Estimating the reproducibility of Experimental Philosophy

Florian Cova

Centre Interfacultaire en Sciences Affectives, Université de Genève, Switzerland

Address for correspondence: Dr. Florian Cova, Swiss Center for Affective Sciences, Campus Biotech, CISA – University of Geneva, Chemin des Mines, 9, 1202 Geneva, Switzerland

florian.cova@gmail.com

Tel: +41 22 379 98 09

Fax: +41 22 379 06 10

Brent Strickland

Département d'Etudes Cognitives, Ecole Normale Supérieure, France

Institut Jean Nicod, CNRS, France

stricklandbrent@gmail.com

Angela Abatista

Faculté de Psychologie et des Sciences de l'Education, Université de Genève, Switzerland

Angela.Abatista@unige.ch

Aurélien Allard

Laboratoire des Théories du Politique, Université Paris 8 Vincennes, France

aurelien.ab.allard@gmail.com

James Andow

University of Reading, United Kingdom

Jamesandow@gmail.com

Mario Attie

Department of Philosophy, Yale University, United States

mario.attie@yale.edu

James Beebe

University at Buffalo, United States

jbeebe2@buffalo.edu

Renatas Berniūnas

Department of General Psychology, Vilnius University, Lithuania

renatasberniunas@gmail.com

Jordane Boudesseul

Universidad de Lima, Instituto de Investigación Científica, Lima, Peru

jmj.boudesseul@gmail.com

Matteo Colombo

Tilburg center for Logic, Ethics and Philosophy of Science, Tilburg University, Netherlands

M.Colombo@uvt.nl

Fiery Cushman

Department of Psychology, Harvard University, United States

cushman@fas.harvard.edu

Rodrigo Diaz

University of Bern, Switzerland

rodrigo.diaz@philo.unibe.ch

Noah N'Djaye Nikolai van Dongen

Tilburg University, Netherlands

n.djaye@gmail.com

Vilius Dranseika

Department of Logic and History of Philosophy, Faculty of Philosophy, Vilnius University, Lithuania

vilius.dranseika@fsf.vu.lt

Brian Earp

Departments of Philosophy and Psychology, Yale University

brian.earp@gmail.com

Antonio Gaitán Torres

Departamento de Humanidades, Universidad Carlos III de Madrid, Spain

agaitan@hum.uc3m.es

Ivar Hannikainen

Pontifical Catholic University of Rio de Janeiro

ivar.hannikainen@gmail.com

José V. Hernández-Conde

Department of Linguistics and Basque Studies, University of the Basque Country, Spain

jhercon@gmail.com

Wenjia Hu

Lawrence University, United States

wenjia.hu@lawrence.edu

François Jaquet

Centre Interfacultaire en Sciences Affectives, Université de Genève, Switzerland

Francois.Jaquet@unige.ch

Kareem Khalifa

Philosophy Department, Middlebury College, United States

kkhalifa@middlebury.edu

Hanna Kim

Washington and Jefferson College, United States

hkim@washjeff.edu

Markus Kneer

University of Zurich, Switzerland

markus.kneer@gmail.com

Joshua Knobe

Yale University, United States

joshua.knobe@yale.edu

Miklos Kurthy

University of Sheffield, United Kingdom

mkurthy1@sheffield.ac.uk

Anthony Lantian

Laboratoire Parisien de Psychologie Sociale, UPL, Université Paris Nanterre, France

anthony.lantian@parisnanterre.fr

Shen-yi Liao

Department of Philosophy, University of Puget Sound, United States

liao.shen.yi@gmail.com

Edouard Machery

Department of History and Philosophy of Science, Center for Philosophy of Science, University of Pittsburgh, United States

machery@pitt.edu

Tania Moerenhout

Department of Philosophy and Moral Sciences and Department of Family Medicine and Primary Health Care, University of Ghent, Belgium

tania.moerenhout@ugent.be

Christian Mott

Yale University, United States

cjmott@gmail.com

Mark Phelan

Lawrence University, United States

phemark@gmail.com

Jonathan Phillips

Department of Psychology, Harvard University, United States

phillips01@g.harvard.edu

Navin Rambharose

Lawrence University, United States

navin.rambharose@lawrence.edu

Kevin Reuter

Institute of Philosophy, University of Bern, Switzerland

kevin.reuter@philo.unibe.ch

Felipe Romero

Tilburg University, Netherlands

F.Romero@uvt.nl

Paulo Sousa

Queen's University Belfast, United Kingdom

paulo.sousa@qub.ac.uk

Jan Sprenger

Center for Logic, Language and Cognition, Department of Philosophy and Educational Sciences, University of Turin, Italy

jan.sprenger@unito.it

Emile Thalabard

Sciences, normes, décision (FRE 3593), Université Paris-Sorbonne, France

emilio.thalabard@laposte.net

Kevin Tobia

Yale University, United States

kevin.tobia@yale.edu

Hugo Viciana

Juan de la Cierva Research Fellow, Instituto de Estudios Sociales Avanzados (IESA-CSIC)

Hviciana@iesa.csic.es

Daniel Wilkenfeld

History and Philosophy of Science & Center for Philosophy of Science, University of Pittsburgh, United States

dawilk@gmail.com

Xiang Zhou

University of Pittsburgh

xiangz@pitt.edu

Acknowledgments

This project could not have been possible without the financial support of multiple organizations. Florian Cova's work on this project was supported by a grant from the Cogito Foundation (Grant No. S-131/13, "Towards an Experimental Philosophy of Aesthetics").

Brent Strickland's work was supported by two grants from the Agence Nationale de la Recherche (Grants No. ANR-10-IDEX-0001-02 PSL*, ANR-10-LABX-0087 IEC).

Matteo Colombo, Noah van Dongen, Felipe Romero and Jan Sprenger's work was supported by the European Research Council (ERC) through Starting Grant. No. 640638 ("Making Scientific Inferences More Objective").

Rodrigo Diaz and Kevin Reuter would like to acknowledge funding from the Swiss National Science Foundation, Grant No. 100012_169484.

Antonio Gaitán Torres and Hugo Viciana benefited from funding from the Ministerio de Economía y Competitividad for the project "La constitución del sujeto en la interacción social" (Grant No. FFI2015-67569-C2-1-P & FFI2015-67569-C2-2-P).

José Hernández-Conde carried out his work as a Visiting Scholar at the University of Pittsburgh's HPS Department. He was financially supported by a PhD scholarship and mobility grant from the University of the Basque Country, and by the Spanish Ministry of Economy and Competitiveness research project No. FFI2014-52196-P. His replication research was supported by the Pittsburgh Empirical Philosophy Lab.

Hanna Kim's work was supported by the Pittsburgh Empirical Philosophy Lab.

Shen-yi Liao's work was supported by the University of Puget Sound Start-up Funding.

Tania Moerenhout carried out her work as a Visiting Researcher at the Center for Bioethics and Health Law, University of Pittsburgh, PA (Aug 2016-July 2017).

Aurélien Allard, Miklos Kurthy, and Paulo Sousa are grateful to Rashmi Sharma for her help in the replication of Knobe & Burra (2006), in particular for her help in translating the demographic questions from English to Hindi.

Ivar Hannikainen and Florian Cova would like to thank Uri Simonsohn for his help in discussing the meaning and best interpretation of p-curves.

Finally, we would like to thank all the authors of original studies who accepted to take the time to answer our questions, share their original material and data, and discuss the results of our replication attempts with us.

Estimating the Reproducibility of Experimental Philosophy

Abstract

Responding to recent concerns about the reliability of the published literature in psychology and other disciplines, we formed the X-Phi Replicability Project (XRP) to estimate the reproducibility of experimental philosophy (osf.io/dvkpr). Drawing on a representative sample of 40 x-phi studies published between 2003 and 2015, we enlisted 20 research teams across 8 countries to conduct a high-quality replication of each study in order to compare the results to the original published findings. We found that x-phi studies – as represented in our sample – successfully replicated about 70% of the time. We discuss possible reasons for this relatively high replication rate in the field of experimental philosophy and offer suggestions for best research practices going forward.

1. Introduction

Over the last several years, impressive efforts have been made to estimate the reproducibility of various empirical literatures. Notable examples include the Open Science Collaboration's (OSC) attempt to estimate the reproducibility of psychological science (Open Science Collaboration, 2015), the Reproducibility Project's analogous initiative for cancer biology (Nosek & Errington, 2017), meta-scientist John Ioannidis's modeling efforts in biomedicine and beyond (e.g., Ioannidis, 2005) and a 2015 estimate produced by the Board of Governors of the Federal Reserve System for research in economics (Chang & Li, 2015). Although there is ongoing debate about what the optimal replication rate[1] should be for a given field in light of trade-offs between, e.g., innovation and confirmation (Gilbert, King, Pettigrew, & Wilson, 2016; Makel & Plucker, 2014), many scientists regard the estimates that have been generated—less than 50% in each of the above cases—as worryingly low. For example, a survey of 1,576 scientists conducted by *Nature* revealed that 52% percent thought there was a "significant" reproducibility crisis (Baker,

[1] Meaning, the ratio of published studies that would replicate versus not replicate if a high-quality replication study were carried out.

2016). A smaller percentage, 3%, thought there was no crisis, while 38% thought there was a "slight" crisis and 7% were unsure. What is not a matter of controversy, however, is that these replication initiatives have generated much-needed discussions among researchers about the state of their sciences. Aspects being put under the microscope include the reliability and effectiveness of common research designs, statistical strategies, publication practices, and methods of peer review (Benjamin et al., 2017; Earp & Wilkinson, 2017; Findley, Jensen, Malesky, & Pepinsky, 2016; Lakens et al., 2018; Locascio, 2017; Young, Ioannidis, & Al-Ubaydli, 2008). Meanwhile, promising ideas for improvement—including the recent push toward norms of pre-registration—are now gaining traction among leading scientists (Chambers & Munafò, 2013; Munafò et al., 2017; Nosek et al., 2018; but see Lash & Vandenbroucke, 2012; Scott, 2013).

One field that has yet to see such an initiative take place is experimental philosophy. As a new academic movement that aims to supplement the classic 'armchair' approach of analytic philosophy with empirical research, experimental philosophy—x-phi for short—uses the data-driven methods characteristic of the social sciences to make progress on the sorts of questions that have traditionally been studied by philosophers. Traditionally, experimental philosophers have focused on the empirical study of philosophically relevant intuitions, including factors that shape them and psychological mechanisms that underlie them (Knobe et al., 2012; Knobe & Nichols, 2008; Machery, 2017a). However, there have recently been calls to go beyond this restrictive conception focused solely on intuitions, to a more inclusive conception that is more reflective of the breadth of work in the field (Cova et al., 2012; O'Neill & Machery, 2014; Rose & Danks 2013). A more comprehensive definition of experimental philosophy, then, could be the use of empirical methods to put to test key premises of philosophical arguments. These premises need not only involve claims about people's intuitions, but could also involve testable assumptions about people's attitudes, behaviors, perceptions, emotional responses to various stimuli, and so on. Experimental philosophy is thus inherently interdisciplinary and can often yield insights about 'how the mind works' that may be of interest to other fields (Knobe, 2007, 2016)

Insofar as x-phi overlaps with other disciplines that study how the mind works, such as cognitive science or social psychology, one might expect that its empirical output should be approximately as replicable as research in those other areas. According to the OSC estimate

concerning psychology, there was some variation in reproducibility depending on sub-field. Papers published in more 'cognitive' journals, such as the *Journal of Experimental Psychology: Learning, Memory, and Cognition*, reportedly replicated at rates of 48-53%, while papers published in the more 'social' journals, such as the *Journal of Personality and Social Psychology*, replicated at rates of 23-29% (Open Science Collaboration, 2015). Since x-phi research explores both 'cognitive' and 'social' questions depending on the nature of the philosophical premise being tested, one possible prediction is that its findings should replicate somewhere in the middle of those estimated ranges, that is, roughly in the vicinity of 35%. If so, we would have good reasons to doubt the reliability of most results gathered by experimental philosophers. How trustworthy, then, is the published literature in our field?

1.1. The need for 'direct' replication

To answer this question, 'direct' replications are needed (Doyen, Klein, Simons, & Cleeremans, 2014). Direct replications—often contrasted with 'conceptual' replications—are replications that attempt to follow the design and methods of an original study as closely as possible in order to confirm its reported findings. Conceptual replications, by contrast, involve making a deliberate change to one or more aspects of the original design or methods, often to explore issues surrounding generalizability (Crandall & Sherman, 2016; Hendrick, 1990; Schmidt, 2009; for a different take on the relation between direct and conceptual replications, however, see Machery, 2017b). But such 'replications' may not be sufficient to identify likely weaknesses or potential errors in the published literature (Earp, in press). As Doyen et al. (2014, p. 28) note:

> The problem with conceptual replication in the absence of direct replication is that there is no such thing as a "conceptual failure to replicate." A failure to find the same "effect" using a different operationalization can be attributed to the differences in method rather than to the fragility of the original effect. Only the successful conceptual replications will be published, and the unsuccessful ones can be dismissed without challenging the underlying foundations of the claim. Consequently, conceptual replication without direct replication is unlikely to [provide meaningful evidence about the reliability of the] underlying effect.

Fortunately, experimental philosophers have not been blind to such issues. Until recently, Joshua Knobe and Christian Mott curated the "Experimental Philosophy Replication Page," a webpage dedicated to collecting all direct replications of experiment philosophy findings (be they published or unpublished).[2] As of November 2017, the page identifies 99 direct replications of experimental philosophy studies, with 42 of these having been classified as unsuccessful replications. Using these data as the basis for an estimate, the replication rate for experimental philosophy would be 57.6%. Although this is higher than the estimate for psychology derived by the OSC (2015), it is still not very encouraging.

But such an estimate would be misleading. Studies that appear on the Replication Page are those that have attracted the interest—or suspicion—of the researchers who attempted to replicate the studies. By contrast, there is likely to be little motivation to replicate a finding that is relatively unsurprising or intuitively robust, which in turn would lead to an exclusion bias against the plausibly more replicable findings. Thus, it is doubtful that studies on the Replication Page constitute a representative sample of experimental philosophy studies. Further support for this view comes from the fact that cross-cultural studies and gender comparisons are vastly overrepresented on the Replication Page, accounting for 41 replications out of 99 (41.4%), a rate that is far beyond the true proportion of such studies (see Knobe, 2016).

Given such considerations, a large and representative sample of studies should be selected and assessed for their (direct) replicability. To accomplish this task, we took inspiration from prior replication initiatives such as the OSC project in psychology, and established the *X-Phi Replicability Project* (XRP), a coordinated effort involving more than 40 researchers from 20 replication teams across 8 countries tasked with conducting and interpreting high-quality direct replications of a wide-ranging sub-set of x-phi studies. Our goal was to derive an accurate estimate of the reproducibility of results obtained by experimental philosophers.

1.2. Interpreting replications

We begin with a note of caution. Scientists have long understood the importance of replicating each other's work: it is not enough for you to report that you ran an experiment and obtained

[2] http://experimental-philosophy.yale.edu/xphipage/Experimental%20Philosophy-Replications.html

certain results; I should be able to run the same experiment and obtain the same results, if I am to be justified in placing confidence in what you reported (Schmidt, 2009). But this is clearly an oversimplification. Even under the best of circumstances, one can never run *exactly the same* experiment that another scientist ran: at the very least, time will have passed between the original experiment and the replication. Moreover, the materials, methods, setting, and background conditions may differ to some degree as well, despite one's best efforts to keep these functionally the same (Collins, 1975; Earp & Trafimow, 2015; Stroebe & Strack, 2014). A more accurate characterization of the follow-up study, then, is that it should be *sufficiently similar* along these and other relevant dimensions that one can meaningfully compare its results to those of the original study. In like manner, the results themselves should be *sufficiently similar* to the original that one can be justified in concluding—however tentatively—that it is the same basic phenomenon being observed, notwithstanding the existence of random variation, statistical noise, measurement error, and so on.[3]

Keeping this in mind, for purposes of estimation we needed to decide for each replication study whether it counted more in favor of, or against, the original reported finding: that is, whether it should be classed as a 'successful' or an 'unsuccessful' replication. There is no single or definitive way to do this (Maxwell, Lau, & Howard, 2015; Open Science Collaboration, 2015). Rather, as with data derived from any study, one must take into consideration a number of factors in order to decide what those data can reasonably be taken to show. Our approach was to use three different methods for designating a replication attempt as a success or a failure, and to report an overall reproducibility estimate based on each method. We will briefly describe these methods in turn:

(a) **Were the replication results statistically significant**? For the present research, we defined 'statistically significant' as a p-value less than .05, following the currently

[3] In practice, it can be hard to determine whether the 'sufficiently similar' criterion has actually been fulfilled by the replication attempt, whether in its methods or in its results (Nakagawa & Parker, 2015). It can therefore be challenging to interpret the results of replication studies, no matter which way these results turn out (Collins, 1975; Earp & Trafimow, 2015; Maxwell, Lau, & Howard, 2015). Thus, our findings should be interpreted with care: they should be seen as a starting point for further research, not as a final statement about the existence or non-existence of any individual effect. For instance, we were not able to replicate Machery et al. (2004), but this study has been replicated on several other occasions, including in children (Li, Liu, Chalmers, & Snedeker, 2018; for a review, see Machery, 2017, chapter 2).

conventional default standards for Null Hypothesis Significance Testing (NHST). However, we must emphasize that the exclusive use of the p-value in a single study to draw inferences about the existence of an effect is controversial (Amrhein & Greenland, 2017; Benjamin et al., in press; Trafimow & Earp, 2017). Thus, p-values should serve as just one piece of information out of many such pieces in a robust, flexible, and context-sensitive inferential process (American Statistical Association, 2016; Lakens et al., 2018; McShane, Gal, Gelman, Robert, & Tackett, 2017; Murtaugh, 2014). Moreover, the use of p-values as a criterion for success is especially dubious when applied to studies reporting null results (Boyle, in press), thus calling for alternate ways of assessing replication success.

(b) **Subjective assessment of the replicating team**. Although a subjective judgment may seem less reliable than a hard-and-fast decision procedure like NHST, this approach has certain advantages. As noted, a single p-value is only one piece of information in an overall judgment about what the data show (American Statistical Association, 2016). By asking our researchers to register their overall subjective judgment about whether the effect replicated, therefore, they were able to take into consideration the 'wider picture' concerning, e.g., facets of study design, methodological details, aspects of the underlying theory as they bear on prediction, and so on.

(c) **Comparison of the original and replication effect size**. The theoretical significance of an effect does not depend only on its existence but also on its size (Cumming, 2013). What counts as a successful replication on the p-value criterion might not always count as a satisfactory replication from a theoretical point of view (see Box 1). Thus, one can also estimate the success of one's replication attempt by comparing the original effect size to the replication effect size. Because sample sizes of replication studies were typically larger than those of original ones, and because calculation of confidence intervals (CIs) for original effect sizes were not always possible (due to a lack of information), we decided to draw this comparison by investigating whether the original effect size fell within the 95% CI of the replication effect size.

Box 1. **What counts as a successful replication? The importance of effect sizes and theory**

Whether something counts as a successful replication depends in part on what the theoretical significance of a given effect-size estimate is. For example, Nichols and Knobe (2007) once argued that the negative emotional reactions elicited by certain actions might impact our judgments about free will and moral responsibility in a (theoretically) significant way, and that this might in turn explain why people are prone to attribute free will and moral responsibility to deterministic agents on some occasions but not others. In their original study, shifting from a 'low-affect' to a 'high-affect' action raised the rate of moral responsibility attributions from 23% to 64%, thus changing participants' modal answer. However, in a meta-analysis based on several unpublished replications, Feltz and Cova (2014) found that, although there was indeed a significant effect of affect, this effect was very small and accounted for only 1% of the variance in participants' answers. Thus, though Nichols and Knobe's effect might be seen as having been 'successfully replicated' according to the p-value criterion, the smaller effect size estimate from the meta-analysis of replications stands in tension with their original theoretical conclusions, as the original authors acknowledge (Knobe, personal communication).

Based on these three criteria, the X-Phi Replicability Project aimed to evaluate the reproducibility of experimental philosophy. The first step was to select a representative sample of studies.

2. Method

2.1. Study selection

Selected studies. 40 studies were selected for replication. For each year between 2003 and 2015 (included), three papers were selected: one as the most cited paper for this year, and two at random (except for 2003, for which only two papers were available). This yielded a total of 38 studies, to which we added 4 additional studies in case some of the originally selected studies proved too challenging to replicate. Out of these 42 studies, we were ultimately able to attempt to replicate 40.

Selection history. To establish an exhaustive, non-arbitrary list of experimental philosophy papers, we began with the papers indexed on the *Experimental Philosophy Page* (http://experimental-philosophy.yale.edu/ExperimentalPhilosophy.html), a resource commonly used by experimental philosophers to make their papers publicly available, and the most

comprehensive available collection of experimental philosophy papers.[4] However, an initial search through this database revealed that a non-trivial number of papers fell well outside of "experimental philosophy" as we have described it above and as it is typically understood, including papers about, e.g., pragmatic abilities in people with autism spectrum disorder (De Villiers, Stainton, & Szatmari, 2007) or the way people choose to punish norm violators in real-life situations (Clavien et al., 2012).

To narrow our choice down and prevent the inclusion of such outliers, we supplemented our preliminary approach with a list of 35 scientific journals. The list was established by XRP coordinators Florian Cova and Brent Strickland by excluding journals from the *Experimental Philosophy* website that were not known for publishing "experimental philosophy" papers as defined earlier, and then systematically collecting the reference of every paper from the remaining journals that contained at least one empirical study (i.e., a study involving the collection of data from participants or any other source in order to test a given hypothesis).

From this set of papers, we retained only those that were published between 2003 and 2015. The upper limit was set by the fact that study selection took place in Spring 2016. The lower limit was set by the fact that experimental philosophy papers only began to be regularly published starting in 2003, mostly in the wake of Joshua Knobe's two seminal papers on intentional action and side-effects (Knobe, 2003a, 2003b).[5] At the end of this process, our list of potential papers contained 242 references that met the following criteria: (i) featuring on the "Experimental Philosophy Page", (ii) being published in one of the 35 journals we identified, and (iii) being published between 2003 and 2015.

To generate our sample, we selected three papers per year between 2004 and 2015 included (for 2003, we selected two papers, as there were only two available for that year, as noted). The first paper was the *most cited* paper of the year (according to Google Scholar) and

[4] Note that this page is basically a mirror of the "Experimental philosophy" category of the *Philpapers* database.

[5] Despite two important studies published in 2001 (Greene et al., 2001; Weinberg, Nichols & Stich, 2001), no experimental philosophy paper is to be found for 2002.

the second and third were selected *at random*. This yielded a total of 38 papers selected for replication.[6]

Our next step was to evaluate individual studies in terms of their feasibility for being replicated. We identified four studies as being more demanding than others on practical grounds on the basis that they required access to a special population (Machery et al., 2004, requiring Chinese participants; Knobe and Burra, 2006, requiring Hindi-speaking participants; Lam, 2010, requiring Cantonese speakers; and Zalla and Leboyer, 2011, requiring individuals with high-functioning autism). Because we could not ensure that replication teams would have the wherewithal to conduct these replications in the available time, a second, plausibly more feasible, study was selected as a potential replacement—either at random if the original paper was selected at random; or the second most-cited paper of the year if the original was the most cited.[7] When both the 'demanding' replication and its more feasible replacement were conducted on time, we decided to include both results in our final analysis. In the end, although we were able to conduct a replication of Machery et al. (2004) and Knobe & Burra (2006), no replication team had the resources necessary to replicate Lam (2010) or Zalla and Leboyer (2011). We thus were left with 40 studies to replicate. The list of all papers (and studies) selected for replication can be found in Appendix 1.[8]

2.2. Assignment of papers to replication teams

[6] There was some initial debate about whether to include papers reporting negative results, that is, results that failed to reject the null hypothesis using NHST. We decided to do so when such results were used as the basis for a substantial claim. The reason for this was that negative results are sometimes treated as findings within experimental philosophy. For example, in experimental epistemology, the observation of negative results has led some to reach the substantive conclusion that practical stakes do not impact knowledge ascriptions (see for example Buckwalter, 2010; Feltz and Zarpentine, 2010; Rose et al., in press). Accordingly, papers reporting 'substantive' negative results were not excluded.

[7] Note, however, that the more 'demanding' paper that was originally selected was not discarded from our list, but remained there in case research teams with the required resources agreed to replicate these studies.

[8] It should be noted that two other papers were replaced *during* the replication process. For the year 2006, Malle (2006) was replaced with Nichols (2006), given that the original paper misreported both the results and statistical analyses, making comparison with replication impossible. For the same year, Cushman et al. (2006) proved to be too resource-demanding after all and was replaced by Nahmias et al. (2006).

The recruitment of replication teams (RTs) took place mostly between October and December 2016. This involved an invitation for contributions that was included in a call for papers for a special issue of the *Review of Philosophy and Psychology* devoted to the topic of replication. The call for papers was subsequently posted on various relevant websites; prominent researchers within the experimental philosophy community were also directly contacted and invited.

Once RTs committed to joining the replication project, they were sent the full list of papers that had not yet been assigned. RTs were invited to estimate the number of replications they could feasibly undertake, and to identify all papers from the available list they could not handle, either because (i) they did not have the resources necessary, or (ii) this would have involved some conflict of interest. Based on these constraints, papers were then randomly assigned to RTs.

2.3. Pre-replication procedure

For each paper, RTs were first asked to fill out a standardized pre-replication form (see Appendix 2). On this form, they were asked to identify the study they would replicate in the paper (in case the paper contained several studies, which was often the case). RTs were instructed to select the first study by default, unless they had a good reason not to (e.g., the first study was only a pilot, or suffered from clear methodological shortcomings that were corrected in later studies). The reason for this instruction was that many experimental philosophy papers present their most striking findings in the first study, with later studies being devoted to controlling for potential confounds or testing for more specific explanations of these results.[9]

Next, RTs were asked to report certain information about the study they selected to replicate. First and foremost, they were asked to identify the study's main hypothesis (or to choose one hypothesis when several equally important hypotheses were tested within the same study). They were then asked to report what statistical analyses were employed to test this hypothesis and the results of these analyses (when no statistical analysis was reported, which occurred several times for early experimental philosophy studies, RTs were asked to reconstruct

[9] In this respect, our methodology differed from the OSC's methodology, which instructed replication teams to focus on the papers' last study.

the appropriate statistical test). When possible, RTs were asked to compute the corresponding effect size and 95% confidence interval.

RTs were also asked to answer a few additional questions about the original study. Questions were about (i) the size and nature of the original sample, (ii) the presence or absence of a selection procedure, and (iii) whether the original paper contained all of the information necessary to properly conduct the replication.

Finally, RTs were asked to compute the sample size needed for their replication. For studies reporting significant results, the replication sample size was computed on the basis of the original effect size, assuming a power of 0.95. Because initial effect size estimates in the literature tend to be inflated due to publication bias (Anderson, Kelley & Maxwell, 2017; Button et al., 2013), we elected to use a higher than usual power assumption (typically .80) so that we would be able to detect even smaller effects that nevertheless do exist. For studies reporting null results (see footnote 6), RTs were instructed to use at least twice the reported sample size, given that the results might have been due insufficient power in the original study.

Completed pre-replication forms were then sent to Florian Cova for approval. Once the forms were approved, RTs were instructed to pre-register their replication on the Open Science Framework (https://osf.io/), using the Pre-Registration form of the Replication Recipe (Brandt et al., 2014). Following best practices (Grens, 2014), RTs were also advised to contact authors of the original study to ensure the greatest fidelity along all relevant dimensions between the original study and the replication. Most original authors agreed to help the RTs, and we thank them for their contribution.

2.4. Post-replication procedure

After running the replication study, RTs were asked to fill out a post-replication form (see Appendix 3). The post-replication form asked RTs to report the procedure and results of their replication study as they would in a normal research paper. Then, they were asked to report about their study the same kind of information they reported about the original study in the pre-replication form (effect size, 95% CI, size and nature of their sample). Finally, RTs were asked

to report their own subjective assessment about whether they successfully replicated the original result.

Once the post-replication form was completed, replication teams were instructed to upload it, along with the all relevant data and documentation, to the corresponding OSF depository, and to register their results using the post-registration form of the Replication Recipe (Brandt et al., 2014) if possible.

Details for all individual replications can be accessed online through the X-Phi Replicability Project main OSF page (osf.io/dvkpr).

2.5. Replication teams (RTs)

Overall, 20 RTs (involving 40 persons) took part in the replication project. Once the data were collected, an additional project member was recruited (Brian Earp) to aid with interpretation and theoretical framing, as well as drafting various sections of the manuscript. Research teams from 8 countries (Brazil, France, Lithuania, Netherlands, Spain, Switzerland, United Kingdom, United States) were involved.

3. Results

40 studies were repeated one time each in an attempt to replicate the originally reported results. Studies came from several different sub-areas of experimental philosophy: 8 from Action Theory, 1 from Aesthetics, 4 from Causation, 5 from Epistemology, 8 from Free Will, 8 from Moral Psychology, 1 from Philosophy of Language, 2 from Philosophy of Mind, 3 uncategorized.

The average *N* was 215.1 (*SD* = 542.3) for original studies and 206.3 (*SD* = 131.8) for replication.[10] However, the mean for the original studies was biased by an extremely high *N* for one study with 3422 participants (Hitchcock & Knobe, 2009). In fact, the median *N* was 85 for original studies and 183 for the replication studies. In 32 studies out of 39 that used participants

[10] *N*s were computed not from the total *N* recruited for the whole study but from the number of data points included in the relevant statistical analysis.

(excluding Reuter, 2011, that used internet hits as data points), the replication *N* was greater than the original *N*. Overall, mean *N*s for original studies tended to increase over time, going from an average of 57.5 in 2003 to 162 in 2015.

Both original and replication studies made ample use of convenience samples, but there were differences between the two. The original studies, particularly in the early years of experimental philosophy, tended to use university students: out of 39 studies, 25 used student samples, 6 used Amazon Mechanical Turk (MTurk) workers, 4 used other online samples, and 4 used pedestrians recruited from the street. On the contrary, replication studies tended to focus on online samples: out of 39 studies, 29 used MTurk workers, 6 used other online samples, and 4 used university students. This difference in populations comes with the possible disadvantage of lowering replication rates—insofar as the original findings were dependent upon a particular population—but simultaneously allows for an assessment of generalizability (see below).

Out of 32 studies reporting a significant result and for which we could perform the relevant power analysis, 26 had a power superior or equal to 0.80, and 18 had a power superior or equal to .95 (assuming the original study's effect size). The average power was 0.88 (*SD* = 0.14).[11]

To assess the successful replication rate, we used three different criteria as described earlier: (i) the RT's subjective assessment, (ii) p-values and statistical significance, and (iii) comparison of the original and replication effect sizes.

3.1. Replication team's subjective assessment

We first examined RTs' subjective assessment of whether they had successfully replicated the original results. Out of 40 replications, 31 were considered to be successful replications by the RTs that conducted them, yielding a successful replication rate of 77.5% by this metric. The replication rate was 78.4% (29 out of 37) for original studies presenting significant results, and 66.7% (2 out of 3) for original studies presenting null results.

[11] For this analysis, studies for which power > 0.99 were counted as power = 0.99.

3.2. p-values

We then assessed replication success using the p-values obtained by the RTs. For original studies presenting statistically significant results, a replication was considered successful when $p < .05$ and the effect went in the same direction as the original effect. The 3 studies presenting null results were excluded from this analysis, given the difficulty of assessing such results using NHST (Boyle, in press).

By these criteria, the overall successful replication rate was 78.4% (29 out of 37).

3.3. Comparison of original and replication effect sizes

As a final criterion for successful replication, we compared the original and replication effect sizes. First, when possible, original and replication effect sizes were converted to a common *r* effect size, and 95% CI interval were computed for both. This was possible when the corresponding statistical test was either (i) a Chi-square test with $df = 1$, (ii) a Student's or Welch t-test, (iii) a correlation test, or (iv) an ANOVA with $df_1 = 1$. When this was not possible, alternate effect sizes and 95% CIs were used (such as RMSEA for Structural Equation Modelling). When the replication obtained an effect that went in the opposite direction to the original effect, replication effect sizes were coded as negative. Effect sizes and 95% CI for replication are presented in Figure 1.

For studies reporting statistically significant results, we treated as successful replications for which the replication 95% CI was not lower than the original effect size.[12] For studies reporting null results, we treated as successful replications for which original effect sizes fell inside the bounds of the 95% CI.

We were able to calculate (i) the original effect size and (ii) the replication 95% CI for 34 studies out of 40 (32 original studies reporting significant effects, 2 reporting null effects).

[12] For studies reporting statistically significant results, we counted studies for which the original effect size was *smaller* than the replication 95% CI as successful replications on the ground that, given the studies' original hypotheses, a greater effect size than originally expected constituted even more evidence in favor of these hypotheses. Of course, theoretically, this need not always be the case, for example if a given hypothesis makes precise predictions about the size of an effect. But for the studies we attempted to replicate, a greater effect size did indeed signal greater support for the hypothesis.

Details of the results are presented in Table 1. Overall, according to this more stringent criterion[13], the overall successful replication rate was 24 successful replications out of 34 (70.6%).

[13] As pointed out by a reviewer on this paper, this criterion might even be considered *too* stringent. This is because, in certain circumstances in which no prediction is made about the size of an effect, a replication for which the 95% CI falls below the original effect size might still be considered as a successful replication, given that there is a significant effect in the predicted direction. Other ways of assessing replication success using effect sizes might include computing whether there is a statistical difference between the original and replication effect size (which would present the disadvantage of rewarding underpowered studies), or considering whether the replication effect size fell beyond the lower bound of the 95% CI of the original effect size (which returns a rate of 28 successful replications out of 34 original studies, i.e. 82.4%). Nevertheless, we decided to err on the side of stringency.

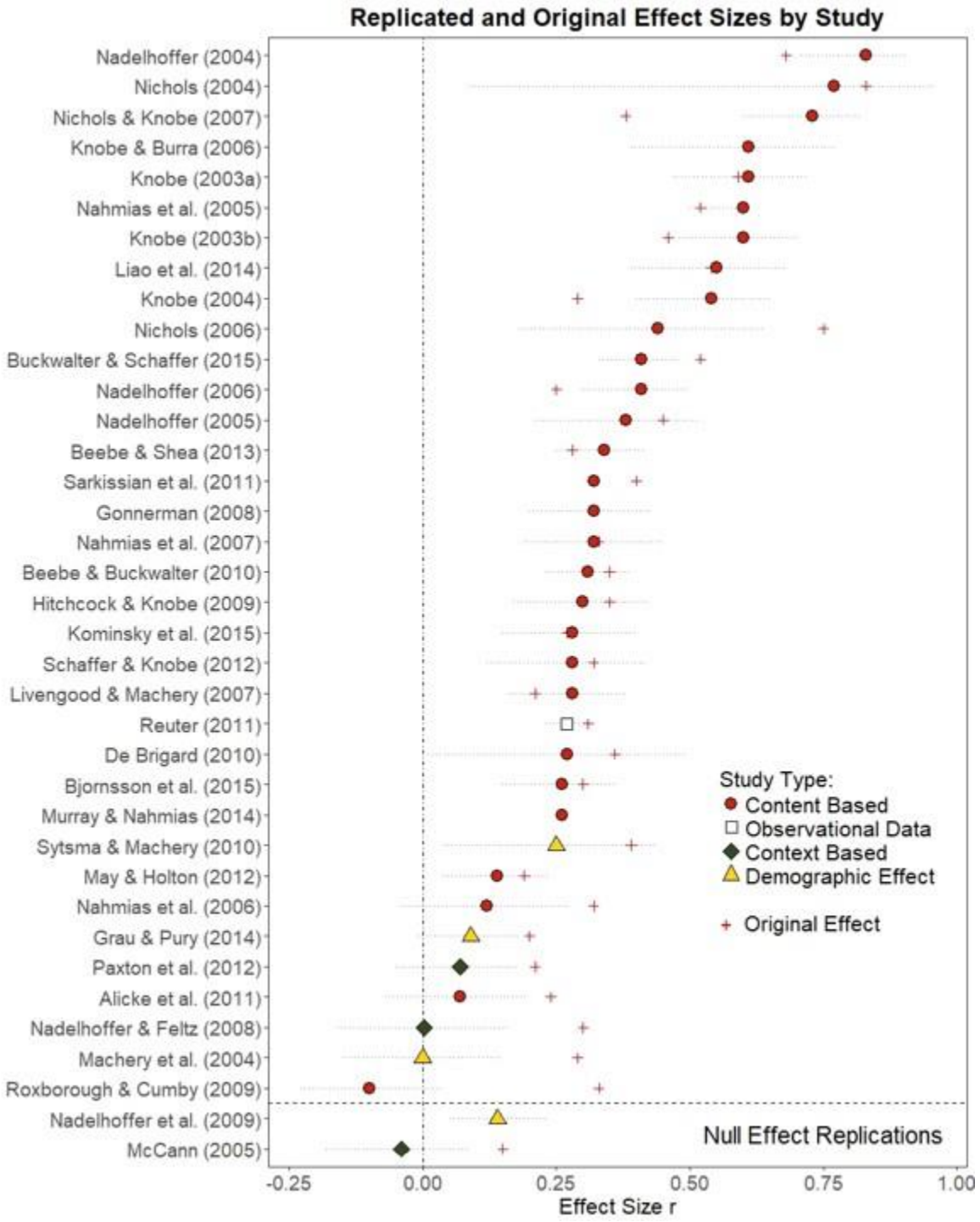

Figure 1. Original effect size, replication effect size and replication 95% CI for each study. For descriptions of "Content Based," "Observational Data," "Context Based," and "Demographic Effect" see Section 4.3.

Original effect size is __ the replication 95% CI	*Below*	*Within*	*Over*
Significant effects	**5**	**18**	9
Null effects	0	**1**	1

Table 1. Results for the comparison of the original effect size with the replication 95% CI. Bold numbers indicate replications that count as successful.

Of note, when focusing on studies originally reporting statistically significant results, it seemed that only 9 out of 32 (28.1%) overestimated their effect size compared to the replication estimate (assuming that the latter is more accurate). For these 32 studies, the average original *r* effect size was 0.39 (*SD* = 0.16), while the average replication *r* effect size was 0.34 (*SD* = 0.24) (see Figure 2). The effect size for this difference was small ($t(62) = 0.85$, $p = .40$, $d = 0.21$, power = 0.22), suggesting that original effect sizes were not much larger than replication effect sizes.

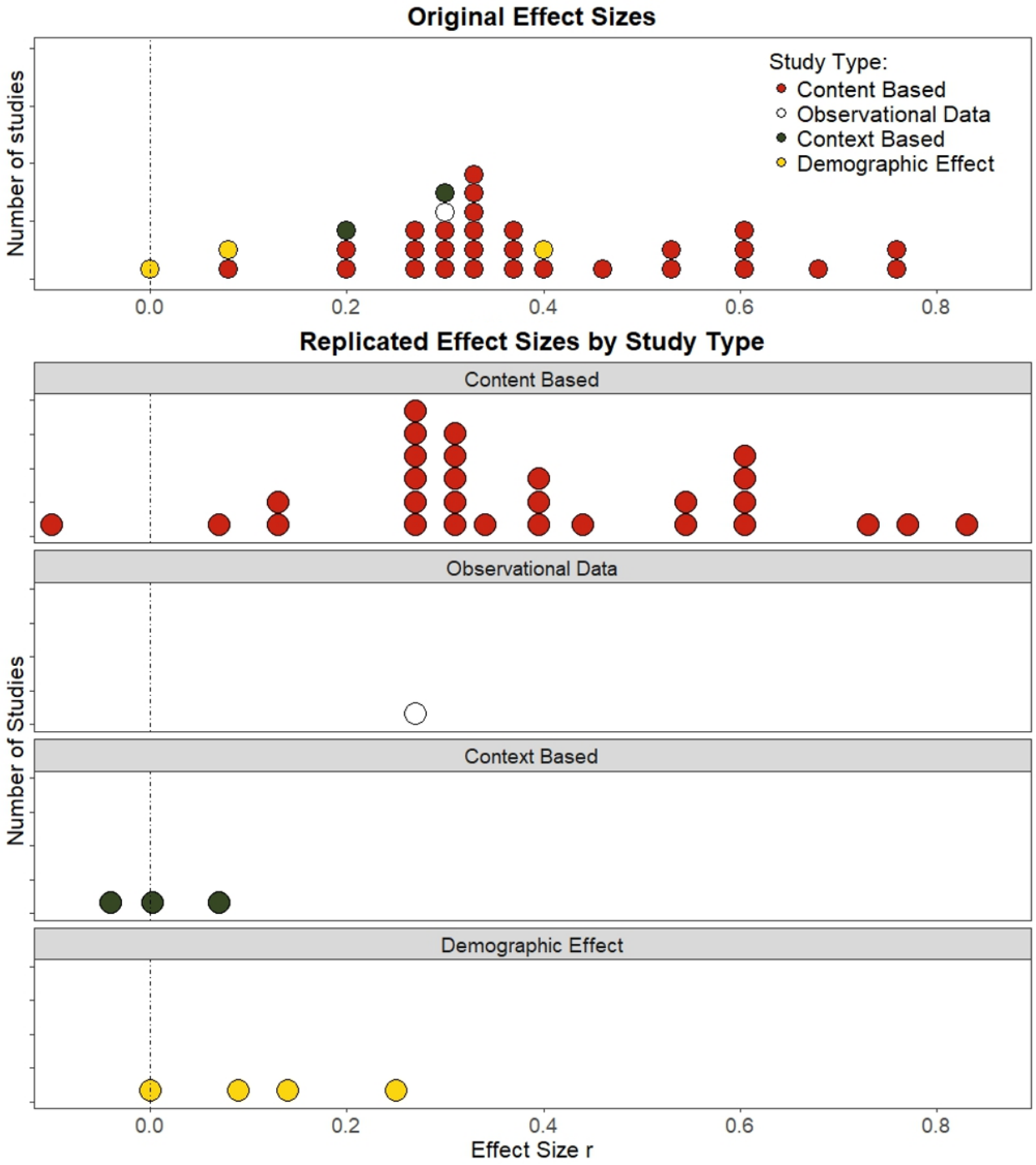

Figure 2. Effect sizes (correlations) for original and replication studies. Replication studies are sorted by type of studies (observational, content-based, context-based, or demographic).

3.4. Most cited vs. Random studies

In our study selection, we picked studies both at random and among the most cited ones. One reason for this procedure was that we wanted our estimate to be representative of both experimental philosophy at large (random selection) and the kinds of effects people are more likely to discuss when thinking about experimental philosophy (most-cited papers). Overall, the papers we selected as most cited had a greater number of citations per year ratio than papers we selected at random (M = 30.7, SD = 18.0 vs. M = 8.4, SD = 6.1; $t(38) = 5.835$, $p < .001$, d = 1.93).[14]

	Subjective assessment	*P-values*	*Effect sizes*
Most cited (N=14)	64.3%	64.3% (9 out of 14)	54.5% (6 out of 11)
Random (N=26)	84.6%	87.0% (20 out of 23)	78.3% (18 out of 23)
TOTAL	77.5%	78.4% (29 out of 37)	70.6% (24 out of 34)

Table 2. Replication rates according to three criteria (subjective assessments, *p*-values, and effect size comparisons) for most cited and randomly selected studies.

Table 2 summarizes the replication rates according to all three criteria for both most-cited and randomly selected studies. Overall, the replication rate for most-cited studies (subjective assessment = 64.3%) was lower than the replication rate for randomly selected studies (subjective assessment = 84.6%). However, a logistic regression did not reveal citation rates to be a significant predictor of success (measured through subjective assessment) (OR = -0.97, p = .18). Thus, due to the small size of our sample, it is not possible to determine with confidence whether this reflects an actual trend or is simply the product of random variation.

3.5. Effect of publication year on replication success

There was no evidence of an effect of publication year on replication success (as measured by p-values or RTs' subjective assessment), OR = 0.99, t = -0.14, p =.89.

[14] This analysis was done on the basis of Google Scholar's citation count (as of March 23rd, 2018).

3.6. Generalizability of results obtained on convenience samples

As mentioned above, within our sample, most original studies used pedestrians or university students as convenience samples, while most replications used online survey participants (mostly MTurk workers) as convenience samples. This allows us to assess the generalizability of results obtained from such samples. Among our studies, we identified 24 in which the original sample was either a pedestrian (4 out of 24) or university student (20 out of 24) sample and the replication sample an online sample. Out of these 24 studies, 20 successfully replicated (according to RTs' subjective assessment), a replication rate of 83.3%. Thus, it seems that most original findings based on convenience samples such as pedestrians or university students could be generalized to online samples (Casler, Bickel, & Hackett, 2013).

3.7. Summary

Overall, our three criteria converge on the conclusion that the reproducibility rate of experimental philosophy studies, as estimated through our sample, is greater than 70%. Moreover, the analysis of effect sizes for studies originally reporting significant effects suggests that most of them did not overestimate their effect sizes compared to replications.

4. Potential explanations for the relatively high replication rate

Recall that, for the OSC attempt to estimate the reproducibility of psychological science, the replication rate was 36.1% - 47.4% depending on the measure, which is much lower than the roughly 70% replication rate we observed for x-phi studies. How are we to explain our finding that x-phi results seem to replicate at a far higher rate than results in psychological science? In the following sub-sections, we explore several different (though not mutually exclusive) answers.

4.1. Larger effect sizes

The OSC attempt found that effect sizes were good predictors of an effect's replicability (Spearman's rank-order correlation of 0.277 for the original effect size and of 0.710 for

replication effect sizes). Thus, the higher replicability rate of experimental philosophy results might be explained by those results' being characterized by larger effect sizes.

For original effect sizes, the OSC reports an average *r* effect size of 0.403 (*SD* = 0.188). This is in fact higher than our average original *r* effect size (*M* = 0.38, *SD* = 0.16). But the initial estimates—at least for the psychology studies—were most likely inflated due to publication bias, relatively small sample sizes, and other factors (Anderson, Kelley, & Maxwell, 2017; Button et al., 2013). Let us assume that effect size estimates derived from replication studies are on average more accurate than those reported in original studies, due to the interaction of publication bias and statistical regression to the mean (Trafimow & Earp, 2017). In this case, replication effect sizes were actually higher for x-phi studies (*M* = 0.33, *SD* = 0.23), compared to psychology studies (*M* = 0.20, *SD* = 0.26). Since the most-cited and random x-phi studies did not differ in either original, $t(32) = 0.30$, $p = .77$, or replication effect size, $t(35) = 0.18$, $p = .86$, the large average effect among the sample of x-phi studies is not likely due to oversampling from highly-cited publications. This suggests that the true effect sizes reported in x-phi may tend to be on average larger than those in reported in psychology studies. This, in turn, would increase the relative likelihood of effects from x-phi studies replicating.

However, we should note that, at least among the studies we replicated, effects were especially large in the early years of experimental philosophy but have tended to get smaller over time. Indeed, publication year correlated negatively with effect size (converted to *r*) whether looking at original reports, $r(31) = -.36$, $p = .040$, or replication data, $r(34) = -.37$, $p = .025$ (see Figure 3), even when excluding studies that were originally reported as null results (original, $r(30) = -.42$, $p = .017$; replication, $r(32) = -.44$, $p = .009$). One possible explanation for this trend is that later studies tend to be attempts to elaborate on initial findings by decomposing them into constituent parts, as illustrated by the trolley literature (Cova, 2017) or the literature on the side-effect effect (Cova, 2016). Another possibility is that it is increasingly unlikely over time that one will observe a large effect that had previously gone unnoticed. However, such possibilities would best be explored by analyzing the effects of publication year on the population of experimental philosophy studies as a whole, which is not something we are able to undertake based on our sample.

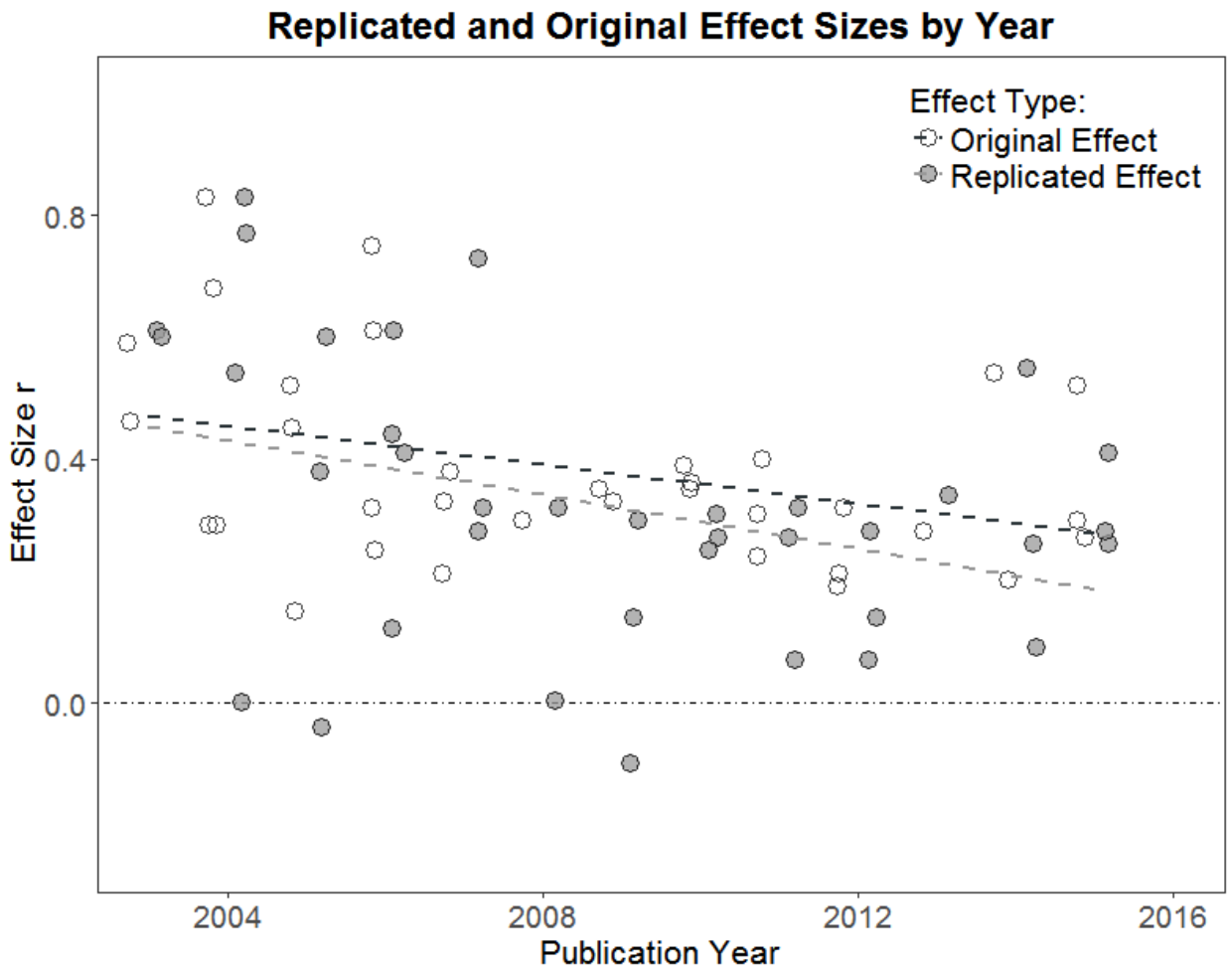

Figure 3. Original and replication effect sizes per year.

4.2. Cost of studies

Another explanation for the higher replicability rate for experimental philosophy compared to psychology could be that x-phi studies are, on average, 'easier' to run – in large part by being less costly. Indeed, many experimental philosophy studies are simple surveys that can be relatively quickly and inexpensively administered.

This feature might explain the higher replication rate in two ways. First, 'easier' studies might lead to larger sample sizes, which in turn might lead to higher-powered studies. To test for this hypothesis, we compared sample sizes in our sample to typical sample sizes in social-personality psychology. According to Fraley and Vazire (2014), median sample sizes in the latter field range from 73 to 178, depending on the journals. As we saw, the median *N* for our studies

was 85, which falls within this range. Moreover, assuming a typical effect size or $r = .20$, Fraley and Vazire found that the power of the typical social-personality psychology study was below the recommended 80% and even reached 40% for certain journals. Using a similar method, we computed power assuming an effect size of $r = .20$ for original x-phi studies for which a r effect size could theoretically be computed (34 out of 40). The average power was 0.5 ($SD = 0.28$) and only 7 studies out of 34 reached a power > .80. Thus, if the easiness of running experimental philosophy studies explains our higher replication rate, it is not because it allowed our original studies to be higher-powered than typical psychology studies.

However, there is a second way in which 'easiness' might explain the higher replicability rate: because there is relatively little cost (in terms of time and resources) in running an x-phi study, experimental philosophers can recruit more participants per condition, double-check their results by re-running the study if they are uncertain about any findings, and subject their results to scrutiny by others, who can in turn easily run their own replications. By contrast, the more time- or resource-intensive it is to obtain data, the more distressed a researcher may feel about failing to get something 'publishable' out of the effort. This in turn could promote so-called Questionable Research Practices (Fiedler & Schwartz, 2016; John, Loewenstein, & Prelec, 2012) which may increase the likelihood of committing a Type 1 error (Simmons, Nelson, & Simonsohn, 2011).

To test this second 'easiness' hypothesis, we rated our 40 studies according to how easy to run we perceived them to be. Scores ranged from 0 to 2. One 'difficulty' point was awarded to studies that were *not* simple surveys that could have potentially been run online (for example, studies that involved a cognitive load task and as such required an in-lab setting, such as Greene et al., 2008). An additional 'difficulty' point was awarded to studies that required an unusual population and so could not be run using a convenience sample (for example, cross-cultural studies comparing specific populations, such as in Machery et al., 2004). In the end, no study received a score of 2: 36 studies received a score of 0, and 4 a score of 1. This highlights the relative 'easiness' of running x-phi studies in general. As expected, the replicability rate for 'difficult' studies was lower than the rate for 'easy' studies: 50% (2 out of 4) compared to 80.6% (29 out of 36).

What about psychology studies? To complete the comparison, we went back to the list of studies replicated by the OSC project and selected 99 of them that (i) were included in the final OSC analysis and (ii) made the results of the replication available. We then rated them in the same way as we rated the x-phi studies. Overall, out of 99 studies, 17 received a score of 0, 70 a score of 1, and 12 a score of 2. This suggest that psychology studies were indeed more 'difficult' to run on average, which might factor into the difference in replication rate between experimental philosophy and psychological science. However, within the OSC project, the replicability rate was not much higher for 'easy' studies (43.8%, 7 out of 16), compared to 'medium' (38.2%, 26 out of 68) and 'difficult' studies (36.4%, 4 out of 11), which suggests that other factors than 'easiness' might be at play.

4.3. Type of effects

Why else, then, might our replication rate have been so much higher? Another hypothesis is that the high replication rate for x-phi studies might be due to the *kind of effect* studied by experimental philosophers. Indeed, the studies selected for replication in our project can be organized into four main categories:

1) *Observational studies:* These are studies that do not involve data collected in an experimental setting in which independent variables are under the direct control of the experimenter, but rather make use of other kinds of data (e.g. instances of linguistic expressions in a corpus as in Reuter, 2011).
2) *Content-based studies:* These are studies that focus on how participants perform a certain task or react to certain stimuli (e.g., how intentional they find an action to be), and how their behavior is determined by the content of the task or stimuli. Experimental manipulation in these studies typically focuses on changing certain properties of the task or the content of the stimuli and testing whether this change affects participants' responses (e.g., changing the side effect of an action from 'harming the environment' to 'helping the environment' and seeing how this affects participants' judgments of an agent's intention, as in Knobe, 2003a).

3) *Context-based studies:* These are studies that keep the content of a task or stimulus constant but explore how participants' reactions can be changed by manipulating the context and the way in which the content is presented (e.g., presenting the stimuli with or without cognitive load as in Greene et al., 2008; presenting the same vignette in a first- versus third-person framing as in Nadelhoffer & Feltz, 2008).
4) *Demographic effects:* These are studies that keep both the content of the stimulus and/or task and the context in which it is presented constant, but explore how participants' answers can be shaped by differences in the participants themselves (e.g., cross-cultural comparisons such as in Machery et al., 2004; correlations between character traits and philosophical intuitions as in Nadelhoffer, Kvaran & Nahmias, 2009).

In investigating the effect of kind of study on the replicability of experimental philosophy, we tested two related hypotheses. The first is that most x-phi studies fall into the second category: they study how participants' reactions to a given stimulus (vignette) are shaped by properties of the stimulus itself (its content). The second is that, at least within our sample, effects of the second kind (content-based) are less fragile than effects of the third (context-based) and fourth (demographic effects) kinds. Indeed, context-based effects are often dependent on the participant's attention, and her or his ignorance of the manipulation (Cesario, 2014), while demographic effects are threatened by intra-group variability (Heine et al., 2002).

To test these hypotheses, we first classified our 40 studies as falling within one of these four categories: 1 fell into the observational category, 31 fell into the *content-based* category, 4 into the *context-based* category, and 4 into the *demographic effect* category.[15] These results support the first hypothesis: experimental philosophy studies seem to be mostly *content-based*, focusing on how (a change in) the content of a given stimulus (typically a vignette) impacts participants' reactions.

We next tested the second hypothesis, asking whether content-based studies are *more* replicable than the others. Table 3 sums up the replication rate (based on RTs' subjective assessment) for each category (excluding the *observational* category, for which we only had one

[15] In a previous version of this manuscript, we reported 30 *content-based* studies and 5 *demographic effects*. However, helpful commentaries from readers, including Wesley Buckwalter, led us to revise our classification for Nichols (2004).

data point). For our sample at least, it does appear that content-based studies have a higher replication rate when compared to context-based and demographic-based studies.[16] They also tended to have larger effect sizes (see Figure 2).

Type of effect	*Replication rate*	*Average original effect size*	*Average replication effect size*
Content-based	90.3%	0.41 (0.17)	0.39 (0.21)
Context-based	25.0%	0.22 (0.08)	0.01 (0.06)
Demographic effect	25.0%	0.29 (0.10)	0.12 (0.10)

Table 3. Replication, average original effect size and replication effect size for each category of studies

Of course, this conclusion pre-supposes that context-based and demographic-based studies make up a greater proportion of studies in traditional psychological science than in experimental philosophy. To determine whether this is really the case, we went back once again to the list of 99 OSC studies we selected, and categorized them in the same way we categorized x-phi studies. We ended up with 34 content-based studies, 44 context-based studies, 16 demographic-based studies, 4 observational studies, and 1 that was uncategorized. Thus, content-based studies played a less important role in psychological science than in experimental philosophy ($\chi^2(1, N = 139) = 19.62$, $p < .001$). Moreover, the replication rate for content-based studies was 64.5%, while it was 20.5% for context-based studies and 31.3% for demographic-based studies.

Thus, the difference in replication rates between experimental philosophy and psychological science might be explained by the different kinds of effects they typically investigate: while experimental philosophy focus mostly on robust effects triggered by changes

[16] A low replication rate for demographic-based effects should not be taken as direct evidence for the nonexistence of variations between demographic groups. Indeed, out of 3 demographic-based effects that failed to replicate, one was a null effect, meaning that the failed replication found an effect where there was none in the original study.

in the content of the very stimulus participants are asked to react to, traditional psychological science tends to focus more on subtle effects wherein participants are led to react differently to a given stimulus by external changes. This contrast might be heightened by the fact that many of the content-based effects investigated by experimental philosophers are effects that can be accessed to some extent introspectively. For example, Dunaway, Edmonds and Manley (2013) found that philosophers were able to predict *a priori* some central results in experimental philosophy. In this respect, parts of experimental philosophy might be compared to works in linguistics, and derive their reliability from the fact that some effects are robust enough to be introspectively assessed (see Sprouse & Almeida, 2017).

4.4. Differences in culture and practice

Finally, it might be that differences in replication rates could be explained by differences in academic cultures and research practices. Based on such perceived differences, Liao (2015) predicted a higher replication rate for experimental philosophy studies before the launch of the XRP. As philosophers, Liao noted, x-phi researchers might be more sensitive to certain methodological questions, such as what counts as strong evidence for a given claim; moreover, they might have a more welcoming attitude toward replication – in part due to the typically low cost of running x-phi studies, as mentioned above – and be more transparent in some of their research practices.[17]

These perceived characteristics of the practice and culture of experimental philosophy might have contributed to the relatively high replication rate by discouraging questionable research practices. Although these claims are hard to test directly, a few indicators provide indirect support. First, as noted, published effect sizes for x-phi studies appeared to be only slightly (and non-significantly) overestimated as compared to effect sizes in the replication attempts: ratio of mean-replication to mean-original effect size = .88, paired $t(31) = 1.67$, $p = .11$. Second, when researchers p-hack, the resulting distribution of p-values below .05 tends to be flat

[17] Possible reasons for such transparency might be that (i) experimental philosophy is still a smaller academic community where individual researchers are likelier to be well known to each other and thus able and willing to hold each other accountable, and (ii) research resources (such as online survey accounts) used to be shared among researchers in the early days of the field, thus making questionable research practices more difficult to obscure (see Liao, 2015).

or even leftward skewed (Simonsohn, Nelson & Simmons, 2014), whereas the p-curve for our target set of x-phi findings revealed a substantial rightward skew (see Figure 4), with few p-values in the .025 - .05 range. Finally, recent research by Colombo and colleagues (2017) found that the rate of statistical reporting inconsistencies was lower in experimental philosophy than in others parts of behavioral science. In any case, Liao (2015) does seem to have been right with his prediction, and we cannot exclude the possibility that the higher observed replicability of x-phi findings compared to psychology findings might reflect particular cultural values and research practices within the field.

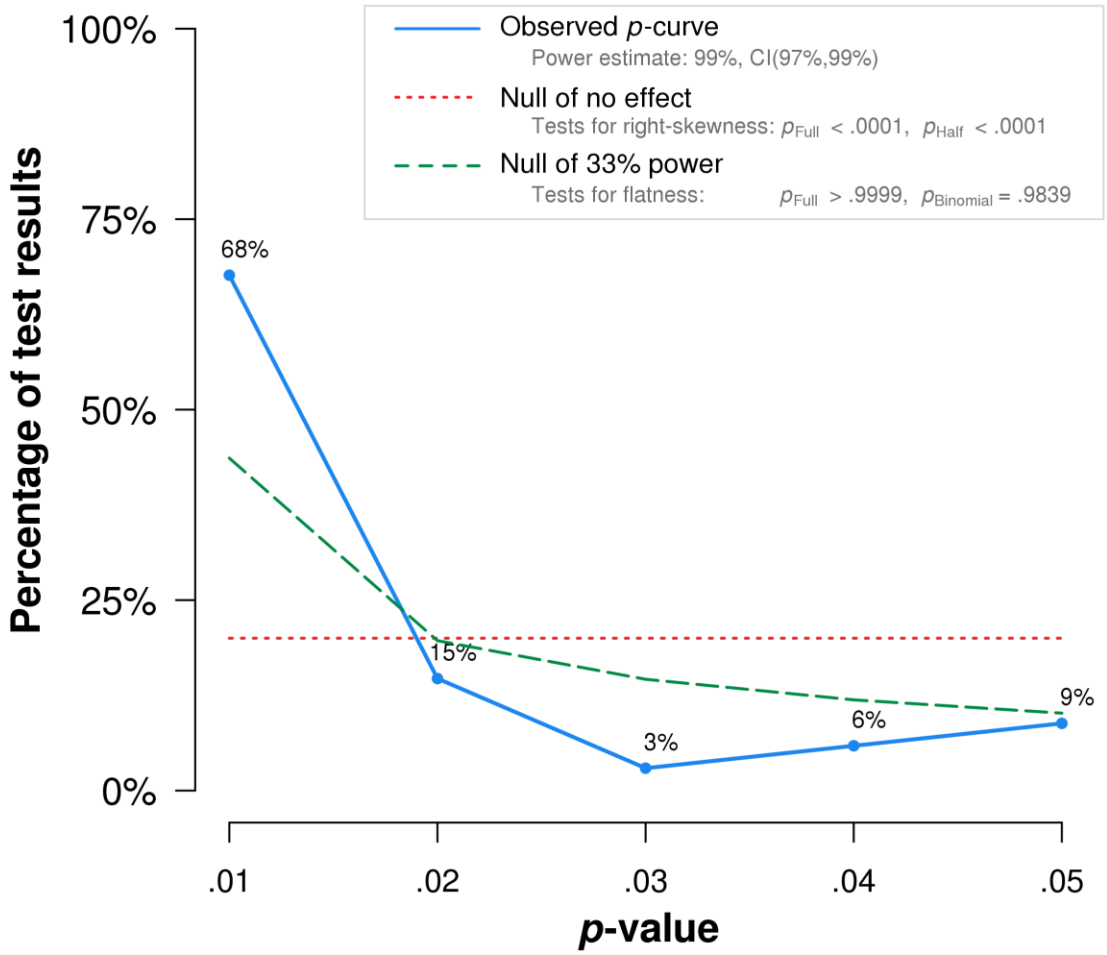

Figure 4. Distribution of *p* values corresponding to target effects in original publications, generated by the p-curve app (www.p-curve.com; see Simonsohn, Nelson & Simmons, 2014). Three studies reported insufficient information to calculate precise *p* values, and therefore are excluded. Two other *p* values (> .05) were not displayed.

One such cultural value might be a greater tolerance or even appreciation among experimental philosophers for negative or null results. As many have argued, the systematic non-publication of null results – which contributes to the so-called file-drawer effect – is a leading factor in increasing the proportion of false positives in the literature and thus of non-replicable

effects (Earp, 2017; Franco, Malhotra, & Simonovits, 2014; Rosenthal, 1979). In our experience, experimental philosophers tend to have a more positive attitude toward null results: they take null results from adequately powered studies to have some evidential value, and indeed some key findings in experimental philosophy are based on failures to reject the null hypothesis (which might explain why 10% of the studies we sought to replicate were null results, while studies with null results only constituted 3% of OSC's original pool). Moreover, null results that are clearly or at least plausibly due to weaknesses in the study design can be discarded without too much anguish: as noted, x-phi studies tend to be fairly easy as well as inexpensive to run, such that there is little incentive to 'tease' an ultimately dubious finding out of a data set for the sake of publication. Instead, one can simply run another, better-designed study, only submitting for publication results in which one has high confidence (ideally because one has already replicated them in one's own lab).

In fact, compared to 'traditional' psychologists, experimental philosophers may be less susceptible to such 'publish-or-perish' pressures in general. First, it is presumably far easier to abstain from publishing the (dubious) results of a study that took a few days or weeks to run – as is common in x-phi research – than a study that took many months to run at potentially great cost. And second, experimental philosophers may not need to publish data-driven papers in order to maintain or advance their careers in the first place. In their capacity as philosophers, at least, they may have ample opportunities to publish papers without any data—i.e., dealing 'purely' with theoretical issues—and the publication pressure is generally lower in philosophy. Taken together, these and the above-mentioned factors might create field-specific norms and practices that decrease the likelihood of false positives proliferating throughout the literature. Finally, although we do not have direct evidence of this, it is possible that philosophy journals are on average less reluctant than psychology journals to publish null results. If so, this would diminish problems associated with the file-drawer effect, thus reducing the proportion of non-replicable effects.[18]

[18] One more cynical explanation would simply be that experimental philosophers are less well versed in into statistics, and that certain questionable research practices are only available to those who have sufficient skills in this area (i.e., the ability to take advantage of highly complex statistical models or approaches to produce 'findings' that are of questionable value).

5. Conclusion

In this project, our goal was to reach a rough estimate of the reproducibility of experimental philosophy studies. We sampled 40 studies from the experimental philosophy literature, and drew on the resources of 20 separate research teams from across 8 countries to undertake a high-quality replication of each one. Based on three different classification systems, we converged on an estimated replication rate situated between 70% and 78%. This means that, roughly, the replication rate for experimental philosophy would be 3 out 4.

This appears to be good news for experimental philosophy. As a new field, it has been subjected to criticism from skeptical quarters, including the claim that it is little more than *bad psychology*—an upstart enterprise run by philosophers who mimic the methods of behavioral science without fully mastering or even understanding them (Cullen, 2010; Woolfolk, 2013). In the wake of the replication crisis, this line of thought gave rise to the *companion-in-guilt* argument: if experimental philosophy is just bad psychology, and if psychology suffers from a serious replication problem, then we should expect experimental philosophy to fare even worse (see Liao, 2015). Indeed, the replication crisis in psychology has sometimes been framed as a limitation of—or argument against—experimental philosophy (see Loeb & Alfano, 2014, section 5.1).[19]

In this context, the results of the current replication initiative appear to provide a strong, empirically-based answer to these criticisms. In particular, our observed replication rate of over 70% seems to undermine pessimistic inductions from low replicability rates in psychology and other behavioral sciences to presumed replication rates in experimental philosophy. It also calls into question the idea of x-phi being mere 'amateurish' psychology, suffering from the same shortcomings and methodological issues as the latter, only worse. Simply put, such a characterization of experimental philosophy is inconsistent with our findings.

Of course, these results should not be taken as invitation for experimental philosophers to rest on their laurels and no longer worry about methodological issues in the behavioral sciences. As long as we are uncertain of the reason behind experimental philosophy's high replication rate,

[19] For example, as of November 2017, the Wikipedia page for "Experimental Philosophy" dedicates a large part of its "Criticisms" section to the "Problem of Reproducibility," arguing that "a parallel with experimental psychology is likely."

we cannot reasonably infer that future experimental philosophy studies will meet the same success. That said, we have considered a number of potential factors: the apparently larger typical effect sizes in x-phi studies, the lower cost of running survey-based experiments, the different kinds of manipulations characteristic of x-phi research (e.g., content-based vs. context-based), and perceived cultural norms discouraging the use of questionable research practices while encouraging greater transparency and acceptance of null results. Each of these explanations makes a different prediction: for example, if the high replication rate of experimental philosophy depends on the size of the effects it typically investigates, then we would need to adjust our practice as experimental philosophy begins searching for more subtle and smaller effects. If it is due to experimental philosophy's focus on easy-to-run, content-based studies, then a similarly high rate should not be taken for granted as more complex, context-based studies begin to become more widespread. And finally, if it stems from values and practices that are specific to the field, then we should try to maintain and foster this positive culture. The current project, which could not have been possible without the contribution of so many dedicated researchers willing to engage in a good-faith collective enterprise to examine the strengths and weaknesses of their science, might be one important step in this direction.

OSF Repository

Details, methods and results for all replications can be found online at https://osf.io/dvkpr/

Softwares

Most of the analyses reported in this manuscript were conducted using the R {compute.es} and {pwr} packages (Champely, 2018; Del Re, 2015). We are also indebted to Lakens' R2D2 sheet (Lakens, 2013).

Appendix 1. List of studies selected for replication

(Crossed-out studies are studies who were planned for replications but did not get replicated.)

***2003**

Most cited: Knobe, J. (2003a). Intentional action and side effects in ordinary language. *Analysis*, *63*(279), 190-194. [Study 1] (Content-based, successful, osf.io/hdz5x/)

Random: Knobe, J. (2003b). Intentional action in folk psychology: An experimental investigation. *Philosophical Psychology*, *16*(2), 309-324. [Study 1] (Content-based, successful, osf.io/78sqa/)

***2004**

Most cited: Machery, E., Mallon, R., Nichols, S., & Stich, S. P. (2004). Semantics, cross-cultural style. *Cognition*, *92*(3), B1-B12. (Demographic effect, successful, osf.io/qdekc/)

- *Replacement:* Knobe, J. (2004). Intention, intentional action and moral considerations. *Analysis*, *64*(282), 181-187. [Study 1] (Content-based, successful, osf.io/ka5wv/)

Random 1: Nadelhoffer, T. (2004). Blame, Badness, and Intentional Action: A Reply to Knobe and Mendlow. *Journal of Theoretical and Philosophical Psychology*, *24*(2), 259-269. (Content-based, unsuccessful, osf.io/w9bza/)

Random 2: Nichols, S. (2004). After objectivity: An empirical study of moral judgment. *Philosophical Psychology*, *17*(1), 3-26. [Study 3] (Content-based, successful, osf.io/bv4ep/)

***2005**

Most cited: Nahmias, E., Morris, S., Nadelhoffer, T., & Turner, J. (2005). Surveying freedom: Folk intuitions about free will and moral responsibility. *Philosophical Psychology*, *18*(5), 561-584. [Study 1] (Content-based, successful, osf.io/4gvd5/)

Random 1: McCann, H. J. (2005). Intentional action and intending: Recent empirical studies. *Philosophical Psychology*, *18*(6), 737-748. [Study 1] (Context-based, null effect, successful, osf.io/jtsnn/)

Random 2: Nadelhoffer, T. (2005). Skill, luck, control, and intentional action. *Philosophical Psychology*, *18*(3), 341-352. [Study 1] (Content-based, successful, osf.io/6ds5e/)

***2006**

Most cited: ~~Cushman, F., Young, L., & Hauser, M. (2006). The role of conscious reasoning and intuition in moral judgment testing three principles of harm. *Psychological Science*, *17*(12), 1082-1089.~~

- *Replacement:* Nahmias, E., Morris, S. G., Nadelhoffer, T., & Turner, J. (2006). Is incompatibilism intuitive? *Philosophy and Phenomenological Research*, *73*(1), 28-53. [Study 2] (Content-based, unsuccessful, osf.io/m8t3k/)

Random 1: Knobe, J., & Burra, A. (2006). The folk concepts of intention and intentional action: A cross-cultural study. *Journal of Cognition and Culture*, *6*(1), 113-132. (Content-based, successful, osf.io/p48sa/)

- *Replacement:* ~~Malle, B. F. (2006). Intentionality, morality, and their relationship in human judgment. *Journal of Cognition and Culture*, *6*(1), 87-112.~~
- *Replacement:* Nichols, S. (2006). Folk intuitions on free will. *Journal of Cognition and Culture*, *6*(1), 57-86. [Study 2] (Content-based, successful, osf.io/8kf3p/)

Random 2: Nadelhoffer, T. (2006). Bad acts, blameworthy agents, and intentional actions: Some problems for juror impartiality. *Philosophical Explorations*, *9*(2), 203-219. (Content-based, successful, osf.io/bv42c/)

***2007**

Most cited: Nichols, S., & Knobe, J. (2007). Moral responsibility and determinism: The cognitive science of folk intuitions. *Nous*, *41*(4), 663-685. [Study 1] (Content-based, successful, osf.io/stjwg/)

Random 1: Nahmias, E., Coates, D. J., & Kvaran, T. (2007). Free will, moral responsibility, and mechanism: Experiments on folk intuitions. *Midwest studies in Philosophy*, *31*(1), 214-242. (Content-based, successful, osf.io/pjdkg/)

Random 2: Livengood, J., & Machery, E. (2007). The folk probably don't think what you think they think: Experiments on causation by absence. *Midwest Studies in Philosophy*, *31*(1), 107-127. [Study 1] (Content-based, successful, osf.io/7er6r/)

***2008**

Most cited: 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. (Context-based, unsuccessful, but with deviations from the original procedure, see osf.io/yb38c)

Random 1: Gonnerman, C. (2008). Reading conflicted minds: An empirical follow-up to Knobe and Roedder. *Philosophical Psychology*, *21*(2), 193-205. (Content-based, successful, osf.io/wy8ab/)

Random 2: Nadelhoffer, T., & Feltz, A. (2008). The actor–observer bias and moral intuitions: adding fuel to Sinnott-Armstrong's fire. *Neuroethics*, *1*(2), 133-144. (Context-based, unsuccessful, osf.io/jb8yp/)

***2009**

Most cited: Hitchcock, C., & Knobe, J. (2009). Cause and norm. *The Journal of Philosophy*, *106*(11), 587-612. (Content-based, successful, osf.io/ykt7z/)

Random 1: Roxborough, C., & Cumby, J. (2009). Folk psychological concepts: Causation. *Philosophical Psychology*, *22*(2), 205-213. (Content-based, unsuccessful, osf.io/5eanz/)

Random 2: Nadelhoffer, T., Kvaran, T., & Nahmias, E. (2009). Temperament and intuition: A commentary on Feltz and Cokely. *Consciousness and Cognition*, *18*(1), 351-355. (Demographic effect, null effect, unsuccessful, osf.io/txs86/)

***2010**

Most cited: Beebe, J. R., & Buckwalter, W. (2010). The epistemic side- effect effect. *Mind & Language*, *25*(4), 474-498. (Content-based, successful, osf.io/n6r3b/)

Random 1: ~~Lam, B. (2010). Are Cantonese-speakers really descriptivists? Revisiting cross-cultural semantics. *Cognition*, *115*(2), 320-329.~~

- *Replacement:* Sytsma, J., & Machery, E. (2010). Two conceptions of subjective experience. *Philosophical Studies*, *151*(2), 299-327. [Study 1] (Demographic effect, successful, osf.io/z2fj8/)

Random 2: De Brigard, F. (2010). If you like it, does it matter if it's real? *Philosophical Psychology*, *23*(1), 43-57. (Content-based, successful, osf.io/cvuwy/)

***2011**

Most cited: Alicke, M. D., Rose, D., & Bloom, D. (2011). Causation, norm violation, and culpable control. *The Journal of Philosophy*, *108*(12), 670-696. [Study 1] (Content-based, unsuccessful, osf.io/4yuym/)

Random 1: ~~Zalla, T., & Leboyer, M. (2011). Judgment of intentionality and moral evaluation in individuals with high functioning autism. *Review of Philosophy and Psychology*, *2*(4), 681-698.~~

- *Replacement:* Reuter, K. (2011). Distinguishing the Appearance from the Reality of Pain. *Journal of Consciousness Studies*, *18*(9-10), 94-109. (Observational data, successful, osf.io/3sn6j/)

Random 2: Sarkissian, H., Park, J., Tien, D., Wright, J. C., & Knobe, J. (2011). Folk moral relativism. *Mind & Language*, *26*(4), 482-505. [Study 1] (Content-based, successful, osf.io/cy4b6/)

***2012**

Most cited: Paxton, J. M., Ungar, L., & Greene, J. D. (2012). Reflection and reasoning in moral judgment. *Cognitive Science*, *36*(1), 163-177. [Study 1] (Context-based, unsuccessful, osf.io/ejmyw/)

Random 1: Schaffer, J., & Knobe, J. (2012). Contrastive knowledge surveyed. *Noûs*, *46*(4), 675-708. [Study 1] (Content-based, successful, osf.io/z4e45/)

Random 2: May, J., & Holton, R. (2012). What in the world is weakness of will? *Philosophical Studies*, *157*(3), 341-360. [Study 3] (Content-based, successful, osf.io/s37h6/)

***2013**

Most cited: Nagel, J., San Juan, V., & Mar, R. A. (2013). Lay denial of knowledge for justified true beliefs. *Cognition*, 129(3), 652-661. (Content-based, successful, osf.io/6yfxz/)

Random 1: Beebe, J. R., & Shea, J. (2013). Gettierized Knobe effects. *Episteme*, *10*(3), 219. (Content-based, successful, osf.io/k89fc/)

Random 2: Rose, D., & Nichols, S. (2013). The lesson of bypassing. *Review of Philosophy and Psychology*, *4*(4), 599-619. [Study 1] (Content-based, null effect, successful, osf.io/ggw7c/)

***2014**

Most cited: Murray, D., & Nahmias, E. (2014). Explaining away incompatibilist intuitions. *Philosophy and Phenomenological Research*, *88*(2), 434-467. [Study 1] (Content-based, successful, osf.io/rpkjk/)

Random 1: Grau, C., & Pury, C. L. (2014). Attitudes towards reference and replaceability. *Review of Philosophy and Psychology*, *5*(2), 155-168. (Demographic effect, unsuccessful, osf.io/xrhqe/)

Random 2: Liao, S., Strohminger, N., & Sripada, C. S. (2014). Empirically investigating imaginative resistance. *The British Journal of Aesthetics*, *54*(3), 339-355. [Study 2] (Content-based, successful, osf.io/7e8hz/)

***2015**

Most cited: Buckwalter, W., & Schaffer, J. (2015). Knowledge, stakes, and mistakes. *Noûs*, *49*(2), 201-234. [Study 1] (Content-based, successful, osf.io/2ukpq/)

Random 1: Björnsson, G., Eriksson, J., Strandberg, C., Olinder, R. F., & Björklund, F. (2015). Motivational internalism and folk intuitions. *Philosophical Psychology*, *28*(5), 715-734. [Study 2] (Content-based, successful, osf.io/d8uvg/)

Random 2: Kominsky, J. F., Phillips, J., Gerstenberg, T., Lagnado, D., & Knobe, J. (2015). Causal superseding. *Cognition*, *137*, 196-209. [Study 1] (Content-based, successful, osf.io/f5svw/)

Appendix 2. Pre-replication form

Reference of the paper: …

Replication team: …

---

*Which study in the paper do you replicate? …

*If it is not the first study, please explain your choice: …

*In this study, what is the main result you will focus on during replication? Please give all relevant statistical details present in the paper: …

*What is the corresponding hypothesis? ...

*What is the corresponding effect size? …

*Was the original effect size:

- Explicitly reported in the original paper
- Not explicitly reported in the original paper, but inferable from other information present in the original paper
- Not inferable from information present in the original paper.

*What is the corresponding confidence interval (if applicable)?

*Was the original confidence interval:

- Explicitly reported in the original paper
- Not explicitly reported in the original paper, but inferable from other information present in the original paper
- Not inferable from information present in the original paper.

*From which population was the sample used in the original study drawn? (Which country, language, students/non-students, etc.)

*Was the nature of the original population:

- Explicitly reported in the original paper

- Not explicitly reported in the original paper, but inferable from other information present in the original paper
- Not inferable from information present in the original paper.

*What was the original sample size (N): …

*Was the original sample size:

- Explicitly reported in the original paper
- Not explicitly reported in the original paper, but inferable from other information present in the original paper
- Not inferable from information present in the original paper.

*Does the study involve a selection procedure (e.g. comprehension checks)? (YES/NO)

*If YES, describe it briefly: …

*Were all the steps of the selection procedure (including, e.g., comprehension checks):

- Explicitly reported in the original paper
- Not explicitly reported in the original paper, but inferable from other information present in the original paper
- Not inferable from information present in the original paper.

*Overall, would you say that the original paper contained all the information necessary to properly conduct the replication (YES/NO)

*If NO, explain what information was lacking: ...

Power analysis and required sample size:

(Please, describe briefly the power analysis you conducted to determine the minimum required sample size. If the original effect is a null effect, just describe the required sample size you obtained by doubling the original sample size.)

Projected sample size:

(Please, describe the actual sample size you plan to use in the replication.)

Appendix 3. Post-replication form

Reference of the paper: …

Replication team: …

Methods

Power analysis and required sample size:

(Please, describe briefly the power analysis you conducted to determine the minimum required sample size. If the original effect is a null effect, just describe the required sample size you obtained by doubling the original sample size.)

Actual sample size and population:

(Describe the number of participants you actually recruited, and the nature of the population they are drawn from. Indicate whether the number of participants you actually recruited matched the one you planned on the OSF pre-registration. Describe briefly any difference between the population you drew your sample from and the population the original study drew its sample from.)

Materials and Procedure:

(Describe the procedure you employed for the replication, like you would in the Methods section of a paper. At the end, indicate all important differences between the original study and replication, e.g. language,)

Results

Data analysis - Target effect:

(Focusing on the effect you singled out as the target effect for replication, describe the results you obtained. Then describe the statistical analyses you performed, detailing the effect size, the significance of the effect and, when applicable, the confidence interval.)

Data analysis - Other effects:

(If the original study included other effects and you performed the corresponding analyses, please, describe them in this section.)

Data analysis - Exploratory Analysis:

(If you conducted additional analyses that were absent from the original study, feel free to report them here. Just indicate whether they were planned in the OSF pre-registration, or exploratory.)

Discussion

Success assessment:

(Did you succeed in replicating the original result? If applicable, does the original team agree with you?)

References

Alfano, M. & Loeb, D. (2017). Experimental Moral Philosophy. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy (Fall 2017 Edition)*. Retrieved from https://plato.stanford.edu/archives/fall2017/entries/experimental-moral/

American Statistical Association. (2016). *American Statistical Association statement on statistical significance and p-values. American Statistical Association*. Retrieved from http://www.amstat.org/asa/files/pdfs/P-ValueStatement.pdf

Amrhein, V., & Greenland, S. (2017). Remove, rather than redefine, statistical significance. *Nature Human Behaviour*. https://doi.org/10.1038/s41562-017-0224-0

Anderson, S. F., Kelley, K., & Maxwell, S. E. (2017). Sample-size planning for more accurate statistical power: a method adjusting sample effect sizes for publication bias and uncertainty. *Psychological science*, *28*(11), 1547-1562. https://doi.org/10.1177%2F0956797617723724

Baker, M. (2016). Is there a reproducibility crisis? *Nature*, *533*(1), 452–454.

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk, R., … Johnson, V. E. (in press). Redefine statistical significance. *Nature Human Behaviour*. https://doi.org/10.1038/s41562-017-0189-z

Boyle, G. J. (in press). Proving a negative? Methodological, statistical, and psychometric flaws in Ullmann et al. (2017) PTSD study. *Journal of Clinical and Translational Research*.

Brandt, M. J., IJzerman, H., Dijksterhuis, A., Farach, F. J., Geller, J., Giner-Sorolla, R., … van 't Veer, A. (2014). The replication recipe: what makes for a convincing replication? *Journal of Experimental Social Psychology*, *50* (Supplement C), 217–224. https://doi.org/10.1016/j.jesp.2013.10.005

Buckwalter, W. (2010). Knowledge isn't closed on Saturday: A study in ordinary language. *Review of Philosophy and Psychology*, *1*(3), 395-406. https://doi.org/10.1007/s13164-010-0030-3

Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., & Munafò, M. R. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*(5), 365-376. https://doi:10.1038/nrn3475

Casler, K., Bickel, L., & Hackett, E. (2013). Separate but equal? A comparison of participants and data gathered via Amazon's MTurk, social media, and face-to-face behavioral testing. *Computers in Human Behavior*, *29*(6), 2156-2160. https://doi.org/10.1016/j.chb.2013.05.009

Cesario, J. (2014). Priming, replication, and the hardest science. *Perspectives on Psychological Science*, *9*(1), 40–48. https://doi.org/10.1177/1745691613513470

Chambers, C., & Munafò, M. (2013, June 5). Trust in science would be improved by study pre-registration. *The Guardian*. Retrieved from http://www.theguardian.com/science/blog/2013/jun/05/trust-in-science-study-pre-registration

Champely, S. (2018). Package 'pwr'. Retrieved from http://cran.r-project.org/package=pwr

Chang AC, Li P (2015) Is Economics Research Replicable? Sixty Published Papers from Thirteen Journals Say "Usually Not", *Finance and Economics Discussion Series 2015-083*. (Board of Governors of the Federal Reserve System, Washington, DC).

Clavien, C., Tanner, C. J., Clément, F., & Chapuisat, M. (2012). Choosy moral punishers. *PloS One*, *7*(6), e39002. https://doi.org/10.1371/journal.pone.0039002

Collins, H. M. (1975). The seven sexes: a study in the sociology of a phenomenon, or the replication of experiments in physics. *Sociology*, *9*(2), 205–224. https://doi.org/10.1177/003803857500900202

Colombo, M., Duev, G., Nuijten, M. B., & Sprenger, J. (2017, November 17). Statistical Reporting Inconsistencies in Experimental Philosophy. Retrieved from https://osf.io/preprints/socarxiv/z65fv

Cova, F. (2012). Qu'est-ce que la philosophie expérimentale ? In Cova, F., Dutant, J., Machery, E., Knobe, J., Nichols, S. & Nahmias, E. (Eds.), *La Philosophie Expérimentale*, Paris: Vuibert.

Cova, F. (2016). The folk concept of intentional action: Empirical approaches. In W. Buckwalter & J. Sytsma (Eds.), *A Companion to Experimental Philosophy*, pp. 121-141. Wiley-Blackwell.

Cova, F. (2017). What happened to the Trolley Problem? *Journal of Indian Council of Philosophical Research*, *34*(3), 543-564.

Crandall, C. S., & Sherman, J. W. (2016). On the scientific superiority of conceptual replications for scientific progress. *Journal of Experimental Social Psychology*, *66* (Supplement C), 93–99. https://doi.org/10.1016/j.jesp.2015.10.002

Cullen, S. (2010). Survey-driven romanticism. *Review of Philosophy and Psychology*, *1*(2), 275-296. https://doi.org/10.1007/s13164-009-0016-1

Cumming, G. (2013). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. Routledge.

Del Re, A. C. (2015). Package "compute.es". Available from https://cran.r-project.org/web/packages/compute.es/compute.es.pdf [accessed 08 April 2018]

De Villiers, J., Stainton, R. J., & Szatmari, P. (2007). Pragmatic abilities in autism spectrum disorder: A case study in philosophy and the empirical. *Midwest Studies in Philosophy*, *31*(1), 292-317. doi:10.1111/j.1475-4975.2007.00151.x

Doyen, S., Klein, O., Simons, D. J., & Cleeremans, A. (2014). On the other side of the mirror: priming in cognitive and social psychology. *Social Cognition*, *32* (Supplement), 12–32. https://doi.org/10.1521/soco.2014.32.supp.12

Dunaway, B., Edmonds, A., & Manley, D. (2013). The folk probably do think what you think they think. *Australasian Journal of Philosophy*, *91*(3), 421-441.

Earp, B. D. (2017). The need for reporting negative results – a 90 year update. *Journal of Clinical and Translational Research*, *3*(S2), 1–4. http://dx.doi.org/10.18053/jctres.03.2017S2.001

Earp, B. D. (in press). Falsification: How does it relate to reproducibility? In J.-F. Morin, C. Olsson, & E. O. Atikcan (Eds.), *Key Concepts in Research Methods*. Abingdon: Routledge.

Earp, B. D., & Trafimow, D. (2015). Replication, falsification, and the crisis of confidence in social psychology. *Frontiers in Psychology*, *6*(621), 1–11. https://doi.org/10.3389/fpsyg.2015.00621

Earp, B. D., & Wilkinson, D. (2017). The publication symmetry test: a simple editorial heuristic to combat publication bias. *Journal of Clinical and Translational Research*, *3*(S2), 5-7. http://dx.doi.org/10.18053/jctres.03.2017S2.002

Feltz, A., & Cova, F. (2014). Moral responsibility and free will: A meta-analysis. *Consciousness and Cognition*, *30*, 234-246. https://doi.org/10.1016/j.concog.2014.08.012

Feltz, A., & Zarpentine, C. (2010). Do you know more when it matters less? *Philosophical Psychology*, *23*(5), 683-706. https://doi.org/10.1080/09515089.2010.514572

Fiedler, K., & Schwarz, N. (2016). Questionable research practices revisited. *Social Psychological and Personality Science*, *7*(1), 45-52. https://doi.org/10.1177/1948550615612150

Findley, M. G., Jensen, N. M., Malesky, E. J., & Pepinsky, T. B. (2016). Can results-free review reduce publication bias? The results and implications of a pilot study. *Comparative Political Studies*, *49*(13), 1667–1703. https://doi.org/10.1177/0010414016655539

Fraley, R. C., & Vazire, S. (2014). The N-pact factor: Evaluating the quality of empirical journals with respect to sample size and statistical power. *PloS One*, *9*(10), e109019. https://doi.org/10.1371/journal.pone.0109019

Franco, A., Malhotra, N., & Simonovits, G. (2014). Publication bias in the social sciences: unlocking the file drawer. *Science*, *345*(6203), 1502–1505. https://doi.org/10.1126/science.1255484

Gilbert, D. T., King, G., Pettigrew, S., & Wilson, T. D. (2016). Comment on "Estimating the reproducibility of psychological science." *Science*, *351*(6277), 1037–1037. https://doi.org/10.1126/science.aad7243

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. https://doi.org/10.1126/science.1062872

Grens, K. (2014). The rules of replication. Retrieved November 8, 2017, from http://www.the-scientist.com/?articles.view/articleNo/41265/title/The-Rules-of-Replication/

Hendrick, C. (1990). Replications, strict replications, and conceptual replications: are they important? *Journal of Social Behavior and Personality; Corte Madera, CA*, *5*(4), 41–49.

Heine, S. J., Lehman, D. R., Peng, K., & Greenholtz, J. (2002). What's wrong with cross-cultural comparisons of subjective Likert scales? The reference-group effect. *Journal of Personality and Social Psychology*, *82*(6), 903-918. DOI: 10.1037//0022-3514.82.6.903

Hendrick, C. (1990). Replications, strict replications, and conceptual replications: are they important? *Journal of Social Behavior and Personality*, *5*(4), 41–49.

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, *2*(8), e124. https://doi.org/10.1371/journal.pmed.0020124

Ioannidis, J. P. A. (2014). How to make more published research true. *PLOS Medicine*, *11*(10), e1001747. https://doi.org/10.1371/journal.pmed.1001747

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. https://doi.org/10.1177/0956797611430953

Knobe, J. (2016). Experimental philosophy is cognitive science. In J. Sytsma & W. Buckwalter (Eds.), *A Companion to Experimental Philosophy* (pp. 37–52). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118661666.ch3

Knobe, J., & Nichols, S. (2008). *Experimental Philosophy*. Oxford University Press.

Knobe, J., Buckwalter, W., Nichols, S., Robbins, P., Sarkissian, H., & Sommers, T. (2012). Experimental philosophy. *Annual Review of Psychology*, *63*(1), 81–99. https://doi.org/10.1146/annurev-psych-120710-100350

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. *Frontiers in Psychology*, *4*, 863.

Lakens, D., Adolfi, F. G., Albers, C., Anvari, F., Apps, M. A. J., Argamon, S. E., … Zwaan, R. (2017). Justify your alpha: a response to "Redefine statistical significance." *PsyArXiv*. https://doi.org/10.17605/OSF.IO/9S3Y6

Lash, T. L., & Vandenbroucke, J. P. (2012). Should preregistration of epidemiologic study protocols become compulsory? Reflections and a counterproposal. *Epidemiology*, *23*(2), 184–188. https://doi.org/10.1097/EDE.0b013e318245c05b

Li, J., Liu, L., Chalmers, E., & Snedeker, J. (2018). What is in a name?: The development of cross-cultural differences in referential intuitions. *Cognition*, *171*, 108-111. https://doi.org/10.1016/j.cognition.2017.10.022

Liao, S. (2015). The state of reproducibility in experimental philosophy. Retrieved from http://philosophycommons.typepad.com/xphi/2015/06/the-state-of-reproducibility-in-experimental-philosophy.html

Locascio, J. (2017). Results blind science publishing. *Basic and Applied Social Psychology*, *39*(5), 239-246. https://doi.org/10.1080/01973533.2017.1336093

Machery, E. (2017a). *Philosophy within its proper bounds*. Oxford: Oxford University Press.

Machery, E. (2017b). What is a replication? Unpublished manuscript.

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. https://doi.org/10.1177/1745691612460688

Maxwell, S. E., Lau, M. Y., & Howard, G. S. (2015). Is psychology suffering from a replication crisis? What does "failure to replicate" really mean? *The American Psychologist*, *70*(6), 487–498. https://doi.org/10.1037/a0039400

McShane, B. B., Gal, D., Gelman, A., Robert, C., & Tackett, J. L. (2017). Abandon Statistical Significance. *arXiv preprint*. arXiv:1709.07588.

Munafò, M. R., Nosek, B. A., Bishop, D. V. M., Button, K. S., Chambers, C. D., Sert, N. P. du, … Ioannidis, J. P. A. (2017). A manifesto for reproducible science. *Nature Human Behaviour*, *1*(21), 1–9. https://doi.org/10.1038/s41562-016-0021

Murtaugh, P. A. (2014). In defense of p-values. *Ecology*, *95*(3), 611–617. https://doi.org/10.1890/13-0590.1

Nakagawa, S., & Parker, T. H. (2015). Replicating research in ecology and evolution: feasibility, incentives, and the cost-benefit conundrum. *BMC Biology*, *13*(88), 1–6. https://doi.org/10.1186/s12915-015-0196-3

Nosek, B. A., & Errington, T. M. (2017). Reproducibility in cancer biology: making sense of replications. *eLife*, *6*, e23383. https://doi.org/10.7554/eLife.23383

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences*, 201708274.

O'Neill, E. & Machery, E. (2014). Experimental Philosophy: What is it good for? In Machery, E. & O'Neill, E. (Eds.), *Current Controversies in Experimental Philosophy*, New York: Routledge.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716. https://doi.org/10.1126/science.aac4716

Rose, D., & Danks, D. (2013). In defense of a broad conception of experimental philosophy. *Metaphilosophy*, *44*(4), 512-532. doi:10.1111/meta.12045

Rose, D., Machery, E., Stich, S., Alai, M., Angelucci, A., Berniūnas, R., ... & Cohnitz, D. (in press). Nothing at stake in knowledge. *Noûs*.

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, *86*(3), 638–641. https://doi.org/10.1037/0033-2909.86.3.638

Schmidt, S. (2009). Shall we really do it again? The powerful concept of replication is neglected in the social sciences. *Review of General Psychology*, *13*(2), 90–100. https://doi.org/10.1037/a0015108

Scott, S. (2013, July 25). Pre-registration would put science in chains. Retrieved July 29, 2017, from https://www.timeshighereducation.com/comment/opinion/pre-registration-would-put-science-in-chains/2005954.article

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359-1366. https://doi.org/10.1177/0956797611417632

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). P-curve: a key to the file-drawer. *Journal of Experimental Psychology: General*, *143*(2), 534.

Sprouse, J., & Almeida, D. (2017). Setting the empirical record straight: Acceptability judgments appear to be reliable, robust, and replicable. *Behavioral and Brain Sciences*, *40*.

Stroebe, W., & Strack, F. (2014). The alleged crisis and the illusion of exact replication. Perspectives on Psychological Science, 9(1), 59–71.

Trafimow, D., & Earp, B. D. (2017). Null hypothesis significance testing and Type I error: the domain problem. *New Ideas in Psychology*, *45*, 19–27.

Weinberg, J. M., Nichols, S., & Stich, S. (2001). Normativity and epistemic intuitions. Philosophical topics, 29(1/2), 429-460.

Woolfolk, R. L. (2013). Experimental philosophy: A methodological critique. *Metaphilosophy*, *44*(1-2), 79-87. https://doi.org/ 10.1111/meta.12016