What lessons does the "replication crisis" in psychology hold for experimental economics?

3 What lessons does the "replication crisis" in psychology hold for experimental economics?

Nicholas Bardsley

3.1. Introduction

In recent years, psychology has come under considerable critical scrutiny concerning the soundness of its experimental results. There has been concern over outright academic fraud, following Diederik Stapel's public exposure in 2011. Stapel was a leading social psychologist who had worked at the Universities of Amsterdam, Groningen and Tilburg, with a prolific research and publication record. He was found to have fabricated data, and a scandal ensued receiving significant national and international media attention. Consequently, he lost his post and became an academic and social outcast (Bhattacharjee, 2013). An investigating commission set up by three Dutch universities found that 55 papers should be retracted and evidence indicating fraud in 10 further articles, from the corpus of 137 Stapel publications (Levelt *et al.*, 2012). At the time of writing (November 2016) 58 of his publications are documented as retracted by the website "Retraction Watch".

In many cases of academic fraud, retracted papers continue to be cited many years after their retraction (Bornemann-Cimentiet *et al.*, 2016), exacerbating the pollution of a field. Of perhaps much greater import than spectacular but rare cases such as Stapel's, however, is the likelihood that careers built upon fraud are indicative of wider failures of a discipline to enforce appropriate critical and academic standards (Levelt *et al.*, 2012). Relatedly, there has also been concern over more subtle but plausibly more pervasive problems of selection bias, publication bias, selective presentation of data, small samples, data mining and so on. The problems are a combination of "questionable research practices" by researchers and "questionable publication practices" by journal editors and reviewers. Attempts to replicate results plausibly provide one indicator of the health of an experimental discipline, facilitating detection of spurious findings regardless of their origin. Though problems in academic processes extend to a variety of fields, including for example medical studies (Ioannidis, 2006), psychology has been subject to particular scrutiny, with several recent systematic replication efforts. The results are generally seen as problematic, with low rates of replication and effect sizes that fall far short of those of the original studies. This chapter provides an overview of these replication initiatives, and considers in comparison the situation obtaining in experimental economics.[1]

Replication efforts themselves have sometimes proved controversial, given alleged publication incentives to disconfirm the original results, and the high stakes involved given that authors' professional reputations may be tied to the outcomes. Two positions are analysed that have featured significantly in discussion. The first states that authors ought to be extensively consulted during the design phase of a replication study in order that there is sufficient similarity between procedures (hereafter the "authorial inclusion issue"). It is argued that the methods sections of experimental reports are typically too lacking in detail to enable adequate similarity to be achieved.

[1] It is not the author's intention to draw a sharp distinction between "behavioural" and "experimental" economics, but to the extent there is a set of research and publication practices distinct from those in experimental psychology, centred around incentivised experiments and a close relationship between laboratory and theory, the latter term seems more applicable.

The second holds that objections to a replication study raised after its results are known have diminished force, and might be taken to imply that peer review of replication studies should only take place *ex-ante*.

Finally, consideration is given to the links between economics experiments and theory, which may mark a key difference concerning the situation in the two fields. The issue is approached from the perspective described in Bardsley et al. (2010) where a critique was developed of model-implementing experiments. It is proposed that the kinds of theory-proximity prevalent in experimental economics may favour replications with limited confirmatory power.

The following section deals with some preliminaries, including concepts of replication and the role of repeatability in accounts of science. Section 4 gives an overview of replication efforts in the two fields. Sections 5 and 6 consider the authorial inclusion demand and CARKing issue respectively. Section 7 then considers a difference between the generality of experimental economics and psychology designs, namely the theory-centric character of the former, and its bearing on replication. Section 8 concludes.

3.2. Preliminary considerations

3.2.1 Direct and conceptual replication

An important distinction has been drawn between two types of replication: direct and conceptual. Other schema have been proposed but the remaining distinctions seem less fundamental. Schmidt (2009) provides a useful discussion, and his definitions are followed here. A direct replication can be defined as "repetition of an experimental procedure". A conceptual replication consists of "repetition of a test of a hypothesis or result of earlier research work with different methods" (Schmidt, 2009 p91). Schmidt goes on to note that the desirable characteristics of a replication depend on which goal is being pursued, and that there is a variety of possible goals. These include checking that some result obtains, checking its robustness, checking for its boundary conditions or generality and seeing whether it, or its theoretical interpretation, can be confirmed using different procedures entirely.

For direct replication, the general aim is often to check that a certain result in fact obtains, and is not the result of one or another kinds of failure of experimental validity. On some views, for example Brandt et al. (2014), a direct replication should seek to mimic the original as closely as possible concerning every detail of the original setting. At the ideal limit, only the sample drawn, and academic personnel involved would vary, along with some difference in timing. A failure to replicate will presumably either signify Type 1 error, academic malpractice or perhaps rapid social change.

In practice, working understandings of direct replication seem far looser than this, and it is perhaps unclear why one would wish to check only for this very narrow set of problems. It seems perhaps impossible to repeat exactly all bar three aspects of a behavioural or cognitive experiment, and this is generally seen as no bad thing since it allows for the robustness of results to be explored. However, the more aspects are varied, the more this opens the door to a predictable reaction to a failed replication attempt, that it introduced problematic deviations from the original design, which explains the absence of the original findings. We consider such objections further in section 4, where we consider which kinds of deviation from the original design are sensible in direct replication. For Schmidt, checking for generalisability by running a design on another category of participant, or another sample of stimuli, is also a case of direct replication.

In contrast, conceptual replication seeks to confirm a particular hypothesis by implementing different procedures and conditions. The aim is to reproduce the same phenomena abstracted from their original setting. This requires a different experimental procedure. Schmidt (2009) gives the example of Rosenthal and Fode's (1963) study of experimenter effects, in which students were given different information about the abilities of rats they were given to train in maze tasks. In reality the information was randomly allocated, but those rats labelled as higher ability completed the mazes faster than those labelled as lower ability. The phenomenon of an experimenter effect was reproduced using teachers and school pupils, for example in Rosenthal and Jacobson (1963), a completely different experimental setting, but with a similar manipulation of expectations about the trainee's capacities. Here teachers were given false information about scores on an IQ test, with the false information being randomly allocated. Subsequent use of the same IQ test demonstrated larger gains for the pupils who had been arbitrarily deemed more successful in the initial test. The very different setting and sample confirms the underlying hypothesis that an experimenter's expectancy can influence the behaviour of experimental subjects.

A basic example of conceptual replication in experimental economics is, arguably, Battalio *et al.'s* (1985) demonstration of risk aversion in rats, which necessarily uses a completely different experimental set up to the corresponding (monetary gamble-based) tasks which demonstrate this for people. The authors set up prospects for the rats using two levers, which released food in a probabilistic manner. Lever pairs were implemented offering either a high chance of a smaller quantity of food, or a smaller chance of a larger amount of food, with the same expected (mean) quantity. The rats generally preferred the former. In this way the authors were also able to test for, and find, the common ratio effect in rats. That is, preferences switch from a safer to a riskier gamble when the probability components of each are scaled down by a common factor.

Recent replication initiatives both in psychology and experimental economics have concentrated on direct replication, whereas previous replication activity in social sciences has arguably tended towards conceptual replication. The confirmatory power of conceptual replication is high, but comes at a price. For if a conceptual replication fails, it may be relatively unclear what we learn from this failure. Conversely, if a direct replication succeeds, it has limited confirmatory power, the more limited the closer the design is to the original. But failures of direct replication are correspondingly informative.

3.2.2 The role of repetition in science: essential in theory but not in practice?

Arguably, the replication crisis sits uneasily with classical accounts of scientific activity offered in the history and philosophy of science. On empiricist views, following for example Hume (1739-40), science is a process of collating repeated observations, and distilling from these invariant regularities, or laws, where "constant conjunctions" of events are evident. On such views there is nothing particularly special about a replication study, it merely adds to the stock of available observations. However, the emphasis here on repetition seems to imply that there should be many instances of corroborating observation before something can be accepted as a scientific finding.

In contrast, realists (for example Bhaskar, 1978) see experimentation as revealing causal powers or tendencies, which can be manifested by just a single act of experimental closure. From this point of view the role of repetition is to confirm a previous finding in a controlled setting, that is, providing confirmation that the setting was adequately isolated and controlled (Greenwood, 1982,

p227-31). Causal powers or tendencies are normally hidden from observation because so much is going on simultaneously in any naturally-occurring setting. On this view the observations that constitute scientific results are necessarily outcomes of rather unusual situations. Very few confirming instances may be required for a finding to count as a result, but they are still required. Consequently, initial replications or failures to replicate may be highly significant, but additional direct replications have little scientific value once a result has been established.

The empiricist account seems difficult to reconcile descriptively with the situation obtaining in scientific publishing in many fields. The situation seems rather to be that many findings are accepted as results with relatively few, or even no, direct replications. Indeed, the "news value" of disappointing replication studies results has stemmed from the fact that a body of findings with non-provisional status in the relevant field has low reproducibility the first time direct replications are tried. This non-provisional status of non-replicated results does not seem to be confined to psychology. In the "Handbook of Experimental Economics Results" for example, explicit inclusion criteria were that the authors found the results in question important, not that they had been repeatedly observed (Plott and Smith, 2008, preface). The extent of corroboration is frequently addressed in the volume nonetheless. On the other hand, even if the realist account holds, any non-replicated findings ought really to count within a given field as mostly provisional, not *bona fide*, results.

Both camps, arguably, tend to offer an idealised view of science, abstracting from human and institutional failings. Such defects are factors which the replication movement must in contrast consider as central. For if the correct explanation of low replicability is shortcomings in research and publication practices, the same flawed processes would normally govern replication attempts. And secondly, the broader aim is not only to measure replicability but to improve it. As will be outlined, replication initiatives have also sought to innovate to improve the integrity of the review and publication process. We note however that further controversy, beyond the scope of this chapter, has centred on whether critique based on unsuccessful replication should be conducted outside of peer review processes, via social media or blogs. See for example Fiske (2016) and, ironically, the online commentary to the digital edition of that article. The issue hangs partly on the extent and speed with which practices can be improved. On the one hand there is no quality control over potential insinuations or allegations of malpractice, which can cause unjustified reputational damage, in self-published media. On the other hand, if journal publication processes lack sufficient integrity it seems unlikely that sufficient critique will take place in peer-reviewed formats alone (Coyne, 2016).

4. Overview of recent replication studies

Key attributes of recent substantial replication efforts are shown in Table 1 debajo de. Works are included which bring together many attempts at direct replication of a range of results, in one outlet, following a common approach and sampling regime. Stand-alone studies and attempts to replicate a single finding or a single author's work are excluded. For psychology the studies comprise a journal special issue with pre-registered designs, a panel data study deploying numerous labs to study 13 effects across each of 36 samples, and an attempt to replicate 100 findings across independent samples.

The "replication rate" is the reported proportion of effects that replicated with a significant effect in the same direction, at the 5% level, as the original. Whilst this is not the only or necessarily best indicator of replicability, it is perhaps the main "headline" figure. Klein et al. (2014) is published

in Social Psychology 45(3) but merits inclusion in its own right because it seeks to replicate a set of findings repeatedly using numerous samples, a "many labs" approach. The "effect size quotient" is the mean effect size of the replication studies divided by the mean effect size of the original studies, and is another widely-cited indicator of replicability. A replication protocol counts here as pre-registered if prospective replication attempts were subject to peer review and an editorial decision taken either by journal editors or project organisers before data were collected.

Study	Field	Replication Rate	Effect Size Quotient	N (samples)	Pre-registered	Target selection
Open Science Collaboration (OSC), 2015	psychology	36%	0.49	100	No	Quasi-random
Klein *et al*., 2014	psychology	77%	1.26	36	Yes	Non-random
Special Issue contributions to *Social psychology* vol. 45(3), 2014	psychology	8%	--	13	Yes	Non-random
Camerer *et al.*, 2016	economics	61%	0.66	18	No	Non-random

Table 1: recent initiatives on direct replication in psychology and experimental economics

Note

1. The replication rate for *Social psychology* 45(3) is based on the author's collation of the results reported in each study. One study is excluded from this count because of equivocal results.

A notable feature of these studies is the wide variation in the replication rate, which ranges from 1 in 13, to 10 in 13 successes. This is a huge range which seemingly gives the lie to the simple "headline" view that psychology experiments are unlikely to replicate. However, we should also ask what the replication rate *ought* to be in the absence of questionable research practices. For there will inevitably be false positives in any corpus of experimental studies, arising from sampling variation alone. It can be shown that the probability of a replicated rejection of a null hypothesis equals

$$\frac{\alpha^2 p + \beta^2 (1-p)}{\alpha p + \beta (1-p)}$$

where β is the power of a trial, α is the threshold significance level applied and p is the probability that any given hypothesis selected for investigation is true (Appendix).[2] Thus, if α = 0.05, p=0.5 (an uninformative prior) and all trials had 90% power, a commonly-used benchmark for experimental quality, one would expect a replication rate of approximately 86%. So estimated

[2] I owe this point and the derivation in the Appendix to Kelvin Balcombe (personal communication). The calculation assumes that power is equal in original and replication designs. Rates of replicated rejection of null hypotheses should be higher if original designs are of lower power.

replication rates in Table 1 are uniformly lower than they seemingly ought to be, but variable. It is worth noting that this is sensitive to the value of *p*, however. For expected replication of significant results to be less than 80% requires p>0.7, a field in which researchers study unlikely effects.

Large variation is also apparent for effect sizes. OSC (2015) reports lower effect sizes in the replication studies, whereas Klein et al. (2014)'s figures imply that on average (mean) they are larger, albeit using a different measure of effect size. For this study effect sizes are reported in Cohen's d units, whilst the other studies report Pearson's r. Also, the mean effect size is influenced by extreme values; the corresponding ratio of median effect sizes is ~1 (Klein *et al.* 2014, Table 2, cols. 1 and 4.) For the other papers in *Social psychology* vol. 45(3), calculation of an overall effect size quotient would require work with the data from each paper. Ironically, though perhaps inevitably, these large replication efforts used quite different procedures, which may help explain the variation.

None of the studies in Table 1 used a pure random sampling method. In OSC (2015), in which the effect selection protocol is described as "quasi-random" sampling, authors could choose which of an available set of studies to propose to replicate. The available studies initially consisted of the first 20 articles in 3 leading psychology journals for 2008. The last reported finding in each study was the target to be replicated. In this manner the set of studies available at any particular time was restricted to reduce selection bias. As the set depleted, the project team helped allocate remaining studies to replicators. But it remains the case that initially authors could choose which study to attempt to replicate. Assuming there is greater news value in producing an unsuccessful replication, this could result in a lower replication rate than would occur under random sampling and assignment. In that case one might choose a study which one expects not to replicate and these expectations might, even subliminally, influence the conduct of the research. Teams might also apply motivated by scepticism of a particular study, or even, conceivably, personal and professional grudges, selection effects which would again tend to lower the replication rate.

The problem just outlined seems exacerbated in the case of *Social Psychology* issue 45(3). Here, authors self-proposed a study to replicate, and then submitted a research design for peer-review. If the research design was deemed of high enough quality, the replication attempt proceeded. There is more scope for selection effects to operate than in OSC (2015) because the set of targets for replication was also determined by prospective replicators.

In Klein et al.'s (2014) "many labs" study, in contrast, with the much higher replication rate, the sampling procedure to select target effects is not fully documented. But four selection criteria are specified: suitability for online presentation, length of a trial, simplicity of design and diversity of effects (*ibid.*, p143). *A priori*, it seems reasonable to expect short and simple trials to replicate better than long, complicated ones, if only for the mundane reason that there are fewer details that have to be correctly implemented.

The first lesson to draw for experimental economics, then, is that it is not clear from the above whether the results of Camerer *et al.* (2016) suggests a better or worse situation than that obtaining in psychology, because of the mixed approaches and results for the latter field. Even if, that is, one were to accept the 61% replication rate reported as a good estimate. If we ignore the non-random selection of studies, a conventional 95% confidence interval for this rate ranges from 39% to 84%, whilst that for OSC (2015) ranges from 27% to 48%. For the latter, if we restrict attention to the replications which had sample sizes sufficient for at least 90% power, we obtain an improved 46% replication rate, and if we further restrict to the top journal it increases to 52% (author's calculations, using the publicly available dataset). Here again, however, the findings to be replicated were not randomly selected and assigned. Only between-subject designs were used, and

the author's declared most important finding selected. The allocation of target findings to teams is not reported. Thus, while this study is a landmark contribution to the field, systematic replication efforts in experimental economics are still at a very initial stage. It seems desirable for a random selection and assignment protocol to be used, which would give a better idea of the general direct replicability of results. However, as Gilbert *et al.* (2016) point out, in commentary on OSC (2015), this would presuppose a defensible definition of the population of studies to be sampled from.

Regarding stand-alone direct replication studies, few of these seem to have been conducted in experimental economics. A past editor of the journal *Experimental Economics* has informed the author that during his tenure not a single direct replication article had been submitted, despite it being the policy of the journal to welcome such submissions.[3] The *Journal of the Economic Science Association* has also solicited replications but with disappointing response. This does not imply that the field is bereft of replication, since it is relatively common for researchers to include a baseline treatment which directly replicates a progenitor study. Such studies are often described as "follow-up" studies.

However, it is possible that in these cases the inclusion of the baseline replication may also be with the motivation to provide credibility for the subsequent, original, manipulation. In that case, the assumption seems to be that the original findings *ought* to be reproduced, which is not the case in a pure replication study.[4] In this mode of operation, a field may conceivably be prone to "bad equilibria", in which an effect that is not actually robust may acquire the status of an expected outcome.

A second lesson for the field, then, is that more is required than journal policy statements to elicit replication attempts from authors. Pre-registered designs, in which peer review and a provisional publication decision take place before a design is implemented, may increase authors' confidence of being able to publish replications of findings which have standing in the field. The fears which are countered are rejection because of lack of originality in case of a positive result, and because of perceived reputational damage to other researchers, in the case of a null or negative result. The other researchers in question may also be reviewers or associates of reviewers, a prospect potential replicators will doubtless be wary of. Combining a special journal issue with pre-registered designs as with *Social Psychology* 45(3) seems a good strategy, and it if could be combined with randomised effect selection and assignment, would be considerably strengthened with regard to obtaining a general picture of how likely results are to hold.

5. Which factors can be varied in direct replication?

5.1 The authorial inclusion issue

In response to the enfolding controversy, Nobel laureate Daniel Kahneman issued a call for a "new etiquette" for replication studies:

> "Authors should be guaranteed a significant role in replications of their work. …
>
> In the myth of perfect science, the method section of a research report always includes enough detail to permit a direct replication. Unfortunately this seemingly

[3] Personal communication with Professor Tim Cason.

[4] According to Collins (1991), replication disputes are always concerned to establish what the results of an experiment *ought* to be. This seems consistent with the correct factual outcome being a logically prior matter for resolution, however, which Collins seems at pains to deny.

> reasonable demand is rarely satisfied in psychology, because behaviour is easily affected by seemingly irrelevant factors. For example, experimental instructions are commonly paraphrased in the methods section, although their wording and even the font in which they are printed are known to be significant.
>
> … the original author should have detailed advanced knowledge of what the replicator plans to do. [Since] authors will generally be more sensitive than replicators to the possible effects of small discrepancies of procedure."
>
> Kahneman (2014)

Kahneman goes on to argue that the burden of justification should lie on replicators in the event that they deviate from any experimental detail advocated by the originator in pre-trial correspondence.

This intervention seems to have been especially contentious. Perhaps one reason for this is because it puts power squarely back in the hands of original authors, since it increases the burden of planning and preparation for would-be replicators and increases the scope for adversarial confrontation, between researchers with probably unequal status, prior to a replication attempt. Since that objection concerns power relations, it is not epistemic however. That is, it does not follow from this consideration that authorial involvement is not necessary or best scientifically for replication studies. Since the power relations aspect is plausibly negative, a factor which is likely to suppress valid critique, we need to consider whether there is a credible epistemic counterpoint.

Kahneman's plea actually echoes an earlier argument by Schmidt (2009). Schmidt gives an example from the natural sciences of a paper on laser construction. Apparently, several authors attempted to reproduce the laser in question following the methods and procedures set out in the paper, but were unable to do so without help from the original author. Schmidt concludes that there is often a substantial element of tacit knowledge involved in producing a result, which it is not possible to set out completely in a scientific paper, and that it is reasonable to conclude that the same is likely in the social sciences. On this basis he advocates authorial involvement by the original author in a direct replication study, but specifically for cases where the set-up is complex.

Both authors argue from a premise that methods sections of journal articles lack sufficient information to conduct direct replications, and conclude that *ex-ante* authorial involvement is necessary (at least in some cases) to supply the missing information. However, Schmidt goes on to point out a consequent danger that experimental artefacts in the original study (we should note, other than Type 1 errors) will be reproduced in the replications.

That danger seems to this author especially likely given the kinds of thing that Kahneman actually mentions as missing. It seems there are epistemic asymmetries between the laser case and social science experiments which lie behind Schmidt's concern, that require analysis. Firstly, in the former case, the result is not necessarily in doubt, since once a laser has been produced it can be independently inspected and tested. In the latter case, in contrast, all we have is a reported finding. Secondly, in the case of the laser, one can well imagine that there are details with a physical basis that have to be right, even if these are not yet fully understood by the inventor. For example, a laser inventor might conceivably recommend a particular type of glass, even if they did not understand fully why this seemed superior. It would seem to be a different matter if the original author were to communicate that, for example, the laser should only be constructed on a Tuesday evening, after everyone involved has danced the Tango. If *that* were the missing information, perhaps one could

reasonably conclude that the originator has happened on the result by chance, and does not really understand what is going on in the supposedly controlled setting.

A strong authorial inclusion norm therefore seems potentially counter-productive from the point of view of perpetuating possible experimental artefacts. There may be some special cases that are analogous to Schmidt's laser example where complex technical details have not been described in sufficient detail to enable replication, where authorial consultation is needed. In general, however, as commentators have argued, if methods and procedures sections are typically inadequate, a logical response is for authors to write them better, and for editors to enforce this, on pain of a replicator being unable to reproduce the results.[5] This means that all relevant details should be included if one wishes others to be able to corroborate the result. The mere mention by an author of a particular experimental detail does not make it relevant. But what exactly counts as relevant?

5.2 Relevant and Irrelevant Details

The notion that there are seemingly irrelevant details that have to be right to reproduce certain psychology results is iterated in Stapel's (2014) confessional autobiography. Stapel writes, of an earlier and supposedly innocent period in which he had been reproducing other authors' results,

> "My colleagues from around the world sent me piles of instructions, questionnaires, papers, and software.
> Now I saw what was going on. In most of the packages there was a letter, or sometimes a yellow Post-It note stuck to the bundle of documents, with extra instructions:
> "Don't do this test on a computer. We tried that and it doesn't work. It only works if you use pencil-and-paper forms."
> "This experiment only works if you use 'friendly' or 'nice'. It doesn't work with 'cool' or 'pleasant' or 'fine'. I don't know why."
> "After they've read the newspaper article, give the participants something else to do for three minutes. No more, no less. Three minutes, otherwise it doesn't work."
> "This questionnaire only works if you administer it to groups of three to five people. No more than that."
>
> Stapel (2014, p69)[6]

It seems that something sets apart these kinds of detail from *bona fide* missing technical details, and puts them instead on a par with the Tuesday Tango. A potential candidate for a demarcation criterion here is that the suspect requirements are irrelevant given the state of knowledge. Both Tuesdays and the Tango are social constructions with no basis in Physics, and so are strictly theoretically irrelevant to producing a physical effect or working apparatus. Font choice, whilst it may have certain psychological effects, seems similarly orthogonal to (for example) the availability heuristic, cognitive dissonance, or a hypothesised relationship between moral judgements and cleanliness. Also, any explanation of how font selection might mediate a purported treatment effect

[5] See for example Wilson (2014) and the ensuing online exchange.

[6] It is possible that Stapel exaggerates the role of such mystery factors in order to make his fraudulent activity seem less extreme. Stapel also asks what the alleged sensitivity to mystery factors implies about the scientific value of the purported results, but does not provide an answer (Stapel, 2014 p72).

would have to deal with the fact that the font selected for the experimental instructions will be the same in each experimental treatment.

For these reasons it seems almost certain that if any of the effects just mentioned required use of a certain font for written instructions, publication of this fact would appear to undermine the result as originally described. For it would appear to be impossible to explain the influence of font in a manner that preserves unaltered either generality for the purported result, or any theoretical explanations of it considered in the relevant papers. Historians of science have argued that experimental results are unlikely to achieve significant status within a field without a theory which explains how the results came about. In the absence of this they are likely to remain just numbers. According to realists, the theory in successful cases sets out a generating mechanism (Pawson and Tilley, 1997). Mystery factors like the Tuesday Tango could play no useful role in such a theory, serving instead as a "ghost in the machine".

This demarcation criterion, which we may term epistemic considerations, also seems relevant to deciding which of the *specified* details of a design are essential to repeat in order to effect a direct replication. It is uncontroversial that repetition of all details except the sample drawn is practically impossible. This seems impossible even with extensive authorial involvement, since without a photographic memory the original author will not remember such details as the exact colour of the carpet in a laboratory, the course of the weather through each experimental session, or how many times the experimenter and his assistants happened to blink. Nonetheless it presumably makes sense to speak of conducting "the same" experiment, as this and cognate expressions are used routinely in replication literature. This paradox, that the same experiment is implemented with many details changed, seems to be an instance of the old philosophical puzzle of Theseus's ship. The ship in question remains the same vessel despite each of its timbers being replaced over time. How is this possible?

An influential answer to this question comes from philosopher Robert Nozick, in his *Philosophical Explanations* (Nozick, 1981). Nozick explains that if someone regards Theseus's ship as unchanged we are weighting factors highly other than its material substance in our conception of its identity. Such factors might include who owns it and its general physical form, for example. The identity of something is thus akin to a weighted sum of distance scores over several types of characteristic.[7] Which weights apply may be different in different circumstances. To an archaeologist interested in dating the construction of the ship, the fact that all the timbers have changed is crucially important. To the Athenians who allegedly maintained the ship in a seaworthy condition for centuries as an act of thanks to the Gods, this factor was assigned far less weight than its physical continuity and form.

In the context of direct replication, it seems that criteria of identity for an experiment derive in part from theoretical considerations. Thus, for example, if a particular explanation is proposed for an experimental result, the procedural details employed in the original study should be consistent with this. A replication study should give equal chance for the explanations proposed to operate as did the original. This may also enable some "gap filling" to take place in relation to missing procedural details. To take an economic example, if a purported explanation for a finding comes from the theory of finitely repeated games, one may infer that subjects were informed about the existence and timing of a final round. If this matter is not explicitly clarified in procedures, as it ought ideally to be, and full (verbal plus written) instructions are not available, this is a reasonable

[7] See also Glass (2000) who discusses the identity issue and Nozik's solution in the context of meta-analysis.

inference.[8] But it will not be of interest which font was used in the instructions unless, as seems improbable, this features in a theory explaining the finding, and it will not be a serious objection that a different font was used.

Under theoretical considerations far more can be included than the specific social science theories referred to as being tested or providing explanations of results, however. There is also the general theory of experimentation and experimental traditions informed by a theoretical background in particular social sciences. A more general "theory" of experimentation includes such aspects as which factor(s) were changed between treatments, which factors operative in a naturally occurring setting were excluded from the experimental setting, the content of information, graphical or display factors, sampling procedure and subject pool, allocation to treatment, time and place of experiment, communication protocol, anonymity protocol, and "blinding" procedures. In economics experiments considerations which would normally be relevant also include the nature and level of incentives, one shot or repeated setting, and so on.

If we vary factors that are theoretically *ir*-relevant and obtain the same results, this form of replication has greater confirmatory power than an exact replication would. In common parlance amongst behavioural economists, we see that the finding is more "robust". The pronouncement by Nosek and Lakens (2014, p7) seem apposite here:

"A direct replication is the attempt to duplicate the conditions and procedure that existing theory and evidence anticipate as necessary for obtaining the effect ... Successful replication bolsters evidence that all of the sample, setting, and procedural differences presumed to be irrelevant are, in fact, irrelevant."

This coincides with the demarcation criterion of epistemic considerations, with the proviso that the factors mentioned as evidenced are not ones that are ruled out in (well-established) theory. Another anecdote from Stapel (2014) may be illuminating here. Stapel claims inside knowledge that a particular, prized, effect in Social Psychology does not replicate under the conditions as described by researchers in their reports. "Terror Management Theory" (TMT) posits that individuals normally go about in a state of suppression of knowledge of their own death (Solomon *et al*., 1991). Consequently effects are observable when experimental subjects are asked to reflect on their own death, in particular, bold, assertive statements are likely which serve to restore the participant's sense of wellbeing and security. Stapel claims that an unnamed research team tried repeatedly and unsuccessfully to replicate key findings in this literature, and that the difference between themselves and a laboratory which did produce the phenomena lay in the experimenters' "alternative" subculture orientation, in particular their clothing. Specifically, in the latter laboratory, the experimenters tended to wear black clothes, heavy metal t-shirts and so on, rather than well-ironed shirts, ties, suits and so on.

The story stretches credulity, particularly because the evidence for TMT derives from a large set of research teams, not one laboratory (Burke *et al.*, 2010). But supposing that it were true, for the sake of argument, this detail would not count as epistemically irrelevant, because it suggests a closely-related priming effect. Interacting with the nihilistically-dressed researchers and assistants, that is, may provide subtle cues of mortality which enable the experimental manipulation to be more effective. However, at the current state of knowledge, no such priming effect has been posited in TMT studies, so authors would be fully justified in not implementing this detail in a replication

[8] It is not an infallible inference, since ideas from the theory of finitely repeated games may have traction in settings which do not perfectly instantiate the way such games are modelled.

attempt. They would also be justified, it seems, in reporting a failure to replicate even if they knew about the nihilistic clothing anecdote, and even if they had strong private grounds for believing it to be true, on the preceding discussion. Thus, prior to anything being published on it, the clothing detail would not count among the conditions that the existing evidence identifies as necessary. On the other hand, continuing to produce the result with the aid of nihilistic clothing without reporting the matter would be detrimental to the field, because this omission would make the effect seem more robust than it really is. In psychologists' parlance we would be suppressing knowledge of the boundary conditions for the finding in question.

Depending on our goals, however, and following Schmidt (2009), what counts as an irrelevant factor may change. Nosek and Lakens' formulation (above) of what can be changed leaves on the face of it considerable latitude for factors which might or might not be varied in a direct replication. The original instructions will include many factors ("Good morning. Welcome to the experiment…") that existing theory and evidence suggest can be changed. It is normal practice in direct replication, notwithstanding this, including for the studies in Table 1, to use the original instructions and materials if these are available. This has the virtue of minimising omission of anything that is possibly epistemically relevant. It probably also reflects a relatively narrow goal, however, to confirm that something is there worthy of the attention of the research community, with relatively less emphasis on robustness checking.

A different goal obtains if we wish to see whether a result generalises beyond its initial context, and this may mean different criteria of identity for "the same" experiment. It may well be defensible to argue that existing theory and evidence point to a particular context as being necessary to produce a result, but we wish to go beyond this by applying the same procedure elsewhere or differently. In doing so, however, it seems we are not necessarily in the realm of conceptual replication, where the more variation is applied to test the same hypothesis or reproduce the same phenomenon the better, and we will still be essentially repeating a procedure in a new setting. To take an economic example, we might ask if the result of dictator game giving is robust. To have a variety of designs all of which found positive transfers in the dictator game, all of which used different instructions, gives us greater confidence in this result than if everyone had used the same instructions verbatim. Or one might have found that, for example, members of a particular religious group give more than subjects in general in the dictator game. Existing evidence might, at that point, be taken to suggest that use of this particular group was necessary to produce the effect, but one would still be amply justified in using a different denomination to see if the result generalised, and in reporting this as "the same" experiment for this purpose.

6. HARKing and CARKing

"HARKing" stands for "Hypothesizing After the Results are Known" and is a conventionally (albeit not universally) regarded as poor academic practice. The arguments for and against HARKing are thoughtfully discussed by Kerr (1998) who argues that the costs of HARKing clearly outweigh the benefits. "CARKing" stands for "Critiquing After the Results are Known", and seems to have been coined by Nosek and Lakens (2014). By analogy, it means presenting a criticism of a design as one that you would have made in advance of the results being known, which is in fact a criticism made in reaction to unwelcome results. A CARKing criticism therefore seems to be one that the reviewer would not have made had the results been favourable, whether the critic is conscious of this fact or not.

For a regular research report, as opposed to a replication study, it is usually seen as making a crucial difference to the validity of statistical testing whether the hypothesis is advanced prior to the knowledge of the data or not. Kerr (1998) considers objections to HARKing from both classical and Bayesian statistics viewpoints. Many of the objections from a classical perspective amount to the charge that it increases the likelihood of Type 1 error beyond the specified p-value of a reported test statistic. To put this more strongly, virtually any real dataset will contain some pattern that is "significant" according to some statistical test if one collects enough covariate and outcome data, and so can be misrepresented after the fact as testing and confirming a hypothesis posited *ex-ante*. Under HARK-ing, then, one is almost guaranteed a high enough test statistic even if there are no genuine effects behind the data. This makes a nonsense of hypothesis testing since the real p-value should in that case, by definition, be close to 1.[9]

From a Bayesian statistics perspective, the objection Kerr (1998) relates is not temporal but substantive. That is, HARKing gives rise to *ad-hoc* theorising in the sense that nothing but the data to hand are likely to support the hypothesis. In that case, by definition, it has a very low prior probability. By disguising the *post-hoc* nature of a hypothesis, authors conceal the likelihood that it is *ad-hoc* from the reader and so exaggerate its plausibility.

No general parallel argument to the classical objection applies, it seems, in the case of CARKing, however, since it is not an exercise in pattern finding or rationalisation, but in criticism of a piece of research. Let us suppose for clarity that the reviewer is not testing a new hypothesis, which would amount to HARKing. *When* a criticism is advanced is not strictly relevant, in itself, to the soundness of a criticism. Implicit in the concern over CARKing, therefore, seems to be the view that criticisms prompted by "motivated reasoning" are particularly likely to be spurious. Since all acts of reasoning are motivated, what this refers to is reasoning motivated other than by a desire to further the scholarly objectives of the research field, such as the quest for knowledge. Most obviously, the original author or their associates, followers or admirers may want to protect a result to limit reputational damage, or perhaps to protect ongoing and future research projects, if a finding has not been reproduced. Since it is not possible to exactly replicate a study, imperfectly-motivated reviewers will predictably insist that some detail ought to have been implemented, and will have a menu of options to choose from.

This seems to mirror the Bayesian objection to HARKing. It is not the timing of the criticism *per se* that is worrying but the likelihood that, if an objection is made that would not have been made prior to the results being known, it is likely to be an objection with little to recommend it. In short, CARKing is probably carping. In an ideal scenario where editors have time to dispassionately evaluate reviewers' criticisms of a study, this might be of little consequence, since reasons motivated or not can still be independently evaluated in terms of evidential support and logic. But in the messy, real world where busy editors lack sufficient time for this, they are likely to take reviewers' advice to a significant extent on trust. As most academics know, who have bothered to argue with an editor. Thus, concern about CARKing seems pragmatically reasonable.

For an earnest reviewer to guard against CARKing ("Would I have made that criticism if I had not seen the results?") seems to imply an overly-exacting standard of self-knowledge. Pre-registering designs (in general) must help with both CARKing and HARKing. In particular, it is simply not possible to CARK as a reviewer if one does not get to see the results, and so attention is more

[9] By analogy. every time one strikes a match, the pattern of flame produced is unique in the history of the universe. The ex-ante probability of that particular pattern is vanishingly small, but one is guaranteed to produce some pattern that has negligible likelihood.

likely focussed on the epistemically important details of the design. The same menu of differences between the original and replication is available for the reviewer as material for objections, but at Nosek and Lakens argue (2014, p138), insistence on minor details risks trivialising a result. Ex-ante, that is, even an imperfectly-motivated reviewer faces a trade-off between protecting a cherished result and allowing the study to find that seemingly unimportant details do not matter.

Under pre-registration, a provisional decision is made on publication prior to data collection, and normally the only subsequent check is that the data was collected as planned, for example with sufficient sample size to achieve the pre-specified statistical power.

However, does it follow from the motivated reasoning concern, and the availability of pre-registration as a reviewing protocol, that criticism of a replication study made after the fact should never preclude publication? I think it does not. An example, which was the subject of heated controversy, will help to clarify the matter. The example is not introduced with a view to taking sides on the dispute over the particular study but to illustrate the existence of kinds of concern which are legitimate and can only arise after the data are known.

One of the findings targeted for replication in *Social Psychology* 45(3) was that judgements of an ethical nature tend to be less extreme if the person making them is clean, as reported in Schnall *et al*. (2008). In the original study, conducted in the UK and published in *Science*, participants either washed their hands (treatment) or did not (control) after experiencing disgust (implemented by a video screening) and before being asked to make moral judgements using a rating scale. A priming manipulation also found the same effect: when primed with the concept of physical cleanliness, subjects made more lenient judgements. The replication attempt (Johnson *et al*., 2008) used the same materials as the original study, and sample sizes far larger, but failed to reproduce either effect at a statistically significant level. The effect size (Cohen's *d*) estimated in each case is almost zero, but as seems the norm in replication studies, the authors conclude only that the true effect sizes "are substantially smaller than the estimates generated from the original … studies" (Johnson et al. 2014), and that researchers in the field should therefore use very large samples. This might be read as a joke, since the sample sizes for an effect size close to zero would be truly enormous to achieve a design with adequate power.[10] With the reported $d = 0.01$, for example, a 2-tailed, 2 sample t-test would require more than 400,000 subjects in total to achieve power of 90%. Good luck with obtaining funding for that!

On the face of it, this seems a very informative replication failure. Since the authors used the original materials, all of the epistemically relevant factors would appear to be held constant, including any flaws in the original design, alongside many irrelevant factors. In the authors' response to this replication (Schnall, 2014), however, it is pointed out that there were many more responses in the replication study close to the ceiling of the rating scale than obtained in the original study. Schnall argues that respondents in the USA may exercise more severe moral judgements than their UK counterparts on the scale used, for the particular stimuli deployed. The subsequent debate focussed on whether or not this made observing the purported effect too unlikely, whether observations near the ceiling can be dropped without too much detriment to the analysis and so on. But the replicators also invoked the dangers of CARKing as a response to Schnall's observation.

Whether or not the replicators are right in this specific case that the statistical analysis still goes through in the absence of the observations close to ceiling, it is generally the case that having

[10] Other seemingly modest pronouncements in this genre include the statement that the conditions necessary to produce the effect in question are "not yet fully understood".

too many observations at the ceiling of a scale could preclude any effect being observed. At the limit, with 100% of observations in the treatment group at the ceiling, one could only observe an effect in one direction or a null result. There is also no reason *a priori* to expect moral judgements to be equally severe across different populations. So this kind of objection must be valid in principle. Further, it is inappropriate to invoke a prohibition on CARKing here since this kind of flaw cannot possibly be identified prior to the data being collected. The flaws in question, as a general class, are features of the data that either render it implausible or that undermine the proposed analysis. Consider another, hypothetical, example. A replication study that collected an income variable from participants, and then reported that they were all billionaires, has implausible data, is not a credible piece of research, and ought not to be published to the extent that it relies on unbiased measurement of income. A strict adherence to a CARKing prohibition, and quality control only prior to data collection, will allow any such cases to go through, and might even contribute to poor quality data collection.

Since editors do exercise an element of quality control even with pre-registered designs after the data has been collected, it seems arbitrary to this author to rule out checks on data quality. As a formal extra step in a replication protocol this would increase time and effort necessary to complete the publication process, however. Assuming this could be done, additional measures to guard against the re-introduction of problematically-motivated reasoning might be necessary. For example, reviewers might be instructed not to comment on design features in an ex-post review, except in so far as they explain a data quality problem.

7. Closeness to theory in experimental economics – a double-edged sword?

We should perhaps be less surprised if findings turn out not to be repeatable, the less plausible or well specified is the theory set out whereby conditions A should produce effect or phenomenon B. In some of the replication failures from psychology, the theory supporting the core hypothesis seems under-developed. For example, consider the purported result that sensations of physical warmth affect people's tendency to exhibit selfishness as opposed to altruism. This finding was reported experimentally by Williams and Bargh (2008) and a replication attempt was conducted by Lynott *et al.* (2014), which failed to reproduce the effect. There is a theory section in the original study developed to motivate the research, glossed by the replicators as follows: "The basic idea is that physical feelings of warmth translate to greater interpersonal warmth."

As a theory, this seems to involve a pun on "warmth", since in the latter expression warmth is a linguistic metaphor for either an emotional state or a personality disposition. If there is a general body of knowledge implying specific causal connections between the two it should be invoked. Although motivating literature is presented in the original article, it takes a more impressionistic form that this. Evidence is cited that physical warmth is important in infant contact with early carers and subsequent healthy development; it is asserted that because of such contact that physical warmth and psychological warmth should be associated mentally, and finally that the insula is involved in processing sensations of both interpersonal and physical warmth (Williams and Bargh, 2008 p606).[11] Thus, arguably, the theory reads more as a chronology of factors that led to the formulation of a hypothesis rather than as a purported mechanism which would demonstrably

[11] The weakness of the neurological component of this sequence lies in the fact that specific parts of the brain, including the insula, are typically implicated in a wide range of cognitive and affective processes. See Poldrack (2006) for critical discussion.

produce the hypothesized effect, as realists would have it,[12] or even as a set of premises that logically imply the hypothesis.

Similar comments apply for the hypothesis linking moral judgements and cleanliness. At a risk of painting too clear and simple a picture, in economics, in contrast, theory development carries much prestige and experimental work usually has very close ties to theory. It would be tempting to assume, then, that economics is in a better position in this respect, and that therefore a better replication record can be expected to follow.

In one respect this seems unexceptional. There is an emphasis on deductive derivation of hypotheses from theories in modern economics which, plausibly, serves to discipline researchers against loose theorising. In other respects, however, it is debatable whether common certain kinds of tie between theory and experiment are necessarily healthy or good news for replication.

To see this, consider that most designs implement a setting that resembles in some way the ontology of economic theory. For example, rational choice theory is populated by independent decision makers, whose choice problems are represented as mathematical structures, with well-defined probability and utility / value components, for example. In the laboratory, independent decision making is implemented by having subjects sit and make decisions at workstations, blind to what others are doing. (There are also good statistical reasons for doing this.) And they typically choose (in parametric choice experiments) between lotteries with monetary payoffs and precise probabilities. In terms developed by Cubitt (2005), whilst a theory may have an *intended* domain which is very broad and inclusive of settings where ambiguity or radical uncertainty reign and communication is free, its legitimate *testing* domain may still include highly stylised settings. What matters for testing is whether the theory makes predictions for that setting, and this is a matter of what can be deduced from the theory. Using choice problems which resemble the relevant theories' representations helps, it seems, to make sharp predictions. If the theory fails in that setting, arguably, this is a particularly bad failure, since it should be more likely to work in a setting that resembles the theory. The downside to this is that if the theory performs well in such a setting, we still lack evidence for how well it performs in its intended domain.

Given a preference for working in such stylised environments, it is likely that the kinds of replication that take place in follow-up studies tend towards direct replication. Grether and Plott's (1979) famous study of the robustness of preference reversals, for example, consisted of repeating the P vs $ bet choice and valuation tasks, characterised by monetary prizes and well defined probabilities, in a variety of modified settings. However, this is a choice, and it is also possible to run economics experiments closer to the intended domain. Ball et al. (2012) for example, study (and find) classic preference reversals in tasks which have neither known probabilities nor known consequences. Arguably this sits closer to the conceptual replication end of the spectrum, despite the fact that choice and valuation task procedures are repeated.

However, a second common way in which experimental economists stick to theory is by "implementing" an economic model as closely as possible, in order to test the theory associated with the model. Bardsley *et al*. (2010) argue that this approach commits the "fallacy of misplaced concreteness", originally posited in another context by A.N. Whitehead (1925). This fallacy occurs when an abstraction is mistaken for the concrete thing that it is ostensibly derived from or

[12] Realists would presumably not endorse the 'as-if' interpretation of economic theories, nor the conscious production of theories which only sustain an 'as if' reading, however.

represents. For an economic model is essentially a mathematical structure that is used in the course of theorising *about something else*. The theory is not *about* the model.

Attention should therefore be paid to implementing the something else, that which that the theory is about, rather than the model. Only in this way will the theory make predictions about the experimental environment, and therefore only in this way can the theory be tested or usefully developed by experiment. A classroom experiment which maximally implements the assumptions of, for example, Cournot duopoly theory, is one in which there is no firm in the natural language sense of the word. For this natural language is not part of the formal model, despite the fact that the theory forms part of "theory of the firm" or "industrial organisation". There are limits to implementation, however, if subjects really are to choose how to behave. What is really being studied in such cases, according to Bardsley *et al*.'s (2010) critique, is that which is not implemented, namely assumptions about motivation and interaction. By implementing the model, therefore, what one actually studies is a behavioural game, and what one actually tests is a joint hypothesis comprising a game theoretic solution concept and motivational assumptions.

A model-implementing experiment can clearly be subject to direct replication, since any procedure can be repeated, with irrelevant details such as the instruction fonts, wall colours and so on changed, and can also be found to be robust (or not) to levels of incentives, variations in samples, degree of repetition and so on. But it seems that it is difficult to specify what would constitute a conceptual replication. To stick with the same example, if we aim to reproduce a result about Cournot duopoly in a radically different setting, by implementing the same model, it is not clear what can constitute the different setting. For if the model has to be implemented this determines the setting, and the same model then implies essentially the same setting, and there seems to be no real room for manoeuvre. If we introduced potential competition on quality, for example, this contradicts the Cournot model. A different model would concern a different theory and a different set of hypotheses.

Alternatively, under the interpretation of these experiments as actually constituting decision- and game-theoretic experiments, it is relatively straightforward to specify what would count as a conceptual replication. A test of Nash equilibrium can be carried out in a huge range of strategic situations. To take one finding, there is evidence of convergence to Nash equilibrium in several settings involving repetition and random re-matching. One might try to reproduce this convergence using animals and children, using non-monetary rewards for example. And indeed this has been attempted with some success (*inter alia* Lee *et al.* 2004; Sanabria and Thraikill, 2009; Sher *et al.*, 2014). Arguably it does not make sense for a finding only to admit of direct replication attempts.

The theory-centric character of many economics experiments therefore seems to be double-edged. On the one hand, it is plausible to expect a greater degree of direct replicability in experimental economics than perhaps obtains in psychology (on average), given the consequent uniformity of experimental designs and the logically tight derivation of hypotheses from formally specified theories. On the other hand, the more stylised and less rich are the experimental designs, the less we are likely to learn about how the results generalise to settings that the theories are intended to apply to. The more extreme version of closeness to theory, implementing theoretical models in the laboratory, seems to preclude the possibility of conceptual replication. This may be an indicator of a conceptual mistake underlying this approach, but the problem disappears if we reinterpret the designs as research exercises in the behavioural, as opposed to applied economics, domain.

8. Conclusions

Systematic replication efforts have only recently been conducted in psychology and are still more recent and nascent in experimental economics. Although the indications are that the replicability in either field is less than what would obtain under ideal conditions of scholarship, research and publishing, the extent of this problem is actually currently unclear in either field. There is wide variation in replication rates in the recent literature. No general case was found for authorial involvement in direct replication efforts, though it may be necessary in some cases if potentially relevant information has been omitted. There are legitimate concerns over CARKing, in that it may consist in motivated reasoning which journal editors as a practical matter may not be in a good position to guard against. On the other hand a complete prohibition on data-based criticism seems unjustified. There appears to be a perceived lack of incentives to conduct direct replications, which may stem partly from motivated reasoning issues on the part of reviewers. Pre-registered designs and journal special issues help with the incentives to replicate problem. But there would appear to be enhanced value in randomising selection of findings to replicate and in assignment of each target finding to a replication team, to give a clearer idea of the general extent of replicability in a field. Concerning which details should be replicated, this chapter argues that what counts are the goals of the replication and epistemic features of the original designs, relative to which some details will be irrelevant.

A key characteristic of experimental economics that may to some extent distinguish it from experimental psychology is the tight relationship most designs have to theory. It seems that experimental economics may in one sense be well served by such proximity, in terms of the avoidance of under-developed theorising supporting a hypothesis. But sticking closely to environments which are isomorphic to economic theory arguably comes at a cost in terms of the informativeness of designs about a theory's intended domain, which can be seen as a potential deficit in conceptual replication. When this closeness to theory consists of implementing an economic model, it seems that conceptual replication is precluded. This seems to be an indicator that the approach involves conceptual error, as has been argued on methodological grounds by Bardsley *et al.* (2010).

Appendix

Definitions

p	the probability that a null hypothesis selected for investigation is true
T	the event that the null hypothesis is True
~	Not
r	the event that an original trial rejected H_0
R	the event that H_0 is rejected in a new trial
α	significance threshold
β	power of replication design

By definition $P(T) = p$, and by assumption the original and new trials are independent.
Under precise replication of the experiment

$P(r|T) = P(R|T) = \alpha$

$P(r|\sim T) = P(R|\sim T) = \beta$

$P(R) = P(R, T) + P(R, \sim T) = P(R|T)P(T) + P(R|\sim T)P(\sim T) = \alpha p + \beta(1-p)$

With no biases in reporting,

$P(R) = \alpha p + \beta(1-p) = P(r)$

$$P(T|r) = \frac{\alpha p}{\alpha p + \beta(1-p)} = P(T|R)$$

$$P(\sim T|r) = \frac{\beta(1-p)}{\alpha p + \beta(1-p)} = P(\sim T|R)$$

Rates of replicated rejection of H_0 are therefore given by

$$\begin{aligned} P(R|r) &= P(R, T|r) + P(R, \sim T|r) \\ &= P(R|T, r)P(T|r) + P(R|\sim T, r)P(\sim T|r) \\ &= P(R|T)P(T|r) + P(R|\sim T)P(\sim T|r) \\ &= \frac{\alpha^2 p + \beta^2(1-p)}{\alpha p + \beta(1-p)} \end{aligned}$$

References

Ball, L., Bardsley, N. and Ormerod, T. (2012). Do preference reversals generalise? Results on ambiguity and loss aversion. *Journal of Economic Psychology*, 33, 48-57.

Bardsley, N., Cubitt, R., Loomes, G., Moffatt, P., Starmer, C. and Sugden, R. (2010). *Experimental economics: Rethinking the Rules*. New Jersey: Princeton University Press.

Battalio, R.C., Kagel, J.H. and MacDonald, D.N. (1985). Animals' choices over uncertain outcomes: some initial experimental results. *American Economic Review*, 75, 597-613.

Bhaskar, R. (1978). *A Realist Theory of Science.* Hassocks: Harvester Press.

Bhattacharjee, Y. (2013). The mind of a con man. *The New York Times*, 26th April.

Bornemann-Cimenti, H., Szilagyi, I.S. and Sandner-Kiesling, A. (2016). Perpetuation of retracted publications using the example of the Scott S. Reuben Case: Incidences, reasons and possible improvements. *Science and Engineering Ethics*, 22, 1063-1072.

Brandt, M.J., Ijzerman, H., Dijksterhuis, A., Farach, F., Geller, J., Giner-Sorolla, R. , …, and van't Veer, A. (2014). The replication recipe: what makes for a convincing replication? *Journal of Experimental Social Psychology*, 50, 217-224.

Burke, B. L., Martens, A., Faucher, E. H. (2010). Two decades of terror management theory: a meta-analysis of mortality salience research. *Personality and Social Psychology Review*, 14, 155–195.

Camerer, C.F. *et al.* (2016). Evaluating replicability of laboratory experiments in economics. *Science*. DOI: 10.1126/science.aaf0918

Collins, H. (1991). The meaning of replication and the science of economics. *History of Political Economy*, 23, 123-142.

Coyne, J. (2016). Replication initiatives will not salvage the trustworthiness of psychology. *BioMed Central Psychology*, 4:28.

Cubitt, R.P. (2005). Experiments and the domain of economic theory. *Journal of Economic Methodology*, 12, 197-210.

Fiske, S.T. (2016). A call to change science's culture of shaming. *Association for Psychological Science Observer*, 29(9), 5-6.

Gilbert, D.T., King, G., Pettigrew, S. and Wilson, T.D. (2016). Comment on "Estimating the reproducibility of psychological science". *Science*, 351, 1037.

Glass, G.V. (2000). Meta-Analysis at 25. Retrieved Aug 15 2016 from http://www.gvglass.info/papers/meta25.html

Greenwood, J.D. (1982). On the relation between laboratory experiments and social behaviour: causal explanation and generalisation. *Journal of the Theory of Social Behaviour*, 12, 225-249.

Grether, D. and Plott, C. (1979). Economic theory of choice and the preference reversal phenomenon. *American Economic Review*, 69, 623-638.

Hume, D. (1739-40 [2007]) *A Treatise of Human Nature*, ed. D.F. Norton, M.J. Norton. Oxford: Clarendon Press.

Ioannidis, J.P.A. (2016). Why most published research findings are false. *Public Library of Science – Medicine*, 2, e124.

Johnson, D.J., Cheung, F. and Donellan, M.B. (2014). Does cleanliness influence moral judgements? A direct replication of Schnall, Benton and Harvey (2008). *Social Psychology*, 45, 209-215.

Kahneman, D. (2014). A new etiquette for replication. *Social Psychology*, 45, 310-311.

Kerr, N.L. (1998). HARKing: hypothesizing after the results are known. *Personality and Social Psychology Review*, 2, 196-217.

Klein, R.A., Ratliff, K., Vianello. M., Adams, A.B.Jr., Bahník, S., Bernstien, N.B., … , and Nosek., B.A. (2014). Investigating variation in replicability: a "many labs" replication project. *Social Psychology*, 45, 142-152.

Lee, D., Conroy, M.L., McGreevy, B.P. and Barraclough, D.J. (2004). Reinforcement learning and decision making in monkeys a competitive game. *Cognitive Brain Research*, 22, 45-58.

Levelt, (2012). Levelt, W. J. M., Drenth, P., & Noort, E. (Eds.). (2012). Flawed science: The fraudulent research practices of social psychologist Diederik Stapel. Tilburg: Commissioned by Tilburg University, the University of Amsterdam and the University of Groningen.

Lynott, D. Corker, K.S., Wortman, J., Connell, L., Donnellan, M.B., Lucas, R.E. and O'Brien, K. (2014). Replication of "Experiencing physical warmth promotes interpersonal warmth" by Williams and Bargh (2008). *Social Psychology*, 45, 216-223.

Nosek, B.A. and Lakens, D. (2014). Registered reports. A method to increase the credibility of published results. *Social Psychology*, 45, 137-41.

Nozick, R. (1981). *Philosophical Explanations*. Cambridge MA: Harvard University Press.

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

Pawson, R. and Tilley, N. (1997). *Realistic Evaluation*. London: Sage.

Plott, C.R. and Smith, V.L. (2008). *Handbook of Experimental Economics Results. Volume 1.* Amsterdam: North Holland Press.

Poldrack, R.A. (2006). Can cognitive processes be inferred from neuroimaging data? *Trends in Cognitive Sciences*, 10, 59-63.

Rosenthal, R. and Fode, K.L. (1963).The effect experimenter bias on the performance of the albino rat. *Behavioral Science*, 8, 183-189.

Rosenthal, R. and Jacobson, L. (1963). Teachers' expectancies: Determinants of pupils' IQ gains. *Psychological Reports*, 19, 115-8.

Sanabria, F. and Thraikill, E. (2009). Pigeons (*Columba livia*) approach Nash equilibrium in experimental matching pennies competitions. *Journal of the Experimental Analysis of Behavior*, 91, 169-183.

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, 90-100.

Schnall, S., Benton, J. and Harvey, S. (2008). With a clean conscience: cleanliness reduces the severity of moral judgements. *Psychological Science*, 19, 1219-1222.

Schnall, S. (2014). Commentary and rejoinder on Johnson, Cheung and Donnellan (2014a). Clean data: statistical artefacts wash out replication efforts. *Social Psychology*. Online article; DOI: 10.1027/1864-9335/a000204

Sher, I., Koenig, M. and Rustichini, A. (2014). Children's strategic theory of mind. *Proceedings of the National Academy of Sciences*, 111, 13307–13312.

Solomon, S., Greenberg, J., Pyszczynski, T. (1991). A terror management theory of social behavior: The psychological functions of self-esteem and cultural worldviews. *Advances in Experimental Social Psychology*, 24, 93-159.

Stapel, D. (2014). Faking science: a true story of academic fraud. Translated by N.J.L.Brown. Available at https://errorstatistics.files.wordpress.com/2014/12/fakingscience-20141214.pdf

Williams, L.E. and Bargh, J.A. (2008). Experiencing physical warmth promotes interpersonal warmth. *Science*, 322, 606-607.

Wilson, A. (2014). Psychology's real replication problem: our methods sections. Online article at http://psychsciencenotes.blogspot.co.uk/2014/05/psychologys-real-replication-problem.html accessed 01.08.2016

Whitehead, A.N. (1925). *Science and the Modern World.* Cambridge: Cambridge University Press.