Reproducibility Essay

Trent Henderson

2022-09-16

The work of Renner et al. (2018)[1] sought to investigate the effects of acute alcohol consumption on self-ratings and the ratings of observers on foreign language skills. This essay discusses aspects of the study that would make it more possible to conduct a replication study.

The methods section of the study failed to report the following, which would inhibit the conduction of a replication study:

- Poster materials for participant recruitment were not provided (p. 117) which may limit a replication study's ability to recruit a similar sample
- Instructions to observers who rat3e the audio recordings were not supplied (p. 118)
- The list of standardized questions for cases when participants needed less time (p. 118) were not supplied
- The list of thirteen arithmetic problems were not supplied (p. 118), meaning an exact replication cannot be conducted
- It is unclear if all participants waited together or if there was indeed some overlap at all prior to engagement with the experiment (p. 119) as the study just notes that testing took place between 1.00 pm and 4.00 pm
- The location of the study was described as a "laboratory visually resembling a pub" (p. 119) but no visual evidence was provided to enable an exact replication (i.e., photograph) — this is important as the overall setting of the experiment may itself influence the behaviour and perceptions of participants (and the experimenter)
- The exact instrumental music that participants listened to after consumption of the drink for fifteen minutes prior to the language task was not shared (p. 119), so an exact replication could not be performed
- The type of headphones used to listen to the instrumental music were also not noted (p. 119) — while not a large deal, there could be some influence on the study if noise cancelling headphones were used versus non-noise cancelling in a potential replication study
- Experimenter scripts were not provided, meaning the process of welcoming the participant, providing written and verbal information about the study, and the offering of the drink (p. 119) cannot be reproduced exactly — this primer of study information is important to be the same to ensure participant expectations are the same in a potential replication

In addition to the points above, there are other aspects of reproducibility worth discussing. Importantly, the article did not provide open access to the data nor code used for analysis. This limits the ability for other researchers to engage in computational reproducibility of the results and raises questions about the transparency of the study. However, despite this criticism, there were elements of reproducibility that this study handled well. The study blinded participants to the experimental condition (low dose of alcohol or control beverage with no alcohol) they were assigned to, and both the experimenter who conducted the study with them and the two native Dutch speakers who conducted observer ratings were also condition-blind. Blinding is critical to this study for several reasons: (i) blinding of participants mitigates the placebo effect, whereby un-blinded participants' expectations affect what they report; (ii) blinding of the experimenter who runs the experiment with participants is crucial because knowledge of the manipulation condition may

[1] Renner F, Kersbergen I, Field M, Werthmann J. Dutch courage? Effects of acute alcohol consumption on self-ratings and observer ratings of foreign language skills. J Psychopharmacol. 2018 Jan;32(1):116-122. doi: 10.1177/0269881117735687. Epub 2017 Oct 18. PMID: 29043911.

(consciously or unconsciously) make them react differently to participants based on condition (e.g., encourage participants that their Dutch skills are strong, resulting in a self-fulfilling prophecy), thus introducing a source of bias; and (iii) blinding of observers is important because knowledge of the manipulation condition may influence their perceptions and introduce bias in their scoring. The study also does not appear to have engaged in hypothesising-after-results-are-known (HARKing). The authors specify a discrete set of four hypotheses (p. 117), each grounded in prior research and sensible real-world beliefs (such as Hypothesis 1 which posited that those who consumed alcohol would rate their performance higher). Further, not all the results for the hypotheses returned statistically significant results and it is this lack of exceedingly favourable results which makes it unlikely that the authors engaged in HARKing or even *p*-hacking more broadly.