Love and Power: Grau and Pury (2014) as a Case Study in the Challenges of X-Phi Replication

Edouard Machery, Christopher Grau & Cynthia L. Pury

Review of Philosophy and Psychology

ISSN 1878-5158

Rev.Phil.Psych.
DOI 10.1007/s13164-020-00465-x

Springer

Review of Philosophy and Psychology
https://doi.org/10.1007/s13164-020-00465-x

Love and Power: Grau and Pury (2014) as a Case Study in the Challenges of X-Phi Replication

Edouard Machery[1,2] · Christopher Grau[3] · Cynthia L. Pury[4]

Published online: 07 March 2020

Abstract
Grau and Pury (*Review of Philosophy and Psychology, 5*, 155–168, 2014) reported that people's views about love are related to their views about reference. This surprising effect was however not replicated in Cova et al.'s (in press) replication study. In this article, we show that the replication failure is probably due to the replication's low power and that a metaanalytic reanalysis of the result in Cova et al. suggests that the effect reported in Grau and Pury is real. We then report a large, highly powered replication that successfully replicates Grau and Pury 2014. This successful replication is a case study in the challenges involved in replicating scientific work, and our article contributes to the discussion of these challenges.

1 Doing Replication Correctly

1.1 Challenges in Conducting Replications

While replications have been rarely published, and probably rarely conducted, for decades (Makel et al. 2012), for the last seven or eight years psychologists have been intensely concerned with assessing the replicability of experimental results. This development has been rich in lessons: First and foremost, we have learned that a surprisingly large proportion of experimental results, including influential, widely celebrated experimental results, fail to replicate. But we have also learned a lot about the challenges and perils of conducting replications. In this section, we elaborate on these, with a special focus on replications' power and sample size.

✉ Edouard Machery
machery@pitt.edu

[1] Department of History and Philosophy of Science, University of Pittsburgh, Pittsburgh, PA, USA

[2] Center for Philosophy of Science, 1108CL, Pittsburgh, PA 15260, USA

[3] Department of Philosophy, Clemson University, Clemson, SC, USA

[4] College of Behavioral, Social, and Health Sciences, Clemson University, Clemson, SC, USA

Replications can fail for many reasons, and, when a replication failure can be chalked to one of these reasons, little can be learnt from the failure. We will call such reasons "replication underminers." While we do not plan to be exhaustive here, the following replication underminers have turned out to be important:

- Limited generalizability of the original finding: A replication can fail because the original study and its replication sample from two distinct populations, and because the effect holds only from the original population (Machery Forthcoming).
- Limited validity of the manipulations: A replication can fail because a manipulation that was effective in the past or is effective in one population is not effective anymore when the replication is conducted or is not effective in another population.[1]
- Failure of measurement invariance: A replication can fail because measurement is not invariant across times or populations.
- Insufficient power: A replication can fail because it has insufficient power.

While these replication underminers have all been extensively discussed, the fourth one is of particular importance for our purposes. If the power of the replication is low, then the probability that the replication will successfully reject the false null hypothesis is low even if the original effect was genuine, and a replication failure would not be informative. Psychologists have naturally been aware of the importance of the power of replications, and have strived to do highly powered replications. Because the power of a given test is a function of the sample size, the size of the true effect, and the significance level, psychologists mostly increase power by increasing the sample size. Psychologists conducting replication have typically determined the sample size required to obtain a sufficient power to detect the effect size reported in the original experiments (e.g., at least .8 in Open Science Collaboration 2015; a .9 power to detect 75% of the original effect in Camerer et al. 2018). The problem with this approach, however, is that the effect sizes reported in the original experiments are typically substantially inflated: When experiments are insufficiently powered, statistical significance is obtained by capitalizing on chance, and the significant effect sizes (the only ones to be published) are larger than the population effect sizes. As a result, the true power of the replication is lower, possibly substantially, than the power computed on the basis of the published effect size.

The large replication efforts conducted over the last few years illustrate the issues just discussed. Open Science Collaboration (2015) attempted to replicate 100 studies in cognitive and social psychology. Sample size was determined so as to have at least an .80 power (with an average of .92) to detect an effect size equal to the one reported in the original study. Shockingly, only 39% of the studies were successfully replicated (when successful replication is decided on the basis of statistical significance at the .05 significance level). The effect sizes observed in the replications were almost always smaller than the effect sizes originally reported. When Camerer et al. (2018) attempted to replicate the studies published in *Science* and *Nature*, they found that slightly more

[1] In this case, the effect may exist in both the original population and the population of the replication, but may not be observed in the replication.

than 60% of the studies successfully replicated and that the average effect size was only 46.2% of the original average effect size.

The explanation of the apparent low replicability of the studies published in psychology is no doubt complex: Some of the original studies that failed to replicate were innocent false positives while others certainly resulted from p-hacking or the flexibility of data analysis (Simmons et al. 2011). However, some replication failures probably also resulted from the insufficient power of the replication despite psychologists' efforts to run highly powered replications. Etz and Vandekerckhove (2016) reanalyzed 72 of the 100 replication attempts reported in Open Science Collaboration (2015). They computed the Bayes factors for the 72 original studies (chosen because of their univariate data analysis) and their replications. Bayes factors were computed by contrasting the null model and the information unit prior (a normal distribution centered at 0 with a standard deviation equal to 1). They found that many of the replications (as well as many of the original findings) provided little evidence for or against the null model because replications' sample sizes are not large enough to undermine the null model decisively. As they put it (2016, 1), "the apparent failure of the Reproducibility Project to replicate many target effects can be adequately explained by overestimation of effect sizes (or overestimation of evidence against the null hypothesis) due to small sample sizes and publication bias in the psychological literature."

1.2 Replication and Meta-Analytic Estimates

It has long been known that low powered literatures can create apparent inconsistencies: Some studies may report significant results by sheer chance, while others may fail to obtain significant results because of their low power (e.g., Schmidt 1996, 2010). Counting only the number of significant and non-significant results would thus suggest disagreement among studies that may in reality substantially agree: Their effect sizes may be comparable (although some may be significant, while others wouldn't) or they may at least be in the same direction.

To prevent being misled by the appearance of disagreement among studies due to low power, meta-analysts have suggested aggregating their results by means of metanalytic tools: Instead of counting the number of significant and non-significant results, meta-analysts propose, scientists ought to estimate the population effect size by aggregating data across studies. Failed replications that may seem to undermine the original results may in fact provide further support for their veracity. Schmidt (1996) gives the following startling example of 21 studies examining whether a single-item test predicts job performance. Because of the small sample size of each study, only 8 of them results in a significant correlation: The literature seems ripe with disagreement and the predictive validity of this single-item test is dubious. However, all the correlations are positive, and in this sense the studies agree. A meta-analytic aggregation of the correlation coefficients shows indeed that there is a genuine correlation between the single-item test and job performance.

Metaanalysts' observation about low powered scientific literatures and their meta-analytic recommendation should be extended to the contemporary debate about the replicability of psychology. If replications are insufficiently powered—a not uncommon situation if the re-analysis of OSC 2015 by Etz and Vandekerckhove (2016) is on the right track—determining replication success and assessing replicability by means of

statistical significance may be misleading, and a meta-analytic approach may be a necessary corrective.

2 Assessing the Replicability of Experimental Philosophy: Perils and Challenges

2.1 The X-Phi Replicability Project

Inspired by Open Science Collaboration (2015), Cova and colleagues (in press) attempted to assess the replicability of experimental philosophy. They selected 40 studies chosen as follows: "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)." Two studies were added for a total of 40 studies. Sample size was determined so as to obtain a .95 power to detect the effect reported in the original studies (although the final average power was .88). Surprisingly, a larger proportion of studies replicated than in similar projects in psychology: 78% of these articles successfully replicated, as assessed by the significance criterion (Fig. 1). In addition, the average effect size for the replications was similar to the average effect size in the original studies.

Additional analyses examined which type of study is less likely to replicate. Cova et al. (in press) compared the papers reporting demographic effects and context effects to the papers reporting "content effects" (i.e., studies investigating how judgments vary when the content of the vignettes is changed), and argue that the former are less likely to replicate (our emphasis): "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."

This secondary analysis is philosophically significant. The negative program of experimental philosophy, which uses experimental-philosophy results to undermine the use of cases in philosophy, appeals primarily to studies reporting demographic and context effects (Machery 2017). If these studies are not trustworthy, the empirical basis of the negative program crumbles, and critics of the method of cases in philosophy are left without an argument.

2.2 Grau and Pury (2014)

In "Attitudes Towards Reference and Replaceability" (2014) Grau and Pury described the results of their study comparing people's intuitions concerning the linguistic reference of proper names with attitudes regarding the replacement of original entities with potential duplicates (including potential duplicates of loved ones). Grau and Pury's study was inspired by Robert Kraut's (1986) essay "Love *De Re,*" which proposed an analogy between the nature of the bond lovers often form to those they love and the manner in which many linguistic terms (particularly proper names) refer. Several philosophers have noted that love often seems to involve an historic tie to an individual such that the love can survive changes in the qualities of the beloved and is not necessarily proportional to any such changes (Nozick 1974): The beloved is taken to be, in important sense, irreplaceable. Kraut proposed a parallel between this sort of

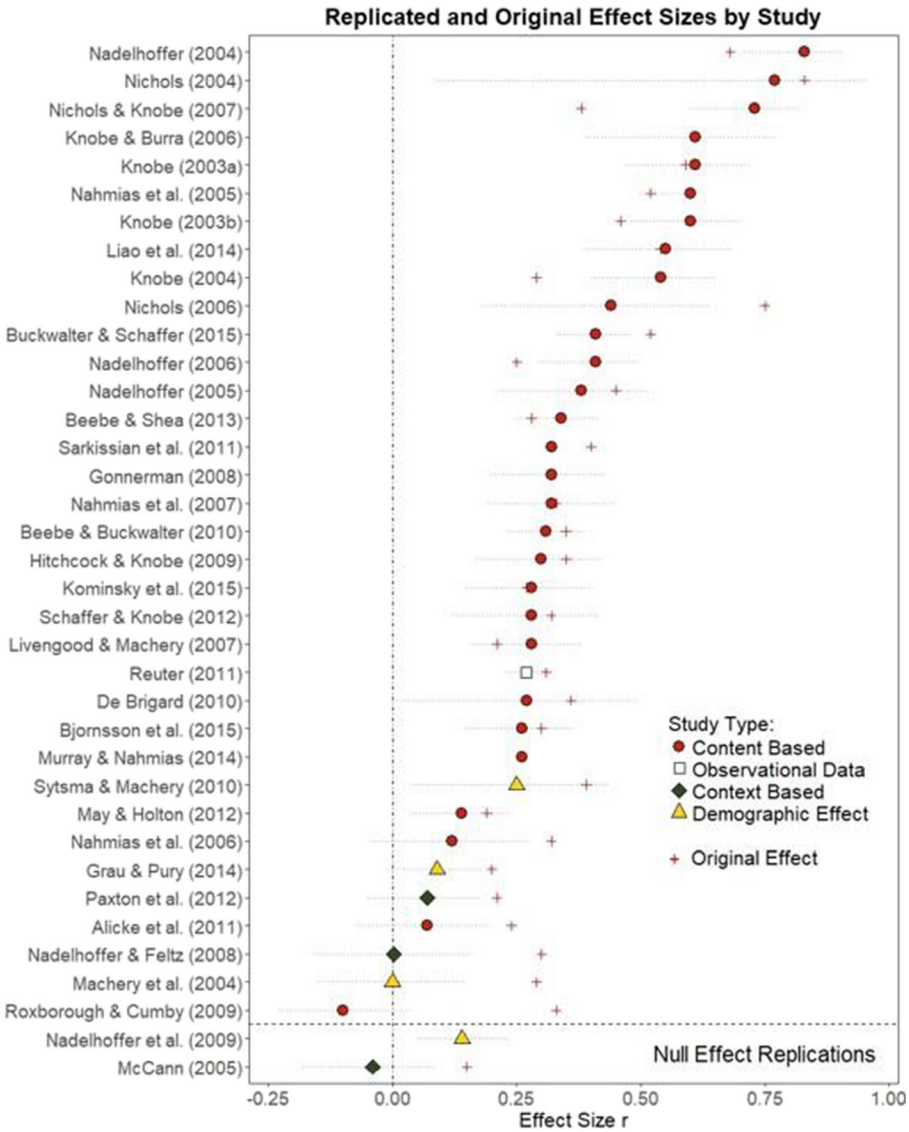

Fig. 1 Results of the 40 Replications in Cova et al. in press (from Cova et al. in press)

"rigid" attachment in love and the manner which proper names "rigidly designate", i.e. Saul Kripke's contention that proper names attach to objects in an historical manner such that reference cannot not switch to an even exactly similar object (Kripke 1980). Kraut summarized his analogy as follows (Kraut 1986, 427):

> It is usually agreed that Lisa's history and origin are essential to her. If so, a name that uniquely refers to Lisa is not properly applicable to any possible object with a history and origin different from Lisa's. And, analogously, a love that is genuinely directed toward Lisa does not get directed toward any object with a history

> and origin different from hers. It is that that confers upon love the property of being directed toward Lisa. And it is that that makes love, at least love of individual persons, historical.

While Kraut wrote as though Kripkean rigid designation is simply the true theory of reference, empirical work on the issue suggests that individuals vary in the degree to which they hold a Kripkean view (that names rigidly refer in virtue of a particular causal history) or alternatively a "descriptivist" view by which a cluster of properties attributed to a name determines its referent (this is sometimes explained in terms of a "description" that uniquely picks out the object to which a name refers) (e.g., Machery et al. 2004; Machery et al. 2009; Machery et al., 2010; Machery et al. 2015; Sytsma et al. 2015; Beebe and Undercoffer 2015, 2016). With this variation in mind, Grau and Pury examined whether people's attitudes regarding reference correspond to attitudes regarding replaceability. In their original study, 162 participants completed an online questionnaire asking them to consider how appropriate it would be to feel the same way about a perfect replica of a loved one, as well as other related questions about replaceability, and a series of questions regarding reference (using the Gödel/Schmidt scenario utilized by Machery and others). The results appeared to provide support for Grau and Pury's key hypothesis that "participants who endorse a Kripkean, rather than a Descriptivist view of linguistic reference should report that it is less appropriate to feel the same way toward a replaced loved one" (Grau and Pury 2014, 160). Participants who previously had endorsed Kripkean reference ($n = 96$) rated loved ones as less replaceable on two different measures than participants who had previously endorsed Descriptivist reference ($n = 66$, $t(160) \geq 2.27$, $p \leq 0.02$, $\eta 2 \geq 0.03$; Table 1).

2.3 The Failed Replication of Grau and Pury

Grau and Pury (2014) was one of the 40 studies replicated by Cova et al. (in press). The replication was conducted by Vilius Dranseika and Renatas Berniūnas (https://osf.io/xrhqe/). As instructed by the leaders of the x-phi replication study, they computed

Table 1 Ratings by Descriptivists (n = 66) and Kripkeans ($n = 96$) of the Liklihood and Normative Appropriateness of the Participant Feeling the Same Way About a Duplicate

Duplicated Target	Linguistic Reference	Likelihood Rating					Normative Appropriateness Rating				
		m	*sd*	*t*(160)	*p*	*eta*^{2}	*m*	*sd*	*t*(160)	*p*	*eta*^{2}
Loved One	Descriptivist	3.1	2.2	2.27	.02	.031	3.3	2.2	2.58	.01	.040
	Kripkean	2.4	1.8				2.5	2.0			
Pet	Descriptivist	3.3	2.2	2.17	.03	.029	3.8	2.1	2.21	.03	.030
	Kripkean	2.7	1.9				3.0	2.1			
Acquiantance	Descriptivist	4.7	1.9	1.03	.31	.007	4.4	2.1	1.25	.21	.010
	Kripkean	4.3	2.0				4.0	2.0			
Shoes	Descriptivist	5.3	1.7	1.13	.26	.008	5.7	1.5	1.52	.13	.014
	Kripkean	4.9	2.0				5.3	1.9			

the sample size required to get a power equal to .95 given the effect reported in the original study, and they found that they needed 383 participants. They recruited 405 participants in Lithuania, from which 18 were excluded, for a final sample size of 387 participants. Dranseika and Berniūnas compared the ratings of normative appropriateness of the participant feeling the same way about a duplicate of the loved one by Descriptivists ($n = 287$) and Kripkeans ($n = 100$), and they failed to find a difference between them at the .05 level ($t(385) = 1.69$, $p = .092$, $d = .196$). Their conclusion was as follows: "We did not succeed in replicating the original results ($p = 0.092$) in Lithuanian language in a sample of Lithuanian students." Dranseika and Berniūnas's findings were reported in Cova et al. (in press).

Dranseika and Berniūnas also conducted a follow-up study that used a modified version of the Gödel case with two different questions in order to identify Descriptivists and Kripkeans (also available at https://osf.io/xrhqe/). The first question follows the first replication and Grau and Pury (2014). They recruited 216 participants in Lithuania, from which 8 were excluded, for a final sample size of 208 participants. Dranseika and Berniūnas compared again the ratings of normative appropriateness of the participant feeling the same way about a duplicate of the loved one by Descriptivists ($n = 140$) and Kripkeans ($n = 68$), and they failed to find a difference between them at the .05 level for both ways of identifying Descriptivists and Kripkeans (first way: Descriptivists ($n = 140$) and Kripkeans ($n = 68$); $t(206) = 1.69$, $p = .103$, $d = .242$; second way: Descriptivists ($n = 53$) and Kripkeans ($n = 155$); $t(206) = .172$, $p = .864$, $d = .027$).

2.4 Challenges to the Replication of Grau and Pury

As we noted in Section 1, failed replications are not always informative, as happens when participants are sampled from another population and the effect is expected to only hold in the original population or when the power of a replication is low. Furthermore, when the power of a replication is low, replication success can be usefully analyzed by computing a meta-analytic estimate of the population effect size. Concerns about power apply straightforwardly to the failed replication of Grau and Pury.[2] First, the power of the replication of Grau and Pury is low for many possible effect sizes. If the effect size is equal to $d = .2$ (a small effect size according to Cohen's (1992) classification), for instance, the power of the experiment is only .4 (computer using G*Power 3.0, Faul et al. 2007): That is, if the null hypothesis is true, the probability of rejecting it is only 40%.[3] If so, one is more likely to get at the truth, if the null hypothesis is true, by throwing a coin! Figure 2 reports the power curve of the replication of Grau and Pury for effect sizes between 0 and .4 (the effect size reported in Grau and Pury (2014) that was used to determine the sample size of the replication). In light of the low power for many possible effect sizes, we should refrain from interpreting the replication failure as indicating that Grau and Pury's results are false positives.

Second, while the effect reported is not significant at the .05 level, it is in the expected direction (see Fig. 1): It does not reach significance because the effect observed in the

[2] Other aspects of the replication might explain the different results (on the importance of these considerations, see Trafimow and Earp 2016): For instance, the vignettes were translated and the participants were from a different country.

[3] The power of the follow-up study is even lower.

replication is substantially smaller than the effect originally reported, a common situation as we have seen. Thus, instead of assessing replicability by means of statistical significance, it is better to compute the meta-analytic estimate, combining the original and replication results.[4] We did not include the follow up study by Dranseika and Berniūnas since it is not an exact replication, but the results of their first analysis are in line with the metaanalytic estimate. The original effect size is $d_O = .381$, the replication effect size is $d_R = .196$.[5] Following a fixed effects metanalytic approach, we computed a weighted r for the original study and its replication (Goh et al. 2016): $r_A = .126$ (95%CI = [.042; .208]) or equivalently $d_A = .254$. The average effect size is significantly different from zero.

Far from suggesting that Grau and Pury's (2014) results do not replicate, the replication failure is more plausibly viewed as resulting from the low power, and, if anything, aggregating the two studies suggests that the effect is genuine. However, the conclusions of meta-analyses can be undermined by questionable research practices in the meta-analyzed studies (Nelson et al. 2018). To confirm the conclusion of our meta-analysis, we decided to go beyond this reanalysis of the data reported by Cova et al. by conducting a genuinely highly powered replication of Grau and Pury (2014).

3 Replication of Grau and Pury

3.1 Methods

Participants were recruited on Amazon Turk, and paid $.33 for their participation in this study. The study was preregistered on OSF (https://osf.io/h32zd) and approved by the IRB of the University of Pittsburgh.[6] The data can be accessed at the OSF link. To obtain a power of .8 assuming a small effect size d = .2, an alpha of .05, and an equal distribution between the two conditions, we determined that a sample size of 394 per condition was needed. Assuming that 15% of participants on Amazon Turk fail the attention and comprehension check, we estimated that our sample size should be at least 905. Assuming a 2 to 1 ratio between Kripkeans and Descriptivists on the basis of past results, we determined that 295 Descriptivists and 591 Kripkeans were needed. Assuming again a 15% failure rate, our total sample size should be 1018.

1018 participants started the survey; 843 completed it; 738 passed attention checks (see below), with 715 of those also indicating they were fluent in English, answering all of the relevant study questions, and stating that they had not participated in a similar study in the past.

Of the 715 participants, 56% ($n = 397$) indicated they were female and 44% ($n = 318$) indicated they were male. All were either American or residing in the United States. Reported highest educational attainment was high school ($n = 98$, or 14%), some college ($n = 171$, or 24%), college degree ($n = 326$, or 46%), grad or master's degree

[4] Meta-analyses are often done on many studies, but can also be done on a few, indeed a couple of studies (Goh et al. 2016).

[5] d_O differs from the effect size reported in Grau and Pury 2014 ($\eta^2 = .4$ corresponding to $d = .408$) and used in the replication of Grau and Pury to compute the desired sample size and to assess whether the observed effect size of the replication was included in the confidence interval of the original study.

[6] Unfortunately, preregistration only included the materials used in the experiment, but not the sample size and exclusion criteria.

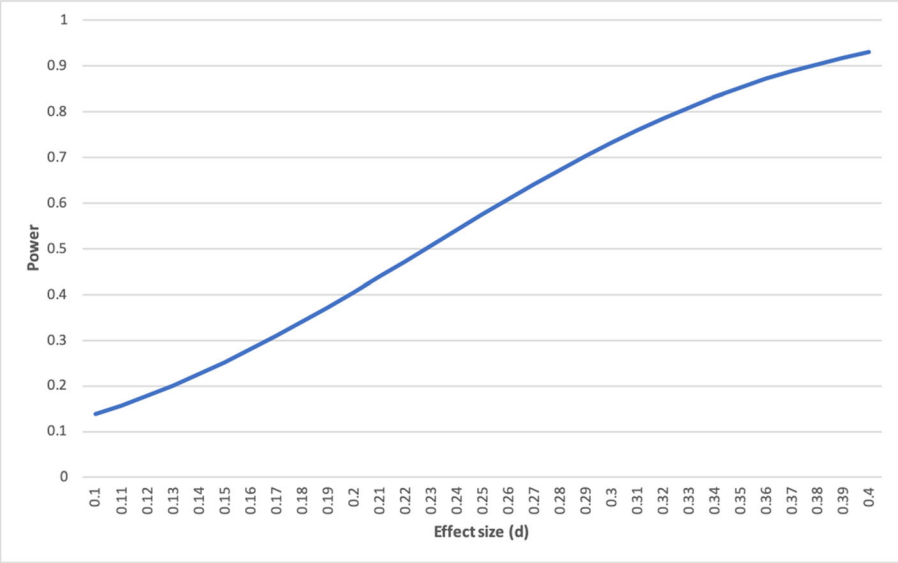

Fig. 2 Power of the Replication of Grau and Pury (2014) for Various Effect Sizes

($n = 105$, or 15%) and PhD ($n = 15$, or 2%). Age ranged from 19 to 81, with a mean age of 37.8 ($sd = 12.1$). Approximately half of participants ($n = 363$, or 51%) had taken at least one philosophy class, while 352 (49%) had not. Of those who had taken at least one course, only 54 (15%) had taken more than one or two courses; therefore, for the purposes of this study participant responses will be analyzed based on some philosophy coursework ($n = 363$) versus no philosophy coursework ($n = 352$).

3.2 Materials and Procedure

We used an online survey replicating Grau and Pury's (2014) questionnaires with additional attention check items built in. One attention check item directed participants to answer two questions in a particular way (e.g., "[P] lease copy and paste the words 'I have read the instructions' (with the quotation marks) in the box labelled 'Any comments or questions?'"), and an additional attention check item verified that participants had read the content of the Gödel question. The text read "John is quite good at mathematics, and he can accurately describe the content of the incompleteness theorem"; The question then asked whether John can or cannot describe the content of the incompleteness theorem.

As in Grau and Pury (2014), participants were classified as either Kripkean or Descriptivist based on their answer to Machery et al.'s (2004) Gödel question:

> Suppose that John has been told in college that Gödel is the man who proved an important mathematical theorem, called "the incompleteness of arithmetic." John is quite good at mathematics, and he can accurately describe the content of the incompleteness theorem. He attributes this theorem to Gödel as the discoverer. But this is the only thing that he has heard about Gödel. Now suppose that Gödel was not the author of this theorem. A man called "Schmidt," whose body was

found in Vienna under mysterious circumstances many years ago, actually did the work in question. His friend Gödel somehow got hold of the manuscript and claimed credit for the work, which was thereafter attributed to Gödel. Thus, Gödel has been known as the man who proved the incompleteness of arithmetic. Most people who have heard the name "Gödel" are like John; the claim that Gödel discovered the incompleteness theorem is the only thing they have ever heard about Gödel.

They were then asked to answer the following question:

When John uses the name 'Gödel', is he talking about:

(A) the person who really discovered the incompleteness of arithmetic? or.

(B) the person who got hold of the manuscript and claimed credit for the work?

For analysis, participants answering A were classified as Descriptivist; participants answering B classified as Kripkean.

Participants then completed replaceability questions for the same targets as Grau and Pury (2014): an unspecified pair of shoes, a favorite pair of shoes, a pair of shoes that were a gift from a friend, a pet one is fond of, a (human) loved one, and a person with whom one has a purely transactional relationship. Unlike Grau and Pury, all participants answered questions about all targets, including all types of shoes.[7] Questions about feelings toward the replacement were the same, including the target questions about the Likelihood ("On a scale from 1 to 7, with (1) being 'not at all likely' and (7) being 'extremely likely', how likely would you be to feel the same way about [the target] as you felt about the original?") and Appropriateness ("On a scale from 1 to 7, with (1) being 'not at all appropriate' and (7) being 'extremely appropriate', how appropriate would it be to feel the same way about [the target]?"). The question about a replaced pet also was followed with the question: "On a scale from 1 to 7, with (1) being 'I strongly disagree' and (7) being 'I strongly agree', to what extent do you agree with the claim 'A pet is like a member of the family'"? Additional questions, not analyzed here, were identical to Grau and Pury and asked about the likelihood and appropriateness of the pet and replaced people feeling the same way about the participant, and about confidence that replicants were identical to the original.

To ensure comparability between results, all effect sizes were translated into the equivalent Cohen's *d* using the spreadsheet developed by DeCoster (2012) and available at http://www.stat-help.com/spreadsheets/Converting%20effect%20sizes%202012-06-19.xls

3.3 Results

Kripkean Versus *Descriptivist Classification:* Overall, 65% ($n = 463$) of participants selected the Kripkean answer to the Gödel question, while 35% ($n = 252$) selected the Descriptivist answer. As in Grau and Pury (2014), Kripkeans were slightly older (mean age = 38.6, $sd = 12.2$) than Descriptivists (mean age = 36.4, $sd = 11.8$, $t(713) = 2.34$, $p = .02$, d = 0.18); Kripkeans and descriptivists did not differ significantly in terms of gender ($\chi 2 = 0.382$, $p = .54$), having taken at least one philosophy course ($\chi 2 < .001$), or believing a pet is like a member of the family ($t(713) = .36$, $p = .72$, d = .03).

[7] This difference was introduced in order to maximize power while keeping the sample size manageable.

Kripkean Versus *Descriptivist Linguistic Reference*: Table 2 presents the ratings of Likelihood and Normative Appropriateness of feeling the same way about a duplicate by Linguistic Reference for each of the duplicate targets. The effect of Linguistic Reference on both ratings of replacement of a loved one replicated the effect reported in Grau and Pury (2014), with Kripkeans making significantly lower ratings than Descriptivists in the likelihood and the appropriateness of responding to the replaced loved one in the same way as the original. These effect sizes ($d = .26$ and $d = .20$, respectively) were similar to those found in the Cova et al. (in press) replication of the original study. While there was a trend for a similar difference in the likelihood of the same response to a pet, it did not reach statistical significance. No other differences were found in the Likelihood or Appropriateness ratings.

Using age and philosophy training as covariates, the relationship between Linguistic Reference and both Likelihood and Appropriateness ratings for a Loved One remained significant ($F(1,711) = 30.28$, $p = .002$, $\eta 2 = 0.013$ for Likelihood, and $F(1,711) = 5.34$, $p = .02$, $\eta 2 = 0.007$ for Appropriateness). Those effect sizes translate to Cohen's d values of .23 and .17, respectively, similar to the effect sizes found without the added covariates.

Further analyses that follow the analyses in Grau and Pury, but are tangential to the replication of the effect of interest are reported in the Appendix.

4 Discussion

4.1 Love and Reference

In contrast to the conclusion drawn by Cova et al. (in press), but in line with the metaanalysis reported in Section 2, we were able to replicate successfully the main finding in Grau and Pury (2014): As predicted, people who have descriptivist leanings, as measured by their judgment about the reference of "Gödel" in the Gödel case, are more likely to feel the same way about the identical duplicate of a lover as they felt about the original lover and they are also more likely to judge it appropriate to do so.

One may object that whether or not we were able to replicate the main finding in Grau and Pury (2014) successfully depends on how success is defined. Because the effect size we observed is much smaller than the effect size reported by Grau and Pury, a critic may insist that our result differs substantially from the main finding in Grau and Pury (2014), as did the one reported in Cova et al.

Replication success can be assessed by different criteria, including whether the replication is statistically significant, whether the confidence interval around the replication's effect size includes the original effect size, and whether the confidence interval around the original study's effect size includes the replication effect size. So which criterion should be used here? First, we simply relied on the main criterion used in Cova et al. (in press), namely significance. Second, which criterion is appropriate depends on the question of interest: If what matters is the effect size, then replication success should be assessed on the basis of confidence intervals; if what matters is the mere existence of a causal relation or of a correlation, then replication success should be assessed on the basis of significance. Grau and Pury were only assessing whether

Table 2 Ratings by Kripkeans ($n = 463$) and Descriptivists ($n = 252$) of the Likelihood and Normative Appropriateness of the participant feeling the same way about a duplicate

Duplicate Target		Likelihood Rating					Appropriateness Rating				
	Linguistic Reference	*m*	*sd*	*t* (713)	*p*	*d*	*m*	*sd*	*t* (713)	*p*	*d*
Loved One	Kripkean	2.5	1.7	3.41	<.01	.26	2.6	1.8	2.61	.01	.20
	Descriptivist	2.9	1.9				3.0	1.9			
Pet	Kripkean	3.0	1.8	1.75	.08	.14	3.2	1.9	1.27	.20	.10
	Descriptivist	3.3	1.8				3.3	1.9			
Acquaintance	Kripkean	4.0	2.0	1.28	.20	.10	3.9	2.0	0.60	.55	.05
	Descriptivist	4.2	2.0				4.0	2.0			
Shoes (Unspecified)	Kripkean	4.7	1.7	.66	.51	.05	5.2	1.6	−0.84	.40	.07
	Descriptivist	4.8	1.7				5.1	1.5			
Shoes (Favorite)	Kripkean	4.5	1.8	.09	.93	0.01	4.8	1.8	.32	.75	.02
	Descriptivist	4.5	1.8				4.9	1.7			
Shoes (Gift)	Kripkean	4.3	1.8	.63	.53	.05	4.6	1.8	−.75	.45	.06
	Descriptivist	4.4	1.8				4.5	1.8			

there was a relation between one's views about love and about reference, and had no hypothesis about effect size. For this reason, significance is a more appropriate criterion for assessing replication success.

In line with Section 2, we report a metanalytic estimate of the effect size, aggregating over the three existing studies (Grau and Pury 2014; Cova et al., in press; and the present study). Following a fixed effects metanalytic approach, we compute a weighted r for these three studies (Goh et al. 2016): $r_A = .116$ (95%CI = [.061; .170]) or equivalently $d_A = .234$. The average effect size is significantly different zero.

While the effect is small according to Cohen (1992), it is real, and extremely striking: it reveals a surprising connection between two apparently disparate domains—what people understand proper names to refer to and their attitude toward the object of their love. People who associate proper names with individuals historically, and thus independently of their characteristics, are more likely to view love as anchored to individuals, independently of what those are like. So to speak, they value a connection to individuals, not to what these individuals are like, and this either when it comes to love or to the interpretation of proper names.

4.2 Lessons about Replication: The Importance of Power and Metaanalysis

Our findings show the importance of interpreting failed replications, such as the failed replication of Grau and Pury (2014) reported in Cova et al., carefully. A failed replication is silent about the veridicality of the original effect when it has a low power for a range of plausible effect sizes: Such low power is a replication underminer. Cova et al. were thus too quick to raise doubts about the reality of the effect reported in Grau and Pury.

How should future replications address this first concern? We recommend two strategies.[8] First, when replicating a study, scientists could focus on the smallest effect size of theoretical interest instead of the effect size reported in the original study. On the basis of the smallest effect size of theoretical interest, scientists would then determine the required sample size to have the desired power. We recommend a power at least equal to .8 when the sample size is planned this way. We followed this strategy in this paper, using the effect size observed in the failed replication as a way to identify on the smallest effect size of theoretical interest. There are two limitations with this approach. First, it is often unclear what the smallest effect size of theoretical interest is. Second, when the smallest effect size of theoretical interest is small, a very large sample size is needed for power to be large enough. (Incidentally, this is why we suggested to set the target power at .8 rather than .9 or .95.)

Alternatively, a two-step procedure can be followed. First, the sample size is determined so as to have a 90% power to detect 75% of the original effect size (alpha at .05). We propose computing the sample size on the basis of 75% of the original effect size to take the inflation of effect size due to low power. No further data is collected if the replication is successful. Second, if the replication is not successful, a follow up replication is run with a sample size set so as to have a 90% power to detect the metaanalytic estimate of the population effect size resulting from the original study and the failed replication.

One may worry that this will make replication efforts at the very least extremely challenging, if not practically impossible. We concede that some experiments may require less stringent requirements given the difficulties involved in collecting data. However, when data are collected on line with on-line pools of participants, much larger sample sizes are perfectly doable, although undoubtedly costly.

Our findings also show the limitations of significance as a way of assessing replication success, particularly when power is low: A failed replication may in fact support the reality of the original effect rather than undermine it. A metaanalytic approach, of the type followed in this article, goes beyond counting positive and negative results: It can reveal that results that seem to speak against one another are in fact consistent. We have argued that the results reported in Cova et al. are indeed consistent with the results of Grau and Pury (2014), and our large replication confirms this claim. While it is beyond the scope of this paper to argue for a particular definition of replication success, we would recommend a pluralistic approach, taking into account several criteria (as was done by Open Science Collaboration 2015).

4.3 Demographic Effects and the Negative Program of Experimental Philosophy

Finally, our findings bear on the proper interpretation of the demographic and context effects in experimental philosophy. Cova et al. attempted to replicate four demographic effects:

- Machery et al. 2004: People in the USA and in East Asia tend to respond differently to the Gödel case.
- Nadelhoffer et al. 2009: Extraversion does not predict compatibilist judgments.

[8] A third strategy is followed by Camerer et al. (2018).

- Sytsma and Machery 2010: philosophers and lay people tend to assign conscious experiences differently.
- Grau and Pury 2014: Kripkeans and Descriptivists tend to understand love differently.

Of those four studies only one—Sytsma and Machery (2010)—successfully replicated.

These results need to be scrutinized closely, however, before drawing any conclusion about the fragility of demographic effects. The failed replication of Machery et al. (2004) is an outlier in a series of successful replications (Machery 2017, Chapter 2), and a recent metaanalysis strongly suggests that judgments elicited by the Gödel case do vary across cultures (van Dongen et al. n.d.). Nadelhoffer et al. claimed, against Feltz and Cokely (2009), that extraversion does *not* predict compatibilist judgments: The failed replication confirmed Feltz and Cokely's original results that people *do* have different intuitions about free will depending on their personality (a point incidentally acknowledged in a footnote by Cova et al.). Feltz and Cokely (2019) also meta-analyze 17 published and 8 unpublished studies (for a total $n = 2811$), and confirm the robustness of this correlation. Using a very large sample ($n = 5268$), Hannikainen et al. (2019) have provided further evidence that extraversion correlates with compatibilism. Finally, we have just shown that Grau and Pury (2014) replicates just fine if the replication is sufficiently powered. We are thus led to a very different conclusion than Cova et al.: Far from showing that demographic effects are "fragile" (Cova et al., in press), the total evidence in fact shows the robustness of experimental philosophers' results reporting demographic variation.

This conclusion is of extreme importance for experimental philosophy: The negative program of experimental philosophy is built on the existence of demographic and context effects (Weinberg, 2007; Machery 2017). Recently, these effects have come under serious attack (Nagel 2012; Knobe 2019), but a closer look at the actual empirical evidence shows that this attack utterly fails (for further discussion, see Stich 2013; Stich et al. n.d.; Knobe n.d.): Demographic results are often robust, and their alleged fragility is often merely due to hasty conclusions.

5 Conclusion

Grau and Pury (2014) reported an extremely surprising effect: People's understanding of love is associated with how they assign reference to proper names. This suggests that people vary in how they value "bare individuals," i.e., individuals independently of what they are like, and this varying valuation manifests itself in domains as disparate as proper names interpretation and love. The reality of this surprising effect was challenged by the failed replication reported in Cova et al. (in press). We have however shown that this replication was underpowered for many plausible effect sizes, and that a metanalytic estimate confirmed the reality of the effect reported in Grau and Pury (2014). Going beyond this reanalysis, we conducted our own, highly powered replication, and showed the reality of the effect. This successful replication confirms Grau and Pury's (2014) claim. It also highlights the importance of a careful interpretation of failed replications as well as the robustness of demographic effects in experimental philosophy.

Appendix

In this appendix we report a series of analysis of the data collected in our paper that extends the analyses done in Grau and Pury (2014).

Type of Replicant: Paralleling Grau and Pury (2014), the effect of type of replicant was tested with a 2 (Love for Replicant: Loved One and Pet versus Acquaintance and Shoes [Unspecified]) × 2 (Humanness of Replicant: Loved One and Acquaintance versus Pet and Shoes) × 2 (Measure of Replaceability: Appropriateness versus Likelihood) repeated measures ANOVA (Table 3). Note that, as in Grau and Pury, Love had the strongest effect (equivalent to d = 2.02, with loved targets (marginal mean = 2.93, *sd* = .06) rated as less replaceable than nonloved targets (marginal mean = 4.48, *sd* = .05). This was followed by a strong effect for Humanness (equivalent to d = 1.15) – again, with human targets (marginal mean = 3.34, *sd* = .06), rated as less replaceable than nonhuman targets (marginal mean = 4.06, *sd* = .05), modified by other smaller effects and interactions. The same pattern and approximate effect sizes remained when Linguistic Reference was added as a between participants variable. Thus, as in Grau and Pury, loved and humanness strongly predicted replaceability.

Sentimentality Unlike in Grau and Pury (2014), all participants were asked about the replaceability of a pair of shoes that were a gift from a friend (High Sentiment) and a favorite pair (Low Sentiment). Thus, we were able to test the results with a 2 (Sentiment: High versus Low) × 2 (Measure of Replaceability: Appropriateness versus Likelihood) repeated measures ANOVA. We found significant main effects of both Sentiment ($F(1,714) = 12.00$, $p = .001$, $\eta 2 = .017$, equivalent d = .26) and Measure ($F(1,714) = 45.32$, $p < .001$, $\eta 2 = .060$, equivalent to d = .51) that were modified by a significant interaction of Sentiment and Measure ($F(1,714) = 13.37$, $p = .02$, $\eta 2 = .018$, equivalent d = .27). Participants gave their highest ratings for the appropriateness of feeling the same way about their favorite shoes ($m = 4.8$, $sd = 1.7$), next for the appropriateness of feeling the same way about gift shoes ($m = 4.6$, $sd = 1.8$), then the likelihood of feeling the same way about their favorite shoes ($m = 4.5$, $sd = 1.8$), with the lowest ratings made for the likelihood of feeling the same way about gift shoes ($m = 4.4$, $sd = 1.8$). Means for the same type of shoes differed significantly based on Measurement (minimum $t(714) = 4.04$, $p < .001$). While participants rated it as more appropriate to feel the same way about replaced favorite shoes compared to gift shoes

Table 3 Within Participants ANOVA results for two Type of Replicant Effects (Loved and Humanness), Measure, and their interactions

Effect	*F*(1, 714)	*p*	*η2*
Love	726.37	0.00	0.504
Humanness	237.42	0.00	0.250
Measure	22.26	0.00	0.030
Love x Humanness	29.89	0.00	0.040
Love x Measure	2.08	0.15	0.003
Humanness x Measure	27.46	0.00	0.037
Love x Humanness x Measure	28.20	0.00	0.038

($t(714) = 4.84$, $p < .001$), there was no significant difference between the likelihood of feeling the same way about the different pairs ($t(714) = 1.49$, $p = .138$). These findings paralleled the between-participant findings of Grau and Pury, finding an effect of sentiment for objects on appropriateness but not likelihood ratings.

Sentimentality for pets was tested by correlating the belief that a pet is like a member of the family with likelihood of feeling the same way about a replaced pet ($r = -.14$, $p < .001$, $r^2 = .018$, equivalent d = 0.27) and appropriateness of feeling the same way about that replacement ($r = -.12$, $p = .002$, $r^2 = .014$, equivalent d = .24). These effects were smaller than in Grau and Pury (2014)'s correlations of −.22 and − .21, respectively, but still statistically significant and in the same direction.

References

Beebe, J.R., and R.J. Undercoffer. 2015. Moral valence and semantic intuitions. *Erkenntnis* 80: 445–466.

Beebe, J.R., and R.J. Undercoffer. 2016. Individual and cross-cultural differences in semantic intuitions: New experimental findings. *Journal of Cognition and Culture* 6: 322–357.

Camerer, C.F., et al. 2018. Evaluating the replicability of social science experiments in nature and science between 2010 and 2015. *Nature Human Behaviour* 2: 637–644.

Cohen, J. 1992. A power primer. *Psychological Bulletin* 112: 155–159.

Cova, F., et al. in press. Estimating the reproducibility of experimental philosophy. *Review of Philosophy and Psychology*.

Etz, A., and J. Vandekerckhove. 2016. A Bayesian perspective on the reproducibility project: Psychology. *PLoS One* 11: e0149794.

Faul, F., E. Erdfelder, A.G. Lang, and A. Buchner. 2007. G* power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods* 39: 175–191.

Feltz, A., and E. Cokely. 2009. Do judgments about free will and responsibility depend on who you are? Personality differences in intuitions about compatibilism and incompatibilism. *Consciousness and Cognition* 18: 342–350.

Feltz, A., and E. Cokely. 2019. Extraversion and compatibilist intuitions: A ten-year retrospective and meta-analyses. *Philosophical Psychology* 32: 388–403.

Goh, J.X., J.A. Hall, and R. Rosenthal. 2016. Mini meta-analysis of your own studies: Some arguments on why and a primer on how. *Social and Personality Psychology Compass* 10: 535–549.

Grau, C., and C.L. Pury. 2014. Attitudes towards reference and replaceability. *Review of Philosophy and Psychology* 5: 155–168.

Hannikainen, I., et al. 2019. For whom does determinism undermine moral responsibility? Surveying the conditions for free will across cultures. *Frontiers in Psychology* 10: 2428.

Knobe, J. 2019. Philosophical intuitions are surprisingly robust across demographic differences. *Epistemology & Philosophy of Science* 56: 29–36.

Knobe, J. n.d.. Difference and robustness in the patterns of philosophical intuition across demographic groups.

Kraut, R. 1986. Love de re. In *Midwest studies in philosophy, vol. X*, ed. P.A. French, T.E. Uehling, and H.K. Wettstein, 413–430. Minneapolis: University of Minnesota Press.

Kripke, S. 1980. *Naming and necessity*. Cambridge: Harvard University Press.

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

Machery, E. (Forthcoming). What is a replication? *Philosophy of Science*.

Machery, E., R. Mallon, S. Nichols, and S.P. Stich. 2004. Semantics, cross-cultural style. *Cognition* 92: B1–B12.

Machery, E., C. Olivola, and M. De Blanc. 2009. Linguistic and metalinguistic intuitions in the philosophy of language. *Analysis* 69: 689–694.

Machery, E., M. Deutsch, J. Sytsma, R. Mallon, S. Nichols, and S.P. Stich. 2010. Semantic intuitions: Reply to lam. *Cognition* 117: 361–366.

Machery, E., J. Sytsma, and M. Deutsch. 2015. Speaker's reference and cross-cultural semantics. In *On reference*, ed. A. Bianchi, 62–76. Oxford: Oxford University Press.

Makel, M.C., J.A. Plucker, and B. Hegarty. 2012. Replications in psychology research: How often do they really occur? *Perspectives on Psychological Science* 7: 537–542.

Nadelhoffer, T., T. Kvaran, and E. Nahmias. 2009. Temperament and intuition: A commentary on Feltz and Cokely. *Consciousness and Cognition* 18: 351–355.

Nagel, J. 2012. Intuitions and experiments: A defense of the case method in epistemology. *Philosophy and Phenomenological Research* 85: 495–527.

Nelson, L.D., J. Simmons, and U. Simonsohn. 2018. Psychology's renaissance. *Annual Review of Psychology* 69: 511–534.

Nozick, R. 1974. *Anarchy, state, and utopia*. New York: Basic Books, Inc..

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, 349 (6251): aac4716.

Schmidt, F.L. 1996. Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. *Psychological Methods* 1: 115–129.

Schmidt, F. 2010. Detecting and correcting the lies that data tell. *Perspectives on Psychological Science* 5: 233–242.

Simmons, J.P., L.D. Nelson, and U. Simonsohn. 2011. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science* 22: 1359–1366.

Stich, S.P. 2013. Do different groups have different epistemic intuitions? A reply to Jennifer Nagel. *Philosophy and Phenomenological Research* 87: 151–178.

Stich, S. P., Rose, D., and Machery, E. (n.d.). Demographic differences in philosophical intuition – a reply to Knobe.

Sytsma, J., & Machery, E. (2010). Two conceptions of subjective experience. Philosophical studies, 151(2), 299-327.

Sytsma, J., J. Livengood, R. Sato, and M. Oguchi. 2015. Reference in the land of the rising sun: A cross-cultural study on the reference of proper names. *Review of Philosophy and Psychology* 6: 213–230.

Trafimow, D., and B.D. Earp. 2016. Badly specified theories are not responsible for the replication crisis in social psychology: Comment on Klein. *Theory & Psychology* 26: 540–548.

van Dongen, N., Colombo, M., Romero, F., and Sprenger, J. (n.d.). Intuitions about the reference of proper names: A meta-Analysis.

Weinberg, J.M. 2007. How to challenge intuitions empirically without risking skepticism. *Midwest Studies in Philosophy* 31: 318–343.