared to an American population, 96% of whom indicated English as their native language. For the exact details, see Turri (2013). Minsun & Yuan (ms) compared EAs and Ws as categorized on the basis of native language on the Car scenario and found no significant difference (Minsun & Yuan, ms). These studies indicate that there may not be a difference between Ws and East/South Asians born and residing in their native countries. However, strictly speaking, this possible difference cannot be ruled out, as we do not have the necessary data.

This discussion, though, is beyond the scope of Weinberg et al. (2001) and consequently the current paper. Weinberg et al. surveyed a population living and studying in the New Jersey area; the target-population was ethnic minorities living in the West and for these they found big differences when compared to individuals of European descent. This is mainly what makes Weinberg et al.'s results so surprising. It would not be too surprising if surveying a population that is very unlike Ws yielded different outcomes, as one is likely to introduce problems of language and task comprehension, amongst others. What gives Weinberg et al. (2001) its surprise factor is that merely surveying participants from

[18] This was pointed out by an anonymous referee.
[19] For lack of a better term, with Gettier-type or Gettier-style scenarios, we broadly refer to cases involving unwarranted or disputed knowledge, including all the cases presented by Weinberg et al. (2001) and discussed in this paper.

different ethnic backgrounds within the same university yielded different intuitive responses. And demographic variables such as ethnicity, most philosophers believe, should not factor into evaluations of philosophical questions.

Conclusion

The aim of this paper was to test the robustness of the results of Weinberg et al. (2001). Despite collecting data from various sources and attaining larger samples in several of the cases, we failed to detect differences on epistemic intuitions between participants from different ethnic backgrounds and socioeconomic statuses. With regard to socioeconomic status, we collected data from 241 individuals on four scenarios surveying epistemic intuitions for which Weinberg et al. (2001) report significant differences (on two of the cases) but failed to find statistically significant differences. Given this data, we do not believe that socioeconomic status by itself has an impact on epistemic intuitions for the cases evaluated in this paper. With regard to ethnicity and epistemic intuitions, even though we collected data in several different settings, we could not replicate the results of Weinberg et al. (2001) on differences among individuals from East Asian, South Asian and Western backgrounds. Given this data, we do not believe that ethnic background has a significant impact on epistemic intuitions.

As mentioned in the Introduction, Weinberg et al. (2001) has been an influential paper, which has received numerous citations. In discussions with other researchers in the field, it often appears that it is an established fact that epistemic intuitions differ among ethnic groups. Our data suggests that this conception needs to be corrected. Despite the important implications of the original paper and despite the debate surrounding the findings of Weinberg et al. (2001) for conducting epistemology as well as philosophy in general, there

had not been exact replication attempts of Weinberg et al. (2001) to test the robustness of the reported results until 2011 when we started out on this project. As mentioned above, since then more work has been done in this area and all studies produced results in line with our findings. One of our initial hopes for sharing our data was that other researchers would study the cases, so that the original results could be verified independently. We further hope that other researchers will find it worthwhile to examine these cases and we hope to have provided a useful reference point with this paper.

Page Intentionally Left Blank

Paper 2: On Gender and Philosophical Intuitions[20]

Failure of Replication and Other Negative Results

Abstract

In their paper titled "Gender and Philosophical Intuition," Wesley Buckwalter & Stephen Stich argue that the intuitions of women and men differ significantly on various types of philosophical questions (Buckwalter & Stich, 2013). Furthermore, men's intuitions, so the authors claim, are more in line with traditionally accepted solutions of classical problems. This inherent bias, so the argument goes, is one of the factors that leads more men than women to pursue degrees and careers in philosophy. These findings have received a considerable amount of attention and the paper is to appear in the second edition of *Experimental Philosophy* (Knobe & Nichols, 2013), which itself is an influential outlet. Given the exposure of these results, we attempted to replicate three of the classes of questions that Buckwalter & Stich review in their paper and for which they report significant differences. We failed to replicate the results using several different sources for data collection (one being the same as in the original article). Given our results, we do not believe the outcomes from Buckwalter & Stich (2013) that we examined for this paper to be robust. That is, men and women do not seem to differ significantly in their intuitive responses to these philosophical scenarios.

[20] The author would like to thank Wesley Buckwalter and Stephen Stich for providing the details of the procedures and methods of their experiments and answering any questions we had. We would also like to thank Donal Cahill for his help with the Moral Sense Test and all class teachers at the University of London who provided us with class time to collect data and all students who participated. This paper was published in *Philosophical Psychology* (Taylor & Francis) in 2014 with some minor changes.

2.1: Introduction and Overview

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

In supporting their claims, the authors review some recent findings in experimental philosophy (Section 3 of their paper) and also present new data for four classical scenarios in which they report men and women to respond differently to survey questions (Section 3.8). The thought experiments in this section (3.8) include the Brain in the Vat, Hilary Putnam's Twin Earth, John Searle's Chinese Room and the Plank of Carneades. These cases are of special interest to Buckwalter & Stich because these are cases that undergraduate students typically encounter early on in introductory philosophy classes. Hence, so the authors argue, if women's responses differ from commonly accepted solutions of philosophical problems, women could be discouraged from pursuing further philosophy courses. In addition to the scenarios of section 3.8, in section 3.2 Buckwalter & Stich present results on compatibilism, physicalism and dualism cases where women and men are also reported to answer questions differently. We attempted direct replications of Sections 3.2 and 3.8 of Buckwalter & Stich (2013) and our results indicate that the outcomes reported by Buckwalter & Stich are not robust.

Furthermore, in Section 3.1 Buckwalter & Stich report differences between men and women for two variations of a Gettier style scenario. We had collected data on four Gettier-type scenarios for another study and analyzed the results to see how women and men answered the questions. Although this is a conceptual replication,[21] we believe the results to be relevant for this paper. Once again, our data showed no difference between the two groups of respondents.

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

To address this issue, we wanted to analyze individuals who had taken at least some philosophy courses but whose views had not been biased by previous study of the cases. We collected information on how many philosophy courses participants had taken and

[21] We distinguish a conceptual replication from a direct replication in that the latter tests a previously reported effect by presenting survey participants with scenarios identical in wording to the original study. By contrast, a conceptual replication tests previously reported effects through scenarios of the same type but not necessarily by using identical cases. In this instance, Section 2.4 offers a conceptual replication where we did not use identical scenarios to those reported by Starmans and Friedman but instead used Gettier-type scenarios that examined similar concepts.

whether they had seen the scenarios before. In this way we could evaluate the answers of respondents who had been interested enough to take at least some philosophy classes and may have pursued philosophy as a career but had not seen and not been familiar with these scenarios.[22] Here again, we failed to detect a difference between men and women.

We do not necessarily disagree with the general approach Buckwalter & Stich attempt to take. Intuitive responses to survey questions may (or may not) differ among men and women for certain problems and this may (or may not) lead more women or men to pursue certain fields and careers. This is an issue that needs to be examined empirically. However, if much rests on the results that Buckwalter & Stich present in section 3 of their paper, then the failure of replication weakens their argument. Buckwalter & Stich suggest that differences in intuitions may be one factor among many that influence career choices in philosophy and if our results are robust, this factor plays a smaller role (if any) than the findings in Buckwalter & Stich (2013) suggest.

Our main aim for this paper was to test to the robustness of the results in Buckwalter & Stich (2013) and to share our results with others, especially researchers who may want to build on the results of Buckwalter & Stich (2013). It is because of this focus that we will keep the discussion on the role and importance of intuitions in philosophical endeavors to a minimum. Furthermore, others have provided a better overview and discussion on the issue than we can present here (Alexander, 2012; Buckwalter & Stich, 2013; Kauppinen, 2007; Knobe & Nichols, 2008; Nadelhoffer & Nahmias, 2007; Nagel, 2012); also, see Section 1.1 of Paper 1.

[22] It may be possible to bias individuals in philosophy courses other than through direct exposure of some cases. Nevertheless, by restricting samples as described, we could at least rule out that participants had been influenced directly.

This paper is structured as follows. In the next section we will examine the classical philosophical scenarios presented in Section 3.8 of Buckwalter & Stich (2013). Specifically, in Section 2.2.1 we present the procedures and methods of data collection. In Section 2.2.2 we present the results for the replication experiments and in Section 2.2.3 we present data on participants who had taken some philosophy classes but who indicated that they had not seen the cases before. In Section 2.3 we examine the scenarios for Compatibilism, Physicalism and Dualism that Buckwalter & Stich describe in their Section 3.2. In Section 2.3.1 we present the replication results and in Section 2.3.2 we examine the sample of respondents with some philosophy background but who were not familiar with the scenarios. In Section 2.4 we present the data for Gettier-type scenarios. In the final section we provide some concluding remarks and a brief discussion on possible reasons for why replication failed.

The data for this paper was collected between April 2011 and June 2012. A first draft of this paper was uploaded to SSRN on October 24, 2012 (Seyedsayamdost, 2012a). First drafts of this paper were sent to conferences starting in July 2012.

2.2: Classical Thought Experiments

For this section we collected data though two different sources. The first was through Amazon's Mechanical Turk (MT) following the methodology in Buckwalter & Stich (2013). For the second data set we ran surveys on SurveyMonkey (SM). We will describe the procedures of data collection for all data sets first and then present the results in the subsequent sub-section. This way we can compare the outcomes more readily.

2.2.1: Procedures and Methods of Data Collection

Mechanical Turk

We tried to follow Buckwalter & Stich's methodology as closely as possible. In the Human Intelligence Task (HIT) description respondents were given some brief information about what the task entailed, the approximate time needed to complete the task and some other information required by MT. Once participants accepted a task, they were shown one of the four scenarios presented in section 3.8 of Buckwalter & Stich (2013). The scenario was followed by a comprehension check question (the same that was asked in the original paper) and a question asking for a response on a seven-point scale. The one difference we made to Buckwalter & Stich's outline is the inclusion of another question asking whether respondents had seen the scenario before. We included this question for two reasons. First, as mentioned in the introduction, we wanted to test participants with a background in philosophy but who were not familiar with these scenarios. Second, Buckwalter & Stich had run these same scenarios on MT and we wanted to be able to exclude respondents who may have had seen these cases in a run conducted by Buckwalter & Stich.

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

SurveyMonkey

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

2.2.2: Results

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

Brain in the Vat: Original Results

The first case that Buckwalter & Stich present in their section 3.8 is the Brain in the Vat scenario. The exact wording is as follows. [23]

[23] All scenarios in this section are taken from Buckwalter & Stich (2013).

> George and Omar are roommates, and enjoy having late-night 'philosophical' discussions. One such night Omar argues, "At some point in time, like, the year 2300, the medical and computer sciences will be able to simulate the real world very convincingly. They will be able to grow a brain without a body, and hook it up to a supercomputer in just the right way so that the brain has experiences exactly as if it were a real person walking around in a real world, talking to other people. The brain would believe it was a real person walking around in a real world, except that it would be wrong. Instead it's just stuck in a virtual world, with no actual legs to walk and with no other actual people to talk to. And here's the thing: how could you ever tell that it isn't really the year 2300 now, and that you're not really a virtual-reality brain? If you were a virtual-reality brain, after all, everything would look and feel exactly the same to you as it does now! George thinks for a minute, and then replies: "But, look, here are my legs". He points down to his legs. "If I were a virtual-reality brain, I wouldn't have any legs really, I'd only just be a disembodied brain. But I know I have legs, just look at them! So I must be a real person, and not a virtual-reality brain, because only real people have real legs. So I'll continue to believe that I'm not a virtual-reality brain."
>
> George and Omar are actually real humans in the actual real world today, and so neither of them are virtual-reality brains, which means that George's belief is true.

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

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

Brain in the Vat: Replication Results

[24] Summaries of data for the original outcomes are taken from Buckwalter & Stich (2013). All tests in the original as well as replication studies are two-sided.

Mechanical Turk

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

Our data for a sample of 114 individuals (58 Female and 56 Male) resulted in a mean score of 5.25 (*SD* = 2.24) for men and a mean score of 5.86 (*SD* = 1.85) for women. We conducted an independent-samples t-test for men's and women's responses which yielded: $t(107) = -1.59$ (equal variance not assumed), $p = 0.115$.[25] Despite a sample that was close to twice as large as that of Buckwalter & Stich, we did not detect a difference at the 10% level.

SurveyMonkey

The overall result for the Brain in the Vat scenario from our SurveyMonkey data is as follows. $N = 100$ (Male = 51, Female = 49). Male: $M = 5.78$, $SD = 1.86$. Female: $M = 5.61$, $SD = 1.82$. An independent-samples t-test comparing men and women yielded: $t(98) = 0.455$), $p = 0.650$.[26]

[25] We will refer to the groups as women/men and female/male interchangeably as female/male is how we asked for gender in the demographic part of our surveys.
[26] See Appendix B for the breakdown of the individual surveys.

What stands out from the three data sets is the high value for women's mean response (6.72) in Buckwalter & Stich (2013). Following is a visual presentation for the outcomes of the three procedures.[27]

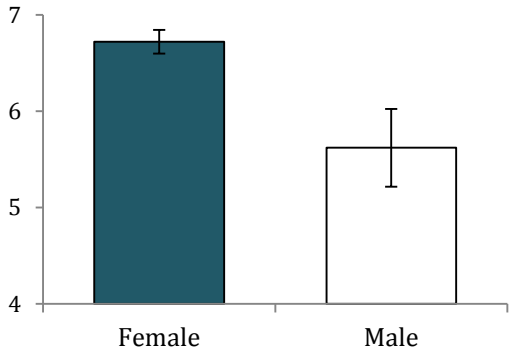

Figure 2.1a: Brain in the Vat – Original[28]

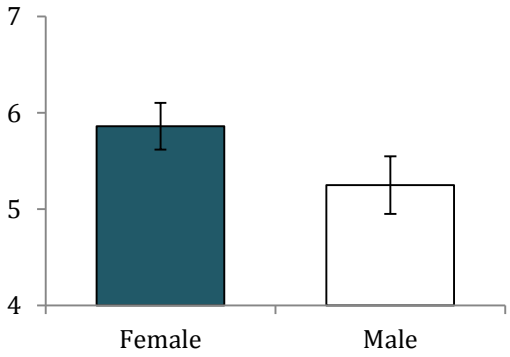

Figure 2.1b: Brain in the Vat – MT

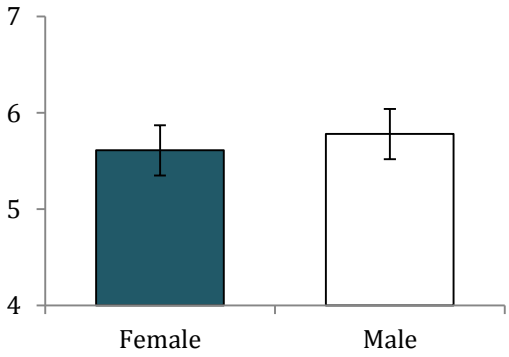

Figure 2.1c: Brain in the Vat – SM

Statistical Power

The statistical power of the Brain in the Vat data set from Buckwalter & Stich (2013) with a sample of 63 participants, an effect size of $d = 0.81$ and an alpha level of 0.05 came out to

[27] We used the same scale ranges for the charts as in Buckwalter & Stich (2013) for better comparison. Also in keeping with Buckwalter & Stich's layout, our error bars represent S.E. +/- 1.

[28] All charts of the original outcomes in this paper are taken from Buckwalter & Stich (2013). The charts have been adapted for clarity and to match the style of the replication charts.

0.87.[29] Using $d = 0.81$ as the population effect size, the replication attempt from the MT and SM datasets achieved a power of 0.99 and 0.98, respectively. Although Buckwalter & Stich attain an adequate power according to the conventional 0.80 mark, our data sets had considerably greater power to detect a difference between men and women, if one existed. If a difference existed between women and men at the 5% level (given a population effect size as drawn from Buckwalter & Stich (2013)), our data would have had a 99% (or 98%) likelihood of detecting it, yet in both data sets replication failed.

Twin Earth

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

> Suppose that elsewhere in the universe there is a planet called "Twin-Earth". Twin-Earth looks exactly like our Earth in virtually all respects. It is populated by twin equivalents to every person and thing here on our Earth, and even revolves around a star that appears to be exactly like our sun.
>
> Oscar grows up here on our Earth, while someone exactly like Oscar, who we can call "Twin-Oscar", lives on Twin-Earth. Oscar and Twin-Oscar both go through life having the same experiences, and both perceive their environment in exactly the same way. They look and act completely alike, and even experience the same emotions.
>
> In fact, there is only one difference between these two planets. The difference is that on Earth the stuff that fills the lakes and rivers and that people and animals drink is H2O, while on Twin Earth, the stuff that fills the lakes and rivers and that people and animals drink is another chemical compound, XYZ, that to the naked eye looks completely indistinguishable from the H2O on Earth. H2O and XYZ also taste exactly the same, and both have the ability to quench thirst and to sustain life.
>
> However, Oscar and Twin-Oscar both live before the development of modern science, and they have no idea about chemistry or molecular composition. When they go for a swim, both Oscar and Twin-Oscar point to the liquid in the lake and

[29] Power analyses were conducted following (Faul et al., 2007).

> call it “water” even though on Earth that liquid is made up of H2O, and on Twin-Earth it is made up of XYZ.

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

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

Participants then entered their response on a seven-point scale where the leftmost option was marked “they mean different things,” the midpoint labeled “in between” and the rightmost option marked “they mean the same thing” (they mean different things = 1, in between = 4, they mean the same thing = 7).

Twin Earth: Original Results

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

Twin Earth: Replication Results

Mechanical Turk

In our MT sample there was no significant difference among men and women, and in fact women had a higher average mean than men.[30] We used the same criteria as in the Brain in

[30] It should be noted that in the original study, women, more than men, tended to give the ‘standard’ or ‘male-centrist’ response. Buckwalter & Stich do not comment on whether this undermines their overall hypothesis to a degree.

the Vat case to exclude participants from analysis: N = 117 (Male = 65, Female = 52). Male: M = 5.22, SD = 2.35. Female: M = 5.46, SD = 2.11. Independent-samples t-test: $t(115) = -0.589$, $p = 0.557$.

SurveyMonkey

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

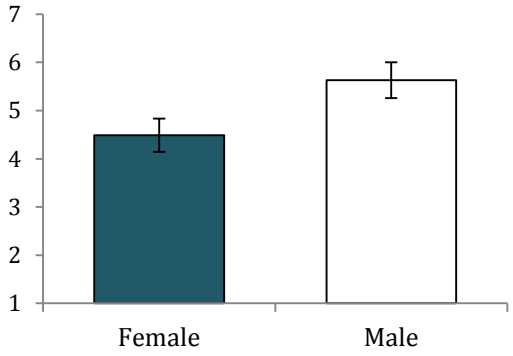

Figure 2.2a: Twin Earth – Original

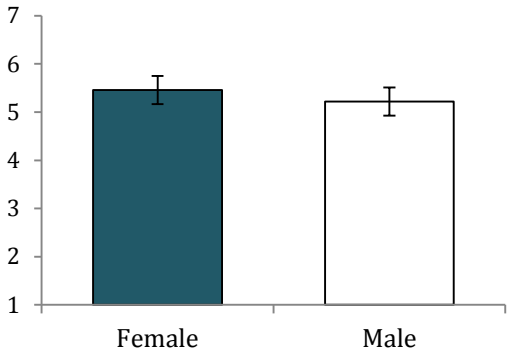

Figure 2.2b: Twin Earth – MT

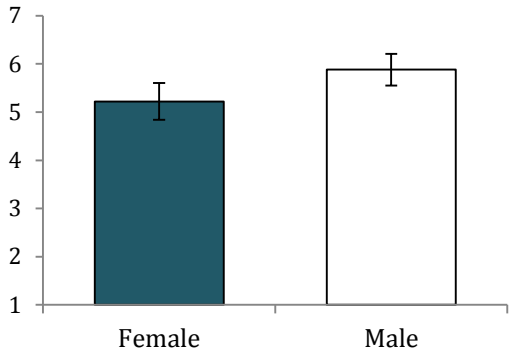

Figure 2.2c: Twin Earth – SM

Statistical Power

The power achieved by Buckwalter & Stich for their Twin-Earth data set equals 0.59. Taking Buckwalter & Stich's effect size as that of the population, our MT and SM samples yielded power values of 0.74 and 0.60, respectively. The MT sample comes close to the 0.80 convention and both of the replication samples achieve a greater power, although the SM sample does so only marginally. The MT sample had considerably greater power than the original study to detect a difference, yet again replication failed.

Chinese Room

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

> Jenny is a native English speaker who can only speak English. She is locked in a room full of boxes of Chinese symbols, together with an instruction manual written in English for manipulating the symbols. People from outside the room send in notes on pieces of paper with Chinese symbols written on them, which unknown to Jenny, are questions in Chinese. Jenny's job is to look through her manual until she finds the symbols that look exactly like the ones written on the pieces of paper. When she finds that string of symbols, the manual will tell her what new string of symbols to write down, and send to the people outside the room.
>
> By following the instructions in the manual, Jenny is able to give the correct answers to the questions. The system consisting of Jenny and the instruction manual that she is using can be thought of as an unusual sort of computer. Jenny gets so good at following the instructions in the manual, that from the point of view of any one outside the room who speaks Chinese, her responses are absolutely indistinguishable from those of Chinese speakers.

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

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

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

Chinese Room: Original Results

Buckwalter & Stich report for $N = 127$ (Male = 54, Female = 73) Male: $M = 4.13$, $SD = 2.47$. Female: $M = 3.25$, $SD = 2.36$ ($d = 0.37$). Independent-samples t-test: $t(125) = 2.05$, $p < 0.05$

Chinese Room: Replication Results

Mechanical Turk

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

SurveyMonkey

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

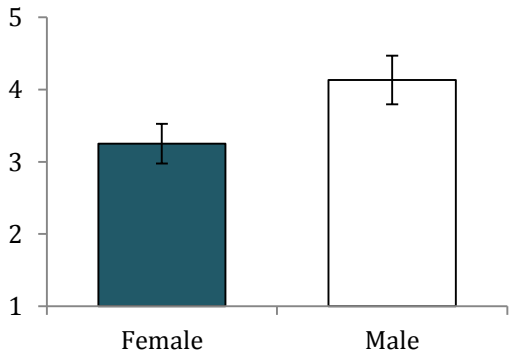

Figure 2.3a: Chinese Room – Original

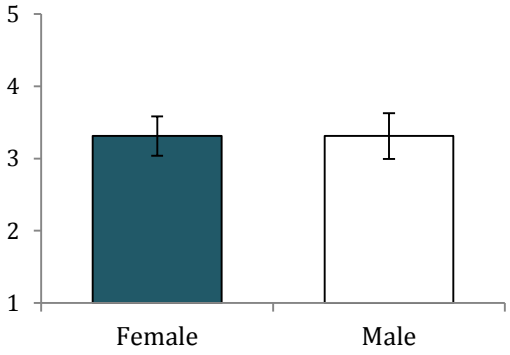

Figure 2.3b: Chinese Room – MT

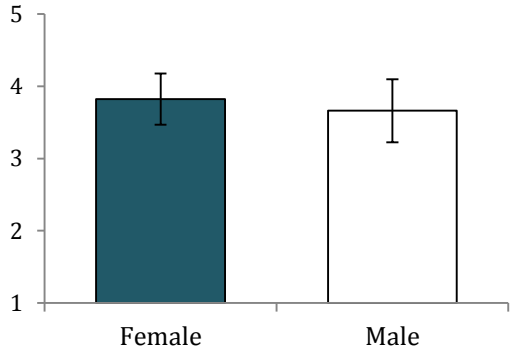

Figure 2.3c: Chinese Room – SM

Statistical Power

The original data set by Buckwalter & Stich achieved a power of 0.53. The replication experiments from MT and SM achieved power values of 0.46 and 0.37, respectively. All three studies were underpowered in this instance.

Plank of Carneades

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

> There are two shipwrecked sailors, Jamie and Ricki. They both see a small plank that can only support one of them and both of them swim desperately towards it.

> Jamie gets to the plank first. Ricki, who is stronger and is going to drown, pushes Jamie off and away from the plank and, thus, ultimately, causes Jamie to drown. Ricki gets on the plank and is later saved by a rescue team.

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

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

Plank of Carneades: Original Results

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

Plank of Carneades: Replication Results

Mechanical Turk

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

SurveyMonkey

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

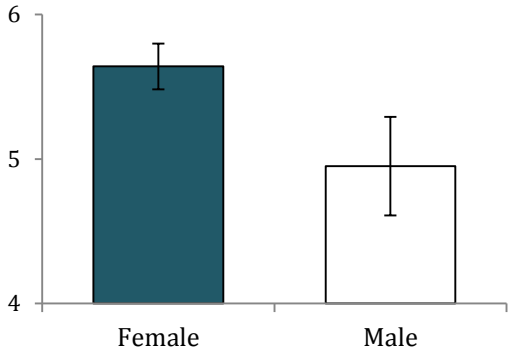

Figure 2.4a: Plank of Carneades – Original

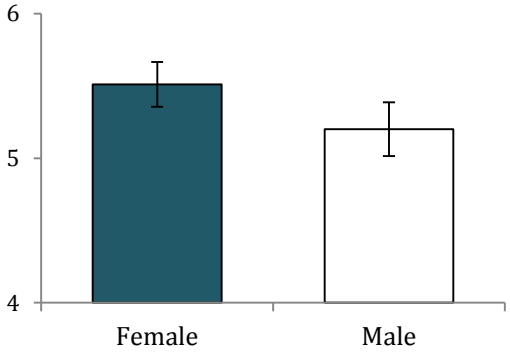

Figure 2.4b: Plank of Carneades – MT

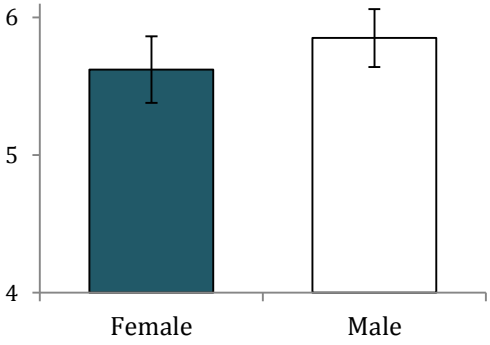

Figure 2.4c: Plank of Carneades – SM

Statistical Power

The statistical power achieved by Buckwalter & Stich was 0.54. The replication samples from MT and SM attained power values of 0.74 and 0.54, respectively. Here again, the MT sample achieved a much higher power than the original study and would have been much more likely to detect a difference, if one existed.

Leaving aside the Brain in the Vat scenario, the average power achieved by Buckwalter & Stich for the other cases was relatively low at 0.55. Given this value, even if an effect actually existed for all three scenarios, the likelihood of having detected all of them with the samples of the original study would have been just somewhat better than 12%. In the great majority of runs we would expect at least one out of the three scenarios to yield a false negative (failure to detect an effect where one exists). Others have pointed out that there is some evidence that Buckwalter & Stich selectively reported experiments that yielded positive results but neglected to mention conditions where women and men showed no differences (Nahmias, 2013). This could explain the somewhat unexpected findings, despite low power values.

Further Analyses

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

2.2.3: Some Philosophy Background but Cases Not Seen Before

As mentioned in the introduction, we believe that the respondents Buckwalter & Stich selected for their analysis is not quite adequate for the purpose of examining the gender gap in professional philosophy. The reason is that anyone who had taken at least one or more philosophy courses was excluded from analysis. This leaves a sample of respondents who never had an interest or perhaps possibility to pursue philosophy in an academic setting.

In the context of Buckwalter & Stich's discussion on who chooses to pursue philosophy as a degree or career, we thought it useful to examine those respondents who had taken some philosophy classes but indicated that they had not seen the scenarios before. This way we wanted to attain a sample of individuals who had been interested and had the possibility to pursue philosophy in an academic setting but who were unbiased by previous (possibly 'male-centrist') discussions of the cases. Filtering in this way would also reduce some amount of noise given that now respondents had more in common in terms of their educational background.

To summarize, the criteria that had to be met for participants to be included in the analysis here were 1) comprehension check was answered correctly, 2) time spent to complete the task was not less than 30 seconds, 3) native language was English, 4) indicated that they had not seen the scenarios before and 5) indicated number of classes were between one and three. In specific, in the demographic section of the surveys we

asked how many philosophy courses respondents had taken and the answer choices provided were '0', '1 to 3', '4 to 6' and '> 6'. This was the same for all surveys with the exception of the Chinese Room scenario where we asked whether participants had taken any philosophy courses and the answer choices were 'Yes' and 'No'. Respondents had to have chosen 'Yes' (in addition to fulfilling the other criteria) to be included in the analysis provided below.

For this group of respondents, again, there were no statistically significant differences between women and men. The data in this section is drawn from the Mechanical Turk data sets. The samples for the SurveyMonkey data were relatively small after filtering in this way. None of the scenarios from SM yielded a significant difference and hence we will not present the outcomes here. We present the summary of the outcomes and graphs for the Mechanical Turk data below.

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

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

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

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

Chinese Room: More than One Philosophy Course (MT)

N = 77 (Male = 32, Female = 45). Male: *M* = 3.22, *SD* = 2.00. Female: *M* = 3.33, *SD* = 1.78. Independent-samples t-test: *t*(75) = -0.264, *p* = 0.792.

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

N = 190 (Male = 99, Female = 91). Male: *M* = 5.39, *SD* = 1.60. Female: *M* = 5.71, *SD* = 1.46. Independent-samples t-test: *t*(188) = -1.438, *p* = 0.152.

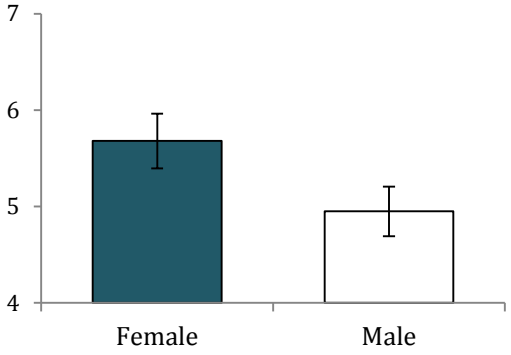

Figure 2.5a: Brain in the Vat - One to Three Philosophy Courses (MT)

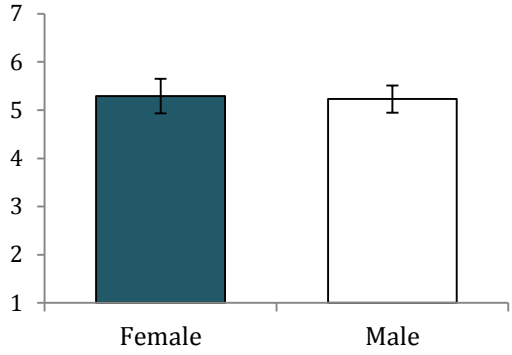

Figure 2.5b: Twin Earth – One to Three Philosophy Courses (MT)

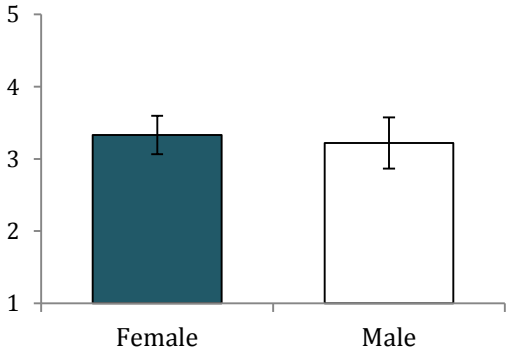

Figure 2.5c: Chinese Room – One to Three Philosophy Courses (MT)

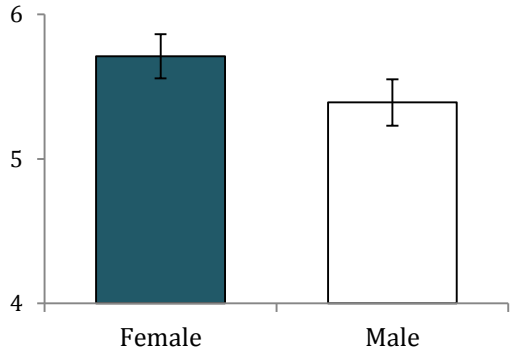

Figure 2.5d: Plank of Carneades – One to Three Philosophy Courses (MT)

Statistical Power

Lacking a better estimate of the population effect sizes, we used the sample effect sizes from Buckwalter & Stich (2013) as presented in section 2.2.2 above. The power values for Brain in the Vat, Twin Earth, Chinese Room and Plank of Carneades came out to 0.99, 0.58, 0.35, and 0.82, respectively. The Twin Earth and especially the Chinese Room samples were underpowered when using the indicated effect sizes. The Plank of Carneades sample had adequate power to detect a difference between women and men, had one existed. The power value for the Brain in the Vat scenario was again very high at 0.99 and yet again we failed to detect a difference.

We will provide a brief discussion of these results in the concluding section of the paper. Next, we will discuss section 3.2 of Buckwalter & Stich (2013) where the authors present results taken from Geoffrey Holtzman on compatibilism, materialism and dualism.

2.3: Compatibilism, Materialism, and Dualism

2.3.1 Results

For the scenarios in this section we collected data through SurveyMonkey. The method of data collection is the same as described in section 2.2.1.

Compatibilism

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

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

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

Replication Results: SM

Using the same filter as Holtzman, we failed to attain a significant difference among men and women. Our sample consisted of 92 participants with 50 of those being female and 42 male. A chi-square test yielded $\chi^2 = 0.652$, $p = 0.419$.[31]

[31] Throughout this paper we will report the results for chi-square tests when none of the cells have an expected count of less than five and will conduct Fisher's exact tests otherwise.

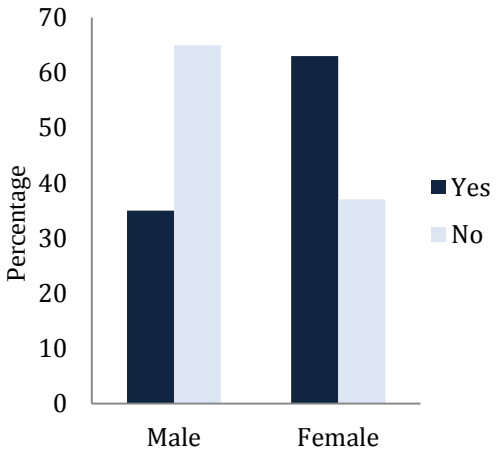

Figure 2.6a: Compatibilism – Original Results

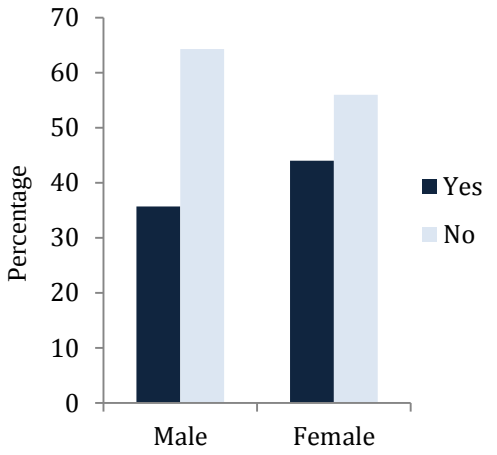

Figure 2.6b: Compatibilism – Replication Results (SM)

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

Physicalism

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

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

Again, the possible answer choices were 'Yes' or 'No'. Buckwalter & Stich report a Fisher's exact test with $p < 0.005$, $d = 0.50$ and $N = 195$ (93 women and 102 men). Thirty-nine percent of male participants answered 'Yes' but only 17% of women answered so.

Replication Results: SM

As before, we excluded from analysis participants who had taken one or more philosophy courses. The data yielded no statistically significant difference among women and men. $N = 101$ (49 Male, 52 Female), Fisher's exact test yielded $p = 0.518$ (one cell had expected count < 5).

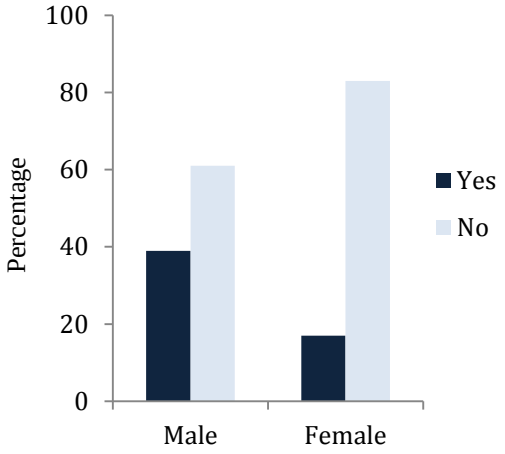

Figure 2.7a: Physicalism – Original Results

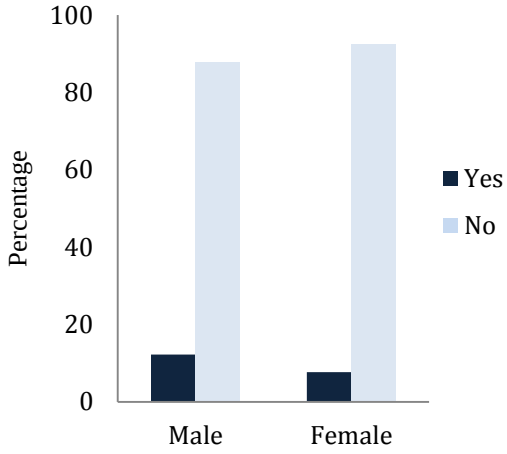

Figure 2.7b: Physicalism – Replication Results (SM)

Dualism

The dualism scenario Holtzman presented to participants reads as follows.

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

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

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

Replication Results: SM

A chi-square test for 137 participants (65 Male, 72 Female) yielded χ^2 = 0.090, p = 0.764.

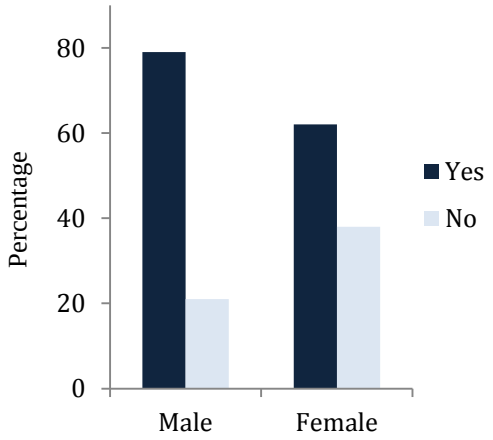

Figure 2.8a: Dualism – Original Results

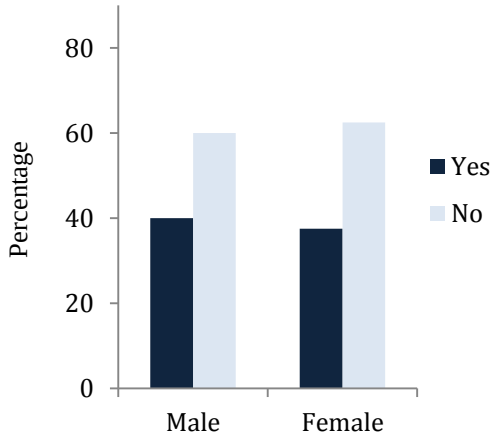

Figure 2.8b: Dualism – Replication Results (SM)

Statistical Power

The power achieved by Holtzman for the Compatibilism, Physicalism, and Dualism cases was 0.97, 0.92, and 0.69, respectively. Taking the sample proportions from Holtzman as the population proportions, the replication samples attained power values of 0.72, 0.65, and 0.56 in the same order. Our samples for all three scenarios were smaller than those of Holtzman. It may be that our data did not provide the necessary power to detect a difference. However, since we shared our results with other researchers, we have been informed that others have also been unable to replicate these outcomes (see thesis

conclusion). In this context, we believe the results of this section to be meaningful, despite the relatively low power achieved.

2.3.2: Some Philosophy Background but Cases Not Seen Before

We ran a similar analysis as in section 2.2.3 where we filtered for respondents who had taken one to three philosophy courses but who indicated that they had not seen the scenarios before (and whose native language was English).

Once again there was no significant difference between women and men on any of the three scenarios though the samples for the Compatibilism and Physicalism cases were relatively small after filtering. We will omit the power analysis because of the small sample sizes. See summary results below.

Compatibilism

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

Physicalism

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

Dualism

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

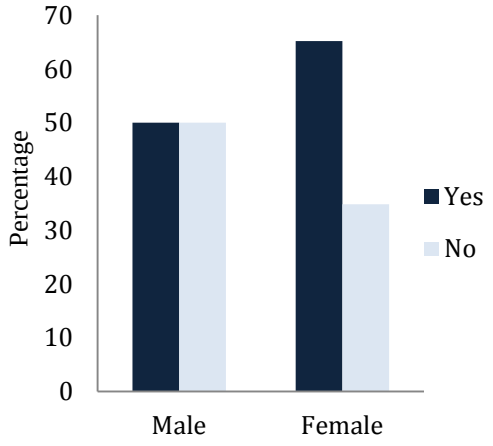

Figure 2.9a: Compatibilism – One to Three Philosophy Courses (SM)

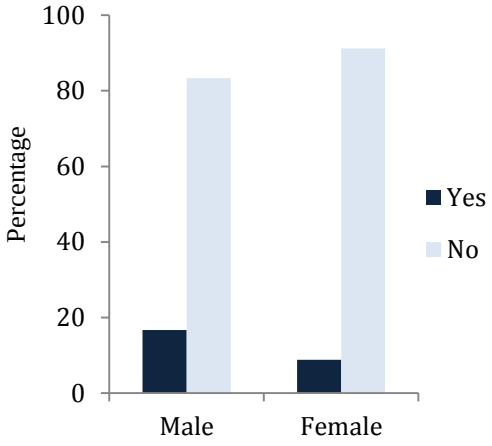

Figure 2.9b: Physicalism – One to Three Philosophy Courses (SM)

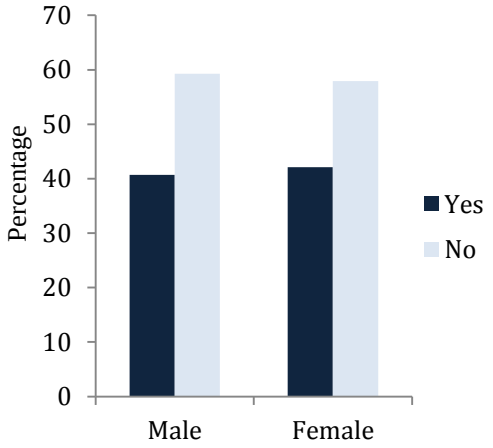

Figure 2.9c: Dualism – One to Three Philosophy Courses (SM)

2.4. Epistemic Intuitions

In section 3.1 of their paper, Buckwalter & Stich present data for experiments conducted by Starmans & Friedman on Gettier-style cases (Starmans & Friedman, 2009). Although we did not collect data on the exact same scenarios, we had conducted surveys on four other Gettier-type questions for a different study (Paper 1). We did not find significant differences among women and men in these experiments. We were interested in examining the exact cases that Starmans & Friedman used; however, upon contacting the authors, we were told that the authors themselves were unable to replicate the outcomes in further studies[32] and hence we did not see a need to carry out direct replications. In addition to Starmans & Friedman's own failed attempt to replicate their experiments, the results presented below offer strong evidence that women and men do not have different epistemic intuitions.

2.4.1: Procedures

For this section we collected data mainly through SurveyMonkey. However, for one of the scenarios (Car) we also collected data in classes at the LSE and online through Harvard University's Moral Sense Test (MST) website.[33] The procedures and methods for data collection for the SurveyMonkey samples were the same as described in Section 2.2.1.

[32] Personal correspondence with Ori Friedman, 5/1/2012.
[33] http://moral.wjh.harvard.edu/index2.html

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

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

2.4.2: Results

Original Results

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

> Peter is in his locked apartment, and is reading. He decides to have a shower. He puts his book down on the coffee table. Then he takes off his watch, and also puts it on the coffee table. Then he goes into the bathroom. As Peter's shower begins, a burglar silently breaks into Peter's apartment. The burglar takes Peter's watch, puts a cheap plastic watch in its place, and then leaves. Peter has only been in the shower for two minutes, and he did not hear anything.
>
> Does Peter really know that there is a watch on the table, or does he only believe it? [34]

[34] Taken from Buckwalter & Stich (2013).

The answer choices available were 'really knows' and 'only believes'. Starmans & Friedman report that whereas 71% of women chose 'really knows' only 41% of men chose this answer ($p < 0.05$, Fisher's exact test).[35]

Starmans & Friedman ran a variation on the above scenario where they changed the gender of the protagonist to female out of concern that this detail may have had an effect on responses and again attained a significant difference with $p < 0.01$ for $N = 112$ (54 men and 58 women); 75% of women answered 'really knows' and only 36% of men answered so. All participants in these experiments were reported to be native English speakers; for further details, see Buckwalter & Stich (2013) and Starmans & Friedman (2009).

Replication Scenarios[36] and Results

Car (SM)

The first scenario we examined was the following.

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

[35] Buckwalter & Stich do not provide the sample size for this experiment.
[36] All Gettier-style scenarios in this section were taken from Weinberg, Nichols & Stich (2001).

For our sample of 105 individuals (54 Male, 51 Female), a chi-square test yielded χ^2 = 0.108, p = 0.742; (minimum expected count 10.9).

Truetemp (SM)

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

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

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

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

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

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

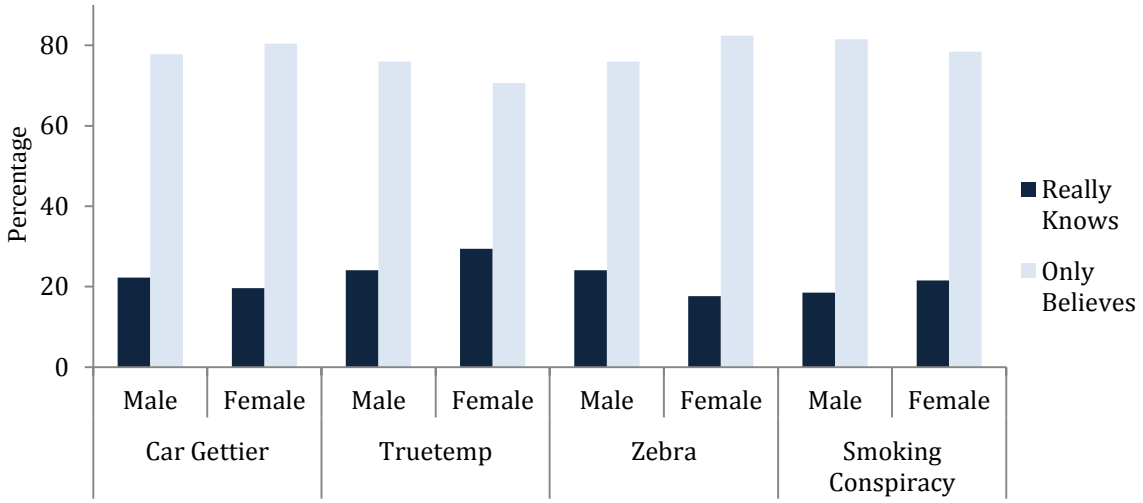

Figure 2.10: Epistemic Intuitions – SM

In addition to the tests above where the only filter used was for English as native language, we also ran two further analyses. In one, we further filtered out respondents who were not of a Western background as there has been a question whether individuals from Western and non-Western backgrounds answer these scenarios differently (Weinberg et al., 2001). Furthermore, in addition to native language and ethnic background filters, we also filtered out individuals whose highest level of education attained was below college. Again, this is because there has been a question whether individuals from different socioeconomic statuses (measured by an education proxy) answer Gettier-type questions differently (Weinberg et al., 2001). None of the tests yielded a significant difference among men and women.

In-Class and MST Data: Car Case

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

In-Class Car Case Results

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

MST Car Case Results

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

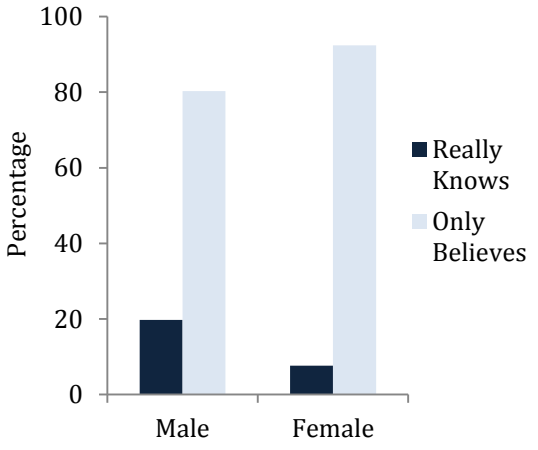

Figure 2.11a: Car Case – In Class

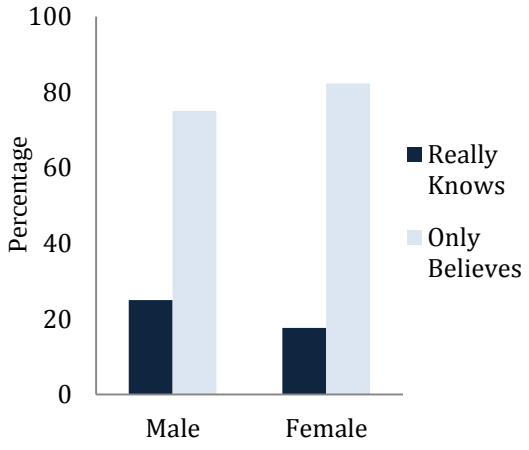

Figure 2.11b: Car Case – MST

Statistical Power

When taking the results of the first run from Starmans & Friedman as the population proportions, the power values attained are as follows. Car Case: 0.85; Truetemp: 0.85; Zebra; 0.85; Smoking Conspiracy: 0.85; In-Class: 0.93 and MST: 0.73. When taking the outcomes of the second procedure from Starmans & Friedman as the population ratios, we attained the following power values. Car Case: 0.98; Truetemp: 0.98; Zebra; 0.98; Smoking Conspiracy: 0.98; In-Class: 0.99 and MST: 0.93. These power values are open to discussion because of the nature of the original experiment; however, they do serve as point of comparison to the original outcomes. These values, together with the fact that Starmans & Friedman themselves could not successfully replicate their experiments, give a very strong indication that women and men do not have different epistemic intuitions on these Gettier-type scenarios.

Miscellaneous Points

There were several other scenarios for which we had collected data throughout the past years and which we examined for differences between women and men that also did not yield any statistical differences. Examples include other compatibilism thought experiments and scenarios testing semantic intuitions; however, for the sake of brevity we will omit a formal discussion and restrict this paper to the cases that were presented by Buckwalter & Stich.

2.5: Discussion and Concluding Remarks

Initially, we were very surprised by the outcomes we attained for this paper. We are not too sure about the reasons for the different outcomes in our experiments and those reported in Buckwalter & Stich (2013) and it is likely that there are different reasons for different studies. A general point that may be worth making is that to some extent the failures of replication presented here are not completely unexpected for the following reason. Experimental philosophy is often described as the study of philosophical questions using the tools of experimental psychology (Alexander, 2012; Knobe & Nichols, 2008; Nadelhoffer & Nahmias, 2007). By importing the methods of experimental psychology, philosophers will likely import the problems of that field and one of the problems that has afflicted experimental psychology for some time is the (likely) high rate of false positive results in the published literature (Bakker et al., 2012; Ioannidis, 2005, 2012; Pashler & Harris, 2012). John Ioannidis has contended that in "several fields of investigation, including many areas of psychological science, perpetuated and unchallenged fallacies may comprise the majority of the circulating evidence" (Ioannidis, 2012, p. 645). Concrete evidence comes from the Reproducibility Project which is an initiative that has set out to test the robustness of findings published in influential psychology journals. At the time of writing, roughly half the articles that were investigated could not be replicated successfully.[37]

The main reasons (although not an exhaustive list) for this low rate of reproducibility are publication bias (aversion to publishing negative results) (Bakan, 1966; Bakker et al.,

[37] http://opensciencframework.org/project/EZcUj/wiki/home

2012; Bozarth & Roberts, 1972; Coursol & Wagner, 1986; Sterling, 1959; Sterling, Rosenbaum, & Weinkam, 1995), aversion by journals to publishing replications of previously reported effects (Bozarth & Roberts, 1972; Makel et al., 2012; Neuliep & Crandall, 1991; Nosek et al., 2012), questionable research practices (QRPs) (Fanelli, 2009; John et al., 2012; Martinson et al., 2005), and the incentive structures of the current research environment (Bakker et al., 2012).

With regard to the specifics, studies that have examined the proportion of positive results of null hypothesis significance testing (NHST) in the published literature have found that between 94% to 97% of articles report positive results (Bozarth & Roberts, 1972; Sterling, 1959; Sterling et al., 1995). These numbers strongly indicate that published results in psychology are not representative of all experiments conducted and in fact about two-thirds of studies approved by institutional review boards (IRBs) and completed go unpublished according to one study (H. Cooper, DeNeve, & Charlton, 1997).

In addition to publication bias, the numbers of direct replications that could serve as a check against false positives are remarkably low. Bozarth & Roberts examined roughly one thousand articles published between 1967 and 1970 and found that less than one percent were replications of previous findings (Bozarth & Roberts, 1972). More recently, Makel and colleagues examined the top 100 psychology journals (according to a five-year impact factor) from the year 1900 onward and found that 1.07% of articles were replications. Of these only 14.0% were direct replications. Close to 82% were conceptual replications and 4.1% contained elements of both (Makel et al., 2012).

The problem that publication bias and lack of replications lead to was summarized very coherently by Sterling already in 1959.

> There is some evidence that in fields where statistical tests of significance are used, research which yields nonsignificant results is not published. Such research being unknown to other investigators may be repeated independently until eventually by chance a significant result occurs–an "error of the first kind"–and is published. Significant results published in these fields are seldom verified by independent replication. The possibility thus arises that the literature of such a field consists in substantial part of false conclusions resulting from errors of the first kind in statistical tests of significance. (Sterling, 1959, p. 30)

The issues of QRPs and structural incentives only exacerbate this problem further (Bakker et al., 2012). As mentioned before, given the adoption by philosophers of the methods of experimental psychology, the emergence of false positives in this new field should not be completely unexpected.

Aside from these general points, any attempts at explaining the different outcomes between the original and replication studies would involve speculation and we would like to keep this to a minimum. We have not seen layouts of the surveys that Buckwalter & Stich and Holtzman ran and hence cannot comment on any differences in presentation that may have led to different outcomes. However, a general problem with Buckwalter & Stich's approach is that the authors asked many researchers to examine their data and naturally those who happened to have differences in their data responded. Others have pointed this out and although true, this explanation is obviously not a satisfactory one for the classical scenarios, as Buckwalter & Stich collected the data themselves. A possible

reason for the difference for the Mechanical Turk experiments could be that depending on when the HITs were published, female and male respondents could have had different motivations for filling out surveys. For example, after working hours women may predominantly complete Mechanical Turk HITs for an alternative source of income and men may complete HITs to pass time, or vice versa. However, given that we collected data through several sources, which yielded similar outcomes, this also may not be a satisfactory explanation.

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

Our main aim for this paper was to test the robustness of the findings in Buckwalter & Stich (2013) and to share our results with other researchers, especially those who may want to build on the reported findings. We believe that we have provided strong evidence that women and men do not differ significantly in their intuitions on the cases examined in this study. Given that Buckwalter & Stich (2013) has already been widely circulated, we hope that this paper can correct some of the misconceptions that may have spread as a result and that readers at the very least view the original findings with some caution. Naturally, we do not believe that our data gives a definitive answer on whether

women and men have different intuitions on the cases examined in Buckwalter & Stich (2013). However, we do hope that our findings will encourage other researchers to carry out independent replications in order to attain a better picture. The importance of the subject matter certainly merits further investigation.

Page Intentionally Left Blank

Paper 3: Instability of Moral Intuitions

Abstract

We examined two papers from the psychology literature that have attracted the attention of experimental philosophers because the findings show moral intuitions to be unstable; hence casting doubt on the usefulness of intuitions for conducting philosophy. These papers are Valdesolo & DeSteno (2006) titled "Manipulations of Emotional Context Shape Moral Judgment" and Zhong, Strejcek, & Sivanathan (2010), titled "A Clean Self Can Render Harsh Moral Judgment."

Both papers report to manipulate moral judgments with relative ease. Valdesolo & DeSteno (2006) report that by merely showing a five-minute comedy video, individuals were significantly more likely to judge it appropriate to sacrifice an innocent bystander to save the lives of others in the footbridge trolley dilemma. Zhong et al. (2010) report that a simple cleanliness prime induced individuals to make harsher moral judgments on a host of social and moral issues.

These findings have several implications; however, the immediate one for philosophers is that intuitions are malleable in ways not obvious to them. That is, when considering moral scenarios such as the trolley dilemma, the intuitions that philosophers have may be distorted by situational factors. Furthermore, intuitions may differ for the same person at different times, depending on slight changes in circumstances.

We attempted to reproduce these findings; however, our replication attempts were without success. This paper is divided in two parts. Part One examines Zhong et al. (2010) and Part Two examines Valdesolo & DeSteno (2006). Each part is self-contained and can be read in isolation from the other without loss of understanding.

Part One: Cleanliness and Moral Judgments

In a paper published in 2010 titled "A Clean Self Can Render Harsh Moral Judgment," Zhong, Strejcek & Sivanathan report that an induced sense of cleanliness makes people's moral judgments on a host of issues harsher than they otherwise would be. Zhong et al. (2010) carried out two relatively simple manipulations. In one experiment participants were asked to cleanse their hands using hand wipes. In another experiment participants were asked to visualize a situation where they found themselves in a clean and pristine condition. Zhong and colleagues report that subjects in these experiments judged morally and/or socially contested issues such as abortion, use of drugs, or pollution significantly more harshly than their counterparts in control conditions.

Because of the surprise factor of the findings, this paper has received attention from various widely-read websites and blogs such as *The Chronicle of Higher Education* and *Wired* with attention grabbing headlines proclaiming that "A clean self is morally obnoxious self" (sic) and "Cleanliness Is Next to Priggishness" (Bartlett, 2010; Jarrett, 2010; Singh, 2012; Solon, 2010). The article has also been featured in the 2011 edition of *Issues in Experimental Psychology* (Zhong, Strejcek, & Sivanathan, 2011).

Aside from attention from popular outlets, these results have also come to the attention of experimental philosophers as further evidence of the instability of intuitions, this time in the moral domain. The argument goes that if intuitive judgments on moral or social issues as contested as abortion depend on the cleanliness of the reader's hands – a

variable that should not factor in judgments on important issues – then the use of intuitions in philosophy should be viewed with skepticism (Stich, 2010).

As mentioned, we could not reproduce these findings. We will proceed by first providing a background to the work of Zhong et al. (2010) in Section 3.1.1. In Section 3.1.2 we present the results of the original and replication studies and Section 3.1.3 concludes this part of the paper.

3.1.1: Background

In setting up their hypothesis, Zhong et al. (2010) draw on previous work demonstrating a connection between physical cleanliness and morality (Zhong & Liljenquist, 2006).[38]

Zhong & Liljenquist (2006)

In one of their experiments, Zhong & Liljenquist asked subjects to visualize a fictional story in which they committed an unethical act. Subsequent to the visualization task, these subjects, according to the authors, showed a significantly stronger preference for cleansing products such as a soap or toothpaste. Zhong & Liljenquist suggest that the act of physical cleansing is a substitute for moral purification.

In another experiment, participants were asked to recall an unethical act they had committed in the past. Following this recall, participants in the experimental condition

[38] The main author of Zhong et al. (2010) was involved in this previous work. This may be of relevance, as discussed in Section 3.1.3.

were given a hand wipe to cleanse their hands, whereas participants in the control condition skipped this step. To test whether cleansing had restored participants' moral self-image, toward the end of the study an experimenter asked whether subjects were willing to take part in another study without compensation. This study, participants were told, was for a graduate student who was desperate to find subjects. Zhong & Liljenquist report that those in the experimental condition (those who cleansed their hands) were significantly less likely to volunteer than participants who had not cleansed their hands. The authors reason that those in the control condition chose to volunteer as a means to restore their moral self-image. Participants in the experimental condition saw no need to volunteer as their self-image had been restored by cleaning their hands.

Overall, Zhong & Liljenquist show in four different experiments that when participants' moral self-image is threatened, they seek acts of physical cleansing as a proxy for clearing their moral sense of self.

Zhong et al. (2010)

Following this approach, Zhong et al. (2010) hypothesize that induced physical cleanliness will increase subjects' sense of moral virtue and this in turn will lead to harsher moral judgments. In the authors' own words, "given the association between cleanliness and moral purity, we suggest that a clean person may not only feel dirt-free, but also morally untainted" and "this elevated sense of moral self can in turn license severe moral judgment" (Zhong et al., 2010, p. 859).

Zhong et al. argue that a connection between physical purity and moral superiority can be observed in real world cases. The examples the authors mention are India's caste system and Nazi Germany's "obsession with hygiene" and the portrayal of targeted groups "as not only physically filthy but morally corrupt" (Zhong et al., 2010, p. 859). In Zhong et al.'s view, these examples are not merely coincidental but rather reflect an underlying psychological connection between moral and physical purity (Zhong et al., 2010).

The authors highlight the importance of their findings in noting that "these results provide unique insight to the social significance of cleanliness and may have important implications for discrimination and prejudice." Furthermore, so the authors, if "members of a ''clean" society perceive those who are different as less moral, then separating and segregating them is more easily justified. This may be part of the mechanism behind the caste system or other more extreme forms of social cleansing" (Zhong et al., 2010, p. 859 and 861).

To summarize, there are two components to Zhong et al.'s hypothesis. The first is that individuals with an elevated sense of moral self make harsher judgments on social and/or moral issues. The second component is that acts of cleansing elevate individuals' sense of moral self.

3.1.2: Results

Zhong et al. (2010) report in three experiments that cleanliness leads to harsher moral judgments. In what follows, we will introduce each experiment separately, provide the results of the original paper and contrast these with our findings.

Experiment 1

Procedures

In the first experiment, participants were randomly assigned either to the experiment (clean) or control condition. For the cleanliness condition, Zhong et al. advised participants that because the equipment in the computer lab was new, participants were required to clean their hands with hand wipes before starting the study. In the control condition this instruction was omitted. The equipment in the lab was in fact new. Participants were asked to judge how moral or immoral the following issues were: adultery, littering, pornography, profane language, smoking, and using drugs. Judgments were recorded on an eleven point scale where the leftmost option was marked "-5 (very immoral)" the midpoint "0" and the rightmost "5 (very moral)." The total sample consisted of 58 individuals who received $5 for their participation.

Our procedures were very similar to those of Zhong et al. (2010). As it happened, the lab where we conducted our experiments – the Behavior Research Lab (BRL) at the London School of Economics (LSE) – had just opened and all equipment and everything else was new. We instructed participants as Zhong et al. did. Fifty-nine participants took part in our study for which they received 5 pounds sterling. We presented the same six

moral/social issues in randomized order and participants judged these on the same elven-point scale that Zhong et al. (2010) used.

Results

For their analysis, Zhong et al. construct a composite score by averaging[39] the judgments on the six issues ($\alpha = 0.77$).[40] The authors report that, as hypothesized, individuals who had cleansed their hands made harsher moral judgments. The experimental group's mean score was -2.62 (*SD* = 1.30) and the control group's mean was -1.85 (*SD* = 1.46). Comparing the control and experimental groups using an independent-samples t-test, the authors report $t(56) = 2.10$, $p = 0.04$.

For our replication data we carried out the same analysis by averaging the six judgments ($\alpha = 0.77$). The cleansing condition yielded $M = -1.70$ (*SD* = 1.79) and the control condition yielded $M = -1.82$ (*SD* = 1.12). An independent-samples t-test produced $t(57) = -0.28$, $p = 0.80$.[41]

There are several things to point out. The first is that the judgments in our control condition were actually harsher than in the clean condition. Although we did not detect a significant difference, the direction of responses was opposite to that of Zhong and colleagues. Second, the judgments in both conditions of the original study were harsher than either condition of the replication. Finally, the variability in judgments for our data

[39] Zhong et al. mistakenly write that they take the sum of the categories instead of the average.
[40] Cronbach's alpha (typically abbreviated as α) is a measure of 'internal consistency' of variables. The more correlated several variables are, the higher the value of α. Aggregating several variables into a single measure is considered acceptable for α greater than 0.70.
[41] Throughout this paper, the reported *p* values are two-sided for the original as well as the replication tests.

was comparable to that of the original experiments. Although the standard deviation in our clean condition (1.79) was higher than in the original study (1.30), the standard deviation of our control condition (1.12) was lower than either condition of the original study.

Experiment 2

Procedures

In the second experiment, Zhong et al. used a different prime. Participants were asked to visualize a short paragraph and copy the text in a field on their computers. Participants were told that they would be asked to recall details of the visualized scenario after some unrelated tasks. The paragraphs are presented below. There was also a control condition where participants skipped the visualization task.

Clean

> My hair feels clean and light. My breath is fresh. My clothes are pristine and like new. My fingernails are freshly clipped and groomed and my shoes are spotless. I feel so clean.

Dirty

> My hair feels oily and heavy. My breath stinks. I can see oil stains and dirt all over my clothes. My fingernails are encrusted with dirt and my shoes are covered in mud. I feel so dirty.

For this experiment, Zhong and colleagues wanted to test a wider range of issues and so in addition to the topics surveyed in Experiment 1, the authors added the following ten: abortion, alcoholic, casual sex, wearing animal fur, homosexuality, masturbation, obesity, pollution, premarital sex, and prostitution. The scale used was the same as in the first experiment.

Zhong et al. drew their sample from a US database[42] of 15,000 participants who had registered to take part in studies: 323 individuals participated.

Our replication procedures were modeled very closely on the original study. The main difference was that we used the Moral Sense Test (MST) website run by the Cognitive Evolution Laboratory at Harvard University to conduct our study.[43] Participants (mostly from the US) visited the website without being solicited and after a brief introduction started the visualization task.

Results

As in their first experiment, Zhong et al. construct a composite variable of the 16 moral/social issues by taking their average (α = 0.88). The outcomes are presented in summary form below, together with the replication results. Cronbach's alpha for the 16 issues in the replication study was 0.89.

[42] The authors do not specify which database.
[43] http://moral.wjh.harvard.edu/

Experiment	Condition	*N*	Mean	*SD*
Original	Clean	323	-1.76	1.13
	Dirty		-1.42	1.14
	Control		-1.49	1.55
Replication	Clean	180	-1.05	1.16
	Dirty		-0.71	1.48
	Control		-1.04	1.72

Table 3.1.1: Results for Experiment 2

Zhong et al. report a statistically significant difference between the clean and dirty conditions with $t(320) = -2.02$, $p = 0.045$ but no difference between the dirty and control conditions with $t(320) = 0.42$, $p = 0.675$.[44]

For the replication study an independent-samples t-test yielded no significant difference between the clean and dirty conditions with $t(93) = -1.24$, $p = 0.22$, nor between the dirty and control conditions with $t(124) = 1.06$, $p = 0.29$. We discuss sample sizes and statistical power for all three experiments in Section 3.1.3.

What stands out again with these results is that the least severe judgment of the original study was harsher than the harshest value of the replication study. Another detail worth noting is that this time the clean condition of the replication study did produce harsher judgments than the dirty condition (though not statistically significant). However, the value of the clean condition was almost identical to the control condition. These outcomes are likely simply due to random variation; the data in Experiment 3 substantiates this.

[44] There seems to be a mistake with the degree of freedom the authors report, as they indicate 320 for both tests, comparing clean and dirty and dirty and control. The overall sample size is given as 323.

Experiment 3

Procedures

The procedures of Experiment 3 were identical to Experiment 2 with the exception that after being primed (same visualization prime as in Experiment 2), participants were asked to rank themselves on the following eight characteristics: Sense of Humor, Intelligence, Moral Character, Creativity, Physical Attractiveness, Fitness, Social Sensitivity, and Leadership. Participants were asked to rank themselves compared to others, where 0 denoted "worse than all others" and 100 denoted "better than all others." After ranking themselves, the survey proceeded as in Experiment 2 with participants evaluating the same 16 moral/social issues as in Experiment 2.

Results

Zhong et al. surveyed 136 individuals and for the self-ranking task report an effect of cleanliness prime on Moral Character; in specific, cleanliness prime yielded mean value $M = 80.44$ ($SD = 15.24$) whereas dirty prime produced $M = 75.03$ ($SD = 15.70$). An independent-samples t-test is reported with $t(134) = 2.03$, $p = 0.045$. Zhong et al. did not collect data for a control condition as Dirty and Control yielded no significant difference in Experiment 2.

The replication experiments did not reproduce these findings. For a sample of 166 participants, priming had no effect on Moral Character ratings. For the clean condition, we attained a mean value for Moral Character of 72.81 ($SD = 15.36$) and a mean value of

70.18 (*SD* = 18.60) in the dirty condition. An independent-samples t-test produced $t(164) = 0.99$, $p = 0.32$.

Zhong et al. report that none of the other measures on which participants ranked themselves yielded a difference between clean and dirty conditions ($|t_s| < .78$, $p_s > .40$). We attained a similar outcome ($|t_s| < 1.53$, $p_s > 0.13$).[45]

For Experiment 3, Zhong et al. once again report an effect of cleanliness on moral judgments. Their findings, together with our results, are presented in summary form below.[46]

In all tables, * denotes $p < 0.05$.

Experiment	Condition	*N*	Mean	*SD*	*t*	*p* value
Original	Clean	136	-2.04	1.28	-2.13	0.04*
	Dirty		-1.59	1.16		
Replication	Clean	166	-0.96	1.12	0.68	0.50
	Dirty		-1.09	1.18		

Table 3.1.2: Results for Experiment 3

Once again, the judgments of the original study are both harsher than the least harsh judgment in the replication study. The original experiments were conducted in Toronto, whereas the replications of Experiment 1 and 3 were carried out in London (at the BRL). The difference in location is unlikely to be the reason for this difference, as data for

[45] We are reporting the greatest absolute *t* value and the smallest *p* value attained.
[46] Cronbach's alpha for our sixteen categories was 0.84. Zhong et al. do not report this value.

Experiment 2 was collected in the U.S. for both the original and replication studies; here, again, the same pattern is repeated.

Once again, not only did we not attain a statistically significant effect of cleanliness on judgments but in our data the direction of this effect was the reverse of Zhong et al.'s. The difference between dirty and clean prime was 0.124 (0.173 when taken in combination with moral character) on the average moral judgment. That is, changing the prime from clean to dirty made judgments harsher by 0.124 (0.173) points on the 11-point scale.

The next analysis Zhong et al. carry out is an ordinary least squares (OLS) regression including moral self-image (Moral Character ratings) and cleanliness prime as independent variables and find that when both are included in the analysis, only moral self-image explained moral judgments (B = -0.018, SE = 0.007, t = -2.73, p = 0.007) but cleanliness prime did not (B = -0.348, SE = 0.208, t = -1.67, p = 0.097).[47] We will provide an explanation of these results below. We attained a similar outcome. In our data, moral character predicted moral judgments (B = -0.019, SE = 0.005, t = 3.61, p = 0.000) but cleanliness did not (B = -0.173, SE = 0.177, t = -0.98, p = 0.327).

Finally, Zhong et al. present data from 1000 bootstrap resamples and arrive at the same conclusion. When mediated for the effect of cleanliness on moral self-image, moral self-image had a significant effect on judgments at 95% confidence level (confidence interval

[47] The standard error is mistakenly given as 208 instead of 0.208 in Zhong et al. (2010).

-0.24 and -0.01). We attained a similar outcome for our replication data with 1000 resamples (confidence interval -0.03 and -0.01).

In addition to the analyses that Zhong et al. provide, we want to present some further details to fill in the picture more on the relationships just discussed.

In an OLS regression with two independent variables, say A and B, when the significance test determines that A did not significantly predict the variability of the dependent variable, this means that in combination with variable B, A does not significantly predict variability of the dependent variable. That is, if B's effect is very strong that comparatively A does not predict much of the variability, A will be determined to be an insignificant predictor. However, since this could be due to the strength of variable B (relative to A), it is not so much telling about A in itself but rather about A in combination with B.

In our case, although in an OLS regression cleanliness prime in combination with moral self-image (Moral Character) did not significantly predict variability of judgments, this does not mean that cleanliness did not significantly predict the dependent variable. To get a better picture, we ran a correlation analysis to test for the effect of cleanliness on judgments in the absence of moral self-image and still cleanliness did not predict judgments significantly ($p = 0.32$). Cleanliness prime by itself only predicted 0.3% of the variability in moral judgments. When including cleanliness and moral character as independent variables, R^2 value equals 0.077 (adjusted R^2 value is 0.065); that is, only

about 8% of the variation in moral judgments can be predicted by the combination of cleanliness prime and moral character.

In Section 3.1.1 we highlighted that there were two components to Zhong et al.'s hypothesis. One was that individuals with an elevated sense of moral self make harsher judgments and the second was that acts of cleansing lead to an elevated sense of moral self. While our data confirms the former, our experiments do not reproduce the latter effect. That is, cleanliness prime had no effect on either moral self-image or moralization (moral judgments). On the other hand, our data confirms that individuals who consider themselves as having an elevated moral character compared to others (regardless of cleanliness prime), make harsher judgments on the social and/or moral issues surveyed in this paper.

3.1.3: Concluding Remarks

Zhong et al. (2010) report in three experiments an effect of cleanliness on moralization. We failed to reproduce all three experiments. In the beginning of this paper we referred to Zhong & Liljenquist (2006) as the basis on which Zhong et al. (2010) built their hypothesis and we pointed out that Zhong was the main author of both papers. Two independent groups had previously attempted to reproduce two of the experiments in Zhong & Liljenquist (2006), but failed to do so (Fayard, Bassi, Bernstein, & Roberts, 2009). In the final remarks to their failed replication, Fayard et al. (2009) conclude.

> Contrary to our expectations, neither Study 1 nor Study 2 replicated Zhong and Liljenquist's (2006) finding that physical cleansing, specifically washing one's hands, contributes to the absolution of guilt. Participants who recalled an unethical deed in Study 1 were no more likely than participants who recalled an ethical deed to choose the antibacterial hand wipe as a free gift, indicating that moral emotions may not induce people to cleanse themselves as a reparative strategy. Furthermore, as Study 2 showed, cleansing did not reduce moral emotions such as guilt in participants who recalled unethical deeds, and it did not significantly reduce volunteerism among participants. (Fayard et al., 2009, p. 27)

These failed replication attempts may point to procedural flaws the main author may have had in conducting the experiments discussed here. It should, nevertheless, also be noted that several other papers have reported a connection between physical cleanliness and morality (Cramwinckel, De Cremer, & van Dijke, 2013; Cramwinckel, van Dijk, Scheepers, & van den Bos, 2013; Gollwitzer & Melzer, 2012; Helzer & Pizarro, 2011; Lee & Schwarz, 2010, 2011). Having noted these studies (failure of replication as well as other reports of the cleanliness-morality connection), we want to point out some of the shortcomings of our study.

Experiment 1 involved the use of hand wipes and we did not use the same brand and type as Zhong et al. (2010). We carried out our experiments in the UK and did not have access to North American brands that Zhong and colleagues used. This may or may not pose a problem. The hand wipes we used, even the supposed non-scented ones, carried a scent and this could have influenced participants. Researchers have reported that scents associated with cleanliness influence moral judgments (Liljenquist, Zhong, & Galinsky, 2010; Tobia et al., 2013). We tried to minimize this effect by wiping the desks and equipment participants used before experimental runs so that any effect of scent that participants in the experimental condition had would also be present in the control

condition. The study for Experiment 3 was conducted following a survey where participants answered questions from various fields of philosophy and completed different kinds of tasks. This could have had an effect on participants.

Statistical Power

Taking the effect sizes of the original study as estimates of the population effect sizes, we calculate power values for Experiments 1, 2, 3, and effect of cleanliness on self-ranking of Moral Character as presented in Table 3.1.3 below.

Experiment	**Power**	
	Zhong et al.	Replication
1	0.55	0.56
2	0.59	0.30
3	0.68	0.72
3 (Moral Char.)	0.58	0.59

Table 3.1.3: Statistical Power

The statistical power for the replication of Experiment 2 was well below that of the original study and these replication results should be considered with caution. However, Experiment 2 was identical to Experiment 3 with the exception of the self-ranking and here the replication had the highest statistical power of any of the studies (original and replication) and our results still point to null findings.

In three of the four studies the replication attempts had greater power than the original studies, yet we still did not detect a statistical difference. However, with the exception of

the replication of Experiment 3 (effect of cleanliness on judgment), none of the studies attained statistical power close to the 0.80 convention.

For our replication we aimed to attain larger sample sizes than the original study and we did this in three of the four studies; however, for a more conclusive result we would have attained larger samples to realize power above 0.80. Although, ideally our statistical power would have been closer to the 0.80 convention, Zhong et al.'s studies are below that mark as well. In fact, given the power of the four original experiments, even if an effect existed in all four cases, Zhong et al. would be expected to detect all effects in only about 13% of cases (product of statistical power of each experiment). Either, Zhong et al. were somewhat lucky or there may be studies that the authors do not report (see Paper 4 for a discussion of this practice), or something else may be the case.

In conclusion to this part of our paper, we would like to quote Fayard et al. (2009) whom we mentioned earlier as the authors who attempted to replicate Zhong & Liljenquist (2006), without success. Their conclusion further captures our own sentiments why we thought the replications reported here to be worthwhile and why we decided to conduct a formal replication. Fayard et al. note.

> Given the notorious "file drawer phenomenon" in which researchers file away null results and non-replications instead of publishing their results, we cannot know how many others have also attempted and failed to replicate Zhong and Liljenquist's (2006) surprising results. Here we report two such attempts and failures, both conducted independently of one another for different reasons. Zhong and Liljenquist's (2006) results carry important theoretical implications, so it is important to publish failed replications such as these so that researchers can

have a clearer picture of the plausibility of research findings. (Fayard et al., 2009, p. 27)

Part Two: Affective State and Moral Judgments

In their paper titled "Manipulations of Emotional Context Shape Moral Judgment," Picarlo Valdesolo and David DeSteno report that participants' moral judgments on the footbridge trolley problem are relatively easily manipulated by merely showing a five-minute Saturday Night Live (SNL) video clip before presenting the moral dilemma.

These results have attained some attention with experimental philosophers because of the ease with which moral intuitions are manipulated. It is not that any intuitions are manipulated but intuitions on a question as grave as killing an innocent bystander by physically pushing him off a bridge in order to save the lives of others.

Some experimental philosophers have taken these findings as yet more evidence that intuitions are not reliable and so should be reduced to a minimum in philosophical practice (Carruthers, Stich, & Laurence, 2008; Stich, 2010). The argument is that if intuitive responses to such important questions are dependent on the affective state of the reader, then intuitions cannot be relied on in getting philosophers to the right answers.

Aside from interest from experimental philosophers, the paper has received wider attention for its implications (discussed in Section 3.2.1). At the time of writing, Valdesolo & DeSteno (2006) has received 276 citations[48] and at least one other paper has modeled its approach on it (Greene, Morelli, Lowenberg, Nystrom, & Cohen, 2008).

[48] http://scholar.google.com/scholar?cites=8009255644369793974&as_sdt=2005&sciodt=0,5&hl=en

We attempted to replicate this experiment using several different video clips; however, we did not succeed. Our findings suggest that individuals' moral judgments on an important question about sacrificing a bystander in order to save the lives of others are more stable than the findings of Valdesolo & DeSteno (2006) imply. We will proceed as follows. In Section 3.2.1 we provide a background to the work of Valdesolo & DeSteno (2006). Section 3.2.2 provides an overview of the experimental procedures of the original and the replication studies. Section 3.2.3 reports the results for mood indicators used in the study. Section 3.2.4 reports the judgments participants made on the moral dilemmas. Section 3.2.5 closes with concluding remarks.

3.2.1: Background

Valdesolo & DeSteno (2006) build on the work of Greene and colleagues (Greene, Nystrom, Engell, Darley, & Cohen, 2004; Greene, Sommerville, Nystrom, Darley, & Cohen, 2001) who studied brain activity in fMRI procedures while individuals evaluated moral dilemmas such as the footbridge trolley scenario.

Greene and colleagues set out to examine the psychological bases as to why people generally consider it acceptable to sacrifice one person to save the lives of five others in the standard trolley case but consider this outcome unacceptable in the footbridge dilemma. In both cases one person is sacrificed to save the lives of five. To take away Greene et al.'s conclusion, the short answer (although incomplete[49]) is that the

[49] Incomplete, because the activity of brain areas associated with processing emotions is more nuanced. See, Greene et al. (2004) for details.

"emotional response is likely to be the crucial difference between these two cases" (Greene et al., 2001, p. 2107).

Greene and colleagues present several findings that Valdesolo & DeSteno draw on. The first is that brain areas associated with emotions (Brodmann's Areas (BA) 9, 10, 31 and 39) showed heightened activity in personal moral dilemmas (e.g. footbridge trolley) as compared to impersonal moral (e.g. bystander trolley) and non-moral scenarios. On the other hand, brain areas associated with cognitive processes (dorsolateral prefrontal cortex (DLPFC)) that include abstract reasoning and problem solving showed heightened activity when considering impersonal moral scenarios.

Among participants who gave the deontological response in personal moral dilemmas such as the footbridge case, areas associated with social-emotional processing (BA) showed more activity when compared to participants who gave the utilitarian response. For these latter participants, areas of the brain associated with cognitive control (DLPFC) showed a comparatively increased activity.

Finally, when making the utilitarian judgment in personal moral dilemmas, participants showed longer reaction times (RT) as compared to when they made the deontological judgment. Greene and colleagues suggest that in order to arrive at the utilitarian response, brain regions responsible for abstract reasoning and cognitive control have to override the negative emotional response. In personal moral dilemmas where participants showed longer reaction times (this being indicative of more difficult moral cases), the anterior

cingulate cortex (ACC) – an area that is activated in cases of cognitive conflict – and the DLPFC showed heightened activity. As before, in cases involving longer reaction times, DLPFC activity was correlated with utilitarian judgments.

The overall picture that Greene et al. paint is that when presented with personal moral dilemmas such as the footbridge trolley case, individuals show heightened and likely competitive activity in different brain regions; one of these brain regions is associated with quick reflexive responses (emotional response) and the other associated with higher cognitive functions. The suggested explanation is that in the footbridge scenario a moral/social principle has to be violated by pushing the person off the bridge and this produces the negative reflexive emotional reaction. On the other hand, in order to arrive at the outcome that maximizes overall wellbeing (saving the five) this reflexive reaction has to be overcome and this step requires cognitive processing. As Greene et al. (2004) describe it, their "results suggest that emotional responses drive individuals to disapprove of personal moral violations […] cognitive control processes can override these emotional responses, favoring personal moral violations when the benefits sufficiently outweigh the costs" (Greene et al., 2004, p. 397).

To summarize the most relevant aspect of Greene et al.'s work for Valdesolo & DeSteno: different brain regions – cognitive and emotional – compete over the appropriate response to personal moral dilemmas. Areas responsible for emotional processing react negatively because of the moral violation that is involved in killing an individual to bring

about a greater good. This negative reaction must be overridden if the utilitarian judgment is to win out.

Valdesolo & DeSteno suggest an alternative path whereby the negative emotional response can be overridden and hence the utilitarian judgment can come about: and this without altering DLPFC activity. Valdesolo & DeSteno suggest to counteract the negative emotional reaction with a positive emotional induction. The importance of this step needs to be highlighted. Although the negative emotional response stems from a moral violation (i.e. pushing the man off the bridge), Valdesolo & DeSteno suggest that a positive emotional induction from a completely different source (an SNL clip) can override the initial negative response. If the negative reaction can be overcome or alleviated in this way, more individuals will make the utilitarian judgment.

And indeed, Valdesolo & DeSteno report that by showing a comedy clip before participants consider the moral dilemma, participants were significantly more likely to judge it appropriate to push the bystander off the bridge, sacrificing the bystander in order to save the lives of five others.

Valdesolo & DeSteno conclude.

> Environment-induced feelings of positivity at the time of judgment might reduce the perceived negativity, or aversion ‘‘signal,’’ of any potential moral violation and, thereby, increase utilitarian responding […] These findings demonstrate that the causal efficacy of emotion in guiding moral judgment does not reside solely in responses evoked by the considered dilemma, but also resides in the affective characteristics of the environment. […] What is clear, however, is that a skilled

> manipulation of individuals' affective states can shape their moral judgments. (Valdesolo & DeSteno, 2006, p. 476 and 477)

3.2.2: Experimental Procedures

Original Study

Although Valdesolo & DeSteno do note that they took the scenario from Thomson (Thomson, 1986), they do not specify exactly what the wording of the footbridge scenario was that they presented to participants; however, the outlines of the scenario are familiar by now. An out of control trolley is headed toward five individuals who do not realize they are in danger. If the trolley is not stopped, these five individuals will certainly die. The only way to stop the trolley is to throw a heavy weight in its path. You are standing on a footbridge under which the trolley will pass. As it happens, there is a large man standing next to you. The only way to stop the trolley is to push this man in front of the trolley, thereby killing him to save the five. The question participants were asked was whether it was appropriate or inappropriate to push the man off the bridge.

Valdesolo & DeSteno tested 79 individuals (38 in control condition). Participants were either shown a five-minute SNL clip in the experimental condition or a five-minute documentary on a Spanish village in the control condition. After watching the clip, participants were presented with the footbridge trolley dilemma, which was embedded in non-moral filler questions. The trolley scenario was presented in three parts and each part was shown on the screen for 15 seconds. This was to ensure that the induction of

positive emotions would not fade and the affective state of participants return to their baseline levels if participants contemplated for too long.

After the trolley scenario, participants saw a mood indicator that was designed to capture affective states. Participants were asked to rate themselves on the variables of Happy, Content, Pleasant and Good on a seven-point scale, ranging from 1 to 7 where 1 was marked "not at all" (happy) and 7 marked "very" (happy).

Replication Study

We faced several difficulties in our replication attempts. The details of the experimental procedures in Valdesolo & DeSteno (2006) are very limited. An exact replication would require much more information than the authors provide. For example, the authors do not mention which video clips they used; they merely state that an SNL clip was used for the experimental condition and a documentary about a Spanish village for the control condition. People familiar with SNL know how wide in scope the sketches can be and how different sketches can be in eliciting one emotion over another. A documentary about a Spanish village may be idyllic and elicit some emotions. This was definitely our concern for our video about Spain as some great landmarks, tourist attractions, and a lot of sun were shown. Furthermore, the original paper does not state under what conditions the experiments were carried out and who the participants were. It would have been useful to know whether the experiment was conducted in a lab setting or whether it was carried out online. It is not indicated whether the participants were students or from the

general population. There are no details about the distractor questions. All of these make it very difficult to replicate the experiment in an exact manner.

We contacted the authors on five occasions; however, only received one response from one of the authors saying that the other author had the details and that we would need to contact him. We used email addresses that were current; both authors had recently signed up for a mailing list of a neighboring laboratory.

The solution we saw was to try several different clips and follow the details of experimental procedures in Greene et al. (2001) and (2004) since Valdesolo & DeSteno make frequent references to these articles.

Replication Procedures

We used two different moral dilemmas in our studies. Out of concern that over the years many individuals may have become familiar with the trolley dilemma, we used a different scenario for some of the studies. We will refer to these scenarios as the Train and Tiger[50] scenarios. The exact wording of these cases as presented to participants is shown below.

Train

> An out of control train is running down a track toward five people who will die if the train is not stopped. You are on a bridge under which the train will pass.

[50] We thank Donal Cahill for providing us with this scenario.

> You can stop the train by dropping a heavy weight in front of it. As it happens, there is a very large man next to you. Your only way to stop the train is to push the man off the bridge and onto the tracks, killing him to save the five.
>
> Pushing the man off the bridge to stop the train would be
>
> Appropriate Inappropriate

In one of the studies (we will point out which one) the question asked was "Would you push the man off the bridge?" and the answer choices where "Yes" or "No" in that order.

Tiger

> You are on a visit to a zoo when you see that the barrier between the tiger enclosure and the viewing deck falls over. There are currently five people standing on the deck and they will undoubtedly be killed by the tiger unless something happens. There is another person standing beside you. Your only option is to push the person into the enclosure. While the tiger is devouring that person, the five others will have time to escape.
>
> Would you push the person into the tiger enclosure?
>
> Yes No.

Data Sets

We collected data through two different sources. The descriptions of the data sets are as follows.

Data Set 1

For the first experimental runs we used the Moral Sense Test (MST) website. Participants visited the site without being solicited and started the survey after some initial instructions and explanations.

These surveys were pilot studies and somewhat exploratory. For these surveys we changed the order of the experimental procedure from Valdesolo & DeSteno (2006). We had some concerns about the mood values that Valdesolo & DeSteno report and wanted to investigate this issue first. Instead of showing the comedy clip first, then presenting the scenario, and subsequently presenting the mood indicator, we presented the mood indicator right after the video clip. We wanted to test what the affective states would be right after the video clip, that is, when in the main experimental runs participants would be answering the trolley question. Since our main focus here was on the mood indicator, we did not time the moral dilemmas to 15s per screen but allowed participants to answer in their own time.

We ran three different positive (comedy) video clips and a neutral one. Two of the comedy clips were from SNL and one was a standup comedy routine by a comedian who had been nominated for several awards.[51] One of the SNL clips was titled "Celebrity Jeopardy: Nicolas Cage, Calista Flockhart and Sean Connery" and the other SNL clip was titled "Marble Columns." The standup comedy routine was by Rhod Gilbert titled "Luggage." The control video clip was a five-minute documentary on Spain that showed some of the landmarks of the country.

[51] https://en.wikipedia.org/wiki/Rhod_Gilbert

We presented the Tiger scenario in some of the surveys in Data Set 1 and in all Train scenarios surveyed for this data set we asked, "Would you push the man off the bridge?" instead of asking about appropriateness.

Data Set 2

These experimental runs were much closer in procedure to Valdesolo & DeSteno (2006). We ran these experiments at the London School of Economics. Some of the participants signed up through the Behavioral Research Laboratory (BRL) and some responded to emails sent out by the philosophy department. We used the same sequence as Valdesolo & DeSteno, i.e. we first presented the video clip, then the moral dilemma (timed to 15s per screen) and finally the mood indicator. We only showed the Train scenario in this data set and asked about appropriateness as in the original paper. We had two positive videos and two neutral ones. One of the positive videos was an SNL clip titled "Celebrity Jeopardy: Rock Star Edition" (RSE) and the other clip consisted of two comedy sketches from the BBC One series "Come Fly with Me" edited together to show as one clip. These were "Penny's Royal Visit" followed by "Tommy's New Job." There was a screen in between the clips with the text 'Clip 2' showing for about two seconds. One of the neutral clips – titled "Material World" – was taken from a BBC documentary. The topic of the documentary was materials that exist in nature such as wood and silk. The other clip was about the effects of deforestation. This clip was somewhat dark, ending with prospects of mass extinction if deforestation was not stopped. We chose this video because none of the other clips yielded as low a mood rating as that reported by Valdesolo and DeSteno.

3.2.3: Results for Mood Indicators

Before we present the outcomes of the trolley dilemmas, we would like to discuss the mood indicators in some detail. The mean mood measure Valdesolo & DeSteno report for their control condition is $M = 2.77$ ($N = 38$) and for their experimental condition $M = 4.57$ ($N = 41$).

Data Set 1

The first experiments we ran were from Data Set 1 and we did not attain a mood rating as low as 2.77 in any of our experimental runs. In fact, we did not come anywhere close to this number.

In Data Set 1, the average mood rating for Celebrity Jeopardy, Marbleopolis, Luggage and the Spain Documentary were respectively, $M = 4.69$ ($SD = 1.44$; $N = 99$), $M = 4.52$ ($SD = 1.45$; $N = 38$), $M = 4.75$ ($SD = 1.46$; $N = 50$), $M = 4.56$ ($SD = 1.48$; $N = 63$)[52]. A one-way ANOVA comparing the means yielded no significant difference ($p = 0.29$).

These results may mean one of two things. Either the positive videos failed to increase mood values or the neutral video was not neutral but instead increased mood ratings.

Data Set 2

[52] In all instances in this paper where we averaged the ratings of the four variables (Happy, Content, Pleasant, Good) to construct a composite variable, Cronbach's alpha was at least 0.7 and typically much higher, in the 0.9 region.

For Data Set 2, the average mood ratings for BBC Comedy, Deforestation, SNL RSE, and Material World were respectively, $M = 5.00$ ($SD = 1.15$; $N = 28$), $M = 4.16$ ($SD = 1.63$; $N = 27$), $M = 4.57$ ($SD = 1.34$; $N = 53$), $M = 4.57$ ($SD = 1.65$; $N = 44$). This time a one-way ANOVA comparing the means yielded a significant difference ($p = 0.00$). Post-hoc analysis (Bonferroni) showed a significant difference between BBC and Deforestation ($p = 0.00$).

Several other comparisons produced results close to significance: BBC Comedy compared to SNL yielded $p = 0.069$; BBC Comedy compared to Material World yielded $p = 0.094$; Deforestation compared to SNL yielded $p = 0.100$ and finally Deforestation compared to Material World yielded $p = 0.117$.

Conclusion on Mood Indicator

Some preliminary conclusions may be in place at this point. The lowest mean rating we attained for any of the video clips was 4.16 (Deforestation), which was not necessarily a neutral clip but rather disheartening.

If we took 4.16 and 1.63 as estimates of the population mean and standard deviation respectively, we would expect to attain a value of 2.77 on the mood indicator in less than 0.1% of cases. The mood value of 2.77 that Valdesolo & DeSteno report is simply not within the range of values we would expect in such a study. And this follows from a comparison to the darkest clip we ran, which produced the lowest mood value we attained.

The next thing we conclude is that it is not trivial to increase mood ratings of participants and any comedy video clip will not do. We chose clips that were described as funny on various websites and listed in rankings of the funniest SNL clips. In other instances we sent clips to students and colleagues and chose the ones that received the best feedback. In all, however, our choices for SNL clips were very limited because of copyright restrictions.

We will provide an analysis of the responses on the trolley cases for both data sets, despite failing to elicit differences on mood values in Data Set 1. The various brain regions (ACC, BA, and DLPFC) that are activated when considering moral dilemmas may nevertheless be affected by the video clips even though it is not captured by the mood indicator.

3.2.4: Results for Trolley Dilemma

We will divide this results section in two parts. The first set of results are presented in Section 3.2.4.1 where we compare the original findings to Data Set 2. These were the closest in procedure and in our data set we also managed to produce statistically significant differences in mood ratings. In Section 3.2.4.2 we present the findings from Data Set 1.

3.2.4.1: Results for Trolley Dilemma – Data Set 2

A summary table of the findings is presented below. In all tables, * denotes $p < 0.05$.

Experiment	***N***	***n***	**Response**				χ^2	***p*-exact**
			Appropriate		Inappropriate			
Original SNL	79	41	10	24.4%	31	75.6%	3.90	< 0.05*
Original Spain		38	3	7.9%	35	92.1%		
BBC Comedy	51	26	5	19.2%	21	80.8%	0.17	0.74
Deforestation		25	6	24.0%	19	76.0%		
SNL (RSE)	95	53	12	22.6%	41	77.4%	0.18	0.80
Material World		42	8	19.0%	34	81.0%		

Table 3.2.1: Judgments for Original and Replication Study (Data Set 2)

An independent-samples t-test comparing mood values of BBC Comedy and Deforestation produced $t(49) = 4.46$, $p = 0.00$. Comparing SNL (RSE) to Material World with an independent-samples t-test produced $t(93) = 0.034$, $p = 0.97$.

The main comparison we present in Table 3.2.1 is between BBC Comedy versus Deforestation and SNL versus Material World because the data for these scenarios was collected in the same procedures. That is, in one experimental run participants were randomly assigned to BBC Comedy or Deforestation and in a procedure that was run at a different time, participants were randomly assigned to SNL or Material World.

Nevertheless, given that BBC Comedy produced mood ratings significantly different at the 10% level from the other two clips (SNL and Material World), we also want to report comparisons between these clips, the assumption being that SNL and Material World did

not have an effect on affective state and only functioned as control conditions. Comparing the BBC Comedy and SNL conditions yielded *p*-exact = 1.000 (χ^2 = 0.120; *p* = 0.729) and comparing BBC Comedy with Material World yielded *p*-exact = 1.000 (χ^2 = 0.000; *p* = 0.985).

There are several things we want to highlight at this point. What stands out in Table 3.2.1 is the low value of 'Appropriate' answer choices for Valdesolo & DeSteno's control condition. This value stands out when compared to all other conditions run in the original as well as replication studies. The next thing that stands out is that the proportion of 'Appropriate' answers is around the 20% mark for all of the other conditions (replication as well as original). Finally, the percentage of participants giving the 'Appropriate' answer was higher for Deforestation than for BBC Comedy. The reason could be that watching a negative clip discussing prospects of mass extinction may lower inhibitions against actively bringing about the death of a stranger. This is speculation and the difference could simply be due to chance variation; however, we want to point this out in case the Deforestation clip (due to its negative nature) may not have tested the effect that Valdesolo & DeSteno investigated. It should also be noted that with this line of reasoning we now have the hypothesis that both negative as well as positive clips can induce more participants to choose the 'Appropriate' answer choice.

3.2.4.2: Results for Trolley Dilemma – Data Set 1

Although we did not attain significant differences for the mood indicators in Data Set 1, we will present the outcomes here for completeness. As a reminder, the mood indicator was presented before the moral dilemmas and the dilemmas were not timed. There were also some differences in other aspects of the surveys. For example, in one of the SNL surveys the train scenario was not randomized in a series of non-moral filler questions but instead always presented as the second question. There were some other differences among the surveys but these were not major and we provide the results below with this qualification in mind.

For data analysis, we will separate the data of all surveys where we presented the Train scenario and all surveys where we presented the Tiger scenario.

Results for Train Scenario

We used all of the clips (Marbleopolis, Celebrity Jeopardy, Luggage, and Spain Documentary) in testing the Train scenario. Comparing the positive to the control conditions yielded no difference on judgments (here judging the question "Would you push the man off the bridge?": Yes/No) with *p*-exact = 0.33 (*N* = 348). For the full details, see Table 3.2.2 below. Averaging the mood variables, for the comedy clips we attained a mean mood value of *M* = 4.54 (*SD* = 1.37; *N* = 246) and for the neutral clips mean equaled *M* = 4.25 (*SD* = 1.38; *N* = 56).[53] An independent-samples t-test comparing the two conditions yielded no difference with $t(300) = 1.42$, $p = 0.16$.

[53] One of the surveys did not collect data on mood indicators, therefore the number of 302 for *N* in the mood comparison instead of 348 that is indicated in the moral judgment comparison.

Results for Tiger Scenario

We presented the Tiger scenario only in two conditions – one experimental and one control – and the video clips used for the experimental and control conditions respectively were Celebrity Jeopardy and Spain Documentary. A comparison between the two conditions yielded no statistical significance on judgments with *p*-exact = 0.35 (*N* = 104). For the full details, see Table 3.2.2 below. The mood rating for the positive clip was *M* = 4.76 (*SD* = 1.41; *N* = 43) and for the neutral clip mean was *M* = 4.49 (*SD* = 1.34; *N* = 64). An independent-samples t-test comparing the two conditions yielded no difference with $t(105) = 1.00$, $p = 0.32$.

The summary results of the Train and Tiger scenarios are presented below.

Scenario	**Condition**	***N***	***n***	**Response**				χ^2	***p*-exact**
				Yes		No			
Train	Experiment	348	294	66	22.4%	228	77.6%	0.41	0.33
	Control		54	10	18.5%	44	81.5%		
Tiger	Experiment	104	39	3	7.7%	36	92.3%	0.55	0.35
	Control		65	8	12.3%	57	87.7%		

Table 3.2.2: Judgments on Train and Tiger Scenarios (Data Set 1)

For the train scenario the proportion of 'Yes' answers to the question 'Would you push the man off the bridge?' was around the 20% mark. In the Tiger scenario this number was considerably lower. Being attacked and devoured by a tiger is likely imagined by participants as more painful than being hit by a train and this likely created a stronger emotional reaction. The imagery of the Tiger case is also more vivid and concrete than the Train case.

3.2.5: Concluding Remarks

For all of the replication studies that used the Train scenario, we were somewhat surprised by the high percentages of the 'Appropriate' answer choices and we were especially surprised by the high percentage of the 'Yes' answer choices when the question asked was whether participants would push the stranger off the bridge. We looked at Greene et al. (2001) for a comparison; however, the sample was too small in that article ($N = 9$) to be a good reference point.

For an approximation, we examined Hauser et al. (2007) who ran a large-scale study online, surveying over 5000 participants from 120 different countries on several trolley type scenarios, including the standard bystander case as well as the footbridge dilemma (Hauser, Cushman, Young, Kang-Xing Jin, & Mikhail, 2007).

The procedures in Hauser et al. (2007) differed in several ways from Valdesolo & DeSteno (2006). First, Hauser and colleagues asked about permissibility, whereas Valdesolo & DeSteno asked about appropriateness. Second, Hauser et al. asked whether the course of action was permissible for a third party (the name of the protagonist in the footbridge dilemma was Frank), whereas Valdesolo & DeSteno asked about appropriateness in a neutral way. Finally, the scenarios in Hauser et al. (2007) were not timed, whereas participants in Valdesolo & DeSteno (2006) had 15 seconds to make their choice.

These differences make comparison between the two studies somewhat difficult. With this in mind, overall, in Hauser et al. (2007) twelve percent of respondents said that it was permissible to push the person off the bridge in order to derail the trolley.

Our first thought for the high percentages of approving answers in the experimental condition of Valdesolo & DeSteno (2006) and in the replication studies was that being timed to 15 seconds may have rushed participants into making a choice and given that 'Appropriate' was on the left and given the convention of reading English from left to right, we suspected that some participants may have selected the first choice they came across (at least a larger proportion than if the question was not timed). However, the percentages from Data Set 1 of the replication study where we did not set a time limit were similar to the timed procedures.

We also expected to have much lower approving answer choices when the question asked participants whether they personally would push the man off the bridge as in Data Set 1. However, the percentages were again comparable to responses in Data Set 2, where we asked about appropriateness in a neutral. We do not have a good explanation for these outcomes.

Sample Size

The study that was closest in procedures and that also produced differences in mood ratings was the comparison between BBC Comedy and Deforestation in Data Set 2.

However, as discussed before, Deforestation may not qualify as a neutral clip but rather be a negative one (although the mood ratings still did not remotely approach the low mood ratings of Valdesolo & DeSteno) and furthermore the sample size was relatively small at $N = 51$ as compared to Valdesolo & DeSteno's $N = 79$. It is possible that we did not have sufficient power to detect a difference; however, this does not seem very likely as the direction of the responses was in reverse of those reported by Valdesolo & DeSteno. That is, the neutral clip had a higher percentage of 'Appropriate' answer choices than the comedy clip. We will not present power calculations here because we did not have the details of procedures for the original study and hence comparison may not be straightforward.

Trolley Problem

One of the problems in conducting a replication of this kind is that the trolley problem has been treated in popular media such as newspapers and blogs (Bakewell, 2013; Brean, 2010; Weiss, 2008). Participants in our surveys may have been more familiar with the dilemma than individuals surveyed by Valdesolo & DeSteno; this is especially the case for participants who voluntarily visit sites like the MST. A trend search for the term "trolley problem" on Google shows an increase in frequency since 2008.

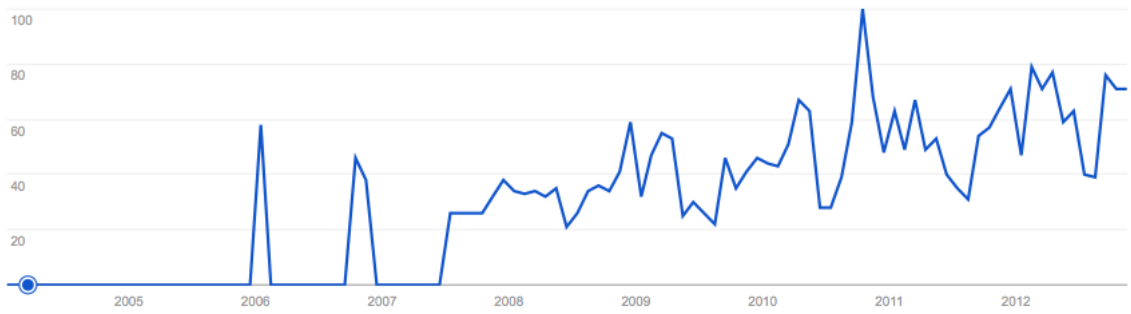

Figure 3.1: Trend Search for Term "Trolley Problem"

Note: The chart depicts relative search volume. The highest volume is designated 100 and subsequently any period that had half that search volume is denoted with 50.

Previous exposure to the scenario may make manipulation of judgments more difficult as individuals may have contemplated on the problem and made up their mind.

Related Studies

As mentioned in the introduction, one paper has modeled its line of investigation on Valdesolo & DeSteno (2006). In a similar way to how Valdesolo & DeSteno (2006) attempt to manipulate brain regions responsible for emotional processing, Greene et al. (2008) attempt to manipulate brain areas responsible for cognitive tasks while participants consider the trolley dilemma.

Greene et al. (2008) gave participants a digit search task where a sequence of numbers scrolled across the screen and every time the number 5 appeared, participants had to press a specific button. This was supposed to increase the cognitive load on areas processing cognitive tasks and hence interfere with moral judgments. Greene et al. (2008)

hypothesized that this would interfere with judgments (lower the frequency of utilitarian judgments) as well as reaction times.[54]

As the authors note,

> utilitarian moral judgments (favoring the sacrifice of one life to save several others) are supported by cognitive control processes, and therefore we predicted that increasing cognitive load by imposing another control-demanding task would interfere with utilitarian moral judgments, yielding increased RT and/or decreased frequency for utilitarian moral judgment." (Greene et al., 2008, p. 1147)

The data did not confirm the hypothesis entirely. While reaction times increased, the frequency of the 'Appropriate' response did not. Greene et al. (2008) conclude that "while load impacted RT, it did not reduce the proportion of utilitarian judgments, as one might have expected based on our theory" (Greene et al., 2008, p. 1151).

The interpretation of these outcomes is not straightforward for Valdesolo & DeSteno's study. On the one hand, Greene et al. (2008) suggests that participants' judgments are not easily manipulated. On the other hand, the findings may be of little relevance to Valdesolo & DeSteno (2006) because the focus is on cognitive and not emotional processes. If one were to assume the reproducibility of Valdesolo & DeSteno (2006), these two papers in combination could weakly suggest that cognitive processes are less open to manipulation than emotional processes.

Conclusion

[54] Reaction time was hypothesized to increase for participants making the utilitarian judgment only, not for those making the deontological judgment.

Although the exact details of the original study were not available to us, following all published procedures and using several different positive as well as control conditions, we could not replicate Valdesolo & DeSteno's (2006) finding. We believe that there are two strong indications that the findings of the original paper are not reliable. First, the mean mood value Valdesolo & DeSteno provide for their control condition (M = 2.77) is an extreme outlier when compared to all other mood ratings and even compared to the lowest value we attained. A rating of 2.77 is extremely unlikely. In a similar way, the percentage of 'Appropriate' answer choices for Valdesolo & DeSteno's control condition also stands out. The only study where this number came about in the replication runs was where we used a different scenario (Tiger), which was much more graphic. Ideally, we would like to obtain the exact details of the original experimental procedures and conduct another replication. Given that this is unlikely, we believe that the effect reported by Valdesolo & DeSteno (2006) needs to be viewed with caution.

Page Intentionally Left Blank

Paper 4: Prevalence of False-Positive Results

After failing to reproduce many of the most cited findings in the experimental philosophy literature (Papers 1 and 2) as well as two papers published in psychology journals (Paper 3), we started to examine the literature – historical as well as current – on the reproducibility of published findings in experimental sciences in general, with a strong focus on psychology. The focus is on psychology because it is the methods of this field that experimental philosophers have adopted to study philosophical questions. The present paper reviews this literature.

Aside from gaining a better understanding of this topic through a literature review, we had another motivation for this study. A straightforward way of assessing the reproducibility of findings in the experimental philosophy literature would be to attempt replication of a representative sample of results. This task lies beyond our resources. However, experimental philosophy is often described as the study of philosophical questions, using the methods of experimental psychology (Alexander, 2012; Knobe & Nichols, 2008; Nadelhoffer & Nahmias, 2007) and by importing the methods of experimental psychology, philosophers will inevitably import some of the problems of that field. The problems of experimental psychology have been discussed for decades and examining these debates will reveal the problems that experimental philosophy is likely to face. There is currently a 'crisis of confidence' in psychology and various other empirical disciplines. By examining the crisis in psychology, it may be possible to

evaluate the state of experimental philosophy as it is practiced today and the likely course it will take (unless changes are made).

The current crisis of confidence in psychology was triggered by two seemingly unrelated events. One was the publication of Bem's paper on extrasensory perception titled "Feeling the Future: Experimental Evidence for Anomalous Retroactive Influences on Cognition and Affect" (Bem, 2011) and the other was a series of high profile cases of fraud involving very esteemed researchers such as Marc Hauser and Diederik Stapel, amongst others.

However, neither of the two above mentioned occurrences – the rare highly improbable result nor the rare cases of high-profile fraud[55] – really lie at the heart of the current crisis of psychology. Instead, what is much more endemic and what seems much more damaging to the field are the likely high rates of published false-positive results.

Various researchers have argued that false positives possibly make up the majority of publications in psychology (Bakker et al., 2012; Ioannidis, 2005, 2012; Pashler & Harris, 2012). Ioannidis has contended that in "several fields of investigation, including many areas of psychological science, perpetuated and unchallenged fallacies may comprise the majority of the circulating evidence." Even more, Ioannidis claims, "the prevalence of unchallenged fallacies may represent even up to 95% (if not more) of the significant findings in some areas of the psychological literature" (Ioannidis, 2012, p. 645 and 650).

[55] Although, it is also debatable whether such cases of fraud are rare, see Sovacool (2008) and Stroebe et al. (2012).

The discipline of psychology is not the only one afflicted with this problem. The rate of reproducible findings is possibly abnormally low in various other empirical fields, most notably the biomedical sciences. Studies such as (Begley & Ellis, 2012b; Diep, 2013; Fang & Casadevall, 2012; Osherovich, 2011; Prinz, Schlange, & Asadullah, 2011a, 2011b; Turner, Matthews, Linardatos, Tell, & Rosenthal, 2008) have led commentators to claim that "evidence from diverse fields suggests that when efforts are made to repeat or reproduce published research, the repeatability and reproducibility is dismal" (Ioannidis, 2012, p. 647).

This paper is structured as follows. Section 4.1 starts by giving some details on the cases of Bem and Stapel as a background to the discussion that follows in the rest of the paper. Section 4.2 gives a discussion of the ways through which false-positive results can enter the literature; the first is through the use of the prevailing statistical and publishing practices and the second is through the use of questionable research practices (QRPs). Section 4.3 discusses some of the reasons why researchers may use questionable practices. The final section concludes with a discussion.

4.1: The Cases of Stapel and Bem

As a background to this paper, we provide a closer account of the Stapel and Bem cases. This is not because these cases are representative of problematic conduct in psychology and also not because these are the most damaging to the field. Instead, the motivation is

to show some extreme cases that can pass through the current publication system and also persist for years in the literature. These cases may be the exceptions; however, according to some researchers, similar practices although in less blatant forms are widespread and possibly the norm in psychology (John et al., 2012).

4.1.1: Stapel – Fraud and Questionable Science

Recently, there have been numerous high profile cases of fraud in psychology (J. Cooper, 2012; Dahlberg, 2012; Levelt, 2012; Oransky, 2012, 2013; Yong, 2012b). There are several reasons for singling Stapel's case out here. First, Stapel was highly regarded by his colleagues as a prolific writer who published in the most prestigious international journals. Second, the universities at which Stapel held positions conducted very thorough investigations (for cases ranging from 1993 until 2011) and made all results public; the main findings being presented by the Levelt committee (Levelt, 2012). This is in contrast to other cases where specific laws prevented universities from making their reports public (Dahlberg, 2012; Oransky, 2013). Furthermore, Stapel himself cooperated with the inquiry to a degree.[56] Third, Stapel's is likely one of the most extreme cases of recent fraud in psychology with some of the most blatant methods used to fabricate results. Thus, it may serve as an extreme of how scientific practice can fail and how researchers can get away for extended periods of time without being exposed by the current standard processes. Finally, the investigation also touches on broader issues such as the

[56] Stapel initially cooperated with the investigations and supplied many details but later stopped due to health concerns (Levelt, 2012).

prevalence of questionable research practices, research culture, and carelessness of Stapel's collaborators and co-authors as well as journals.

As briefly pointed out, Stapel was considered an immensely talented psychologist with a very promising career ahead. He was a prolific writer with publications in many reputable journals, often publishing work in collaboration with international researchers.[57] In 2009 Stapel received the "Career Trajectory Award" from the Society of Experimental Social Psychology which "celebrates scientific contributions made in the early-to-mid stages of a research career" (SESP, 2013). The award has since been retracted. In 2010 Stapel was named dean of the Social and Behavioural Sciences Faculty at Tilburg University. Stapel received more than two million euros in research funding from the Netherlands Organization for Scientific Research.

Stapel used four strategies to falsify data. The first was to straight-out fabricate response sheets to experiments that, although were discussed in group-meetings, were never carried out with subjects. Stapel would tell his students and research assistants that the data was collected in a different laboratory. The second strategy was to carry out experiments with participants as discussed in group-meetings. However, before the data was analyzed by research assistants, Stapel would take the material in for 'inspection'. This gave him an opportunity to alter the data before it was passed on to others. The third strategy Stapel used was to contact researchers at other universities and inform them that he had collected data in the past on experiments that would be of interest. Stapel would explain that he never had the time to analyze this data. This alleged raw data was

[57] All collaborators, including Stapel's PhD students, were cleared of misconduct.

never collected in experiments but was entirely fabricated by Stapel. Stapel would ask his collaborators to analyze the data and to author manuscripts. A final strategy Stapel used was to tell his research group that he had contacts in high schools and that these would gladly collect data in classroom settings in exchange for computer equipment or projectors. In reality these contacts did not exist and Stapel would once again fabricate response sheets.

In many of the cases where research led to publication, Stapel was the only person in charge of data collection. This should have been reason for suspicion, as more senior faculty members typically stay away from the tedious and time consuming task of data collection. Nevertheless, according to one evaluation, Stapel could have eluded discovery and explained away irregularities as mistakes had he only used the first three strategies (Stroebe, Postmes, & Spears, 2012). What gave him away was the fourth approach; after university officials asked to speak with Stapel's contacts in high schools, Stapel had no option but to admit wrongdoing.

On two prior occasions allegations of misconduct had surfaced (brought forth by research students as well as faculty members); however, no action was taken. This time, the whistleblowers waited until they had sufficient evidence and only then contacted university officials. One of the irregularities that the whistleblowers noticed was that the mean age for data collected in high schools came out to 19. Another peculiarity the whistleblowers noticed was that identical lines of data appeared in multiple studies. Furthermore, the effect sizes of Stapel's data were extremely strong and for every study

the data fit the hypothesis perfectly. Upon being contacted by the whistleblowers, university officials this time launched an investigation.

The findings of the investigation were published in 2012 and concluded, amongst other things, that Stapel had used questionable practices as early as 1996, including for his PhD dissertation. From 2002 onward, Stapel shared fabricated data with other researchers. At the time of writing, 46 of Stapel's papers have been retracted and the number is likely to increase to 55. For a full list of articles the committee designated as fraudulent, see (Levelt, 2012).

These papers, so the report, were cases of fraud. However, in addition to fraud, the committee also highlights that many questionable practices were used that did not fall directly under the committee's definition of fraud. The report's definition of fraud is stated as the "fabrication, falsification or unjustified replenishment of data, as well as the whole or partial fabrication of analysis results. It also includes the misleading presentation of crucial points as far as the organization or nature of the experiment are concerned" (Levelt, 2012, p. 17).

In contrast to fraud, some of the examples of questionable research practices highlighted by the report are the following. Experiments were repeated multiple times with minor changes until a significant outcome was achieved. Hereupon the experiment was ended and a positive result recorded without mentioning in the manuscript how many runs had been conducted in total. A further practice consisted of comparing experimental groups

to control groups from other experiments depending on which one yielded a better comparison. Another practice was the unjustified deletion of data points from analysis. This practice can present a gray area, as deleting observations may be legitimate in some instances. For example, an extreme outlier in a reaction time study may be legitimately eliminated from analysis. However, Stapel's approach was not pre-determined and decisions were made on a case-by-case basis. For a more comprehensive list of questionable research practices, see Table 4.2 in Section 4.2.2.

The distinction between fraud and questionable research practices may not be very meaningful as the damage caused by the latter to psychology may in fact be greater. For the most part in this paper as well as Paper 5, we will treat the two as the same; fraud and misconduct being a subset of questionable research practices. As the Levelt report highlights, questionable research practices may be "in principle, equally unacceptable and may, if not identified or corrected, easily lead to more serious breaches of standards of integrity" (Levelt, 2012, p. 57).

In addition to finding fault with Stapel and his collaborators, the committees also express criticism of the field's journals. In interviews the committees conducted, several individuals mentioned that "reviewers encouraged irregular practices" (Levelt, 2012, p. 53). Such irregular practices included suggestions by referees to omit experimental variables in final analyses or to conduct post-hoc pilot studies, which were to be designated as preceding the main experimental run. Furthermore, aesthetic concerns were often given higher priority than truthful reporting. The Levelt report states that "not

infrequently reviews were strongly in favour of telling an interesting, elegant, concise and compelling story, possibly at the expense of the necessary scientific diligence" (Levelt, 2012, p. 53). As a general evaluation, the committees express concern with the entire research environment in which Stapel found himself and refer to a "failure of scientific criticism in the peer community" and a "failure on all levels of the scientific review procedures" (Levelt, 2012, p. 47).

The Levelt report includes some suggestions on how to avoid cases like Stapel's from recurring in the future. Three main areas are identified: replication, transparency and journal standards. One of the criticisms of the report is that Stapel's results were not replicated systematically and in cases were replication did take place, these could not be published because of an aversion by journals to publishing replications. Levelt urges that "far more than is customary in psychology research practice, replication must be made part of the basic instruments of the discipline and at least a few journals must provide space for the publication of replicated research" (Levelt, 2012, p. 58). In order to facilitate replications, the committees advise for more transparency. This includes the detailed descriptions of experimental conduct, making raw data available online as well as experimental materials such as survey sheets and any graphics used.

One worrying conclusion that emerges from the investigation is that Stapel's case may merely be the tip of the iceberg. The report notes that "the Committees have been made aware of several cases of this kind in the Netherlands and abroad, in which much research funding and expensive research time has been wasted" (Levelt, 2012, p. 54).

Overall, the report finds strong words for the research environment as a whole, suggesting that the problems of the field may be systemic. The report notes.

> A byproduct of the Committees' inquiries is the conclusion that, far more than originally assumed, there are certain aspects of the discipline itself that should be deemed undesirable or even incorrect from the perspective of academic standards and scientific integrity […] the critical function of science has failed on all levels. Fundamental principles of scientific method have been ignored, or set aside as irrelevant. In the opinion of the Committees this has contributed significantly to the delayed discovery of the fraud. It is to the credit of the whistleblowers in Tilburg that they did discover these infringements of scientific integrity and took the correct action. (Levelt, 2012, p. 54)

4.1.2: Bem – Feeling the Future

In a paper published in 2011 in the prestigious *Journal of Personality and Social Psychology*, Bem claimed to have evidence for psi or extrasensory perception. Bem presented nine studies with over 1,000 participants to test whether individuals could predict future events and eight of these studies yielded statically significant results (Bem, 2011).

The first experiment of the paper ran as follows. Two pictures of curtains were shown side-by-side on a computer screen. Participants were asked to guess behind which of the curtains there was an object. However, neither the type of object nor its position was determined by the computer program until after individuals made their choice.

If participants were more likely than chance to predict the correct position of the object (behind the left or right curtain), the experiment would provide evidence of precognition or the ability to predict the future. And in fact, Bem found that for certain classes of objects participants were more likely than chance (53.1%; $p = 0.01$) to predict the correct position. Given that the position of the object was not determined by the computer program until after participants made their choice, participants were predicting future events. Bem found effects on classes of objects that were either related to themes of procreation or fight-or-flight responses. Having precognition on these classes of objects, so Bem, gives individuals an evolutionary advantage.

The experiment with the largest effect size ($d = 0.42$) was a recall test (conducted in reverse). Individuals were presented with a list of words and later asked to recall these and enter them in a text box at their computer. After the task was completed, the computer program randomly selected a subset of these words and displayed these on the monitor to participants. Bem found that words that were shown to participants after the recall task, had a higher likelihood of being correctly remembered ($p = 0.002$). That is, future events had an effect on recall performance.

To summarize the main findings of Bem's paper, the results suggest that cause need not always precede effect and that humans may have evolved a capacity for predicting future events. These are certainly important findings that warrant further scrutiny, especially given the strong evidence that Bem presents with over one thousand participants.

Bem's paper received considerable criticism on methodological grounds in various journals (Alcock, 2011; Rouder & Morey, 2011; Shermer, 2011; Wagenmakers, Wetzels, Borsboom, Kievit, & van der Maas, ms; Wagenmakers, Wetzels, Borsboom, & van der Maas, 2011). The paper also received criticism in less formal outlets (Carey, 2011; Shermer, 2011). All of these reactions are important; however, our main interest here is in what happened when a group of researchers attempted to replicate one of Bem's experiments and submitted their manuscript for publication. Ritchie, Wiseman, & French (2012a) took on the task of replicating one of the nine experiments in Bem (2011). They selected the experiment with the largest effect size (outlined above). The replications were conducted in three different laboratories, each with a sample size as large as the original trial. All three replication attempts failed. Ritchie, Wiseman, & French submitted their manuscript to the *Journal of Personality and Social Psychology*, the same journal that had published Bem's original paper. The journal responded that as a matter of journal policy, replication studies would not be considered for publication. This is very revealing of the priorities of one of the most prestigious psychology journals. Hereby, the journal was signaling that it was less concerned about truthful research but instead sensational outcomes.

After the first rejection, three further publication attempts failed (Ritchie et al., 2012b). Finally, the journal *PLOS ONE* accepted the manuscript for publication (Ritchie, Wiseman, & French, 2012a). *PLOS ONE's* approach is to publish articles on their merit and not perceived importance. As the guidelines state, *PLOS ONE* will "publish all papers that are judged to be technically sound. Judgments about the importance of any

particular paper are then made after publication by the readership (who are the most qualified to determine what is of interest to them)" (PLOS, 2013). *PLOS ONE* is a young journal, established in 2006, and represents a novel approach to publishing.

4.2: Sources of False Positives

As previously mentioned, although the cases of Bem and Stapel may be informative, the main concern of this paper is the prevalence of false-positive results in the published literature. There are, very broadly speaking, two ways in which false-positive results enter the published literature. One is through fraud and questionable research practices and the other is as a by-product of prevailing statistical and publication practices.

4.2.1: Prevailing Statistical and Publication Practices

It is often argued that the 5% alpha level typically used in statistical procedures assures that only five percent of positive results in the literature are false positives (Pashler & Harris, 2012). This argument, so Pashler & Harris, is inaccurate because it misses the literature-wide alpha level. Consider the following example as adapted from Pashler & Harris (2012).

Assume that for all experiments carried out in psychology research the prior probability of an effect existing is 10 percent. That is, in 90 percent of cases the null hypothesis is correct. Given an alpha level of 5%, we would expect the false rejection of the null

hypothesis in 4.5% (0.05 x 0.9) of experiments (Type I error). Furthermore, consider that the discipline-wide power (probability of correctly rejecting the null hypothesis when the null is incorrect) of experiments is 0.35. Then, the null will be rejected correctly in 3.5% (0.35 x 0.1) of cases. If all positive results were published, then false positives would make up 56% $\left(\frac{4.5}{4.5+3.5}\right)$ of published results. This number is substantially greater than the 5% alpha level that is supposed to be a stringent safeguard against the introduction of false-positive results into the scientific literature. We provide a closer explanation of these numbers below.

The prior selected in the above example is certainly debatable. What the true prior is in psychology research is difficult to assess (and of course, the lower the prior, the more exciting the result). The average statistical power used in published psychology experiments is estimated to be 0.35 (Bakker et al., 2012). This is calculated from a median sample size of 40 used in experiments (Marszalek, Barber, Kohlhart, & Holmes, 2011; Wetzels et al., 2011) and an average effect size of d = 0.50 (Anderson, Lindsay, & Bushman, 1999; Lipsey & Wilson, 1993; Meyer et al., 2001; Richard, Bond, & Stokes-Zoota, 2003; Tett, Meyer, & Roese, 1994). Generally speaking, there are three components that determine the power of a study. These are the sample size (the greater the sample size the greater the likelihood of finding an effect if one existed); effect size (the greater the effect size, the more likely an experiment will detect an effect if one existed (at a given alpha level and sample size)); and the alpha level (the greater the alpha level, the more likely that an experiment will detect an effect (at that level) given a certain sample and effect size.

Table 4.1 provides different combinations of values of prior and power, as adapted from Pashler & Harris (2012).

Prior probability of effect	**Power**	**Proportion of studies yielding true positives**		**Proportion of studies yielding false positives**		**Proportion of positive results that are false**	
(%)	(%)	(%)	Calculation	(%)	Calculation	(%)	Calculation
10	80	8	0.1*0.8	4.5	(1-0.1)*0.05[58]	36	$\frac{4.5}{4.5 + 8}$
10	35	3.5	0.1*0.35	4.5	(1-0.1)*0.05	56	$\frac{4.5}{4.5 + 3.5}$
50	35	17.5	0.5*0.35	2.5	(1-0.5)*0.05	13	$\frac{2.5}{2.5 + 17.5}$
75	35	26.3	0.75*0.35	1.3[59]	(1-0.75)*0.05	5	$\frac{1.3}{1.3 + 26.3}$

Table 4.1: Combination of Values Yielding False Positives – Pashler & Harris (2012)

Explaining the numbers in Table 4.1 in more detail: with a power of 0.8, if an effect existed, the procedure will detect it in 80 out of a 100 experimental runs. The first row assumes the prior to be 10% and hence an effect that exists will be correctly detected in 8% (0.8 * 0.1) of experimental runs.

The proportion of false positives is calculated as follows. The prior is assumed to be 10% and so in 90% of cases an effect does not exist. With an alpha level of 5%, in five percent of cases where an effect does not exist, a false positive (Type I error) will be recorded. That is, 4.5% (0.9 * 0.05) of cases will be studies that yield false positives.

[58] Alpha level at 5% assumed.
[59] The original paper mistakenly gives this value as 1.6%.

Overall, the proportion of positive results that are false is the percentage of studies that yield false positives (given the prior) divided by the percentage of studies that yield false positives and studies where the null is rejected correctly. For the first row this amounts to $\frac{0.9*0.05}{(0.9*0.05)+(0.8*0.1)}$ or $\frac{4.5}{4.5+8}$ which equates to 36%.

For a 5% false-positive rate to come about, given a power of 0.35, the prior probability of an effect would have to be 0.75. This may be unrealistic, amongst other reasons, because of the high number of studies that are exploratory but for which post-hoc explanations are constructed once positive results are attained (Pashler & Harris, 2012). Exploratory studies do not test a specific hypothesis but rather approach some issues in general strokes and collect data somewhat indiscriminately. These studies will reduce the value of the prior as in most cases there are no effects to be detected. However, often when an effect is detected, researchers will describe their work as having tested a concrete hypothesis (Pashler & Harris, 2012).

In closing this segment, from their discussion of statistical methods used in psychology, Pashler & Harris (2012) conclude that "in summary, our standard statistical practices provide no assurance that erroneous findings will occur in the literature at rates even close to the nominal alpha level" (Pashler & Harris, 2012, p. 533).

Publication Bias

What makes the numbers estimated above likely is the over the years persistent problem of publication bias. Publication bias refers to the preference that positive results receive over negative ones in the publication process. One of the points of contention which has been discussed for decades is the high percentage of positive results of null hypothesis significance testing (NHST) in published papers (Bakan, 1966; Bakker et al., 2012; Bozarth & Roberts, 1972; Sterling, 1959; Sterling et al., 1995). Researchers who have dealt and followed the problem have concluded that "practice leading to publication bias have not changed over a period of 30 years" (Sterling et al., 1995, p. 108).

In a paper published in 1959, Sterling examined four journals each in a different area of psychology and found that out of 294 published papers that used null hypothesis significance testing (81.2 percent of all papers reviewed), 97% of articles achieved positive results at the 5% level (Sterling, 1959). Similarly, in 1972, a study reported a 94% rate of null hypothesis rejection rate at the 5% level for three psychology journals (900 articles reviewed, 86 percent of which used statistical tests) for a three year period between 1967 and 1970 (Bozarth & Roberts, 1972). In 1995, a group of researchers investigated eleven journals of which eight were in the field of psychology and three in clinical and medical journals. Out of 563 psychology articles that used statistical tests (94.3 percent of all papers reviewed) in 1986 and 1987, 95.6% reported rejections of the null hypothesis at the 5% level. The percentage was lower for the clinical journals at 85.4% (456 papers reviewed of which 316 or 69.3% used statistical tests) (Sterling et al., 1995).

These numbers strongly indicate that published results in psychology are not representative of all experiments conducted. An estimate is that roughly two-thirds of studies that are approved by Institutional Review Boards (IRB) and that are completed remain unpublished (H. Cooper et al., 1997). The reasons cited are not always related to publication bias against negative results; however, among studies that attain positive results, 74% are submitted for publication in a journal or book chapter and only 4% of the studies that find negative results are submitted to similar outlets (Cooper, Deneve, & Charlton, 1997).

Another indication that published results are unlikely to be representative of all experiments conducted is the following. As mentioned before, average power in psychology research is calculated to be 0.35 (given sample size, effect size, and alpha). A power of 0.5 means that the likelihood of finding an effect when one exists is 50%. The high percentage of positive results in published papers conflicts with this estimate sharply.

Despite objections to publication bias, which was noted many years ago, the trend seems to be worsening. Fanelli (2012) found that the trend was toward an increase in the proportion of publication of positive results. Examining over 4600 articles from different disciplines published between 1970 and 2007, Fanelli found that the percentage of positive results increased by 22% during this time period. What also stands out is that psychological sciences have the highest proportion of positive results, followed in descending order by Materials Science, Pharmacology and Toxicology, Clinical

Medicine, Biology and Biochemistry, Economics and Business, Molecular Biology and Genetics, Engineering, and Immunology (Fanelli, 2010).

Given that there is an aversion to publishing negative results and also replications, which has also been noted for decades (Coursol & Wagner, 1986; Sterling, 1959), it is very likely that researchers who are not aware of negative results to certain effects will independently carry out experiments testing such effects. Eventually one of these research groups will find a positive result, that is obtain a Type I Error, and submit this finding for publication.

Sterling summarized the problem already in 1959 very coherently.

> There is some evidence that in fields where statistical tests of significance are used, research which yields nonsignificant results is not published. Such research being unknown to other investigators may be repeated independently until eventually by chance a significant result occurs–an "error of the first kind"–and is published. Significant results published in these fields are seldom verified by independent replication. The possibility thus arises that the literature of such a field consists in substantial part of false conclusions resulting from errors of the first kind in statistical tests of significance. (Sterling, 1959, p. 30)

4.2.2: Questionable Research Practices

A common view expressed in academic circles is that although misconduct does occur, it is so rare as to be insignificant (Kennedy, 2006; Koshland, 1987; Kraut, 2011; LaFollette, 2000; Martinson et al., 2005; Reynolds, 2004; Shamoo & Resnik, 2003; Sovacool, 2008). We give a closer account of this view in Paper 5. For here, the question whether this 'bad

apple view' (Sovacool, 2008) is correct has implications on whether substantial changes need to be made in the practice of science or whether minor changes would suffice. The next section explores questionable research practices in detail because many false-positive results are likely to come about as a result of questionable practices. Furthermore, this issue is of importance because much in the current system relies on trust (Koshland, 1987; Stroebe et al., 2012) and given recent developments, an examination whether this trust is deserved may be appropriate.

Already in 1830 Babbage categorized several different practices that constitute questionable research practices and lamented the high prevalence of these practices in his book titled "The Decline of Science in England" (Babbage, 1830). More recently, in order to evaluate whether questionable research practices are the acts of a few bad apples or whether such practices are more common, researchers have surveyed academics about their practices. There are several surveys on this issue; however, in this paper we would like to give a detailed account of John et al. (2012) since this was a study on psychologists only. We will make references to other surveys where appropriate and highlight their conclusions.

For their study, John et al. (2012) surveyed 2155 research psychologists on ten different behaviors that constitute questionable research practices (see Table 4.2 below). A total 5964 researchers were contacted via email of whom 36% responded. All participants received the same questions (in randomized order).

The survey ran two conditions. One was a standard survey asking researchers about their practices. The other condition provided incentives for participants to give truthful answers. This condition is referred to as the Bayesian-truth-serum (BTS) condition. The BTS condition provided incentives for respondents to give accurate answers by making donations to a charity of the participant's choice, if respondents' answers were close to the true outcome as assessed by an algorithm. Participants were instructed of this explicitly. We will omit a detailed explanation of the BTS algorithm, as it is beyond the scope of the current paper and also not too important for our purposes. The BTS condition is simply an effect that the authors report (given the problems of underreporting on surveys of this nature). For more details on BTS, see (Prelec, 2004). In the control condition, a donation was made for each participant regardless of answer choices.

In addition to self-reports, respondents who had engaged in QRPs were also asked to evaluate whether those acts were defensible. The answer choices provided were 'No', 'Possibly' and 'Yes' which were scored with 0, 1, and 2, respectively.

The outcomes for the self-admission rates for both conditions (BTS and control) are presented in Table 4.2 below.

Item	Self-admission rate (%)		Odds ratio (BTS/control)	Two-sided *p* (likelihood ratio test)	Defensibility rating (across groups)
	Control group	BTS group			
1. In a paper, failing to report all of a study's dependent measures	63.4	66.5	1.14	0.23	1.84 (0.39)
2. Deciding whether to collect more data after looking to see whether the results were significant	55.9	58.0	1.08	0.46	1.79 (0.44)
3. In a paper, failing to report all of a study's conditions	27.7	27.4	0.98	0.90	1.77 (0.49)
4. Stopping collecting data earlier than planned because one found the result that one had been looking for	15.6	22.5	1.57	0.00	1.76 (0.48)
5. In a paper, "rounding off" a *p* value (e.g. reporting that a *p* value of 0.054 is less than 0.05	22.0	23.3	1.07	0.58	1.68 (0.57)
6. In a paper, selectively reporting studies that "worked"	45.8	50.0	1.18	0.13	1.66 (0.53)
7. Deciding whether to exclude data after looking at the impact of doing so on the results	38.2	43.4	1.23	0.06	1.61 (0.59)
8. In a paper, reporting an unexpected finding as having been predicted from the start	27.0	35.0	1.45	0.00	1.50 (0.60)
9. In a paper claiming that results are unaffected by demographic variables (e.g., gender) when one is actually unsure (or knows that they do)	3.0	4.5	1.52	0.16	1.32 (0.60)
10. Falsifying data	0.6	1.7	2.75	0.07	0.16 (0.38)

Table 4.2: Questionable Research Practices – John et al. (2012)

Approximately 20% admitted to having stopped data collection prematurely after having attained a significant difference (item 4). Related to this, more than half of respondents in both conditions indicated that they had made data collection dependent on finding significant differences. That is, data was collected and evaluated and based on the outcome either more data was collected or not (item 2). This increases the likelihood of

finding a statistically significant effect when one does not exist. By carrying out multiple tests, the chances of attaining a Type I error are increased considerably.

Simmons et al. (2011) demonstrate concretely how this practice can be used to attain positive results where none exist. In a simulation, Simmons et al. drew two random samples of size $n = 10$ from a normal distribution. A t-test comparing these two samples was conducted. If a statistical significance was detected between the two samples, data collection stopped and a positive result was reported. If no statistical difference was detected, one more data point was added to each sample. This data point was randomly selected from a normal distribution. Now, with sample sizes of $n = 11$, again a t-test of significance was conducted. If a significant difference was attained, data collection ceased and a positive result recorded. Otherwise, these steps continued until either a significant difference was found or a maximum sample size of $n = 50$ was reached (for each sample). The likelihood of attaining a significant difference between two samples (or two conditions) when testing random samples in this way was 22%, which is more than four times the alpha level.

Figure 4.1 displays how *p* values developed in one of these simulations. What is worth pointing out is that merely adding observations does not 'linearly' increase *p* values (when there is no effect). The *p* value will vary, especially in small samples, and potentially run below the five percent threshold (red dotted line).

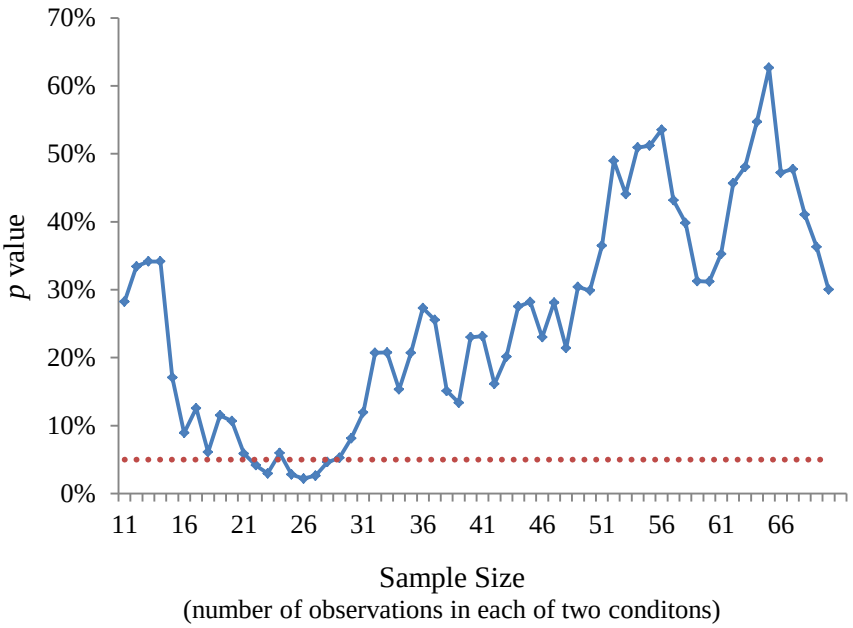

Figure 4.1: *p* Values and Additional t-tests – Simmons et al. (2011)[60]

The authors also ran variations of this simulation where significance was tested after increasing the sample sizes by five, ten, or twenty data points and also by starting with an initial sample size of twenty per condition. The results are displayed in Figure 4.2 below.

[60] This chart was emailed to us by the original authors and is slightly different from the published version. For example, the axes descriptions are different. We furthermore made minimal changes for style and clarity. This version appeared clearer to us and hence the choice to use it. The data points are identical to the published version.

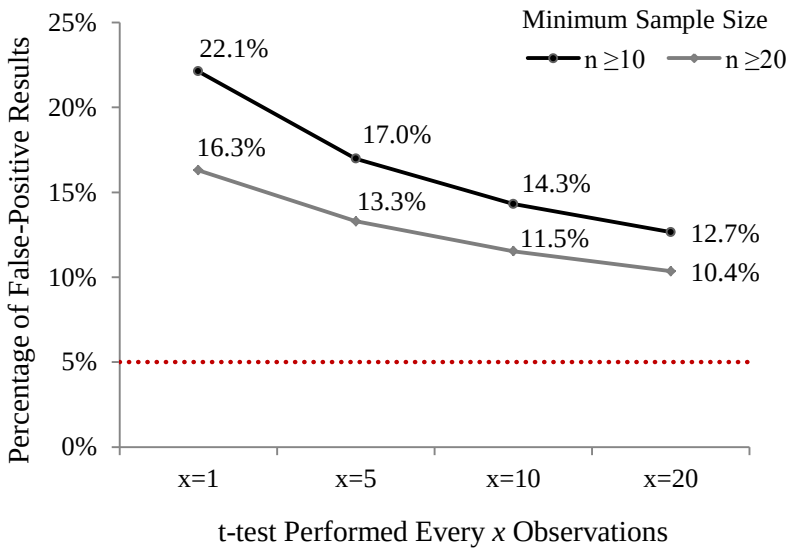

Figure 4.2: False Positives and Additional t-tests – Simmons et al. (2011)[61]

In the best case scenario of this simulation, where the starting sample size was 20 per condition and where significance was tested only one more time after adding 20 data points to each condition, the likelihood of attaining a false-positive result was 10%, or twice the alpha level.

Continuing with the outcomes of Table 4.2, approximately half of respondents in both conditions reported that they had selectively reported studies that worked (item 6). Around 40% indicated that they had excluded observations from their datasets in order to attain a desired outcome (item 7). Circa 65% answered that they did not report all dependent variables of experiments. Failing to report all of a study's dependent measures is problematic because in principle a researcher can run a study with numerous dependent measures and only report those that yield a statistically significant outcome. Including

[61] This chart was emailed to us by the original authors and is slightly different from the published version. For example, the axes descriptions are different. We furthermore made minimal changes for style and clarity. This version appeared clearer to us and hence the choice to use it. The data points are identical to the published version.

many dependent variables increases the likelihood of attaining an effect when one does not exist (Type I error) by chance. This is similar to running an experiment multiple times until a statistical difference is detected, which is then reported in a manuscript without mentioning how many times the experiment was repeated.

What is also interesting is that the defensibility ratings for the items of Table 4.2 were relatively high. It is not the case that respondents were not aware of the problems of these practices. In a follow-up study John et al. asked respondents to rate the defensibility of these practices – without asking whether respondents had engaged in these practices – and the scores were significantly lower. This suggests that researchers understand that these practices are not acceptable; however, they still perform them and in the context of evaluating their own actions possibly make up post-hoc justifications (John et al., 2012).

Finally, respondents were also asked whether they had falsified data and 1.7 percent admitted that they had in BTS. The defensibility rating for this category was low (0.16). It is surprising that close to two percent of researchers admit to having straight-out falsified data.

In all, 94% of respondents admitted to at least one of the QRPs in the BTS condition and 91% did so in the control condition. The mean self-admission rate for the ten practices listed in Table 4.2 was 37% in the BTS condition and 33% in the control condition. John et al. conclude that "across QRPs, […] raw self-admission rates were surprisingly high,

[…] which suggests that these practices may constitute the de facto scientific norm" (John et al., 2012, p. 524).

From their own study, Simmons and colleagues conclude that,

> despite empirical psychologists' nominal endorsement of a low rate of false-positive findings (≤ .05), flexibility in data collection, analysis, and reporting dramatically increases actual false-positive rates. In many cases, a researcher is more likely to falsely find evidence that an effect exists than to correctly find evidence that it does not. […] In fact, it is unacceptably easy to publish "statistically significant" evidence consistent with *any* hypothesis. (Simmons, Nelson, & Simonsohn, 2011, p. 1359)

There have been several other studies surveying researchers' practices. By and large these studies confirm the findings of John et al. (2012). We will highlight four of these here: a *New Scientist* survey, an Office of Research Integrity (ORI) commissioned study conducted in 2005, Martinson et al. (2005) and Fanelli (2009). The reason for highlighting these four is as follows. The *New Scientist* survey was conducted in 1976 and is the oldest of this kind that we are aware of. The second study was commissioned by the ORI and hence carries the authority of that institution. Martinson et al. (2005) is described to be one of the first large-scale studies that directly surveyed researchers from different fields (Martinson et al., 2005). Furthermore, the estimates produced are considered to be conservative (Fanelli, 2009) and hence this may be a good benchmark on practices across disciplines. Finally, Fanelli (2009) is a meta-analysis of 18 surveys similar to those of Martinson et al. (2005) and John et al. (2012).

In 1976 , noting that "science has maintained an ostrich-like attitude about intentional bias for too long" (St James-Roberts, 1976a, p. 482) the *New Scientist* asked its readership to participate in a survey. Two hundred and four readers responded and amongst these, 92% reported of having directly or indirectly witnessed cases of "intentional bias" (St James-Roberts, 1976b).

In the ORI commissioned study, over 2000 principal investigators (PI) of laboratories across all disciplines were asked about misconduct as defined by the ORI.[62] This is a more serious offence than 'intentional bias'. More than seven percent reported that they had suspected cases of misconduct in their own departments alone. For several methodological reasons, the authors of the report consider this number to be a "floor of any generalized estimate" (ORI, 2008, p. 39).

Martinson et al. (2005) surveyed 3247 (1768 mid-career and 1479 early-career) NIH funded researchers on a host of questionable research practices. Overall, one-third of respondents had engaged in at least one of the questionable behaviors. The authors of the study point out that the findings are likely to be conservative estimates. For one, the survey was mailed to researchers and those who had engaged in QRPs are likely to have

[62] The definition of misconduct has not always been straightforward and has changed over time (LaFollette, 2000; Sovacool, 2008). Currently, the Office of Research Integrity (ORI) defines misconduct as,

> fabrication, falsification, or plagiarism in proposing, performing, or reviewing research, or in reporting research results.
> (a) Fabrication is making up data or results and recording or reporting them.
> (b) Falsification is manipulating research materials, equipment, or processes, or changing or omitting data or results such that the research is not accurately represented in the research record.
> (c) Plagiarism is the appropriation of another person's ideas, processes, results, or words without giving appropriate credit. (ORI, 2005)

refrained from participating (non-response bias) and among those who did return the questionnaire, individuals who had conducted QRPs are likely not to have reported all, out of fear of repercussions. Martinson and colleagues conclude from their study that their "findings reveal a range of questionable practices that are striking in their breadth and prevalence" and furthermore that "with as many as 33% of our survey respondents admitting to one or more of the top-ten behaviours, the scientific community can no longer remain complacent about such misbehaviour" (Martinson et al., 2005, p. 737 and 738).

Fanelli (2009) conducted a meta-analysis of 18 surveys on misconduct from various different fields published between 1987 and 2008. The upper bound of the admission rate for having committed QRPs was 33.7 percent and the mean rate was 9.54 percent. When asked about fabrication, falsification, alteration and modification of data, the admission rates ranged from 0.3% to 4.9%, with a mean of 1.97%. However, when the words 'fabrication' or 'falsification' were used explicitly, the mean admission rate dropped to 1.06%. When asked if respondents had observed others falsify, fabricate, alter and modify data, the rates ranged between 5.2% and 33.3% with a mean of 14.12%. Between 6.2% and 72% of respondents indicated that they had observed others engage in questionable research practices. The mean value was 28.53%. When the questions were worded in more general terms such as 'experimental deficiency', 'reporting deficiency' or 'misrepresentation of data' the rates ranged between 12% and 92% with a mean of 46.24%.

Fanelli considers these outcomes to be conservative for two reasons; one having to do with the methodology of self-reports. The other reason is that one of the studies (Martinson et al., 2005) stood out from the rest in that admission rates were uncharacteristically low on some of the questions. This study was also the largest in the sample. When this study was excluded from analysis, the numbers increased considerably. For example, self-admission rate for misconduct increased from two to three percent. Fanelli concludes that "it is likely that, if on average 2% of scientists admit to having falsified research at least once and up to 34% admit other questionable research practices, the actual frequencies of misconduct could be higher than this." (Fanelli, 2009, p. 10).

Another outcome of the study worth pointing out is that the medical, clinical and pharmacological fields had the highest rates of questionable research practices. There may be two reasons for this. One, the influence of money in such research could have an adverse impact; however, at the same time it may be possible that due to increased training to detect and raise awareness of research misconduct, researchers in these fields were more likely to report such occurrences (Fanelli, 2009).

These studies do need to come with a proviso. Some practices that are categorized as questionable in Table 4.2 may be justified. A medical treatment may be effective in preventing a fatal disease to take its course and in such cases it would not be ethically acceptable to continue the study with the knowledge that participants in the control group are being denied a possible treatment. In such instances, premature termination of the

study would certainly not constitute a questionable research practice. In psychology, the concern is that researchers may terminate studies prematurely once a statistical difference has been found, out of concern that with more data, differences that result from chance could disappear.

In cases of inadequate experimental procedures, researchers may not always have the necessary means to construct correct procedures; however, researchers may setup experiments to the best of their possibilities and report the shortcomings openly in manuscripts.

Some observations may justifiably be eliminated from data analysis; an example would be extreme outliers in reaction time studies. However, the criteria for data exclusions need to be pre-determined.

Despite these qualifications, the surveys discussed above may be taken as an indication that the current system is not functioning flawlessly. In the most favorable survey (Martinson et al., 2005), a third of researchers responded that they had committed questionable practices. On the survey that included psychologists only, questionable research practices seemed to be the norm (John et al., 2012; Levelt, 2012). On this, the Stapel Investigation expresses concern by noting that,

> when interviewed, several co-authors who did perform the analyses themselves, and were not all from Stapel's 'school', defended the serious and less serious violations of proper scientific method with the words: that is what I have learned

> in practice; everyone in my research environment does the same, and so does everyone we talk to at international conferences. (Levelt, 2012, p. 48 and 54)

4.3: Reasons for High Prevalence of QRPs

A question that naturally arises is why individuals who dedicate their careers to science, which is prototypically the quest for truth, engage in questionable research practices. As Broad & Wade put it.

> Fraud in science is of course the abnegation of a researcher's fundamental purpose, the search for truth. It is thus an act of considerable moment, and one that is unlikely to be taken without careful consideration of the prevailing attitudes and mores in the laboratory, as well as of the chances of getting caught. (Broad & Wade, 1982, p. 19)

There are likely different reasons for different individuals. We attempt to generalize these in the categories below, namely, incentives, lack of transparency, and lack of accountability.

4.3.1: Incentives

Already in 1961 Reif notes.

> The quest for prestige can cause conflict between the goals of science and the goals of the scientist. […] These are usually the result of conflicts between the requirements of the scientific work proper and the pressure of competition. To the individual scientists they may appear as conflicts between the values inherent in science and more selfish personal values. (Reif, 1961, p. 1957 and 1961)

In a similar vein, Broad & Wade note that,

> scientists are not different from other people. In donning the white coat at the laboratory door, they do not step aside from the passions, ambitions, and failings that animate those in other walks of life […] Not only do careerist pressures exist in contemporary science, but the system rewards the appearance of success as well as genuine achievement. (Broad & Wade, 1982, p. 19)

More recently Nosek et al. (2012) have commented on the conflicting incentives for finding true effects and finding positive effects. Nosek and colleagues maintain that the main problem in the current research environment "is that the incentives for publishable results can be at odds with the incentives for accurate results […] to the extent that publishing itself is rewarded, then it is in scientists' personal interests to publish, regardless of whether the published findings are true" (Nosek et al., 2012, p. 616).

When the aim of experimental practice is to attain a significant *p* value, there are several ways this can be achieved. First, researchers can discover true effects. Second, researchers can fabricate data and create significant *p* values. The first option is difficult, time-consuming, entails uncertainty and can put researchers at a disadvantage to those who employ the second strategy. The second strategy, although it will create many results, potentially interesting and attention grabbing, runs the risk of being uncovered and ending the career of researchers. A third strategy, a middle ground between the first two, would be to use questionable research practices. This strategy avoids the risk of straight-out fraud but still makes attaining significant *p* values much more likely and gives researchers an advantage over their competition.

Researchers who do not engage in such practices are at a disadvantage and so one can expect a race to the bottom. As we showed in Section 4.2, by simply testing for statistical significance twice after increasing the sample size from 20 to 40, a researcher increases the probability of attaining a (false) positive result by 100% (Simmons et al., 2011). This gives researchers who employ questionable practices a considerable advantage over others.

Because of this unfair advantage, commentators have drawn parallels between QRPs in research and performance enhancing drugs in sports, noting that "QRPs are the steroids of scientific competition, artificially enhancing performance and producing a kind of arms race in which researchers who strictly play by the rules are at a competitive disadvantage" (John et al., 2012, p. 524). What further makes QRPs so damaging is that "QRPs, by nature of the very fact that they are often questionable as opposed to blatantly improper, also offer considerable latitude for rationalization and self-deception" (John et al., 2012, p. 524).

Broad & Wade note in 1982.

> History shows that deceit in the annals of science is more common than is often assumed. Those who improved upon their data to make them more persuasive to others doubtless persuaded themselves that they were lying only in order to make the truth prevail. But almost invariably the real motive for the various misrepresentations in the history of research seems to arise less from a concern for truth than from personal ambition and the pursuit, as Darwin put it, of "the bauble fame." (Broad & Wade, 1982, pp. 35-36)

If false-positive results make up a majority of the published findings in psychology, it is to a great extent that competing researchers will use the tools at their disposal to advance their careers. Those who do not engage in QRPs to attain significant *p* values are at a disadvantage and unlikely to persist in their careers. From this perspective, the blame is not to fall exclusively on individual researchers but on the rules of the game (Bakker et al., 2012) or the game they find themselves in as "individual scientists have to work and survive in the system as it exists. Without systemic, structural changes, individual, principled choices . . . may be futile and professionally destructive" (Kerr, 1998, p. 213).

4.3.2: Lack of Accountability and Transparency

Since there are only a few ways of detecting data fabrication (e.g. whistleblowers) (Stroebe et al., 2012), statistical anomalies (Simonsohn, 2013) and almost no methods (other than replications) for detecting individual false-positive results that come about through QRPs, there is very little in the way of holding researchers who use questionable research practices accountable. Stapel himself is to have said that "fraud is too easy, because there are too few control mechanisms in science" (as cited in Stroebe et al., 2012, p. 681) and Stroebe et al. add to this that "people are tempted to commit fraud when the expected rewards are great and punishment is unlikely because the risk of discovery is small" (Stroebe et al. 2012, p. 681).

Currently, science is structured around a system of trust and "any trust-based system, as science is, is open to exploitation" (Stroebe et al., 2012, p. 683). There may be unique

advantages to organizing an enterprise around trust; however, there are also critical shortcomings. Many successful endeavors do not rely on trust and the onus is on those who favor a trust-based system to provide strong reasons why research science should be granted a special role.

On the side of transparency, a difficulty that prevents QRPs from being detected is that very often experimental procedures and raw data are not made available to other researchers. This prevents close examination of published work. In Paper 3 we described the difficulties we had in attaining details of experiments such as experimental designs, stimuli presented, number of questions, and participant population. This prevented us from examining the reference paper more closely and prevented us from composing a manuscript for journal submission.

In closing this section, as an answer to why there seems to be a high prevalence of QRPs, the simple answer is that currently incentives favor these practices and they are facilitated by a lack of accountability and transparency.

4.4: Concluding Remarks

When we started our project there were no good estimates on the reproducibility of published findings in psychology or experimental philosophy. Since then, the efforts by the Reproducibility Project, although still in progress, have produced some revealing results for the psychology literature (Nosek, 2012). This project has started to replicate

all articles published in three major psychology journals for 2008. The project is still in progress. At the time of writing, 24 results have been examined of which 13 failed to replicate and 11 replicated successfully, a false-positive rate of 54%. More than half of the results published in the most prestigious psychology journals turn out to be non-reproducible.

Nosek mentions as one of the motivations for the project that if the outcomes were encouraging, fears of the state of journal publishing in psychology could be put aside. Otherwise, the project would give a better sense of whether changes in the system needed to be made. A very interesting anecdote that Nosek recounts is that after sharing his idea for the project "a senior person in the field ask[ed] [Nosek] not to do it, because psychology is under threat and this could make us look bad" (Carpenter, 2012, p. 1559). Other researchers had also expressed concern that this project may put a bad light on the whole discipline (Carpenter, 2012). These sentiments are very informative. One, these individuals obviously had little confidence in the published results. If they had been confident in the literature, there would have been no need to worry about replication outcomes. These sentiments are also interesting because instead of being concerned that much of the published literature may be unreliable, these individuals were more concerned about the image of the field and by extension their own image.

The issues discussed in this paper are not recent developments and are also not unique to the current crisis. As others have pointed out "crisis is nothing new in psychology" (Giner-Sorolla, 2012, p. 563) and there are several parallels to be found between the

current and the crises of the past. The earliest reference we made to a work discussing questionable research practices was Babbage (1830) titled "The Decline of Science in England." To get a sense of how long lasting this notion of crisis in psychology has been, one of the earlier papers we found addressing a crisis of psychology was published in 1966 (Bakan, 1966). In specific, Bakan refers to the crisis being related to statistical methods used and the prevalence of Type I errors.

In 1975, Anthony Greenwald addresses many of the issues that are part of the current debate on the problems of experimental psychology. In his paper titled "Consequences of Prejudice Against the Null Hypothesis," Greenwald examines the issues of publication bias, continuation of data collection until a desired significance level is achieved, retrospectively declaring findings as hypothesized, including or excluding data from pilot studies in accord with desired outcomes, applying different standards of data analysis when looking to reject the null hypothesis, amongst other problems of the field (Greenwald, 1975).

Greenwald further concludes that "about the only way to demonstrate the existence of Type I errors conclusively is to demonstrate that "established" findings cannot be replicated and that such failures to replicate cannot easily be regarded as Type II errors" (Greenwald, 1975, p. 13). Greenwald then goes on to give several examples of effects that had been so widely accepted in the field that they were presented in many psychology textbooks, which, however, after years of acceptance could not be successfully replicated.

The topic of this paper has been the reproducibility of published findings in the scientific literature, with a strong emphasis on psychology. As we have argued, there is likely to be a serious lack of reproducibility and the Reproducibility Project, although still in progress, confirms this with concrete numbers. Simply because of statistical methods and publishing practices currently prevailing in scientific research, one can expect an overproduction of false-positive results in the published literature. The high prevalence of QRPs exacerbates this problem further.

Given the shortcomings in the current research-publication system, changes in the current organization of science may be warranted. Paper 5 examines this topic.

Page Intentionally Left Blank

Paper 5: Improving the Research-Publication System

In light of the problems discussed in Paper 4, researchers have made various proposals for tackling the shortcomings of the research-publication system – throughout the past decades as well as of late in light of recent developments. This paper reviews some of these proposals and in conclusion offers what we believe to be important components of a sustainable solution.

Before we begin this review, perhaps just as important as considering solutions to shortcomings of the research-publication system, is an emphasis that many researchers (in high positions) dismiss the idea of crisis and maintain that the current system is adequate in regulating scientific practice. In Section 5.1, we give an account of these views. This is important in understanding that there is inertia when it comes to reforming the research-publication system and that any proposal will face opposition. Section 5.2 surveys some solutions that have been suggested over the years. Section 5.3 examines one of these solutions (replications) in detail. The final section concludes with a discussion.

5.1: Rejection of Criticism

We divide this section in four parts, each discussing one reason why advocates of the status quo reject criticism of the current system. The first is simply a belief that no serious flaws exist. The second is trust in the 'self-correcting' nature of science. The

third is confidence in the peer-review publication system. The fourth is the prevalence of conceptual replications.

5.1.1: Refusal to Admit Flaws

A common view expressed is that although misconduct does occur in academic research, it is so rare as to be insignificant (Kennedy, 2006; Koshland, 1987; Kraut, 2011; LaFollette, 2000; Martinson et al., 2005; Reynolds, 2004; Shamoo & Resnik, 2003; Sovacool, 2008). This view has persisted throughout the past decades.

In 1981 at United States congressional hearings on scientific misconduct that followed four high profile cases of fraud in biomedicine, Handler, the then president of the United States National Academy of Sciences, professed that misconduct occurred very rarely (National Academy of Sciences, 1993).

This sentiment was repeated in 1987 in an editorial published by Koshland, the then editor of *Science*, in which he noted that, "we must recognize that 99.9999 percent of reports are accurate and truthful, often in rapidly advancing frontiers where data are hard to collect" (Koshland, 1987, p. 141). Koshland does not provide a reference for the "99.9999 percent" statistic but continues that "there is no evidence that the small number of cases that have surfaced require a fundamental change in procedures that have produced so much good science" (Koshland, 1987, p. 141).

In 2006, again in response to a high profile case of research fraud (the Hwang case), the editor of *Science* noted in a published statement that "fraudulent research is a particularly disturbing event, because it threatens an enterprise built on trust. Fortunately, such cases are rare" (Kennedy, 2006).

The executive director of the Association for Psychological Science published a post in *The Chronicle of Higher Education* in 2011, once again in response to a major case of fraud (the Stapel case), stating that "such egregious cases are rare, and they are harmful to the scientific enterprise. But it's important that they be recognized as the aberrations they are" (Kraut, 2011). With regard to questionable research practices, Kraut continues that "most of these flaws and concerns are undramatic—not the stuff of headlines" (Kraut, 2011). In light of the findings discussed in Section 4.2.2 of Paper 4, this is a very surprising view for the executive director of the Association for Psychological Science to hold.

In response to the Stapel Investigation, the European Association of Social Psychology published a statement rejecting the report, amongst other reasons because it draws "conclusions about a whole, international field of scientific research" by focusing on the "scientific practices and publications associated with one author" (EASP, 2012). Strack, the associate editor of *Psychological Science* commented on the Stapel Investigation,

> if you want an example for "sloppy science", take a closer look at the Levelt report [one of the reports comprising the Stapel Investigation], which is full of sweeping generalizations without clear documentation while neglecting the

> scrutiny to which it subjects and holds up social psychology. I doubt that its claims would pass peer review and editorial scrutiny. (Strack, 2012)

The statements quoted above all followed high profile cases of misconduct. In his historical account of the changing nature of misconduct, Lafollette describes these reactions as common.

> When problems have been uncovered, scientists around the world have initially tended to act much the same. They have characterized the offender as aberrant, argued that the episode is isolated, or attempted to explain it as caused by stress, bad judgment, or moral corruption (or all three)." (LaFollette, 2000, p. 212)

Already in 1982, Broad & Wade observe that,

> each time a new case of scientific fraud breaks into the headlines, the scientific establishment generally responds with one variant or another of the "bad apple" theory. The faker was a psychopath, or under great stress, or otherwise mentally disturbed, this theory goes. Its unspoken implication is that all blame should be put on the erring individual, not on the institutions of science. […] If every smidgeon of fraud can be laid at the door of the poor unhinged, deranged psychopaths who nevertheless managed somehow to infiltrate the research community, clearly there is no need for any change in the institutional mechanism whereby science is said to police itself. (Broad & Wade, 1982, p. 60)

Paper 4 pointed out that whether this bad apple view is accurate has implications on the need for major changes to the research system. Paper 4, furthermore, gave some strong indications that this bad apple view is unlikely to be correct.

5.1.2: Confidence in Scientific Practice – Self-Correction in Science

Very closely related to a refusal to admitting serious shortcomings of the research-publication system is the belief that science is self-correcting. With self-correction commentators broadly mean that published findings are tested by other scientists and results that do not hold are eliminated from the scientific literature and only findings that are reliable will stand scrutiny. This idea is often invoked in discussions on whether fundamental change of the scientific system is necessary.

For example, as Handler describes, although fraud may take place in science, it "occurs in a system that operates in an effective, democratic and self-correcting mode." This, so the argument, makes revelation of fraudulent cases unavoidable (National Academy of Sciences, 1993, p. 91). Koshland makes similar assertions and emphasizes in discussing questionable research that the "cumulative nature of science means inevitable exposure, usually in a rather short time" (Koshland, 1987, p. 141). More recently, the executive editor of *Cognitive Science* and member of the editorial board of *Cognitive Psychology* wrote in response to the fraud committed by Marc Hauser that "science is remarkably self-correcting. […] The field is able to separate the good results from the bad fairly quickly. And that is reassuring" (Markman, 2010).

Self-correction in science is not as straightforward as the above quoted statements make it. Replications, one of the main tools that could verify previous findings, are scarce across various disciplines and especially lacking in psychology. In the rare cases where replications are published in psychology, the median time is four years from date of publication of the original results. Only 10% of replications test effects that are more

than 10 years old (Pashler & Harris, 2012). Since areas of interest change quickly and once the field moves on it is very unlikely that old effects are tested, Pashler & Harris dismiss confidence in self-correction and conclude that "there is every reason to believe that the great majority of errors that do enter the literature will persist uncorrected indefinitely, given current practices" (Pashler & Harris, 2012, p. 535).

In a similar way, Nosek et al. note.

> The myth of self-correction is recognition that once published, there is no systemic ethic of confirming or disconfirming the validity of an effect. False effects can remain for decades, slowly fading or continuing to inspire and influence new research (Prinz et al., 2011). Further, even when it becomes known that an effect is false, retraction of the original result is very rare (Budd, Sievert, & Schultz, 1998; Redman, Yarandi, & Merz, 2008). Researchers who do not discover the corrective knowledge may continue to be influenced by the original, false result. We can agree that the truth will win eventually, but we are not content to wait. (Nosek et al., 2012, p. 619)

As an example from medical publishing (where this issue would seem to be of special importance), in one particular case of fraud where the published paper was retracted, the paper was still cited as a reliable source after 24 years (Korpela, 2010). A study on retractions in biomedicine concludes that although "retractions are on average occurring sooner after publication than in the past, citation analysis shows that they are not being recognised by subsequent users of the work" (Redman, Yarandi, & Merz, 2008, p. 807). For further details and examples of articles that continue to be cited in medical research after retraction, see (Budd, Sievert, & Schultz, 1998; Drury & Karamanou, 2009).

Regarding self-correction in psychology, consider a concrete case. Stapel published internationally in the most prestigious journals; however, none of his findings were revealed as implausible through the standard scientific processes. Maintaining his confidence in self-correction, Kraut (the executive director of the Association for Psychological Science) states that, "it is also worth noting that Stapel was caught. True, he did get away with his intellectual crimes for far too long, embarrassingly so, but in the end it was the suspicions of his colleagues and students that exposed him" (Kraut, 2011).

What stands out in this quote is that Kraut seems to include whistleblowing as part of the scientific method. It is true that Stapel's fraud came to light, but it was not because science performed its functions correctly. To make this claim is to stretch the definition of scientific practice.

In reviewing numerous cases of misconduct, Stroebe et al. complain that it is "disconcerting that hardly any of the fraud cases on our list were uncovered by the […] principal mechanisms of self-correction" (Stroebe et al., 2012, p. 677). Nosek similarly points out in discussing the cases of Karen Ruggiero and Marc Hauser that "if the field was truly self-correcting, why didn't we correct any single one of them?" emphasizing that "like Stapel, they were exposed by internal whistle-blowers (Yong, 2012a).

The Stapel Investigation also expresses concern over the functioning of science. We quoted this passage in Paper 4 and repeat it here because of its relevance.

> The urgent question that remains is why this fraud and the widespread violations of sound scientific methodology were never discovered in the normal monitoring procedures in science.
>
> In the case of the fraud committed by Mr. Stapel, the critical function of science has failed on all levels. Fundamental principles of scientific method have been ignored, or set aside as irrelevant. In the opinion of the Committees this has contributed significantly to the delayed discovery of the fraud. It is to the credit of the whistleblowers in Tilburg that they did discover these infringements of scientific integrity and took the correct action. (Levelt, 2012, p. 53 and 54)

5.1.3: Peer-reviewed Publication

Another argument proponents of the status quo make as to why current publication practices provide sufficient safeguards against the entry of questionable research into the scientific literature is that the manuscript review process prevents papers of low quality from being published.

Loscalso (2012), for example, states.

> The many layers of review a manuscript receives in parallel with and beyond peer review, including discussion at [..] editorial board meeting[s], careful review by associate editors, and rigorous statistical review […] while not eliminating the risk of publishing data that are irreproducible in papers that are later retracted, clearly offers the care necessary to minimize this risk. (Loscalzo, 2012, p. 1213)

A first indication that the manuscript review process does not provide sufficient safeguards against the publication of fraudulent research, let alone, questionable research practices comes from the case of Stapel, who published very prolifically and also in the most esteemed journals of his field.

The Stapel case showed that peer review and journal procedures certainly do not minimize the risk of accepting fabricated findings into the published literature. As the Investigation noted, Stapel "published in nearly all the respected international journals in his field. It was extremely rare for his extraordinarily neat findings to be subjected to serious doubt" (Levelt, 2012, p. 48).

The Levelt committee (one of the committees that was part of the Stapel Investigation) further finds harsh words for the peer review system.

> The Committees can reach no conclusion other than that from the bottom to the top there was a general neglect of fundamental scientific standards and methodological requirements.
>
> This certainly also applies to the editors and reviewers of international journals. Furthermore, many journals insist prior to publication on authors filling in forms in various variants guaranteeing correct research procedures and availability of data and survey material. Authors evidently frequently fail to comply (see among others Wicherts et al., 2006). The journals perform no further monitoring of this requirement. (Levelt, 2012, p. 53)

These issues are by no means limited to psychology research. A study of the biomedical field concludes that,

> reviewers have no time and no resources to reproduce data and to dig deeply into the presented work. As a consequence, errors often remain undetected. Adding to this problem, many initially rejected papers will subsequently be published in other journals without substantial changes or improvements. (Prinz et al., 2011b)

In 2006, after a high profile case of fraud on cloning, *Science* published a statement in which it acknowledged that the review process is not designed to detect fraud by stating that "fraud is unlikely to be eliminated completely through the process of scientific publishing, and truth in science ultimately depends upon confirmation" (Kennedy, 2006).

From a historical perspective, this is not the first time that journal practices have come under question. In the context of the 1980s crisis, "because misconduct had so often come to light after publication in a journal, questions also began to be raised about the reliability of peer review, accuracy of editorial scrutiny, and integrity of the scientific literature overall" (LaFollette, 2000, p. 213).

The manuscript review process is not only insufficient in preventing questionable research from entering the published literature, as it currently stands, journals may often be part of the problem rather than the solution.

On a general level, among the parties that are involved in the publication process – researchers, universities, funding bodies, and journals – the latter seems to be the least impacted by false-positive results and fraudulent research. In cases of fraud, researchers face severe sanctions and universities' reputations suffer. In cases of false positives, funding bodies waste resources that could be allocated to more productive projects. Journals, on the other hand, occupy a somewhat special position. Journals may have incentives to publish questionable research as long as it increases readership and impact factor. When it comes to fraud, journals are considered to be victims, although this

masks the fact that journals play an active role in reviewing research and making publication decisions. As Simonsohn notes, "journals should be embarrassed when they publish fake data, but there's no stigma. They're portrayed as the victims, but they're more like the facilitators […]. I'd like journals to take ownership of the problem and start working towards stopping it" (Simonsohn, 2012).

There have been various reports of journals encouraging behavior that is questionable at best. The Stapel Investigation, for example, notes.

> Co-authors also reported more than once in interviews with the Committees that reviewers encouraged irregular practices. For instance, a co-author stated that editors and reviewers would sometimes request certain variables to be omitted, because doing so would be more consistent with the reasoning and flow of the narrative, thereby also omitting unwelcome results. Reviewers have also requested that not all executed analyses be reported, for example by simply leaving unmentioned any conditions for which no effects had been found, although effects were originally expected. Sometimes reviewers insisted on retrospective pilot studies, which were then reported as having been performed in advance. In this way the experiments and choices of items are justified with the benefit of hindsight.
>
> Not infrequently reviews were strongly in favour of telling an interesting, elegant, concise and compelling story, possibly at the expense of the necessary scientific diligence. It is clear that the priorities were wrongly placed. It is surely simple to post all the information of relevance to an article on a website and to provide an explicit reference in the article. (Levelt, 2012, p. 53)

Apart from the Stapel Investigation, others have also expressed dismay over the review process. One researcher has complained that in the submission process reviewers ask for results to be "novel" or "interesting" but not necessarily true (Yong, 2012a). Another researcher has in part blamed the crisis of false positives in psychology on the demand by

journals to present "slightly freak-show-ish" results and the fact that "high-impact journals often regard psychology as a sort of parlour-trick area" (Yong, 2012a).

Given how journals currently operate, it is in their interest to publish papers even if there is suspicion about the reliability of findings. Journals lose little by publishing novel, highly unlikely effects that are not replicable but on the other hand stand to lose in various ways if they do not publish such papers. The inclusion of such papers in an issue will increase citations received and hence increase journal impact factor, which is generally regarded as an (if not the most) important indicator of journal quality. Since replications are scarce and so the likelihood of uncovering non-reproducible findings is small (see Paper 4 and Section 5.3.1 below), the best strategy for journals appears to be to publish novel and interesting effects regardless of reproducibility. Journals compete amongst each other and a paper that is rejected can always find its way to a competing journal that may be willing to publish, regardless of questionable practices (Prinz et al., 2011b).

Given that there are no strong penalties against journals for publishing papers containing questionable research practices but potentially great gains in citations received, it is natural for journals to publish such papers. In light of these incentive problems, it would be plausible to include a measure of replicability of articles when ranking journals. We took a closer look to see whether issues of reproducibility enter journal ranking for the remainder of this sub-section.

The dominant forms of determining journal rank are impact factor and more recently Eigenfactor/PageRank. Impact factor is the average number of citation articles in a journal receive in a given period of time. Eigenfactor assigns scores to journals according to incoming citations, giving more weight to citations from journals that have higher ratings. For details on this iterative approach, see (Bergstrom, West, & Wiseman, 2008). These two methods of ranking journals do not take reproducibility of findings into account.

Aside from purely quantitative methods of journal ranking such as impact factor and Eigenfactor, there have also been some attempts at more qualitative forms of evaluation using peer-review (Pontille & Torny, 2010).

We examined two recent cases of journal ranking formulations – the Australian Research Council (ARC) June 2008 draft ranking and the European Science Foundation (ESF); European Reference Index for the Humanities (ERIH) 2007-2008 (Pontille & Torny, 2010) – and although there are some detailed descriptions of what makes a good journal, the issue of reproducibility of published results is absent.

Both formulations have formal set of criteria that need to be met at a minimum to be included in the ratings. For example, the ESF states.

> All journals included must fulfil normal international academic standards, i.e. selection of articles is based on an objective review policy. [...] The journals must fulfil basic publishing standards (i.e. ISSN, timeliness of publication, complete

> bibliographic information for all cited references, full address information for every author (Pontille & Torny, 2010, p. 7)

Apart from minimum standards, the rankings of journals are described in detail.

> Typically an A* journal would be one of the best in its field or subfield in which to publish and would typically cover the entire field/subfield. Virtually all papers they publish will be of a very high quality. These are journals where most of the work is important (it will really shape the field) and where researchers boast about getting accepted. Acceptance rates would typically be low and the editorial board would be dominated by field leaders, including many from top institutions.
>
> […]
>
> The majority of papers in a Tier A journal will be of very high quality. Publishing in an A journal would enhance the author's standing, showing they have real engagement with the global research community and that they have something to say about problems of some significance. Typical signs of an A journal are lowish acceptance rates and an editorial board which includes a reasonable fraction of well known researchers from top institutions. (ARC, 2010)

We searched the documents for keywords 'replicability', 'reproducibility', 'replication', 'retraction', 'fabrication', 'fraud', 'misconduct' and found none of these included in the criteria. For the full text of these documents, see (ARC, 2010; ERIH, 2007).

There are some criteria that would protect against the publication of fraudulent findings or false positives, such as the requirement for a paper of being of "very high quality." However, these descriptions do not address reproducibility explicitly.

Given that reproducibility is considered central to scientific conduct (Cohen, 1994; Collins, 1992; Francis, 2012; Lamal, 1991; Nosek et al., 2012; Popper, 2002), these ranking formulae (especially the quantitative methods) do not seem to be concerned with

the ranking of scientifically qualitative work but in general simply popularity or, more generously, impact. The incentive structures that these measures create are not necessarily conducive to publication of true findings but merely publication of findings that receive the most attention.

5.1.4: Conceptual Replications

Proponents of the status quo often make the argument that conceptual replications, which are more frequent than exact replications, ensure the reliability of published results (Pashler & Harris, 2012).

Researchers distinguish between many different kinds of replication studies. Gomez et al., for example, identify 18 different types, which they then narrow down to three categories. These three are direct or exact replications, conceptual replications, and replications that use elements of both (Gómez, Juristo, & Vegas, 2010). Direct replications repeat a reference experiment as closely as possible, thereby verifying aspects of the original study. Conceptual replications attempt to reproduce an effect from a reference experiment by using different methods, conditions, or stimuli. What conceptual replications test is how generalizable a reference effect is: does the effect still hold when an experiment is conducted on a different population, using different stimuli of the same type, or by slightly changing the presentation of stimuli?

Makel et al. found that roughly 1% of papers in the top 100 psychology journals (measured by a five-year impact factor) of the past 100 plus years (from 1900 onward) were replications. Of these, 81.9% were conceptual replications, only 14% were direct replications and 4.1% included elements of both. Conceptual replications were more likely than direct replications to succeed (82.8% compared to 72.9%), although this difference was not statistically significant (Makel et al., 2012).

There are some strong incentives for researchers to carry out conceptual rather than exact replications. When conceptual replications succeed, that is, when the original experiment together with a novel variation succeeds, this is considered by researchers and journals to be a novel finding. This makes successful conceptual replications publishable. Direct replications, regardless of whether they are successful or not, have a much lower chance of being published.

However, when a conceptual replication fails, it is not necessarily informative on the robustness of the original experiment; the original result may simply not be as general and extendable as the conceptual replication attempted (Pashler & Harris, 2012). An effect may be present in some settings but not others and changes in some variables may cancel out the effect. Or the effect may simply not be strong enough to withstand additional noise. A conceptual replication, then, cannot verify the data of a reference experiment because differences in outcomes can be attributed to these additional changes (Makel et al., 2012). Nosek and colleagues paraphrase Braude in noting that a "successful conceptual replication [is] issued as evidence for the original result; a failed

conceptual replication is dismissed as not testing the original phenomenon (Braude, 1979)" (Nosek et al., 2012, p. 619).

As Pashler & Harris emphasize, "the unavoidable conclusion is that a sound assessment of a controversial phenomenon should focus first and foremost on direct replications of the original reports and not on novel variations, each of which may introduce independent ambiguities" (Pashler & Harris, 2012, p. 534). Nosek further clarifies that, "psychology would suffer if it [conceptual replication] wasn't practised but it doesn't replace direct replication. To show that 'A' is true, you don't do 'B'. You do 'A' again." Given that conceptual replications can only verify and not falsify, "conceptual replication allows weak results to support one another" (Yong, 2012a).

5.2: Solutions to the Problem of False Positives

Many solutions have been suggested and many of these may have to be implemented in order to improve the current system. Most (if not all) of the suggestions are not mutually exclusive and hence every point may be considered carefully. However, we believe that various suggestions that have been made will not be sufficient to improve the false positive crisis – at least not in isolation. This follows from a historical view as some of the solutions have been suggested for a long time. Furthermore, the proposals that have been made are all plausible, but implementation of sustainable solutions that last longer than the current crisis is the more difficult part.

Training Programs

Various commentators have suggested that more researchers be given training in recognizing and avoiding questionable research practices. One of the main responses in a 2008 ORI survey on how misconduct could be prevented was the proposal to train researchers on the subject (ORI, 2008). This seems like a logical place to start since researchers need to be aware that certain practices constitute questionable behavior. This may especially be necessary when procedures that are questionable have become the norm, as seems to be the case in some fields of psychology (John et al., 2012).

However, there are questions whether such training programs are effective at all (Funk, Barrett, & Macrina, 2007). Others have noted that such campaigns are not going to be effective if the right incentives are not in place (Nosek et al., 2012). What also speaks against this solution is that in surveys researchers show that they understand that certain practices are questionable, yet, they still report committing them (John et al., 2012), also, see Paper 4.

Transparency

We described in Paper 3 the difficulties we had in replicating one of the articles because the paper did not provide all the necessary details and the authors did not reply to our inquiries. This is one of the reasons why we have not prepared this work for publication. Making raw data, processed data, experimental materials, etc. available allows for easier examination and proofing of published results. Online storage is practically free and so

all materials could be shared at close to no cost. Journals whose papers are only accepted in these formats could be designated in a specific way and also receive higher ratings.

The Stapel Investigation uncontroversially notes that "it follows from the fundamental principles of openness and controllability that research procedures must be described in a way that allows for accurate replication of a given experiment" (Levelt, 2012, p. 51). The report continues that,

> research data that underlie psychology publications must remain archived and be made available on request to other scientific practitioners. This not only applies to the dataset ultimately used for the analysis, but also the raw laboratory data and all the relevant research material, including completed questionnaires, audio and video recordings, etc. (Levelt, 2012, p. 58)

Transparency has also been the main focus of the Center for Open Science (COS, 2014; Nosek et al., 2012). In general, there are likely to be few objections to increased transparency.

Strengthening the Position of Whistleblowers

Given that two of the recent prominent cases of misconduct (Hauser and Stapel) were uncovered by whistleblowers, some have concluded that the best way to prevent misconduct is to strengthen the position of whistleblowers and that "rather than changing the incentive system, the most efficient and effective approach is to improve fraud detection" (Stroebe et al., 2012, p. 683).

With regard to increasing protections for whistleblowers, there is likely to be little disagreement. A survey of 4000 researchers showed that the majority of respondents believed that reporting suspected cases of misconduct would be followed by retaliation and other negative consequences (Swazey, Anderson, & Louis, 1993). An ORI commissioned study of 68 actual whistleblowers showed that 70 percent faced adverse consequences subsequent to blowing the whistle (Frankel, 2000).

Although strengthening the role of whistleblowers may play a role in correcting scientific practice, it is unlikely to correct the problem of false positives that come about through questionable research practices and the use of prevailing statistical practices. This solution seems mainly directed at cases of obvious fraud. The blatant cases of misconduct that consist of fraud and fabrication that are more likely the targets of whistleblowing, probably make up a relatively small contribution to the problem of false positives. Furthermore, relying too much on whistleblowing may create an atmosphere that may not only be uncomfortable but at the same time unfavorable to collaboration.

Accountability

With a view to increasing accountability, some have suggested criminalizing scientific misconduct (Goldberg, 2003; Kuzma, 1992; Redman & Caplan, 2005; Sovacool, 2005, 2008). An immediate argument against this is that the bar of proof for criminal liability is much higher than the research community may be able to set itself (Stroebe et al., 2012). A further point that complicates the matter is that there needs to be the right

competence for prosecution and hence prosecutors may forgo the more complex cases of misconduct (Kuzma, 1992).

We do not want to dismiss this option, especially if the alternative is the "undermining of public confidence in an important public institution" (Kuzma, 1992). To us, however, the strongest argument as to why this option should be considered is that there is no reason why academic researchers should be held to a different standard from professionals in other fields who commit fraud to advance their careers. However, as with some of the other suggested solutions, making misconduct criminally liable will likely be a solution for only the most blatant cases.

Professional Associations

A suggestion we have heard from several colleagues is to have professional associations such as the American Psychological Association (APA) play a more active role. The parallel here is with medical or legal associations where doctors or lawyers have to defend their conduct before a board. The power of these associations stems from granting licenses that permit professionals such as doctors or lawyers to practice; this model could be translated to academic research.

We have doubts about the effectiveness of this solution, or rather about the willingness of these associations to play a more active role. In the 1981 congressional hearings previously mentioned, the chairman of the Investigations and Oversight Subcommittee of the House Science and Technology Committee, stated that "I cannot avoid the conclusion

[…] that one reason for the persistence of this type of problem [fraud in science] is the reluctance of people high in the science field to take these matters very seriously" (Broad & Wade, 1982, p. 11). A 2008 study by the ORI concluded that "our study calls into question the effectiveness of self-regulation" (Titus, Wells, & Rhoades, 2008). In his historical overview of the issue Lafollette writes.

> Looking back, we can also see that the scientific associations and organizations failed to respond swiftly enough to the calls for development of ethics codes and comprehensive ethics training programs […]. Had there been a concerted effort by all the major societies, similar to that which the Society for Neuroscience has undertaken, many of the harshest provisions of the regulatory apparatus could have been avoided. (LaFollette, 2000, p. 214)

Bornstein notes that,

> in spite of years worth of criticism and suggestions for improvement (e.g. Crandall, 1986; Mahoney, 1987) no commitment to improvement is being made by journals and associations. Indeed, direct challenges to associations which publish journals have been answered only by "Well, it's really not so bad." (Crandall, 1987). The fact that the scientific "establishment" has not acted in an area as easy and basic as replication simply confirms the unlikelihood that they will act about the more difficult flaws in the system. (Bornstein, 1991, p. 89 and 90)

Given these evaluations, ideally, solutions to the problem of false positives in science would be independent of professional associations.

Publication Bias

As noted in Paper 4, there have been calls for decades to correct the problem of publication bias. Greenwald, in 1975, suggests that "support for the null hypothesis must

be regarded as a research outcome that is acceptable as any other" (Greenwald, 1975, p. 16). Greenwald's suggestion is principally that research be evaluated by its quality, namely variables such as procedural correctness, sample sizes, statistical methods used and the insight the paper conveys and not merely by its outcomes. This notion is highlighted again in 1995 by Sterling in his suggestion that peer review be blind to the outcome of studies but instead be judged by the quality of research (Sterling et al., 1995). More recent calls along the same lines come from Bakker et al. (2012).

We agree with these views; however, our concern is that making calls for change is not sufficient. Many calls have been made in the past but unless there are 'structural' changes or changes in incentives, these calls will not be heeded as they have not been in the past.

Information

One aspect that separates the current from all previous crises is that the tools for information sharing and more generally information technology has changed considerably. Projects such as the Psych File Drawer (PFD, 2012) aim to take advantage by building a platform for sharing information on attempted replications. The website describes itself as an "Archive of Replication Attempts in Experimental Psychology" (PFD, 2012) where researchers can log their replication attempts of original articles.

This is a new effort and at the time of writing there are only 19 logged replication attempts. The main problem that concerns us is that replications (like any other study)

are very time consuming and perhaps more so than the original experiment. Without receiving credit for such activity as in citation counts, researchers may be reluctant to invest time and effort in contributing to this effort. The logged findings on the Psych File Drawer are not part of the published literature and so the value of the credit researchers receive is not clear.

We do not want to be negative about this project. Researchers' contribution to this website may be taken into account in job and grant applications as well as promotion considerations. This (if in fact the case) may provide sufficient incentives for researchers to contribute. We hope that the number of recorded replications on the site increases; however, regardless of the future development of this project, we believe that a platform for centralizing information is integral to any sustainable solution (see Section 5.4 for more details).

Signed Statement

One proposal we want to present because of its simplicity is merely that authors sign the following statement in their articles.

> We report how we determined our sample size, all data exclusions (if any), all manipulations, and all measures in the study. (Simmons, Nelson, & Simonsohn, 2012, p. 1)

This assures that authors did not manipulate the sample size with an eye to *p* values, that all experimental conditions were reported and that no variables were excluded; put differently, authors assure that they are not mining, cooking or concealing (Fanelli, 2009).

Typically, a requirement for proving fraud is to show intent to deceive. If researchers engage in questionable research practices and yet include this statement, intent to deceive is obvious.

The shortcomings of this solution are that it is voluntary and despite the solution's simplicity, Simmons and colleagues already report that there has been resistance by researchers to adopt this approach (Simmons et al., 2012). A further downside is that this approach is again dependent on trust and as Paper 4 showed, reliance on trust may not be warranted, given the incentives of the current research environment. Finally, the biggest weakness we see here is that this solution is not a defense against false positives that come about as a result of publication bias, Type I errors, and currently used statistical practices.

Further Solutions

There have been a plethora of other suggestions and it would not be possible to examine all of these in detail. We mention some of them for reference. An argument that has been made is that the aesthetic standards in scientific (specifically psychological) research be changed. Given that currently too many manuscripts are submitted to a limited number of journals, it is not sufficient to present well conducted studies but articles need to tell a clear and compelling story (Giner-Sorolla, 2012). This creates pressures to attain certain *p* values (i.e. a *p* value of 0.051 needs to be amended before being submitted to a journal), omit facts about cases where the effect under investigation could not be detected, or HARKing, hypothesizing after the results are known (Kerr,

1998). Other suggestions that have been made are that small studies not be considered definitive, that reporting conventions be improved, that alternative statistical tests and approaches be considered, and that there be a clear distinction between exploratory and confirmatory research (Bakker et al., 2012; Wagenmakers, Wetzels, Borsboom, van der Maas, & Kievit, 2012).

All of the above suggestions may have their place in correcting the problem of false positives in psychological science. However, among all of these possible solutions, replications may have an elevated status. This elevated status of replications is best described by the following description.

> Current controversies about professional standards and practices within psychological science at first glance involve a hodge-podge of issues, including potentially defective statistical methods, publication bias, selective reporting, and data fabrication. Nevertheless, these issues are related in a deeper sense: All flawed research practices yield findings that cannot be reproduced by studies that are specifically designed to repeat an earlier study's procedures. Such "replications" allow researchers to separate findings that are trustworthy from findings that are unreliable. A scientific discipline that invests in replication research is therefore immunized to a large degree against flawed research practices. At present, however, psychological research is rarely explicitly confirmed or disconfirmed by replications. (Koole & Lakens, 2012, p. 608)

The next section examines replications more closely.

5.3: Replications

Theoretical descriptions of scientific conduct typically place a high value on the practice of replications and reproducibility of findings. However, this importance granted to replications in theory, does not generally translate into practice.

Consider the following descriptions on the role of replications and replicability in science.

> Only when certain events recur in accordance with rules or regularities, as is the case with repeatable experiments, can our observations be tested – in principle – by anyone. We do not take even our own observations quite seriously, or accept them as scientific observations, until we have repeated and tested them. Only by such repetitions can we convince ourselves that we are not dealing with a mere isolated 'coincidence', but with events which, on account of their regularity and reproducibility, are in principle inter-subjectively testable. (Popper, 2002, p. 23)

> How do scientists establish that they have made a discovery that should be a new part of the public domain? Press scientists and in the last resort they will defend the validity of their claims by reference to the repeatability of their observations or the replicability of their experiments. […] Repeatability, or replicability (I will use the terms interchangeably), is the touchstone of common sense philosophy of science. […] Replicability, in a manner of speaking, is the Supreme Court of the scientific system. […] Replication is the scientifically institutionalized counterpart of the stability of perception. […] Thus the acceptance of replicability can and should act as a demarcation criterion for objective knowledge. (Collins, 1992, p. 18 and 19)

> One characteristic that is commonly said to distinguish the scientific method from other approaches to knowledge is its objectivity. Replication has been said by textbook writers to be a critical test of objectivity (e.g., Chaplin & Krawiec, 1979) and to be "at the heart of any science" (Hersen & Barlow, 1976, p. 317). Replication would seem to underlie the self-correction which is presumed to be another characteristic of the scientific method. Replication is necessary because our knowledge is corrigible. (Lamal, 1991, p. 31)

> Throughout all of science and especially for fields that depend on statistical data analysis, leading researchers emphasize that experimental replication is the final arbiter in determining whether effects are true or false. (Francis, 2012, p. 585)

Braude considers replications so important that he designates it as a "demarcation criterion between science and nonscience" (as cited in Nosek et al., 2012, p. 618).

These are just some of the examples on the importance of replications to the scientific process and many more can be found. However, at the same time that the extreme importance of replications and replicability is professed, one will paradoxically also find the following observations.

> A field that replicates its work is rigorous and scientifically sound, but researchers who conduct those replications are looked down on as bricklayers and not advancing knowledge. (Makel et al., 2012)
>
> There is a vague sense of disrespect for someone who is interested in doing replications. This sense of vagueness rapidly disappears when one attempts to publish the replication. Prime journals will reject it, usually with an explanation that the paper is not a contribution to new knowledge. Replications are often second-class citizens in the social science literature. (Hendrick, 1991, p. 42)
>
> One important and ironic support for the common sense view is that replication of others' findings and results is an activity that is rarely practiced! Only in exceptional circumstances is there any reward to be gained from repeating another's work. […] Thus, though scientists will cite replicability as their reason for adhering to belief in discoveries, they are infrequently uncertain enough to need, or to want, to press this idea to its experimental conclusion. For the vast majority of science replicability is an axiom rather than a matter of practice. (Collins, 1992, p. 19)
>
> There is a reluctance among social scientists in general, and among psychologists in particular, to invest time, money and energy in replication studies. This reluctance is at least partially based on publication policies of most journals in the social sciences. Journal editors clearly prefer to publish reports which show new findings. This preference, severely hurts the possibility of publishing studies which "merely" replicate results of earlier findings. The outcome is that replication research […] is rarely carried out today. Instead, it is assumed that a research result once found has continuous validity and generality. (Amir & Sharon, 1991, p. 53)

> It is evident that replication is not an essential ingredient in the cookbook of academic science. Certainly, it is added for flavoring every now and then, but that is about all. […] Replication in science is a philosophical construct, not an everyday reality. (Broad & Wade, 1982, p. 81 and 86)
>
> The lack of interest in replication is striking given its centrality to science. (Nosek et al., 2012, p. 618)
>
> Students and academics face tight constraints on time and resources, only a fool would spend effort trying to report mistakes rather than burying them or repeating someone else's work rather than promoting one's own. (Giner-Sorolla, 2012, p. 564)

Considering that theoretical descriptions of the scientific process put a high value on replications, yet in practice this activity is relegated as inferior in psychology, one is led to draw the conclusion that psychology, as it currently stands, is not concerned about proper scientific practice but something else. An (uncharitable) interpretation may be that some fields of psychology may instead be concerned about entertainment value. This view matches the previously cited comment that journals expect "slightly freak-show-ish" results and that top journals consider psychology a "parlour-trick area" (Yong, 2012a).

This sentiment of the supposed inferiority of replications has led to an extreme scarcity of publication of replications, which is one of the main factors leading to the likely high rate of false-positive results in psychology.

5.3.1: Scarcity of Replications

The scarcity of published replications has been pointed out over many years. Bornstein concludes in a 1991 essay that "in light of the fact that valid, replicable empirical findings are the basis of any field of science, it is – to say the least – somewhat disturbing to learn that replication studies are rarely published in the social science journals" (Bornstein, 1991, p. 72).

A study published in 1972 examined three psychology journals covering 1334 articles over a time period of close to four years (1967 to 1970) and found that fewer than one percent of the articles published were replications of previous work (Bozarth & Roberts, 1972). As referred to earlier, Makel et al. (2012) analyzed the rate of replication studies in the 100 top rated journals according to a five-year impact factor. As mentioned, 1.07% of articles were replications. Of these, close to 82% were conceptual replications, 14% were exact replications and the remaining contained elements of both. More than half of the replications (52.9%) were conducted by the team of researchers that published the original article; this poses clear problems of incentives and conflicts of interest. And indeed, the success rate of replications varied considerably depending on overlap of authorship. Specifically, the rate of successful replications was 91.7% when there was an overlap of authorship as compared to 64.6% when there was no overlap.[63]

The problem of lack of replications has persisted for so long that authors have even pointed out that discussion on the topic surfaces periodically, however, little changes.

[63] A reason for this discrepancy could be that the original researchers are more familiar with experimental procedures (some of which may not be explicit in published articles) and so the original group is more likely to succeed in replicating the reference effect. However, the published procedures are the ones that are relevant for the scientific record and if a group following the published procedures cannot replicate the reference findings, this is nevertheless a failure of replication as far as the scientific record is concerned.

Amir & Sharon highlight in 1991 that "the need for replication research and validation is raised every few years in articles published in leading journals, calling for changes in the research approach taken by psychologists" (Amir & Sharon, 1991, p. 55). The authors continue that despite these calls, little changes in this regard. Also in 1991 Lamal writes that "the case for replications has been made before (e.g., by Campbell & Jackson, 1979; Kazdin, 1982; Sidman, 1960; Smith, 1970; Sommer & Sommer, 1983). Unfortunately, it would be difficult to determine whether such exhortation has had much effect. There is some evidence that it has not" (Lamal, 1991, p. 31).

Given an aversion to publishing replications, a 1991 study set out to capture the views of journal editors on this topic (Neuliep & Crandall, 1991). In a survey of 288 editors, 94% of respondents said that their journals did not encourage replications and 42% said that they had never received exact replications. When these editors were asked whether a new effect was more important or a replication of a previously reported effect, 72% said that a new effect was more important (answer options included "both equally"). When asked whether a new effect was more important or a failed replication of a previously reported effect, 58% opted for new effect and 15% chose failed replication. Twenty-one percent indicated that both were equally important and the rest abstained.

Surprisingly, when asked on whom the burden of proof rested when a replication attempt failed, only 9% said that the burden of proof was with the original researchers and 29% answered that the burden lay with the replicators (24% answered both, 6% neither and 32% provided no answer). This is an unexpected outcome and hints at the low status that

replications are afforded. A possible justification would be that replicators may not be familiar with all the procedures and hence may not have been able to follow the protocol closely. However, this is a post-hoc justification that is not entailed in the survey question (also, see Footnote 63). This view towards failed replications is further disincentive for carrying out replications as researchers who fail to replicate will face some pressure to explain their results.

Neuliep and Crandall present some of the comments of the survey respondents on replications.[64] It is important when reading the below comments to recall that in theoretical descriptions of science, replication and replicability was lauded as one of the cornerstones of the scientific enterprise.

> "Dull"
> "The worst of the modern science/social science publish or perish mentality. People aren't interested in them."
> "Referees tend to judge them to have relatively low priority."
> "They seldom make new contributions to our understanding of the phenomena."
> "There needs to be a reason (conceptual, methodological, or otherwise) for conducting a replication."
> "When do you stop? Is one rep enough, or should we let someone build their career replicating the same study?"
> "Readers feel that replications are redundant and don't reflect cutting edge stuff."
> "Explicit, direct replications are often unnecessary and add little to the field's knowledge."
> "They tend to be boring and not contribute a lot."
> "Direct replication with positive outcomes and without other additional manipulations provide no new information."
> "They add little to advance our understanding of the issues."
> "They are given too much weight."

[64] Reproduced from (Neuliep & Crandall, 1991, p. 88).

It would be misplaced to put the sole blame on journal editors. Editors, to a great extent, follow the norms of their fields and their views are largely reflective of existing standards (Neuliep & Crandall, 1991). Once the prevailing opinions change, editor views and editorial policies will change as well. These values, however, will not change by themselves. There need to be changes in incentives and "until new forces come to play on editors, attempts to publish replications can expect to continue to meet strong editorial resistance" (Neuliep & Crandall, 1991, p. 90).

We want to single out one of the comments from the list provided above by Neuliep & Crandall because it possibly represents a legitimate concern. The comment we are referring to is, "When do you stop? Is one rep enough, or should we let someone build their career replicating the same study?" (Neuliep & Crandall, 1991, p. 88). If research, regardless of outcome (positive or negative results) and perceived importance were to be published (as the journal *PLOS* has made it its policy), a single study could be replicated and published numerous times. A replication would contribute little if there have been ten prior replications all of which were successful. Earlier replications are more valuable than later ones (given equal quality) and so earlier replications would ideally receive credit accordingly. A relative credit system could be worked into co-citation systems, so that incentives are low for conducting the, say, eleventh replication when all previous ones have been successful. Furthermore, platforms such as the File Psych Drawer where information is aggregated could solve the problem of too much space dedicated to replications by presenting summaries of methods, procedures, participants, results, etc.

The problem of too many replications is not a great concern of this paper and to researchers who have looked for solutions to shortcomings of the research-publication system because there never has been a situation where there have been too many published replications. Hence, this hypothetical situation will not be a focus of this paper.

5.3.2: Increasing Replication Rates

Various suggestions have been made on how to increase the rates of replications. Below, we will present some of these.

Student Projects

Given that there exist little incentives for carrying out replications and researchers face constraints on time and resources, one suggestions has been to delegate replications to students in training (Frank & Saxe, 2012). One big problem with this proposal is that it designates replications as an insignificant activity to be assigned to individuals low in academic rank. It is also very likely that replications will be even more difficult to publish if they are considered merely as training ground for students.

Replications require great attention to detail and when replications by students fail, it will be easy to reject these as the work of unskilled researchers. This justification is already used with full-time researchers and established academics (Levelt, 2012), and would be invoked even more with student work.

Journals Dedicated to Replications

One of the solutions that emerged from the 1970s crisis was that each discipline set up a journal entirely dedicated to replications (Lamal, 1991). In 1979 a journal named *Replications in Social Psychology* attempted to put this in practice; however, the journal ceased activity after three volumes.

We mostly disagree with this solution because it separates (at the very least spatially) replications from regular research findings, signaling a difference in status. Ideally, replications would be considered integral to scientific practice and given credit accordingly.

Instead of entirely separating replications to specialized journals, another suggestions has been for journals to assign space to replications of previously published work. This would signal to researchers that replications are encouraged and also rewarded with citations (Bornstein, 1991). Rewarding researchers with citations may be critical in increasing the number of replications.

A system of co-citations may offer sufficient incentives (Koole & Lakens, 2012). What this system entails is that whenever an article is cited, any available replications that exist also receive citations. Additionally, co-citation could provide a summary note of how many replications succeeded and how many failed. For example, this could be as follows, OriginalAuthor, Year, Replication: ReplicationAuthor1, ReplicationAuthor2, ReplicationSummary: 1:1, where one replication succeeded and one failed.

Pre-publication replication

In 1957 Lubin, noticing the lack of replications, suggested that manuscripts submitted for publication already contain replication attempts. These manuscripts would be regarded higher and given priority in publication decisions (Lubin, 1957). Similarly, in 1968 Lykken suggests that "ideally, all experiments would be replicated before publication" although he continues that "this goal is impractical" (Lykken, 1968, p. 159).

Given the recent problems of false positives, it may be worth considering pre-publication replications; or a variation where other laboratories replicate the work. Upon receiving manuscripts, journals could send experiment designs to reviewer labs to run replications. This is not an unreasonable suggestion in light of the fact that Stapel had many multi-study reports, as did Bem (2011). Researchers who want to find multiple false-positive results, will find a way to do so. This solution does seem somewhat impractical and it may slow down the pace of publishing. It would also likely face strong opposition from some researchers, as captured by the statement that "if each researcher had to go back and repeat the literature, the enormously productive rush of modern science would slow to a snail's pace" (Koshland, 1987, p. 141).

A less strict version would be to randomly select a percentage of manuscripts for pre-publication replication. Selecting a small percentage of manuscripts for pre-publication replication may introduce enough checks to ensure compliance to better standards by all

submitting authors, as no one would know whose manuscripts would be selected for replication.

A further advantage of pre-publication replications is that it is pro rather than retroactive. The number of retractions would be reduced and researchers would not invest time following research paths that later turn out to be non-reproducible. Finally, there is another benefit to pre-publication replications. With post-publication replication, the issue of publication bias re-appears, although in reverse. Many laboratories may attempt to replicate any given study and those that fail to replicate will naturally receive attention and be more likely to be published.

Prominent, surprising, controversial, or counter-intuitive results are more likely to be replicated, as a failure to replicate such effects would have better chances of being published. Even if the original effect is robust, given that many groups around the world are likely to attempt replication, by statistical chance alone, some will fail reproduction. Successful replications are likely to be stored away but failures of replication (because of the surprise factor) are likely to be submitted for publication. This is in effect the same as publication bias, just in 'reverse'. In 'reverse' because very often in replication studies, negative findings (failure of replication) are noteworthy and positive findings (successful replications) are non-results.

We see two objections to instituting pre-publication replications. One, as briefly mentioned, the pace of publishing would be slowed down. This may be a reasonable cost

given the prevalence of questionable research practices, false positives in the literature, and cases of fraud. Such an institution would likely improve the quality of published articles and although it may slow down publication, quality would be improved.

Furthermore, although the pace of publishing may be slowed down, the pace of scientific discovery may speed up. Researchers who engaged with Stapel and examined his work and attempted to build on those results wasted a lot of time and resources that could have been spent on other projects.

The second problem is that researchers will have to share their work prior to publication with other labs. This may not be amenable to all researchers because they lose control over information before it is published and other labs could appropriate those ideas and an important competitive advantage may be lost in many cases.

5.4: Concluding Remarks

After discussing many of the problems of the current research-publication system in Paper 4, we provided a review of some of the possible solutions in the current paper.

The earliest reference to a discussion of systemic problems we made in Paper 4 was to Babbage who highlighted the prevalence of questionable research practices in 1830 (Babbage, 1830). In the current paper, we referenced many articles from the 1950s to the present in discussing the problems of (psychological) science; not in order to provide a

historical account, but as an indication of how long these problems have persisted. Without substantial changes, these problems are likely to persist, as they have for decades. The literature reviewed here also shows that the problem is unlikely to be with a few bad apples or the thinking of a generation. The problem is more systemic.

Potential solutions need to be long lasting. In the past, after crises, researchers have emphasized replications and have called for improvement of research practices. Efforts were made and researchers spent energy in setting up new journals; however, these efforts never lasted and the next crisis always followed.

Summary of Necessary Components of a Sustainable Solution

We want to highlight what we believe to be necessary elements that any sustainable solution will incorporate. The way we see it in summary is as follows. Sufficient checks on published findings are currently lacking and this keeps the door open for false positives to flourish in the literature of various fields, especially those that use statistical tests of significance. Replications would offer an adequate check; however, researchers currently do not have the necessary incentives for conducting replications. Increased publication and citation of replication studies would offer such an incentive. One way to increase incentives for replications would be a system of co-citations, where replication studies of a reference paper are automatically co-cited whenever the original (reference) paper receives citation. Increased transparency of experimental procedures would reduce the difficulty in carrying out replications. Finally, the outcome of one replication should not be taken to be definitive. It is important to know the outcomes of the second, third,

fourth, etc. replication. So that all replication attempts are recorded – not only those that have a surprising outcome or those that come first – a platform where information on reproducibility of specific articles and experiments is centralized and archived would also be necessary. In effect, this platform would be a verification system where findings are considered preliminary until a certain number of replications have been conducted.

Importance of the Issue

In concluding this paper, we want to highlight the importance of the issue at hand. Bornstein writes in 1991 that the

> replication process in social science research leaves much to be desired. Because social scientists historically have published relatively few replication studies, the social sciences have retained many qualities of a "preparadigmatic" field (see Kuhn, 1962; Mahoney, 1985). Consequently, social science research is perceived by other scientists (and by members of the public) as being less rigorous, less robust, less replicable and less cumulative than research in other branches of science. (Bornstein, 1991, p. 80)

With current practices, psychology is at a risk of losing credibility, as "efficient and unbiased replication mechanisms are essential for maintaining high levels of scientific credibility" (Ioannidis, 2012, p. 645).

Psychology is not the only field affected by this problem. In pharmaceutical research, Amgen attempted to test the robustness of 53 "landmark" published results of pre-clinical studies and only 11% of the work was replicated successfully. The study concludes that "some nonreproducible preclinical papers had spawned an entire field, with hundreds of secondary publications that expanded on elements of the original observation, but did not

actually seek to confirm or falsify its fundamental basis" (Begley & Ellis, p. 532). Similarly, Bayer HealthCare tested the reproducibility of 67 findings and reported a success rate (fully consistent with the original findings) of replications of only 20-25% (Prinz et al., 2011b).

Both of these reports mention that informally the problem of nonreproducibility is known and discussed among academics and industry. Both also specify that this is not a problem that comes down to a few bad apples but that the problem is systemic. On the topic of questionable research practices, Begley & Ellis go as far as to say that the "academic system and peer-review process tolerates and perhaps even inadvertently encourages such conduct" (Begley & Ellis, p. 533).

The papers mention as reasons for non-reproducibility, issues familiar to the experimental psychology literature; among these are publication bias, statistical methods, pressure to publish, and lack of replications.

One thing that is revealing is that private industry has little confidence in results obtained from academic research. Companies that invest in pharmaceutical projects assume that half of the results from academia are not reproducible (Osherovich, 2011).

These problems may be worse in psychology. Comparing psychology to other fields, a *Nature* article notes.

> These problems occur throughout the sciences, but psychology has a number of deeply entrenched cultural norms that exacerbate them. It has become common practice, for example, to tweak experimental designs in ways that practically guarantee positive results. And once positive results are published, few researchers replicate the experiment exactly, instead carrying out 'conceptual replications' that test similar hypotheses using different methods. This practice […] builds a house of cards on potentially shaky foundations. (Yong, 2012a)

We started out this paper discussing views that reject criticism of the current research-publication system. We reviewed these sentiments as an indication that any proposed changes to the current system will face obstacles. In light of the discussions of Paper 4, we do not have much confidence in these views. Given the shortcomings of the current research-publication system (Paper 4) many proposals for improvement have been made. We reviewed some of these in Section 5.2 and highlighted which components we believe to be fundamental for any sustainable solutions. These included transparency, different incentives, information sharing and check on published findings. Central to these discussions was replication. We would like to end this paper by addressing replications with a quote, whose importance, we think, cannot be emphasized enough.

> The importance of original studies is not their originality per se, but their epistemological force. So also with replications, their importance is in terms of their epistemological import. (Lamal, 1991, p. 32)

Page Intentionally Left Blank

Discussion and Conclusion

Discussion

Reliability of Replication Findings

Experimental philosophy has gained attention by producing surprising results. The aim of Papers 1-3 of this thesis was to test the reproducibility of some of these results. Of all the studies that we conduct and of all the data that we analyzed, most of the reference findings were not reproducible. As a brief summary, without the replication work, the following would be considered genuine effects: individuals of different ethnic backgrounds have different epistemic intuitions (discussed in Paper 1); women and men have different intuitions on various types of philosophical questions (discussed in Paper 2); moral intuitions are easily influenced using certain manipulations (discussed in Paper 3).

The question that naturally arises is whether we could have had flaws in our procedures that brought about the null findings. First, not all of our replication attempts failed and not all of our experiments produced null findings. In Papers 1-3 we reported several positive results (significant effects). We reported all analyses[65] that we conducted (whether in support of our hypotheses or not) for the examined effects and so we are clear of omission of reporting. This made some of the presentations less straightforward and perhaps less elegant; luckily, the journal reviewers did not hold this against us.

[65] There are two studies (Beebe & Buckwalter, 2010; Machery et al., 2004) for which we analyzed the replication data but which we have not reported so far. In both cases we did not attain significant *p* values to reproduce the reference findings. However, we are reluctant to call these failures of replication because our samples were relatively small and furthermore, the direction of the data was similar to that of the original articles.

We collected data on more experiments; however, we have not analyzed these. The reason we have not analyzed all collected data is simply due to time constraints.

One of the problems in experimental psychology that has led to an abundance of false positives is that among studies conducted, researchers often selectively report outcomes of interest. In the context of our work, the equivalent of that practice would be to only mention studies that yielded negative results. By reporting all studies, we are staying clear of this practice.

Second, we followed the procedures of the original articles as closely as possible. In many cases we attained greater statistical power than the original experiments, yet still did not detect the investigated effects. Given the greater power of our studies, our procedures were more likely to reveal effects, had these existed.

Third, the strongest indication that our results are robust is that since we made our work public, other groups have attempted to replicate the original effects we examined and these groups also could not reproduce the reference findings (Adleberg et al., 2014; Minsun & Yuan, ms). Furthermore, Nagel et al. (2013), independently of our work, report a failed conceptual replication of Weinberg et al. (2001) on the effect of ethnicity on epistemic intuitions. Nagel and colleagues also report no gender differences on epistemic intuitions (Nagel et al., 2013). There is another study (Turri, 2013) that can be considered a failed conceptual replication of Weinberg et al. (2001) on the effect of

ethnicity on epistemic intuitions.[66] These independent verifications strongly suggest that our findings are reliable.

In any instance a single replication should not be considered definitive. Ideally, any effect should be verified by different researchers independently. This is one of the reasons that our reports are multi-studies and why we very early on urged other researchers to attempt replications of the original effects (or to replicate the replication findings). Unfortunately, currently replications are only publishable if they have a surprise factor, if they are somehow unexpected. The reason our replications were published is because the findings came as a surprise to those familiar with the literature and also because before us others had not carried out this kind of work.

Trend in Replications

We mentioned in the Introduction that since we made our results public in 2012, a trend in replications has started in experimental philosophy, with findings of other exact replications being made public in 2014 (Adleberg et al., 2014; Kvanvig, 2014; Minsun & Yuan, ms). We would like to show by an example why this has been important. On March 31, 2014 *Episteme* published a paper by Colaço et al. reporting differences in epistemic intuitions depending on age. The study reports that "the intuition that fake-barn cases do count as knowledge is negatively correlated with age; older participants are less likely than younger participants to attribute knowledge in fake-barn cases" (Colaço et al., 2014, p. 199).

[66] As pointed out before, Turri (2013) is not a straightforward conceptual replication of Weinberg et al. (2001). See Footnote 17.

Just one week later, Jon Kvanvig posted an entry on a blog with the results of this paper, remarking that successful replication was not guaranteed. Kvanvig ended the post with the note "Next: attempts to replicate?" (Kvanvig, 2014). In the comments section, several researchers started discussing how to go about replicating the effect and by June 24, 2014, Joshua Knobe reported two failed replications with a combined sample size of over 500 individuals. Colaço et al., themselves, submitted a post to a different blog on the same day, reporting on the failed replications (Colaço, 2014).

Without the trend in replications and without the swift call to replication of Colaço et al. (2014), we would have likely seen much time and space dedicated to explaining why age would influence epistemic intuitions, whether intuitions can be trusted and what this means for philosophical methodology in general. We would have seen competing theories explain this effect, proponents and opponents of IAE (Intuition as Evidence) argue whether this means that IAE needs to be abandoned, and so on. Luckily, the non-reproducibility was established quickly and there was no need for these discussions.

Implications of Failed Replications

The next question we want to address is what the failures of replication that we (as well as now others) reported means for the experimental philosophy movement. For the replications that we conducted, the rate of failed replications to successful replications is disconcerting; however, we did not select papers for replication randomly. A note on what criteria we used to select papers may be in order.

We did not select the weakest papers, that is, papers that were most likely not to reproduce. Our first priority was to test a diverse selection of effects (ethnicity, socioeconomic background, gender, moral intuitions). As discussed in the Introduction, we expected many of the studies to replicate successfully. Buckwalter & Stich (2013) appeared without flaws and the sample sizes were relatively large. Nothing indicated that the results were unreliable. Valdesolo & DeSteno (2006) and Zhong et al. (2010) also reported sufficiently large samples and their procedures also did not indicate any apparent flaws. Weinberg et al. (2001) had been the subject of extensive debate and we believed it likely that it had been previously replicated successfully.

A further reason for choosing the articles discussed in Papers 1-3 was that these had been influential and widely circulated amongst experimental philosophers. If these effects were not reproducible, we believed it important that readers of the literature were aware of it.

Nevertheless, we did have some reasons for believing that the reference papers needed further scrutiny. For example, the sample sizes in Weinberg et al. (2001) were relatively small and this could have opened the door to sampling errors. We also did not see why ethnicity or socioeconomic background by themselves would impact epistemic intuitions. We had some suspicion that language proficiency could have had an effect when examining ethnic background; however, further studies have ruled this out (Minsun & Yuan, ms). Gender differences on scenarios such as Brain in the Vat or Twin-Earth did

not seem very plausible to us in the absence of a strong explanation or in fact any explanation, which the authors of the original study could not provide. As another group has argued, there is evidence that Buckwalter & Stich (2013) conducted several studies and only reported those that yielded statistically significant differences (Adleberg et al., 2014).

The effects tested in this thesis represent a subset of the results of the experimental philosophy literature. What other published and established findings are not reproducible has to be seen from future replications. There may be many or only few other findings that are not reproducible. Paper 4 gave some indication that it should not come as a surprise if it were the former. However, that is an empirical question that needs to be tested. It is not exactly clear how damning the failed replications presented in this thesis are for experimental philosophy. We believe that at a minimum, these failed replications give some reason to be more careful going forward and to devise ways to assure the reproducibility of published findings (see Paper 5).

The Research-Publication System

After presenting our empirical findings in Papers 1-3, Paper 4 reviewed some of the shortcomings of the current research-publication system that allows for the publication of high rates of false-positive results. In Paper 4, we also tried to give a sense of how long lasting these problems have been by citing discussion of the issue going as far back as 1830.

In light of the problems discussed in Paper 4, Paper 5 reviewed some of the solutions that researchers have suggested over the years. We believe that replications will be an important component of any sustainable solution; however, we also believe that incentive structures need to change to make conducting replications more appealing. Researchers in general and philosophers in specific will not simply start conducting more replications merely because of their dedication to science or true results. The last century in experimental psychology and the last decade and a half (the entire lifespan) of experimental philosophy have demonstrated this. The components that we believe to be important for any sustainable solution are: more checks on published findings through increased rates of replications, more incentives to conduct replications through co-citations, centralization of information on replications, and increased transparency in order to facilitate replications.

Conclusion

Various fields in the natural and social sciences currently face a 'crisis of confidence'. This crisis amounts to a pervasiveness of false-positive results in the published literature. To mention just a few, areas that have recently received attention include biomedicine (Begley & Ellis, 2012a; Ioannidis, 2005; Prinz et al., 2011b), economics and political science (Dafoe, 2013; Gherghina & Katsanidou, 2013; Herndon, Ash, & Pollin, 2014) as well as psychology (Pashler & Wagenmakers, 2012).

The importance of the issue seems difficult to overemphasize. In biomedicine, potential treatments may be delayed or completely missed, scarce funding wasted, and in general

the pace of progress slowed as researchers embark on paths that later turn out to be non-replicable (Begley & Ellis, 2012a; Prinz et al., 2011b). In economics and political science, policies may be based on flawed findings, wasting resources and potentially slowing economic growth (Lowrey, 2013). Some fields of psychology are at a risk of losing all credibility as a result of an excess of false-positive findings. The state of publishing in psychology has led one prominent researcher to claim that "the prevalence of unchallenged fallacies may represent even up to 95% (if not more) of the significant findings in some areas of the psychological literature" (Ioannidis, 2012, p. 650). Others who have studied the topic have similarly concluded that the majority of published findings may be false-positives (Pashler & Harris, 2012). Concrete data that supports this estimate comes from the Reproducibility Project (Nosek, 2012).

Recently, philosophers have started using the tools of experimental psychology to study philosophical questions. Experimental philosophy has attracted great attention, essentially for producing results that seem highly counter-intuitive and at the same time question some of the fundamental methods used in philosophy. For this thesis, we set out to systematically replicate some of the findings in the experimental philosophy literature and as it turned out, some of the most cited and attention grabbing papers in the field (Buckwalter & Stich, 2013; Weinberg et al., 2001) turned out to be non-reproducible (Seyedsayamdost, 2014, forthcoming).

This development suggests several things. First, the high occurrence of false positives in psychology and other fields does not seem related to localized issues such as research

culture or a few ‘bad apples’ (Sovacool, 2008). The problem appears to be more systemic and is (most likely) related to the incentive structures and fundamental methods of empirical research as currently practiced, especially for areas that use statistical methods. Experimental philosophy in itself provides a case study or an experiment for this hypothesis. A very young field in its early stages, starting from scratch, quickly ran into the same problems that various fields of psychology had to deal with for decades. As it stands, current research practices keep the door open for non-replicable results; be it simply as a result of standard statistical procedures and publication practices (Ioannidis, 2005; Pashler & Harris, 2012) or questionable research practices (Fanelli, 2009; John et al., 2012; Martinson et al., 2005).

The second point is that experimental philosophy, like other empirical fields, needs a better system to test for robustness of published findings. Replications, the most direct way of verifying published findings, are central for this purpose; however, currently replications are scarce. Theoretical descriptions of scientific practice place a high value on replications and consider reproducibility integral to science (Collins, 1992; Lamal, 1991; Nosek et al., 2012; Popper, 2002); however, in practice replications are often considered inferior to original findings (Collins, 1992; Hendrick, 1991; Makel et al., 2012; Neuliep & Crandall, 1991). In psychology, many agree that replications are critical in lowering the rate of false positives (Amir & Sharon, 1991; Koole & Lakens, 2012; Nosek et al., 2012; Pashler & Harris, 2012; Ritchie et al., 2012b) and calls for more replications have been made frequently during various crises of the past decades.

However, without changes to the incentive structures, these calls have remained unanswered.

In closing this thesis, we draw two conclusions from the work presented in Papers 1 to 5. The first is that the instability of intuitions has been exaggerated by experimental philosophers. Intuitions appear to be more uniform across demographic groups. Whether intuitions should be considered legitimate data points in philosophical theorizing is a different question; however, the argument that intuitions need to be discarded because they depend on arbitrary factors such as ethnicity, socioeconomic background, or gender does not seem tenable anymore.

The second conclusion is that experimental philosophy, like some other empirical fields, needs a better system to test for the reproducibility of published findings. As it stands, current research and publication practices lead to an overproduction of false-positive findings in the published literature. Unless changes are made to the research-publication system, this overproduction is likely to continue: in experimental philosophy as well as other disciplines.

Page Intentionally Left Blank

References

Adleberg, T., Thompson, M., & Nahmias, E. (2014). Do men and women have different philosophical intuitions? Further data. *Philosophical Psychology*(ahead-of-print), 1-27.

Alcock, J. (2011). Back from the future: Parapsychology and the Bem affair. *Skeptical Inquirer, 35*(2).

Alexander, J. (2010). Is experimental philosophy philosophically significant? *Philosophical Psychology, 23*(3), 377-389. doi: Doi 10.1080/09515089.2010.490943

Alexander, J. (2012). *Experimental philosophy: An introduction*. Cambridge, UK ; Malden, MA: Polity.

Alexander, J., Mallon, R., & Weinberg, J. M. (2010). Accentuate the negative. *Review of Philosophy and Psychology, 1*(2), 297-314.

Alexander, J., & Weinberg, J. M. (2007). Analytic epistemology and experimental philosophy. *Philosophy Compass, 2*(1), 56-80. doi: 10.1111/j.1747-9991.2006.00048.x

Amir, Y., & Sharon, I. (1991). Replication research: A "Must" for the scientific advancement of psychology. In J. W. Neuliep (Ed.), *Replication research in the social sciences* (pp. 51-71). Newbury Park: Sage Publications.

Anderson, C. A., Lindsay, A. J., & Bushman, B. J. (1999). Research in the psychological laboratory: Truth or triviality? *Current Directions in Psychological Science, 8*(1), 3-9. doi: 10.1111/1467-8721.00002

ARC. (2010). Tiers for the Australian ranking of journals. Retrieved July 18, 2014, from http://www.arc.gov.au/era/tiers_ranking.htm

Babbage, C. (1830). *Reflections on the decline of science in England, and on some of its causes*: B. Fellowes.

Bakan, D. (1966). Test of significance in psychological research. *Psychological Bulletin, 66*(6), 423-&. doi: Doi 10.1037/H0020412

Bakewell, S. (2013). Clang went the trolley. Retrieved May 1, 2014, from http://www.nytimes.com/2013/11/24/books/review/would-you-kill-the-fat-man-and-the-trolley-problem.html?pagewanted=all&_r=0

Bakker, M., van Dijk, A., & Wicherts, J. M. (2012). The rules of the game called psychological science. *Perspectives on Psychological Science, 7*(6), 543-554. doi: Doi 10.1177/1745691612459060

Bartlett, T. (2010). Cleanliness is next to priggishness. Retrieved May 18, 2014, from https://chronicle.com/blogPost/Cleanliness-Is-Next-to-Prig/24084/

Bealer, G. (1996). A priori knowledge and the scope of philosophy. *Philosophical Studies, 81*(2-3), 121-142.

Bealer, G. (2000). A theory of the a priori. *Pacific Philosophical Quarterly, 81*(1), 1-30.

Beebe, J. R., & Buckwalter, W. (2010). The epistemic side-effect effect. *Mind & Language, 25*(4), 474-498. doi: 10.1111/j.1468-0017.2010.01398.x

Begley, C. G., & Ellis, L. M. (2012a). Drug development: Raise standards for preclinical cancer research. *Nature, 483*(7391), 531-533.

Begley, C. G., & Ellis, L. M. (2012b). Raise standards for preclinical cancer research. Retrieved May 29, 2013, from http://www.nature.com/nature/journal/v483/n7391/full/483531a.html
Bem, D. J. (2011). Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology, 100*(3), 407-425. doi: Doi 10.1037/A0021524
Bengson, J. (2013). Experimental attacks on intuitions and answers. *Philosophy and Phenomenological Research, 86*(3), 495-532.
Bergstrom, C. T., West, J. D., & Wiseman, M. A. (2008). The Eigenfactor™ metrics. *The Journal of Neuroscience, 28*(45), 11433-11434.
Bishop, M., & Trout, J. D. (2005). The pathologies of standard analytic epistemology. *Nous, 39*(4), 696-714.
Bornstein, R. F. (1991). Publication politics, experimenter bias and the replication process in social science research. In J. W. Neuliep (Ed.), *Replication research in the social sciences* (pp. 71-83). Newbury Park: Sage Publications.
Bozarth, J. D., & Roberts, R. R. (1972). Signifying significant significance. *American Psychologist, 27*(8), 774-&. doi: Doi 10.1037/H0038034
Braude, S. E. (1979). *ESP and psychokinesis : A philosophical examination*. Philadelphia: Temple University Press.
Brean, J. (2010). Explaining the 'Trolley Problem'. Retrieved May 1, 2014, from http://news.nationalpost.com/2010/11/27/explaining-the-trolley-problem/
Broad, W., & Wade, N. (1982). Betrayers of the truth: Fraud and deceit in the halls of science: New York: Simon & Schuster.
Buckwalter, W. (2010). Knowledge isn't closed on Saturday: A study in ordinary language. *Review of Philosophy and Psychology, 1*(3), 395-406.
Buckwalter, W., & Stich, S. (2013). Gender and philosophical intuition. In J. Knobe & S. Nichols (Eds.), *Experimental philosophy* (Vol. 2). Oxford: Oxford University Press.
Budd, J. M., Sievert, M., & Schultz, T. R. (1998). Phenomena of retraction: reasons for retraction and citations to the publications. *JAMA, 280*(3), 296-297.
Burnyeat, M. (1990). *The Theaetetus of Plato*: Hackett Publishing.
Cappelen, H. (2012). *Philosophy without intuitions*: Oxford University Press.
Carey, B. (2011, January 5, 2011). Journal's paper on ESP expected to prompt outrage, *New York Times*. Retrieved from http://www.nytimes.com/2011/01/06/science/06esp.html?pagewanted=all
Carpenter, S. (2012). Psychology's bold initiative. *Science, 335*(6076), 1558-1561.
Carruthers, P., Stich, S., & Laurence, S. (2008). The innate mind, vol. III, foundations and the future.
Chudnoff, E. (2011). The nature of intuitive justification. *Philosophical Studies, 153*(2), 313-333.
Cohen, J. (1994). The earth is round (p<. 05). *American Psychologist, 49*(12), 997.
Colaço, D. (2014). More on fake-barn intuitions: Replications of Colaço et al. Retrieved July 9, 2014, from http://philosophycommons.typepad.com/xphi/2014/06/more-on-fake-barn-intuitions-replications-of-colaco-et-al.html

Colaço, D., Buckwalter, W., Stich, S., & Machery, E. (2014). Epistemic intuitions in fake-barn thought experiments. *Episteme, 11*(02), 199-212. doi: doi:10.1017/epi.2014.7

Collins, H. M. (1992). *Changing order : Replication and induction in scientific practice*. Chicago: University of Chicago Press.

Cooper, H., DeNeve, K., & Charlton, K. (1997). Finding the missing science: The fate of studies submitted for review by a human subjects committee. *Psychological Methods, 2*(4), 447-452. doi: Doi 10.1037//1082-989x.2.4.447

Cooper, J. (2012). On fraud, deceit and ethics. *Journal of Experimental Social Psychology*.

COS. (2014). Center for Open Science. Retrieved July 27, 2014, from http://centerforopenscience.org/

Coursol, A., & Wagner, E. E. (1986). Effect of positive findings on submission and acceptance rates - a note on metaanalysis bias. *Professional Psychology: Research and Practice, 17*(2), 136-137. doi: 10.1037/0735-7028.17.2.136

Cramwinckel, F. M., De Cremer, D., & van Dijke, M. (2013). Dirty hands make dirty leaders?! The effects of touching dirty objects on rewarding unethical subordinates as a function of a leader's self-interest. *Journal of business ethics, 115*(1), 93-100.

Cramwinckel, F. M., van Dijk, E., Scheepers, D., & van den Bos, K. (2013). The threat of moral refusers for one's self-concept and the protective function of physical cleansing. *Journal of Experimental Social Psychology, 49*(6), 1049-1058.

Cullen, S. (2010). Survey-driven romanticism. *Review of Philosophy and Psychology, 1*(2), 275-296.

Cummins, R. (1998). Reflection on reflective equilibrium. In M. R. DePaul & W. Ramsey (Eds.), *Rethinking intuition: The psychology of intuition and its role in philosophical inquiry* (pp. 113-129): Rowman & Littlefield Publishers, Inc.

Dafoe, A. (2013). Science deserves better: The imperative to share complete replication files. *Available at SSRN 2318223*.

Dahlberg, J. (2012). Findings of research misconduct. Retrieved September 17, 2013, 2013, from https://www.federalregister.gov/articles/2012/09/06/2012-21992/findings-of-research-misconduct

Deutsch, M. (2009). Experimental philosophy and the theory of reference. *Mind & Language, 24*(4), 445-466.

Deutsch, M. (2010). Intuitions, counter-examples, and experimental philosophy. *Review of Philosophy and Psychology, 1*(3), 447-460.

Diep, F. (2013). Number of published cancer studies that can't be reproduced is shockingly high. Retrieved May 29, 2013, from http://www.popsci.com/science/article/2013-05/half-cancer-scientists-have-been-unable-reproduce-studies-survey-finds

Doris, J. M. (2005). Replies: Evidence and sensibility. *Philosophy and Phenomenological Research, 71*(3), 656-677.

Dowell, J. (2008). Empirical metaphysics: the role of intuitions about possible cases in philosophy. *Philosophical Studies, 140*(1), 19-46.

Drury, N. E., & Karamanou, D. M. (2009). Citation of retracted articles: a call for vigilance. *The Annals of thoracic surgery, 87*(2), 670.

EASP. (2012). EASP statement on Levelt. Retrieved September 28, 2013, 2013, from http://www.easp.eu/news/Statement%20EASP%20on%20Levelt_December_%202012.pdf

ERIH. (2007). ERIH summary guidelines. Retrieved January 6, 2012, from http://www.esf.org/index.php?eID=tx_nawsecuredl&u=0&file=fileadmin/be_user/research_areas/HUM/Documents/ERIH/ERIH%20summary_guidelines_Sept_07.pdf&t=1273323337&hash=0507fa86a5cb6038fa52192539a0e959

Evans, K., Rotello, C. M., Li, X., & Rayner, K. (2009). Scene perception and memory revealed by eye movements and receiver-operating characteristic analyses: Does a cultural difference truly exist? *The Quarterly Journal of Experimental Psychology, 62*(2), 276-285.

Fanelli, D. (2009). How many scientists fabricate and falsify research? A systematic review and meta-analysis of survey data. *Plos One, 4*(5). doi: Artn E5738 Doi 10.1371/Journal.Pone.0005738

Fanelli, D. (2010). "Positive" results increase down the hierarchy of the sciences. *Plos One, 5*(3). doi: ARTN e10068 DOI 10.1371/journal.pone.0010068

Fang, F. C., & Casadevall, A. (2012). Reforming science: structural reforms. *Infect Immun, 80*(3), 897-901. doi: 10.1128/IAI.06184-11

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*(2), 175-191.

Fayard, J. V., Bassi, A. K., Bernstein, D. M., & Roberts, B. W. (2009). Is cleanliness next to godliness? Dispelling old wives' tales: Failure to replicate Zhong and Liljenquist (2006). *Journal of Articles in Support of the Null Hypothesis, 6*(2).

Feltz, A. (2008). Problems with the appeal to intuition in epistemology. *Philosophical Explorations, 11*(2), 131-141.

Feltz, A. (2009a). Experimental philosophy. *Analyse & Kritik, 31*(2).

Feltz, A. (2009b). Experimental philosophy. *Analyse und Kritik-Zeitschrift fur Sozialwissenschaften, 31*(2), 201.

Francis, G. (2012). The psychology of replication and replication in psychology. *Perspectives on Psychological Science, 7*(6), 585-594. doi: Doi 10.1177/1745691612459520

Frank, M. C., & Saxe, R. (2012). Teaching replication. *Perspectives on Psychological Science, 7*(6), 600-604. doi: Doi 10.1177/1745691612460686

Frankel, M. S. (2000). Scientific societies as sentinels of responsible research conduct. *Experimental Biology and Medicine, 224*(4), 216-219.

Funk, C. L., Barrett, K. A., & Macrina, F. L. (2007). Authorship and publication practices: Evaluation of the effect of responsible conduct of research instruction to postdoctoral trainees. *Accountability in research, 14*(4), 269-305. doi: 10.1080/08989620701670187

Gendler, T. S. (2007). Philosophical thought experiments, intuitions, and cognitive equilibrium. *Midwest studies in philosophy, 31*(1), 68-89.

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

Gherghina, S., & Katsanidou, A. (2013). Data availability in political science journals. *European Political Science*.

Giner-Sorolla, R. (2012). Science or art? How aesthetic standards grease the way through the publication bottleneck but undermine science. *Perspectives on Psychological Science, 7*(6), 562-571. doi: Doi 10.1177/1745691612457576

Goldberg, D. (2003). Research fraud: A sui generis problem demands a sui generis solution (plus a little due process). *Thomas M. Cooley Law Review, 20*(47).

Goldman, A. I. (2007). Philosophical intuitions: Their target, their source, and their epistemic status. *Grazer Philosophische Studien, 74*(1), 1-26.

Goldman, A. I., & Pust, J. (1998). Philosophical theory and intuitional evidence. In M. R. DePaul & W. Ramsey (Eds.), *Rethinking intuition: The psychology of intuition and its role in philosophical inquiry* (pp. 179-201): Rowman & Littlefield Publishers, Inc.

Gollwitzer, M., & Melzer, A. (2012). Macbeth and the joystick: Evidence for moral cleansing after playing a violent video game. *Journal of Experimental Social Psychology, 48*(6), 1356-1360.

Gómez, O. S., Juristo, N., & Vegas, S. (2010). *Replications types in experimental disciplines.* Paper presented at the Proceedings of the 2010 ACM-IEEE International Symposium on Empirical Software Engineering and Measurement.

Gopnik, A., & Schwitzgebel, E. (1998). Whose concepts are they, anyway? The role of philosophical intuition in empirical psychology. In M. R. DePaul & W. Ramsey (Eds.), *Rethinking intuition: The psychology of intuition and its role in philosophical inquiry* (pp. 75-95): Rowman & Littlefield Publishers, Inc.

Greene, J. D., Morelli, S. A., Lowenberg, K., Nystrom, L. E., & Cohen, J. D. (2008). Cognitive load selectively interferes with utilitarian moral judgment. *Cognition, 107*(3), 1144-1154.

Greene, J. D., Nystrom, L. E., Engell, A. D., Darley, J. M., & Cohen, J. D. (2004). The neural bases of cognitive conflict and control in moral judgment. *Neuron, 44*(2), 389-400.

Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science, 293*(5537), 2105-2108.

Greenwald, A. G. (1975). Consequences of prejudice against null hypothesis. *Psychological Bulletin, 82*(1), 1-19. doi: Doi 10.1037/H0076157

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

Gutting, G. (1998). "Rethinking intuition": A historical and metaphilosophical introduction. In M. R. DePaul & W. Ramsey (Eds.), *Rethinking intuition: The psychology of intuition and its role in philosophical inquiry* (pp. 3-17). Boston: Rowman & Littlefield Publishers, Inc.

Hauser, M., Cushman, F., Young, L., Kang-Xing Jin, R., & Mikhail, J. (2007). A dissociation between moral judgments and justifications. *Mind & Language, 22*(1), 1-21.

Helzer, E. G., & Pizarro, D. A. (2011). Dirty liberals! Reminders of physical cleanliness influence moral and political attitudes. *Psychological Science, 22*(4), 517-522.

Hendrick, C. (1991). Replications, strict replications, and conceptual replications: Are they important? In J. W. Neuliep (Ed.), *Replication research in the social sciences* (pp. 41-51). Newbury Park: Sage Publications.

Herndon, T., Ash, M., & Pollin, R. (2014). Does high public debt consistently stifle economic growth?: A critique of Reinhart and Rogoff. *Cambridge Journal of Economics, 38*(2), 257-279.
Ichikawa, J. (2012). Experimentalist pressure against traditional methodology. *Philosophical Psychology, 25*(5), 743-765. doi: Doi 10.1080/09515089.2011.625118
Ichikawa, J. (2014). Who needs intuitions? Two experimentalist critiques. In A. Booth & D. Rowbottom (Eds.), *Intuitions* (pp. 232-255): Oxford University Press.
Ioannidis, J. P. A. (2005). Why most published research findings are false. *Plos Medicine, 2*(8), 696-701. doi: ARTN e124 DOI 10.1371/journal.pmed.0020124
Ioannidis, J. P. A. (2012). Why science is not necessarily self-correcting. *Perspectives on Psychological Science, 7*(6), 645-654. doi: Doi 10.1177/1745691612464056
Jarrett, C. (2010). Feeling clean makes us harsher moral judges. Retrieved April 10, 2014, from http://bps-research-digest.blogspot.co.uk/2010/08/feeling-clean-makes-us-harsher-moral.html
John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science, 23*(5), 524-532. doi: Doi 10.1177/0956797611430953
Kauppinen, A. (2007). The rise and fall of experimental philosophy. *Philosophical Explorations, 10*(2), 95-118.
Kennedy, D. (2006). Science editorial statement Retrieved September 15, 2013, 2013, from http://www.sciencemag.org/site/feature/misc/webfeat/hwang2005/kennedy_20060110_transcript.pdf
Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review, 2*(3), 196-217. doi: 10.1207/s15327957pspr0203_4
Knobe, J. (2007). Experimental Philosophy. *Philosophy Compass, 2*(1), 81-92. doi: 10.1111/j.1747-9991.2006.00050.x
Knobe, J., & Nichols, S. (2008). *Experimental philosophy*: Oxford University Press.
Knobe, J., & Nichols, S. (Eds.). (2013). *Experimental philosophy* (Vol. 2). Oxford: Oxford University Press.
Koole, S. L., & Lakens, D. (2012). Rewarding replications: A sure and simple way to improve psychological science. *Perspectives on Psychological Science, 7*(6), 608-614. doi: Doi 10.1177/1745691612462586
Koppl, R. (2011). Against representative agent methodology. *The Review of Austrian Economics, 24*(1), 43-55.
Kornblith, H. (1998). The role of intuition in philosophical inquiry: An account with no unnatural ingredients. In M. R. DePaul & W. Ramsey (Eds.), *Rethinking intuition: The psychology of intuition and its role in philosophical inquiry* (pp. 129-143). Boston: Rowman & Littlefield Publishers, Inc.
Korpela, K. (2010). How long does it take for the scientific literature to purge itself of fraudulent material?: The Breuning case revisited. *Current Medical Research & Opinion, 26*(4), 843-847.
Koshland, D. E. (1987). Fraud in science. *Science, 235*(4785), 141-141. doi: 10.1126/science.3798097

Kraut, A. (2011). Despite occasional scandals, science can police itself. Retrieved September 15, 2013, 2013, from http://chronicle.com/article/Despite-Occasional-Scandals/129997/

Kuzma, S. M. (1992). Criminal liability for misconduct in scientific research. *University of Michigan Journal of Law Reform, 25*(2), 357-421.

Kvanvig, J. (2014). More x-phi on fake barn intuitions. Retrieved July 9, 2014, from http://certaindoubts.com/more-x-phi-on-fake-barn-intuitions/#comments

LaFollette, M. C. (2000). The evolution of the "Scientific Misconduct" issue: An historical overview. *Proceedings of the Society for Experimental Biology and Medicine, 224*(4), 211-215. doi: 10.1111/j.1525-1373.2000.22423.x

Lamal, P. A. (1991). On the importance of replication. In J. W. Neuliep (Ed.), *Replication research in the social sciences* (pp. 31-37). Newbury Park: Sage Publications.

Lee, S. W., & Schwarz, N. (2010). Dirty hands and dirty mouths embodiment of the moral-purity metaphor is specific to the motor modality involved in moral transgression. *Psychological Science, 21*(10), 1423-1425.

Lee, S. W., & Schwarz, N. (2011). Wiping the slate clean psychological consequences of physical cleansing. *Current Directions in Psychological Science, 20*(5), 307-311.

Levelt, W. J. M. (2012). Flawed science: The fraudulent research practices of social psychologist Diederik Stapel *Stapel Investigation* (pp. 104).

Levin, J. (2005). The evidential status of philosophical intuition. *Philosophical Studies, 121*(3), 193-224.

Liao, S. M. (2008). A defense of intuitions. *Philosophical Studies, 140*(2), 247-262.

Liljenquist, K., Zhong, C.-B., & Galinsky, A. D. (2010). The smell of virtue clean scents promote reciprocity and charity. *Psychological Science, 21*(3), 381-383.

Lipsey, M. W., & Wilson, D. B. (1993). The efficacy of psychological, educational, and behavioral treatment - confirmation from metaanalysis. *American Psychologist, 48*(12), 1181-1209. doi: 10.1037//0003-066x.48.12.1181

Loscalzo, J. (2012). Experimental irreproducibility: Causes, (mis)interpretations, and consequences. *Circulation, 125*(10), 1211-1214. doi: 10.1161/circulationaha.112.098244

Löwe, B., Müller, T., & Müller-Hill, E. (2009). Mathematical knowledge: a case study in empirical philosophy of mathematics. *Philosophical Perspectives on Mathematical Practice*.

Lowrey, A. (2013, April 16, 2013). A study that set the tone for austerity is challenged, *New York Times*.

Lu, Z., Daneman, M., & Reingold, E. M. (2008). Cultural differences in cognitive processing style: Evidence from eye movements during scene processing.

Lubin, A. (1957). Replicability as a publication criterion. *American Psychologist, 12*(4), 519-520. doi: 10.1037/h0039746

Lykken, D. T. (1968). Statistical significance in psychological research. *Psychological Bulletin, 70*(3p1), 151.

Machery, E., Mallon, R., Nichols, S., & Stich, S. (2004). Semantics, cross-cultural style. *Cognition, 92*(3), B1-B12.

Makel, M. C., Plucker, J. A., & Hegarty, B. (2012). Replications in psychology research: How often do they really occur? *Perspectives on Psychological Science, 7*(6), 537-542. doi: Doi 10.1177/1745691612460688

Mallon, R., Machery, E., Nichols, S., & Stich, S. (2009). Against arguments from reference. *Philosophy and Phenomenological Research, 79*(2), 332-356.

Markman, A. (2010). Why science is self-correcting. Retrieved June 22, 2013, 2013, from http://www.psychologytoday.com/blog/ulterior-motives/201008/why-science-is-self-correcting

Marszalek, J. M., Barber, C., Kohlhart, J., & Holmes, C. B. (2011). Sample size in psychological research over the past 30 years. *Perceptual and Motor Skills, 112*(2), 331-348. doi: 10.2466/03.11.pms.112.2.331-348

Martinson, B. C., Anderson, M., & De Vries, R. (2005). Scientists behaving badly. *Nature, 435*(7043), 737-738.

Meyer, G. J., Finn, S. E., Eyde, L. D., Kay, G. G., Moreland, K. L., Dies, R. R., . . . Reed, G. M. (2001). Psychological testing and psychological assessment: A review of evidence and issues. *American Psychologist, 56*(2), 128-165. doi: 10.1037//0003-066x.56.2.128

Miellet, S., Zhou, X., He, L., Rodger, H., & Caldara, R. (2010). Investigating cultural diversity for extrafoveal information use in visual scenes. *Journal of Vision, 10*(6), 21.

Minsun, K., & Yuan, Y. (ms). *No cross-cultural differences in Gettier Car Case intuition: A replicaiton study of Weinberg et al. 2001*.

Nadelhoffer, T., & Nahmias, E. (2007). The past and future of experimental philosophy. *Philosophical Explorations, 10*(2), 123-149.

Nagel, J. (2012). Intuitions and experiments: A defense of the case method in epistemology. *Philosophy and Phenomenological Research, 85*(3), 495-527.

Nagel, J. (2013). Defending the evidential value of epistemic intuitions: A reply to Stich. *Philosophy and Phenomenological Research, 87*(1), 179-199.

Nagel, J., Juan, V. S., & Mar, R. A. (2013). Lay denial of knowledge for justified true beliefs. *Cognition, 129*(3), 652-661.

Nahmias, E. (2013). Do Women Have Different Philosophical Intuitions than Men? Responding to Buckwalter and Stich. Retrieved September 1, 2014, from http://philosophyofbrains.com/2013/07/15/do-women-have-different-philosophical-intuitions-than-men-responding-to-buckwalter-and-stich.aspx#comments

National Academy of Sciences, N. A. o. E., and Institute of Medicine. . (1993). Responsible science. Volume II: Background papers and resource documents: The National Academies Press.

Neuliep, J. W., & Crandall, R. (1991). Editorial bias against replication research. In J. W. Neuliep (Ed.), *Replication research in the social sciences* (pp. 31-37). Newbury Park: Sage Publications.

Nichols, S. (2004). Folk concepts and intuitions: From philosophy to cognitive science. *Trends in Cognitive Sciences, 8*(11), 514-518.

Nisbett, R. E., Peng, K., Choi, I., & Norenzayan, A. (2001). Culture and systems of thought: holistic versus analytic cognition. *Psychological review, 108*(2), 291.

Nosek, B. A. (2012). Reproducibility Project. Retrieved September 15, 2013, 2013, from https://opensciencefframework.org/project/EZcUj/statistics
Nosek, B. A., Spies, J. R., & Motyl, M. (2012). Scientific utopia: II. Restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science, 7*(6), 615-631. doi: Doi 10.1177/1745691612459058
Oransky, I. (2012). Retraction three for Dirk Smeesters. Retrieved September 17, 2013, 2013, from http://retractionwatch.wordpress.com/2012/12/01/retraction-three-for-dirk-smeesters/
Oransky, I. (2013). Retraction eight appears for social psychologist Lawrence Sanna. Retrieved September 17, 2013, 2013, from http://retractionwatch.wordpress.com/2013/01/11/retraction-eight-appears-for-social-psychologist-lawrence-sanna/
ORI. (2005). Public health service policies on research misconduct (D. o. H. a. H. Services, Trans.) (Vol. 70). Federal Register.
ORI. (2008). Observing and reporting suspected misconduct in biomedical research.
Osbeck, L. M. (1999). Conceptual problems in the development of a psychological notion of "Intuition". *Journal for the theory of social behaviour, 29*(3), 229-249.
Osherovich, L. (2011). Hedging against academic risk. *Science-Business eXchange, 4*(15).
Pardo, M. S. (2005). The field of evidence and the field of knowledge. *Law and Philosophy, 24*(4), 321-392.
Pashler, H., & Harris, C. R. (2012). Is the replicability crisis overblown? Three arguments examined. *Perspectives on Psychological Science, 7*(6), 531-536. doi: Doi 10.1177/1745691612463401
Pashler, H., & Wagenmakers, E.-J. (2012). Editors' introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspectives on Psychological Science, 7*(6), 528-530. doi: 10.1177/1745691612465253
PFD. (2012). Psych File Drawer. Retrieved July 27, 2014, from http://www.psychfiledrawer.org/
PLOS. (2013). PLOS ONE journal information. Retrieved September 25, 2013, 2013, from http://www.plosone.org/static/information
Pontille, D., & Torny, D. (2010). The controversial policies of journal ratings: evaluating social sciences and humanities. *Research Evaluation, 19*(5), 347-360. doi: 10.3152/095820210x12809191250889
Popper, K. R. (2002). *The logic of scientific discovery*. London: Routledge.
Prelec, D. (2004). A Bayesian truth serum for subjective data. *Science, 306*(5695), 462-466. doi: DOI 10.1126/science.1102081
Prinz, F., Schlange, T., & Asadullah, K. (2011a). Believe it or not: how much can we rely on published data on potential drug targets? *Nature Reviews Drug Discovery, 10*(9), 712-712.
Prinz, F., Schlange, T., & Asadullah, K. (2011b). Believe it or not: how much can we rely on published data on potential drug targets? *Nature Reviews Drug Discovery, 10*(9), 712-U781. doi: Doi 10.1038/Nrd3439-C1
Ramsey, W. (1992). Prototypes and conceptual analysis. *Topoi, 11*(1), 59-70.
Rayner, K., Castelhano, M. S., & Yang, J. (2009). Eye movements when looking at unusual/weird scenes: are there cultural differences? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35*(1), 254.

Rayner, K., Li, X., Williams, C. C., Cave, K. R., & Well, A. D. (2007). Eye movements during information processing tasks: Individual differences and cultural effects. *Vision research, 47*(21), 2714-2726.
Redman, B. K., & Caplan, A. L. (2005). Off with their heads: The need to criminalize some forms of scientific misconduct. *Journal of Law Medicine & Ethics, 33*(2), 345-348. doi: 10.1111/j.1748-720X.2005.tb00498.x
Redman, B. K., Yarandi, H. N., & Merz, J. F. (2008). Empirical developments in retraction. *Journal of Medical Ethics, 34*(11), 807-809.
Reif, F. (1961). The competitive world of the pure scientist: The quest for prestige can cause conflict between the goals of science and the goals of the scientist. *Science, 134*(349), 1957-&. doi: DOI 10.1126/science.134.3494.1957
Reynolds, S. M. (2004). ORI findings of scientific misconduct in clinical trials and publicly funded research, 1992-2002. *Clinical Trials, 1*(6). doi: 10.1191/1740774504cn048oa
Richard, F. D., Bond, C. F., & Stokes-Zoota, J. J. (2003). One hundred years of social psychology quantitatively described. *Review of General Psychology, 7*(4), 331-363. doi: 10.1037/1089-2680.7.4.331
Ritchie, S. J., Wiseman, R., & French, C. C. (2012a). Failing the future: Three unsuccessful attempts to replicate Bem's 'Retroactive facilitation of recall' effect. *Plos One, 7*(3). doi: e33423 10.1371/journal.pone.0033423
Ritchie, S. J., Wiseman, R., & French, C. C. (2012b). Replication, replication, replication. *Psychologist, 25*(5), 346-348.
Rouder, J. N., & Morey, R. D. (2011). A Bayes factor meta-analysis of Bem's ESP claim. *Psychonomic Bulletin & Review, 18*(4), 682-689. doi: 10.3758/s13423-011-0088-7
SESP. (2013). Retrieved February 6, 2013, from http://www.sesp.org/awards.htm
Seyedsayamdost, H. (2012a). On gender and philosophical intuition: Failure of replication and other negative results. *Social Science Research Network*. doi:10.2139/ssrn.2166447
Seyedsayamdost, H. (2012b). On normativity and epistemic intuitions: Failure to detect differences between ethnic groups. *Social Science Research Network*. doi:10.2139/ssrn.2168530
Seyedsayamdost, H. (2012c). On normativity and epistemic intuitions: Failure to detect differences between socioeconomic groups. *Social Science Research Network*. doi:10.2139/ssrn.2190525
Seyedsayamdost, H. (2014). On gender and philosophical intuition: Failure of replication and other negative results. *Philosophical Psychology*(ahead-of-print), 1-32.
Seyedsayamdost, H. (forthcoming). On normativity and epistemic intuitions : Failure of replication. *Episteme*.
Shafir, E. (1998). Philosophical intuitions and cognitive mechanisms. In M. R. DePaul & W. Ramsey (Eds.), *Rethinking intuition: The psychology of intuition and its role in philosophical inquiry* (pp. 59-75): Rowman & Littlefield Publishers, Inc.
Shamoo, A. E., & Resnik, D. B. (2003). *Responsible conduct of research*: Oxford University Press.
Shermer, M. (2011, May 3, 2011). Extrasensory pornception: Doubts about a new paranormal claim. *Scientific American*.

Shieber, J. (2010). On the nature of thought experiments and a core motivation of experimental philosophy. *Philosophical Psychology, 23*(4), 547-564.
Simmons, J., Nelson, L., & Simonsohn, U. (2011). False-positive psychology undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22*(11), 1359-1366.
Simmons, J., Nelson, L., & Simonsohn, U. (2012). A 21 word solution. *Available at SSRN 2160588*.
Simonsohn, U. (2012). The data detective. Interview by Ed Yong. *Nature, 487*(7405), 18-19. doi: 10.1038/487018a
Simonsohn, U. (2013). Just post it: The lesson from two cases of fabricated data detected by statistics alone. *Psychological Science*. doi: 10.1177/0956797613480366
Singh, G. (2012). A clean self is morally obnoxious self. Retrieved April 10, 2014, from http://www.examiner.com/article/a-clean-self-is-morally-obnoxious-self
Smith, N. C. (1970). Replication studies: A neglected aspect of psychological research. *American Psychologist, 25*(10), 970.
Solon, O. (2010). Study shows clean people feel morally superior. Retrieved April 10, 2014, from http://www.wired.co.uk/news/archive/2010-08/27/cleanliness-godliness
Sosa, D. (2006). Scepticism about intuition. *Philosophy, 318*, 633.
Sosa, E. (1998). Minimal intuition. In M. R. DePaul & W. Ramsey (Eds.), *Rethinking intuition: The psychology of intuition and its role in philosophical inquiry* (pp. 257-271): Rowman & Littlefield Publishers, Inc.
Sosa, E. (2007). Experimental philosophy and philosophical intuition. *Philosophical Studies, 132*(1), 99-107.
Sosa, E. (2009). A defense of the use of intuitions in philosophy *Stich and his critics* (pp. 101-112).
Sovacool, B. K. (2005). Using criminalization and due process to reduce scientific misconduct. *American Journal of Bioethics, 5*(5), W1-W7. doi: 10.1080/15265160500313242
Sovacool, B. K. (2008). Exploring scientific misconduct: Isolated individuals, impure institutions, or an inevitable idiom of modern science? *Journal of Bioethical Inquiry, 5*(4), 271-282. doi: 10.1007/s11673-008-9113-6
St James-Roberts, I. (1976a). Are researchers trustworthy? *New Scientist, 71*, 481-483.
St James-Roberts, I. (1976b). Cheating in science? *New Scientist, 72*, 466-469.
Starmans, C., & Friedman, O. (2009). *Is knowledge subjective? A sex difference in adults' epistemic intuitions*. Paper presented at the 6th Biennial Meeting of the Cognitive Development Society, San Antonio, TX. http://www.cogdevsoc.org/prog2009/CDS09Program.pdf
Sterling, T. D. (1959). Publication decisions and their possible effects on inferences drawn from tests of significance—or vice versa. *Journal of the American Statistical Association, 54*(285), 30-34. doi: Doi 10.2307/2282137
Sterling, T. D., Rosenbaum, W. L., & Weinkam, J. J. (1995). Publication decisions revisited: The effect of the outcome of statistical tests on the decision to publish and vice-versa. *American Statistician, 49*(1), 108-112. doi: Doi 10.2307/2684823
Stich, S. (1988). Reflective equilibrium, analytic epistemology and the problem of cognitive diversity. *Synthese, 74*(3), 391-413.

Stich, S. (2001). Plato's method meets cognitive science. *Free Inquiry, 21*(2), 36-38.
Stich, S. (2010). Experimental philosophy and the bankruptcy of "The Great Tradition". Retrieved April 30, 2014, from http://www.cohnitz.net/Frege/Stich2010/presio2/index.html
Stich, S. (2013). Do different groups have different epistemic intuitions? A reply to Jennifer Nagel. *Philosophy and Phenomenological Research, 87*(1), 151-178.
Strack, F. (2012, November 28, 2012). Final Report. Retrieved September 28, 2013, from http://news.sciencemag.org/people-events/2012/11/final-report-stapel-affair-points-bigger-problems-social-psychology
Stroebe, W., Postmes, T., & Spears, R. (2012). Scientific misconduct and the myth of self-correction in science. *Perspectives on Psychological Science, 7*(6), 670-688. doi: Doi 10.1177/1745691612460687
Swazey, J., Anderson, M., & Louis, K. (1993). Ethical problems in academic research: A survey of doctoral candidates and faculty raises important questions about the ethical environment of graduate education and research. *American Scientist, 81*(6), 542.
Symons, J. (2008). Intuition and philosophical methodology. *Axiomathes, 18*(1), 67-89.
Tett, R. P., Meyer, J. P., & Roese, N. J. (1994). Applications of meta-analysis: 1987–1992.
Thomson, J. J. (1986). *Rights, restitution, and risk: Essays, in moral theory*: Harvard University Press.
Titus, S. L., Wells, J. A., & Rhoades, L. J. (2008). Repairing research integrity. *Nature, 453*(7198), 980-982. doi: 10.1038/453980a
Tobia, K., Chapman, G. B., & Stich, S. (2013). Cleanliness is next to morality, even for philosophers. *Journal of Consciousness Studies, 20*(11-12).
Turner, E. H., Matthews, A. M., Linardatos, E., Tell, R. A., & Rosenthal, R. (2008). Selective publication of antidepressant trials and its influence on apparent efficacy. *New England Journal of Medicine, 358*(3), 252-260. doi: 10.1056/NEJMsa065779
Turri, J. (2013). A conspicuous art: Putting Gettier to the test. *Philosophers Imprint, 13*(10), 1-16.
Valdesolo, P., & DeSteno, D. (2006). Manipulations of emotional context shape moral judgment. *Psychological Science, 17*(6), 476-477.
Wagenmakers, E.-J., Wetzels, R., Borsboom, D., Kievit, R., & van der Maas, H. L. J. (ms). *Yes, psychologists must change the way they analyze their data: Clarifications for Bem, Utts, and Johnson (2011).*
Wagenmakers, E.-J., Wetzels, R., Borsboom, D., & van der Maas, H. L. J. (2011). Why Psychologists Must Change the Way They Analyze Their Data: The Case of Psi: Comment on Bem (2011). *Journal of Personality and Social Psychology, 100*(3), 426-432. doi: 10.1037/a0022790
Wagenmakers, E.-J., Wetzels, R., Borsboom, D., van der Maas, H. L. J., & Kievit, R. (2012). An agenda for purely confirmatory research. *Perspectives on Psychological Science, 7*(6), 632-638. doi: 10.1177/1745691612463078
Weatherson, B. (2003). What good are counterexamples? *Philosophical Studies, 115*(1), 1-31.

Weinberg, J. M., Nichols, S., & Stich, S. (2001). Normativity and epistemic intuitions. *Philosophical Topics, 29*(1&2), 429-460.
Weiss, J. (2008). The Batman, the Joker, and the Trolley Problem. Retrieved May 1, 2014, from http://religionblog.dallasnews.com/2008/07/the-batman-the-joker-and-the-t.html/
Wetzels, R., Matzke, D., Lee, M. D., Rouder, J. N., Iverson, G. J., & Wagenmakers, E.-J. (2011). Statistical evidence in experimental psychology: An empirical comparison using 855 *t* tests. *Perspectives on Psychological Science, 6*(3), 291-298. doi: 10.1177/1745691611406923
Wild, K. W. (1938). *Intuition*: Cambridge University Press.
Williamson, T. (2004). Philosphical 'intuitions' and scepticism about judgement. *Dialectica, 58*(1), 109-153. doi: 10.1111/j.1746-8361.2004.tb00294.x
Williamson, T. (2011). Philosophical expertise and the burden of proof. *Metaphilosophy, 42*(3), 215-229.
Wisniewski, E. J. (1998). The psychology of intuition. In M. R. DePaul & W. Ramsey (Eds.), *Rethinking intuition: The psychology of intuition and its role in philosophical inquiry* (pp. 45-59): Rowman & Littlefield Publishers, Inc.
Wright, J. C. (2010). On intuitional stability: The clear, the strong, and the paradigmatic. *Cognition, 115*(3), 491-503.
Yong, E. (2012a). Replication studies: Bad copy. *Nature, 485*(7398), 298.
Yong, E. (2012b). Uncertainty shrouds psychologist's resignation. Retrieved September 17, 2013, 2013, from http://www.nature.com/news/uncertainty-shrouds-psychologist-s-resignation-1.10968
Zamzow, J. L., & Nichols, S. (2009). Variations in ethical intuitions. *Philosophical Issues, 19*(1), 368-388. doi: 10.1111/j.1533-6077.2009.00164.x
Zhong, C.-B., & Liljenquist, K. (2006). Washing away your sins: Threatened morality and physical cleansing. *Science, 313*(5792), 1451-1452. doi: 10.1126/science.1130726
Zhong, C.-B., Strejcek, B., & Sivanathan, N. (2010). A clean self can render harsh moral judgment. *Journal of Experimental Social Psychology, 46*(5), 859-862.
Zhong, C.-B., Strejcek, B., & Sivanathan, N. (2011). A clean self can render harsh moral judgment. In A. Q. Acton (Ed.), Issues in Experimental Psychology: 2011 Edition. Atlanta: ScholarlyEditions.
Zhou, J., Gotch, C., Zhou, Y., & Liu, Z. (2008). Perceiving an object in its context—is the context cultural or perceptual? . *Journal of Vision, 8*(12), 2.

Page Intentionally Left Blank

Appendix

Appendix A

Data Set 1

Participants

In order to determine participants' ethnic backgrounds, we used the relevant questions from Richard Nisbett's demographic instrument. In the interest of keeping the survey short, we did not use all of the questions; for example, we did not ask about SAT or ACT scores or annual salary but instead restricted our questions mostly to those specifically aimed at identifying ethnic background. These included the following questions: self-identified ethnicity (if the response to this was white/Caucasian, we further inquired about specific origin, i.e. Eastern Europe, Middle East/West Asia), native language, place of birth, place of birth of parents and grandparents.

In order to be classified as W, participants had to self-identify as white/Caucasian and be born in the EU (there were no participants from Iceland or Switzerland) or US (there were no participants from Australia or Canada). We furthermore asked about family's background and excluded participants who were of Eastern European and generally non-Western (for example, West Asian) background. At least one parent and two grandparents had to have EU or US as their place of birth. In all, only eight participants did not indicate all their parents and grandparents to have been born in the EU/US.

Excluding these participants from analysis did not change the outcomes and in fact increased *p* values. Two participants mistakenly indicated their birthplace as 1991 (their birth year); we included these among Ws since all their grandparents and parents were born in the EU or US and their native language was English.

In order to be classified as EA or SC, participants had to self-identify as East Asian (China, Korea, Japan) or South Asian (Bangladesh, India, Pakistan – there were no participants from Nepal), respectively, in the ethnicity part of the questionnaire. Furthermore, at most one parent and one grandparent could have been born in the EU or US. Among EAs, none of the participants had any parents or grandparents who were born in the EU or US; among SCs, four individuals had one parent born in EU/US and one participant had exactly one grandparent born in EU/US. As before, excluding these participants did not change the final outcomes and in fact increased *p* values. For the EA sample, nine participants were born in the West (Australia, UK, or US); however, excluding these participants increased *p* values slightly. For the SC sample, out of 35 participants (there was one non-response), 23 were born in the West (all in the UK), leaving only 12 participants born internationally. Comparing the latter sample with the W sample is not very meaningful because of the small sample size. Nevertheless, a comparison did not yield a significant difference, although *p* values did decrease.

Data Set 2

Participants

We tested two scenarios in this data set, namely Truetemp and Conspiracy. For Truetemp, we used the same criteria as in Data Set 1 to categorize participants. There were two Canadian individuals in this sample who were born in Canada with native language English/French whom we included as W. None of the EA and SC individuals had either a parent or grandparent born in the EU or US with the exception of one EA who indicated one grandparent to have been born in the EU/US. This participant was born in China with Chinese as her/his native language. For the EA sample, two participants were born in the West (UK, US); excluding these from analysis did not change significance. Five individuals in the SC sample were born in the West (all EU); excluding these from analysis did not make a difference and *p*-exact remained at 1.000.

For Conspiracy, there was a problem with the database and about half of the participants' birth places were not recorded. We used the other pieces of information to categorize participants. As before, participants had to self-identify as East Asian or South Asian to be included in these categories. None of the EA and SC individuals had any parents or grandparents born in the EU or US with the exception of one individual who indicated that one of their parents and one of their grandparents were born in the EU or US. For the available data, no EA participants were born in the West and in the SC sample only two were born in the West. Excluding these from analysis did not change significance.

Data Set 3

Participants

SM collects demographic information on individuals who sign up to participate. We asked SM to send out invitations to individuals of white/Caucasian background and individuals of Asian background. SM does not classify among different regions of Asia, so we used our own demographic questionnaire to filter for East and South Asian participants. Being in SM's white/Caucasian category did not automatically categorize respondents as Westerners. For example, West Asians who were in SM's white/Caucasian category were not classified as Western. We relied on our own questionnaire to categorize participants; however, we used SM's categorization to narrow down the target audience. Ethnicity was self-identified as a response to the question "how would you describe your ethnic background?" Additionally, we asked for native language and used this information for categorizing participants as well. For example, if someone self-identified as East Asian, however, indicated their native language as Hindi or Vietnamese, this person was classified as South Asian or Southeast Asian, respectively. Only one participant among the white/Caucasian group indicated their native language as non-English (it was German); all others indicated English as their native language.

Data Set 4

Participants

For this data set we used similar criteria to Data Set 3; we used self-identification to categorize participants and further used information on birthplace and native language wherever available to correct for obvious mistakes.

Appendix B

Breakdown of the individual SurveyMonkey Surveys

Brain in the Vat

Survey 1 (longer survey)

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

Survey 2 (shorter survey)

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

Twin Earth

Survey 1

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

Survey 2

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

Chinese Room

Survey 1

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

Survey 2

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

Plank of Carneades

Survey 1

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

Survey 2

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

Appendix C

This appendix contains the analyses as carried out in section 2.2.2 with the exception that participants who indicated that they had seen the scenarios were excluded. The data presented here is from the Mechanical Turk samples. Our samples from the SurveyMonkey data sets were not large enough after filtering.

Brain in the Vat

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

Twin Earth

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

Chinese Room

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

Plank of Carneades

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

Appendix D

Zebra Case

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

REALLY KNOWS ONLY BELIEVES

Conspiracy Case

It's clear that smoking cigarettes increases the likelihood of getting cancer. However, there is now a great deal of evidence that just using nicotine by itself without smoking (for instance, by taking a nicotine pill) does not increase the likelihood of getting cancer. Jim knows about this evidence and as a result, he believes that using nicotine does not increase the likelihood of getting cancer. It is possible that the tobacco companies dishonestly made up and publicized this evidence that using nicotine does not increase the likelihood of cancer, and that the evidence is really false and misleading. Now, the tobacco companies did not actually make up this evidence, but Jim is not aware of this fact. Does Jim really know that using nicotine doesn't increase the likelihood of getting cancer, or does he only believe it?

REALLY KNOWS ONLY BELIEVES

Page Intentionally Left Blank