the WINNOWER

SOCIAL SCIENCES

The replication crisis, context sensitivity, and the Simpson's (Paradox)

PSYCH BRIEF

READ REVIEWS

WRITE A REVIEW

CORRESPONDENCE:
psychbrief@gmail.com

DATE RECEIVED:
July 13, 2016

DOI:
10.15200/winn.146839.98275

ARCHIVED:
July 13, 2016

KEYWORDS:
replication crisis, philosphy of science

CITATION:
Psych Brief, The replication crisis, context sensitivity, and the Simpson's (Paradox), *The Winnower* 3:e146839.98275 , 2016 , DOI: 10.15200/winn.146839.98275

The Reproducibility Project: Psychology (OSC, 2015) was a huge effort by many different psychologists across the world to try and assess whether the effects of a selection of papers could be replicated. This was in response to the growing concern about the (lack of) reproducibility of many psychological findings with some high profile failed replications being reported (Hagger & Chatzisarantis, 2016 for ego-depletion and Ranehill, Dreber, Johannesson, Leiberg, Sul, & Weber, 2015 for power-posing). They reported that of the 100 replication attempts, only ~35 were successful. This provoked a strong reaction not only in the psychological literature but also in the popular press, with many news outlets reporting on it.

But it wasn't without its critics: Gilbert, King, Pettigrew, & Wilson (2016) examined the RP:P's data using confidence intervals and came to a different conclusion. They were looking for "whether the point estimate of the replication effect size [fell] within the confidence interval of the original" (Srivastava, 2016). Some of the authors from the RP:P responded (Anderson et al., 2016) by pointing out some of the errors in the Gilbert et al. paper. Another analysis was provided by Sanjay Srivastava (2016) who highlighted that whilst Gilbert et al. use confidence intervals, they incorrectly define them (calling into question any conclusions they draw). Etz & Vandekerckhove (2016) reanalysed 72 of the original studies' data using Bayes' statistics and found 64% of those 72 studies (both originals and replications) did not provide strong evidence for either the null or the alternative hypothesis. Simonsohn (2016) argued that rather than ~65% of the replications failing, ~30% failed to replicate and ~30% of the replications were inconclusive.*

But there is one response and one explanation for the low replication rate I want to focus on: the context sensitivity of an experiment.

Location, location, location:

Context sensitivity is the idea that where you conduct an experiment has a large impact on it. It is a type of hidden moderator as it is a variable that affects the experiment that usually isn't being directly manipulated or controlled by the researcher. The environment in which you perform the study plays a role in the result and should be considered when conducting a replication. It is argued that you cannot detach the "experimental manipulations... from the cultural and historical contexts that define their meanings" (Touhey, 1981). The context of an experiment is very important in social psychology and it has been studied for years, with evidence that it does shape people's behaviour (for one of many examples, you can look at Fiske, Gilbert, Lindzey; 2010).

van Bavel, Mende-Siedlecki, Brady, & Reinero (2016) argue that context sensitivity partly explains the

poor replication rate of the RP:P. They found that the context sensitivity of a study (rated by 3 students with high inter-rater reliability) had a statistically significant negative correlation with the success of the replication attempt (r=–0.23, P = 0.02). This means the more contextually sensitive the finding was, the less likely it was to replicate. It was still significantly associated with replication success after controlling for the sample size of the original study (which has been suggested to have a significant impact on the success of a replication; Vankov, Bowers, Munafò, 2014). It was not the best predictor of a replication though: the statistical power of the replication and how surprising the replication was were the strongest predictors. They also analysed the data to see whether the discipline of psychology the original study was taken from (either social or cognitive psychology) moderated the relationship between contextual sensitivity and replication success. They did not find a significant interaction (this last point is very important but I'm going to examine it in more detail further on).

So this study appears to show that contextual differences had a significant impact on replication rates and that it should be taken into account when considering the results of the RP:P.

There's no such thing as...

One of the responses to the paper was by Berger (2016). He stated that "context sensitivity" is too vague a concept to be of any scientific use. There are an enormous number of ways that "context" could impact on a finding and to present it as a uni-dimensional construct (as was done in van Bavel et al, 2016) is illogical. Context sensitivity can therefore be used to justify any unexplained variance in psychological results. He calls for a more rigorous and falsifiable definition of context sensitivity (namely lack of theory specificity and heterogeneity) and for researchers to be specific when it comes to the source of the problems e.g. is it variation in the population, location, time-period, etc. He also argues that researchers should a prior predict the heterogeneity and effect directions so we can scientifically evaluate the effect of these hidden moderators.

The hidden variable:

Another problem with the paper was highlighted by Schimmack (2016) and Inbar (in press). When you run the analyses again and properly control for sub-discipline (rather than test for the interaction as was originally done), the significant result van Bavel et al. found disappears (from p=0.02 to p=0.51). They also calculated the correlation **within groups** (so the correlation between context sensitivity and replication success for cognitive psychology studies and for social psychology studies) and again found non-significant results (r = -.04, p = .79 and r = -.08, p = .54 respectively). This suggests context sensitivity only has a significant impact on replication rates when you don't control for sub-discipline (so some disciplines of psychology are more likely to replicate than others). van Bavel has replied to this by arguing you can't control for sub-discipline as it is "part of the construct of interest" (van Bavel, 2016).

Simpson's Paradox:

So how does Simpson's Paradox fit into all this? (Not those Simpsons unfortunately, Edward H. Simpson). Well, this is a perfect example of Simpson's paradox: where a trend is found when groups are combined but disappears or reverses when they are examined separately. The classic example comes from Bickel, Hammel, & O'Connell (1975). They examined the admission rates for graduate school at the University of California, Berkeley for 1973. They appeared to show a gender bias towards men as 44% were admitted whereas only 35% of women were.

But when you examine all of the departments individually they show that 6 of them admitted more women than men (and 4 admitted more men than women). When analysed, this preference for females was shown to be statistically significant. So how does this work? It's because of a third variable: rate of admission within the department. As stated in the article: "The proportion of women applicants tends to be high in departments that are hard to get into and low in those that are easy to get into."

Department	Men		Women	
	Applicants	Admitted	Applicants	Admitted
A	825	62%	108	**82%**
B	560	63%	25	**68%**
C	325	**37%**	593	34%
D	417	33%	375	**35%**
E	191	**28%**	393	24%
F	373	6%	341	**7%**

Table showing the 6 most applied to departments [Wikipedia]

This is exactly the same thing that happened in the van Bavel (2016) paper: the original significant finding (r=–0.23, P = 0.02) disappeared after you controlled the hidden variable of sub-discipline.

So what does all this mean?

The purpose of this post isn't to show that context sensitivity doesn't have an impact on the RP:P (it almost certainly did and it will have an impact on other research). But it does show that the van Bavel paper doesn't tell us how much of an impact this variable has on the RP:P and that we need to be more precise in our language. Unless we are explicit in what we mean by "context sensitivity" and predict what effect it will have before the experiment (and in which direction), it will remain post-hoc hand-waving which doesn't advance science.

Notes:

*This post is not meant to be an exhaustive list of all the various analyses of the responses to RP:P. I highly recommend you go and read as many of the commentaries as possible as they go into greater detail. The one's I list are a selection of them.

References:

Anderson, C.J.; Bahník, S.; Barnett-Cowan, M.; Bosco, F.A.; Chandler, J.; Chartier, C.R.; Cheung, F.; Christopherson, C.D.; Cordes, A.; Cremata, E.J.; Penna, N.D.; Estel, V.; Fedor, A.; Fitneva, S.A.; Frank, M.C.; Grange, J.A.; Hartshorne, J.K.; Hasselman, F.; Henninger, F.; Hulst, M. v.d.; Jonas, K.J. Lai, C.K.; Levitan, C.A.; Miller, J.K.; Moore, K.S.; Meixner, J.M.; Munafò, M.R.; Neijenhuijs, K.I.; Nilsonne, G.; Nosek, B.A.; Plessow, F.; Prenoveau, J.M.; Ricker, A.S.; Schmidt, K.; Spies, J.R.; Stieger, S.; Strohminger, N.; Sullivan, G.B.; van Aert, R.C.M.; van Assen, M.A.L.M.; Vanpaemel, W.; Vianello, M.; Voracek, M.; Zuni, K. Response to Comment on "Estimating the reproducibility of psychological science". *Science*, 351 (6277), 1037.

Berger, D. (2016). There's no Such Thing as "Context Sensitivity". [online] Available at: https://www.dropbox.com/home/Public?preview=Berger-no-context.pdf

Bickel, P.J.; Hammel, E.A.; & O'Connell, J.W. (1975). Sex bias in graduate admissions: Data from Berkeley. *Science* 187 (4175), 398-404.

Etz, A. & Vandekerckhove J. (2016) A Bayesian Perspective on the Reproducibility Project: Psychology. *PLoS ONE,* 11(2): e0149794. doi:10.1371/journal.pone.0149794

Fiske, S.T.; Gilbert, D.T.; & Lindzey, G. (2010). Handbook of Social Psychology (John Wiley & Sons, Hoboken, NJ).

Gilbert, D.T; King, G.; Pettigrew, S.; & Wilson, T.D. (2016). Comment on "Estimating the reproducibility of psychological science". *Science*, 351 (6277), 1037.

Hagger, M.S.; Chatzisarantis, N.L.D.; H.J.E.M., Alberts; & Zwienenberg, M. (2016). A multi-lab pre-registered replication of the ego-depletion effect. *Perspectives on Psychological Science.* [online] Available at: http://www.psychologicalscience.org/redesign/wp-content/uploads/2016/03/Sripada_Ego_RRR_Hagger_FINAL_MANUSCRIPT_Mar19_2016-002.pdf

Inbar, Yoel. (in press). The association between "contextual dependence" and replicability in psychology may be spurious. *Proceedings of the National Academy of Sciences of the United States of America.*

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, 349 (6251): aac4716. doi: 10.1126/science.aac4716. pmid:26315443

Ranehill, E.; Dreber, A.; Johannesson, M.; Leiberg, S.; Sul, S.; & Weber, R.A. (2015). Assessing the robustness of power posing: no effect on hormones and risk tolerance in a large sample of men and women. *Psychological Science,* 26 (5), 653-6.

Sample, I. (2015). Study delivers bleak verdict on validity of psychology experiment results. Guardian. [online] Available at: https://www.theguardian.com/science/2015/aug/27/study-delivers-bleak-verdict-on-validity-of-psychology-experiment-results

Simonsohn, U. (2016). *Evaluating Replications: 40% Full ≠ 60% Empty.* [online] Available at: http://datacolada.org/47

Schimmack, U. (2016). [online] Available at: https://www.facebook.com/photo.php?fbid=10153697776061687&set=gm.1022514931136210&type=3&theater

Tim. *Mere Anachrony: The Simpsons Season One* [online] Available at: https://npinopunintended.wordpress.com/2009/11/14/mere-anachrony-the-simpsons-season-one/

Touhey JC (1981) Replication failures in personality and social psychology negative findings or mistaken assumptions? Personality and Social Psychological Bulletin 7(4):593-595.

van Bavel, J. (2016). [online] Available at: https://twitter.com/jayvanbavel/status/737744646311399424

van Bavel, J.; Mende-Siedleckia, P.; Bradya, W.J.; & Reinero, D.A. (2016). Contextual sensitivity in scientific reproducibility. *Proceedings of the National Academy of Sciences of the United States of America*, 113 (23), 6454-6459.

Vankov, I.; Bowers, J.; & Munafò, M.R. (2014). On the persistence of low power in psychological science. *The Quarterly Journal of Experimental Psychology*, 67 (5), 1037-40.

Wikipedia. (2016). Simpson's paradox. Available at: https://en.wikipedia.org/wiki/Simpson%27s_paradox#cite_note-freedman-10 [Accessed: 12/07/2016]

Yong, E. (2015). How Reliable Are Psychology Studies? The Atlantic. [online] Available at: http://www.theatlantic.com/science/archive/2015/08/psychology-studies-reliability-reproducability-nosek/402466/