Peer Review of “Experiment and Fail”

Sean Devine, Stefan Gaillard

Introduction

The paper ‘Alcohol cues and aggressive thoughts’ reports a failed attempt at reproducing two experiments. The massive shortcomings of the reported reproduction are obvious. For a moment I was tempted to think that the authors, in the form of a standard psychological paper, were presenting a philosophical critique of this type of experiments. In my comment I will try to formulate such a critique in a more straightforward manner.

I will first give a brief and plain description of what happened in the experiment and why it was done, according to the authors. Then I will say a few things about the complexity of producing and reproducing experiments in general, followed by a section on the problems of the specific type of experiments of which this one is a specimen: priming studies, mostly found in the subdiscipline of social psychology and since almost a decade the subject of a vigorous debate among methodologists, philosophers of science, priming researchers, in scientific journals, but also in newspapers, magazines, blogs and on Twitter. I will end with assessing the possibilities and the limits of doing experiments in the human sciences: what can we learn from experiments on alcohol cues if we want to tackle physical, mental and social harm, attributed to the consumption of alcohol?

The Experiment

This is how the experiment went. Sixty people, all students, volunteered to come to the Psychological Health and Well-Being lab on Bishop’s University campus (Sherbrooke, Quebec). Upon arrival they are instructed to take part in a word recognition task: how fast and how accurate are they able to decide whether a string of letters presented to them on a computer screen is a legitimate English word. In total, they get to see 45 letter strings, of which 15 are neutral, 15 are ‘non-words’ and 15 are ‘aggression-related words of a sexual nature’. Prior to the presentation of a target word, a photo is shown of a weapon, an alcoholic drink, or a non-alcoholic drink.¹ Participants have to indicate by pressing on a key whether the letters in their judgment represent a legitimate English word or not.

Now why was this done? The main goal of the experiment was to find out if the performance of the participants would be similar to the results of earlier experiments, carried out by other researchers in 2006 and 2010 respectively. If successful, the new experiment would generate support for the idea that people who see the image of a weapon or an alcohol drink – even for a split second – are influenced (unconsciously and automatically) to choose aggression-related words faster than neutral words in the case of neutral (non-alcohol) images. This idea is in line with the ‘semantic network model of memory’, which suggests that human beings can learn to associate a gun with violence, and alcohol with (sexual) aggression, simply by the frequent, simultaneous occurrence of these phenomena.

In the 2006 and 2010 experiments this type of association indeed was shown, in the present experiment it was not: alcohol cues were detected slower

than non-alcohol cues. So, was the alcohol-aggression hypothesis falsified? Not necessarily, according to the authors, who come up with no less than seven possible causes why their experiment generated different results than the previous ones. In the end they concluded: ‘The replication attempt suffered from many methodological and design-related issues.’ (p.23)

From their account it becomes clear how complicated reproducing a seemingly straightforward experiment actually is. While an experienced experimenter might shake his or her head at such an imperfect specimen of experimental research, in my view it is a very instructive case, precisely because of its faults. It is a specimen of ‘sloppy science’. Normally when authors want to publish a paper that is based on weak research, they cover up shortcomings using methodological decisions and statistical manipulations, but these authors refrained from such a procedure.

The Intricacies of Producing and Reproducing Experiments

Performing an experiment is quite a complicated task. Apart from a theoretical description of the required manipulations, the object and the apparatus, there is the material realization of the actual experiment, which necessitates careful preparation. The complexity of experimenting in general can be illustrated by the following example, which is taken from the book *In and about the world* (1996) by the Dutch philosopher of science and technology, Hans Radder.

Consider [...] an experiment for determining the boiling point of a particular liquid. This liquid is our object under study. Our apparatus consists of a heat source, a vessel, a thermometer, and possibly some supplementary equipment. On the basis of our knowledge of the interaction process between thermometer and liquid, we assume that our readings of the thermometer inform us about the temperature of the liquid. Part of the preparation procedure involves making sure that the liquid in question is pure. This is why it may be necessary first to clean the vessel that will contain the liquid. (Radder, 1996, 11)

Besides guiding the preparation of the object and the necessary equipment, the theoretical description informs us about the staging of the processes of interaction between object and equipment and the processes of detection (i.e. measuring). Finally, the experimental system should be ‘closed’, which means that potential disturbances from the outside² should be identified and controlled;

1. Interestingly, one of the photos shows bottles of the beer brand Corona, which would nowadays have a totally different association and might possibly ruin the experimental setup. Also, for the association to occur subjects would have to know that Corona is a beer; this is not clear from the photo.

2. This can also mean: disturbances by internal processes, for instance affective and emo-

this is also part of the theoretical description.³

This all sounds rather straightforward and self-evident, but experimental procedures are full of hidden presuppositions, as becomes clear when a researcher is given the task to instruct a layperson how to perform a certain experiment. An elaborate and very detailed and precise list of actions, worded in common (nontheoretical) language is needed for this layperson to successfully execute the tasks involved. It moreover requires that the researcher already knows how to perform the experiment.⁴

The notions of theoretical description and material realization are both relevant and helpful to analyse the issues at stake with the reproduction of experiments. Radder distinguishes between three types of reproducibility.⁵ Type 1 is the reproducibility of the material realization of an experiment, which means: (a) it is not dependent on any particular theoretical description, (b) it can be done by laypersons. In type 2, an experiment may be reproduced under a identical theoretical interpretation, which allows slight variations in the material procedures. Type 3 concerns the reproducibility of the result of an experiment, which implies that it is possible to obtain the same experimental result while performing – theoretically and materially – different procedures; this is, in Radder's terminology, a replication of the original experiment. In contrast to type 3, type 1 and 2 require a reproduction of the whole of the experimental process.

In addition to these three varieties of reproduction type, Radder distinguishes four possible types of actors in the reproduction process, or four ranges of reproduction: (1) reproducibility by any scientist or even any human being in the past, present or future, (2) reproducibility by contemporary scientists; (3) reproducibility by the original experimenter, and (4) reproducibility by the lay performers of the experiment. Types and ranges combined, there are thus twelve possible categories in the field of reproduction, which allows a far more sophisticated assessment and categorization than the usual differentiation between 'direct' (or exact) and 'conceptual' reproductions (see below), or the categorization by the Dutch research funder NWO: (1) replication with existing data, (2) replication with new data (and the same research protocol), (3) replication with the same research question, but with a different research protocol and new data.

Is 'Alcohol cues and aggressive thoughts' a reproduction? ▼

In what category can we now place the word decision experiment we are discussing here? First of all, considering Radder's aspect of range, it is a reproduction performed by (more or less) contemporary scientists; the timespan between the original and the reproduction amounts to 14 years. Secondly, considering type, we learn from the description of the experiment that the researchers initially aimed at reproducing the original experiment, using the same protocol: 'Dr. Bartholow generously shared the original target word stimuli and a description of the images he employed, allowing extremely similar material to be used in this study.' (p. 22)⁶ In that case, we would have a reproduction under a fixed theoretical interpretation, i.e. type 2.

However, the reproducers wanted to also study the accessibility of sexually aggressive thoughts and therefore decided to change the sets of target words and images, in order to accommodate to addition of a new variable. Neverthe-

less, they themselves considered their experiment to be a true reproduction: 'the initial research protocol was followed closely and the research question remained unchanged [...] even if the nature of the aggressive words [and the photos! RA] was altered'. (p. 22-23) This claim is not irrelevant since it appears that the financing of the study was dependent on the study being similar (with the same research protocol) to the original study.

In my opinion the claim made by the authors in this respect is debatable or even false: changing target words and images means bringing about a change in the 'apparatus' used, which also implies changes in the 'interaction' with the object and probably also in measurement procedures. One way to establish whether the research protocol is really 'the same', as the reproducers claim, would be to explicate in common language the detailed instructions for the material realization of both the original experiment and the reproduction. This is hardly ever done; usually, and also in this case, the method section in journal articles does not give the reader (and the reproducer-to-be) sufficient information about the actual proceedings to create a 'lay persons instruction'. It would require getting the protocol from the original experimenters, and even then more detailed information might be necessary.⁷

For now, I hold that the reproduction experiment discussed here is at best a replication (cf. Radder), i.e. an attempt at attaining the same experimental result while performing – theoretically and materially – different procedures. In itself, this could be valuable: the significance of a result is stronger when it can be obtained under different experimental processes or, as Hans Radder puts it: 'Abstraction through replication enables us to systematically conceptualize experimental results arising in essentially different situations. As such it constitutes an important step towards theory formation.' (Radder, 1996, 84) Because the reproducers introduced a new variable in their study ('sexually aggressive' instead of 'aggressive'), one could argue that this is not even a replication but a new experiment.

Priming Studies ▼

The reproduction experiment under discussion is a so-called priming experiment, which means that a stimulus is presented that is supposed to subconsciously influence the subjects in the experiment in a systematic way, as measured by their results on a specific task.⁸ In this case the prime consists of photos of a different nature (weapon, alcohol, non-alcohol) and the specific task is word recognition. According to the 'semantic network model of memory', an individual primed by for instance the image of a beer bottle would be prone to choose a sexually aggressive word faster than a neutral one as a 'legitimate English word'.

Since a quarter of a century, this type of experiment has become very popular in social psychology. For the subdiscipline as a whole, it is an attractive

tional processes in human beings.

3. Radder 1996, 12.

4. The intricacies of learning how to do a proper experiment (and to perform reproductions) are described and analysed in Collins 1985.

5. Radder 1996, 16-18.

6. Page references in this concept are to the revised manuscript d.d. xx April 2020

7. See Collins

8. See for an overview of the field of priming studies Derksen, Ch. 9.

type of study because it puts a counterintuitive and for some controversial idea centre stage: in human decision making, volition or free will is far less important than is usually thought; instead, people take many – if not most – decisions automatically, influenced subconsciously by environmental factors. For individual researchers, engaging in the priming tradition opens up a variety of topics to study experimentally, and a possibility to take their share in an almost unlimited market of publication opportunities. Presenting results that are at odds with common sense thinking is considered an asset.⁹

From 2011 onwards however, a fundamental debate has started about the quality of the ever expanding field of priming research, leading to a ‘crisis of confidence’ in experimental social psychology, or at least in the area of priming research. The main allegation was that within social psychology there was an abundance of ‘sloppy science’. Researchers were accused of having ‘fotoshopped’ their raw data by methodological and statistical manipulations or, as a group of methodologists put it: the field of (social) psychology ‘currently uses methodological and statistical ¹⁰strategies that are too weak, too malleable, and offer far too many opportunities for researchers to befuddle themselves and their peers.’¹¹ There is in (social) psychology an overly enthusiastic use of ‘researcher degrees of freedom’, which enables researchers to obtain almost every result they want.

Adding fuel to the upheaval were attempts to reproduce ‘classic experiments’ in social psychology, such as Bargh’s study on the effect of subjects being primed by word references to old age, who would afterwards walk slower toward the exit of the building where the experiment was conducted (as elderly people are supposed to do).¹² The reproducers followed Bargh’s protocol as exact as possible, but nonetheless were not able to produce the same results. Fearing that this outcome would damage ‘his life’s work’, Bargh attributed the reproduction failure to the incompetence of the reproducers.¹³ Several years later, in a more general attempt to do something about the ‘replicability crisis’ and validate psychological research, a massive Reproducibility Project was conducted, leading to a shocking result: no more than one-third of experimental results could be replicated.¹⁴

This ignited a debate on the value of doing what was called ‘exact (or direct) replications’ (following the same protocol) versus so-called ‘conceptual replications’, in which researchers with the same theoretical background as the original experimenters, use more or less different operationalizations to produce similar effects, in order to strengthen and extend the theory at stake. Whereas methodologists seemed to favour ‘direct’ reproductions, social psychologists considered them of little value: ‘In psychological research, there are always a multitude of potential causes for the failure to replicate a particular research finding’.¹⁵

Not surprisingly, these causes mostly had to do with the complexity of experimental work; either the protocol wasn’t precise enough or the reproducer lacked the necessary skills; or there were mediating variables at work, that were thus far unknown. According to Bargh, especially priming studies are very sensitive to this, which is why priming experiments require precise control and great skill from the experimenters.¹⁶ This line of reasoning is at least half a century old. Already in 1968, two respected experimentalists in social psychology, Elliott Aronson and Merrill Carlsmith, wrote:

...] when an attempted replication fails, one must interpret this failure with caution because it is difficult to draw firm inferences. The most we can say is that there was something about the original experiment which was not accurately specified and which seemed to have had an important effect on the results. One obvious but frequently overlooked problem about failures to replicate is that negative results are easily produced by incompetence.¹⁷

In the word decision experiment discussed in this commentary, similar issues are at stake. In fact they all pointed to one and the same methodological problem: the endeavour to ward off disturbing influences from the experimental situation or, in other words, to ‘keep the system closed’. Although this ‘closedness’ must be qualified – you only need to control those influences that are relevant in view of the problem and the aim of the experiment – it implies careful and systematic action: you ‘have to produce and maintain them through active intervention’.¹⁸

The fact that we’re dealing with interventions by the experimenter might also imply that we are creating an unnatural situation that might diminish the external validity of the results. Andrew Collier recounts a study in which a group of young men is isolated and manipulated to establish how power relations between the participants develop. His question is: what conclusions can you draw about what people tend to do outside the lab? This is his answer:

The participants are removed ‘from their families and friends, their jobs and normal leisure activities, even the common light of day. Now of course, all experiments are artificial, but in a good experiment the artifice removes the effects of irrelevant variables on the matter being tested. In this case, it does not remove but introduces such variables, just as putting animals in cages makes it impossible to study their natural behavior. [...] The proper way of studying the effects of power changes and power vacuums on humans would be by studying human behavior in the wild [...] in the open system of history, for instance the history of the French Revolution.’ (Collier, 2005, 332)

In the human sciences the experiment constitutes a miniature social system, in which the subjects are not passive, but respond actively and intelligently to all that goes on in the experimental situation. Using the information they

9. Cf. Strack, 2012.

10. See Hashler and Wagenmakers (2012)

11. Quoted in Derksen, 178

12. Bargh, Chen and Burrows, 1996; see also Derksen 183-188.

13. See Yong (2012a and 2012b).

14. Open Science Collaboration, 2015.

15. Stroebe, Postmes and Spears, quoted in Derksen, 183.

16. Derksen, 187.

17. Aronson and Carlsmith, 21.

18. Radder 1996 120-123, quote on 122.

can get, participants will try to guess what the aim of the experiment is and attune their behavior accordingly (demand characteristics). Probably the most important source of information is the experimenter him- or herself, who might unwillingly influence the responses of the subjects (experimenter bias). For a good reproduction of an experiment, it is important to know how the experimenter operated, whether it was a man or a woman, what the precise wording was of the instruction to the participants, what knowledge the subjects might have or obtain about the goal and hypotheses of the study, and to what degree deception of the subjects was involved in the experimental setup. As noted, this type of detailed information is however hardly ever reported in journal articles, which hampers the attempts of reproducers to establish what precautions are taken to minimize experimenter bias or gauge the influence of demand characteristics.¹⁹

Experimenting with Humans ▼

Given the complexity of procedures and the insecurity of results, why would human scientists perform experiments? Experimenting with human subjects seems to be far more difficult than with physical objects. This is ironic, because originally it was precisely the association with the rigour of the natural sciences that inspired psychologists in the late nineteenth century to adopt the experimental method as their favourite *modus operandi*. Nowadays the usual argument is: ‘by doing experiments you are able to establish causal connections’. This takes the form of ‘if p, then q’, which means there is a ‘constant conjunction’ between p and q, where q is ‘caused’ by p.²⁰ In terms of our experiment: if we present the image of a beer bottle to the subject, (s)he will choose an sexual-aggressive word faster than an alternative.

But what if this does not happen? Should we reject this specific regularity and try to come up with a better one? Not necessarily. There might be one or more methodological flaws in the reproduction attempt, for instance, as in the experiment discussed here: ‘The sample size was probably too small to optimize statistical power’. (p.15) Or ‘Selecting words for sexual aggression was flawed.’ (p.17), ‘The validity of the neutral category of pictures employed remains unclear.’ (p.23) This points to a recurrent problem in experimental social psychology. Already in 1954 Leon Festinger wrote: ‘[...] negative results perhaps reveal only the fact that the experiment was not set up carefully and that the experimenter’s attempted manipulation of the variables was ineffective’.²¹

In addition, there could be a misguided preconception in the research question, for instance the idea that alcohol cues are automatically linked to aggressive thoughts, and not to ‘feeling good’ or ‘having fun with friends’. An indication for this is given in the discussion section, where multiple participants said ‘that sexual violence was not necessarily associated with alcohol for them.’ (p.21) The authors suggest that cultural differences might be at stake here: Multiple participants coming from Europe stated that they were ‘raised with an open-minded attitude towards alcohol’ (p.21)

Another shortcoming might be caused by a central feature of priming experiments: they depend on deception. Deceiving subjects about what is the goal of the investigation has been an important instrument for experimental social psychologists since the 1950s to keep the experimental system closed. But do we really know whether we succeed in deceiving our participants? In the case discussed here, one subject was removed from the sample ‘because it

was clear from the debriefing session that this participant had not understood the computer task properly’. (p. 9-10) This indicates that the task may be interpreted in various ways; how can we be sure that the other participants did not have their own, though maybe less interfering, interpretation of the task? They might even have guessed what the experiment actually was about. This is not far-fetched because the authors themselves admit that, on seeing images of alcoholic and non-alcoholic beverages, some participants suspected that the study was ‘researching the effects of alcohol as it is contrasted with non-alcoholic beverages’ (p.18). That guess is close enough to open the experimental system to a confounding variable.

These examples of possibly confounding interpretations by experimental subjects point to a fundamental issue: participants will have their own interpretation of the nature and goal of the experiment, and this interpretation may influence their responses in a way that is not intended by the experimenter. This is usually referred to as the problem of the ‘double hermeneutic’: researchers have to be aware of both their own interpretation of what’s going on in their research and the interpretation that the subjects have of the experiment they are participating in. If researchers fail to assess properly what their subjects think, their interpretation of the results might be seriously flawed. Here we have a fundamental difference between the natural and the human sciences: the objects of study in natural science disciplines do not interpret the experimental interventions they are subjected to.²² In conclusion, experiments may not be the best way to study people.

Does this mean we have to discard with the idea of causality? Yes, if that means sticking to the principle of constant conjunction of isolated variables. No, if causality implies taking into account the characteristics of the object of study, the generative powers that are typical of human beings, for instance their judgmental sophistication, their use of written language, etc. Yes, people do have automatic responses to situations, but they also have the ability to consciously assess what’s going on and choose their course of action. The range of options is not unlimited, but on the other hand it is not possible to predict the outcome with certainty.

Relevance ▼

The importance of this insight can be illustrated if we return to the experiment at hand. Why would we do research into the effects of alcohol use in human beings? Because ‘harmful use of alcohol results in 3 million deaths worldwide every year’ (p. 3). How do these deaths come about? The authors suggest that ‘intimate partner violence’ is among the primary causes, but this is not explicitly stated. Even if partner violence would not be responsible for all of these 3 million deaths, an investigation into its causes would be most important. The main issue would obviously be: how can we prevent partner violence? And yes, alcohol abuse can be an important causal or facilitating factor, so we would also like to know: how can we prevent alcohol abuse?

Instead, the researchers deflect from these core issues and enter into a differ-

19. See Klein, 2012, 574-575.

20. See Manicas, 2006, Ch. 2; Manicas argues convincingly that causality requires more than a ‘constant conjunction’.

21. Festinger, 1954, 143

22. See Radder, 2019, 228, 230-231

ent debate: what is it, regarding alcohol, that leads to or increases aggression? That it does is 'generally accepted', but 'there is still a debate as to what precisely causes or explains this increase'. (p. 4) Not surprising, 'it is usually best explained through a combination of theories and viewpoints'. That makes sense, but instead of informing the reader about this combined theory, the authors give a brief overview of rivalling theories (or hypotheses). For instance, drinking alcohol increases aggression, 'by anesthetizing the part of our [sic] brain that usually keeps our aggressive impulses under control'. The outcome is of course that people are 'more likely to express aggressive behaviours'. For some researchers, this is too straightforward, and they propose that alcohol consumption increases aggression 'by affecting intellectual functioning and reducing self-awareness'. (p. 5) Finally, it might be 'that people tend to associate aggression and alcohol together, even if only unconscious [sic]', which would increase the likelihood of people behaving in an aggressive manner when they have been drinking.

Here we have arrived at the problem that the researchers set out to tackle: is this unconscious association between alcohol and aggression activated by the 'belief that alcohol consumption has occurred' or, alternatively, are 'alcohol cues alone' sufficient for increasing 'the accessibility of aggressive thoughts'? Within two pages the authors have steered their readers from the urgent issue of preventing 3 million deaths by alcohol each year to a sophisticated psychological issue, which can be solved by conducting an experiment involving alcohol cues and a 'lexical decision task'. The results 'suggest that exposure to alcohol cues without [alcohol] consumption is linked with an increase in aggression related thoughts'. (p. 6).

According to the reproducers, in their closing paragraph 'This line of research should be further pursued as it bears significant importance on [sic] today's society.' (p.23) But is there any practical value in this type of research? Even if the results were as expected, they would not help to tackle the alcohol-and-violation-problem. If we would decide to ban all 'alcohol cues' from the public domain, alcohol use and abuse would continue. So suspicion arises that the references to the actuality of a grave social problem are used as a legitimization for conducting an experiment that has to decide on a specific effect of alcohol primes, namely aggression.

The subject of the paper raises many, obvious questions: why wasn't the research aimed at another, very obvious association: alcohol and pleasure? How and in what circumstances do people learn to associate alcohol with violence and aggression? And when we say people, do we mean both sexes or mainly 'the male of the species'? And can we extinct, for instance by operant conditioning, the automatic association between alcohol and aggression?

Conclusion ▼

The call for more reproductions or replications is worthwhile, but only if the original experiments are solid and theoretically interesting, I would say. Readers may also have doubts about the practical value of this type of study, especially since there is hardly any knowledge of the relation between aggressive thoughts and subsequent behavior (p.22) So, in spite of the laudable intentions of the reproducers to help solve the reproducibility crisis, it would seem that their efforts could only have been of interest within the milieu of priming specialists.

These specialists however would probably not be very pleased with this experiment because of its obvious shortcomings, that are reported with unusual candour by the authors. Why attempt to publish a report like this in the first place? Publication might be instructive to psychology students how not to perform replications, but is that a sufficient legitimation? Anyway, the paper gave me the opportunity to reflect on the various levels of complexity that are involved in conducting priming experiments with human subjects, and maybe help some social psychologists to reconsider their research practice.

Peer Reviews

Reviewer 1 Sean Devine

0) General comments

- The author uses an empirical study (Lebouff, Linden-Andersen, & Carriere, 2020) to provide a broader, critical reflection on replication studies and the use of priming studies. The author raises questions about the validity of using methods from the natural sciences to study humans.

1) The Experiment

- The authors provides a more than adequate overview of the experimental study they are reflecting on.

2) The Intricacies of Producing and Reproducing Experiments

- The author uses mostly older sources to indicate the difficulties in producing and reproducing experiments. This has two advantages, namely that it shows how long these insights have already been around (to no avail) and that it traces the lineage of the debate. It would be nice to provide a brief reflection on more current philosophical and methodological debates and to indicate whether they are substantially different from the older critiques.

3) Experimenting with Humans

- This section could benefit from the author expanding on possible 'solutions' to the problems at hand. For example, could methods from complex systems studies help move psychology forward? Or should psychology go back to a more hermeneutic approach?

4) Conclusion

- I would recommend that the author expand on the question of whether this type of research should be published or not.

Reviewer 2 Stefan Gaillard

0) General comments

- The author reframes and critiques the original article (Leboeuf, Linden-Andersen, & Carriere, 2020) within the context of replication and priming studies in social psychology more broadly. Specifically, the author defines various kinds of possible scientific reproduction, concludes that the original experiment is (at best) a replication, and discusses the merits of such work given the historical context from which priming studies such as this one have emerged.

1) The Experiment

- The author does a good job at explaining the rational, results, and discussion of the original experiment.

2) The Intricacies of Producing and Reproducing Experiments

- This section does a good job at detailing (what I imagine is) a broad discussion in philosophy of science for the reader.

3) Is 'Alcohol Cues and Aggressive Thoughts' a Reproduction?

- A critique is given regarding the need to explicate research methodology in such a way to provide 'lay persons instruction' [...]. This, in many cases, may be lengthy and unnecessary, or even impossible if the techniques used in a given study require technical knowledge of the field (e.g., surgical techniques in animal research, computational techniques in learning models, etc.). While these points are not applicable to the article in question, perhaps the author might consider the role of openly accessible materials over (or in conjunction with) written-description alone in addressing their concern. This critique may furthermore dovetail with JOTE's public position regarding open science practices and be informative for their editorial during future publication. As is, this section reads as though social psychology is primarily investigated through person-to-person interaction, whereas a large proportion of studies are in fact computerised and materials can be directly shared.

4) Priming Studies

- This section provides a strong review of controversies in priming studies and how the emphasis on "closed system" experimental replication maximises potential experimenter bias. It also clearly relates this discussion to the article in question.

5) Experimenting with Humans

- This section does a good job of relating higher-order concepts about experiments in human subjects and the research in question.
- This section concludes with what is in essence a restating of the problem of induction ("but on the other hand it is not possible to predict the outcome with certainty"). To what degree can these problems be addressed with statistical reform? Can advances in statistical inference aid in increasing (or at least better describing) researchers' certainty in their outcomes? What is the trade-off between experimental design and statistical inference?

6) Relevance

- The author questions the practical value of priming research. While they would not be the first to critique social psychology's relevance in public discourse (Berkman & Wilson, 2020)—and indeed this is a critical point—a stronger line of argumentation should be provided to support this claim in the current critique. As is, the author states "[e]ven if the results were as expected, they would not help to tackle the alcohol-and-violation-problem. If

we would decide to ban all ‘alcohol cues’ from the public domain, alcohol use and abuse would continue”. While this is true, it is at least conceivable that other public policy could use priming research as a scientific basis (e.g., nudging practices across the world).

- I am tempted to agree with the author’s conclusions here, but I think that justifying the impracticality of the research merits as much attention as justifying its importance.

7) Conclusion

- “*Why attempt to publish a report like this in the first place? Publication might be instructive to psychology students how not to perform replications, but is that a sufficient legitimization*”
 - Given that this article may appear in the Journal of Trial and Error—a journal based largely on “report[s] like this”—this point may merit further discussion. I leave this to the discretion of the author.