Supplementary material 1: Updating replication value once a replication is conducted

Peder M. Isager, Anna van 't Veer, Daniël Lakens

8/14/2021

Calculating replication value for a meta-analytic estimate

When replications of a replication target have already been performed, we will usually want to combine the information from these replications in our replication value estimate. Similarly, once we have replicated a chosen replication target, we may want to combine the information from the original study and our replication to consider if further replication is warranted, or if it would be better to focus new resources on a different replication target. A straight-forward way to calculate RV_{Cn} based on combined evidence from several studies would be to calculate the meta-analytic variance estimate for the studies.

For a fixed effects meta-analysis, RV_{Cn} can be estimated in the following way:

$$RV_{fixed} = \frac{w(C_S)}{Y+1} SE_M = \frac{w(C_S)}{Y+1}\sqrt{\frac{1}{\sum_{i=1}^{k} W_i}} = \frac{w(C_S)}{(Y+1)\sqrt{\frac{1}{\sum_{i=1}^{k} W_i}}} \tag{SM1-1}$$

where RV_{fixed} is the estimate of replication value, C is the citation count of the original article reporting on the target claim, Y is the number of years since the original article was published, SE_M is the standard error of the summary effect for the fixed effect meta-analysis (see Borenstein et al. 2009, equations 11.4 and 11.5), W is the inverse variance weight of each included study (see Borenstein et al. 2009, equation 11.2), and i denotes a particular study in the set k included in the RV_{fixed} estimate.

Equation SM1-1 can still be used for calculating RV_{fixed} whether or not we want to assume that the standard deviation is equal across all candidates and use only sample size to estimate the standard error for each study. When we make the assumption of equal standard deviations, the equation stays identical, but we must change the variance estimate provided to the inverse variance weight W (see Borenstein et al. 2009, equation 11.2) from $\frac{\sigma^2}{n}$ to $\frac{1}{n}$. The inverse variance weight then simply becomes the sample size, since $\frac{1}{Var} = \frac{1}{\frac{1}{n}} = n$.

In many situations, however, it would be more appropriate to calculate the variance for a random effects meta-analysis, because there is often true effect size heterogeneity which will influence the variance estimate (Borenstein et al. 2009, chap. 13)[1]. For a random effects meta-analysis, RV_{fixed} can be estimated in the following way:

$$RV_{fixed} = \frac{w(C_S)}{Y+1} SE_{M*} = \frac{w(C_S)}{Y+1}\sqrt{\frac{1}{\sum_{i=1}^{k} W_i*}} = \frac{w(C_S)}{(Y+1)\sqrt{\frac{1}{\sum_{i=1}^{k} W_i*}}} \tag{SM1-2}$$

[1] However, when we only allow close replications into the meta-analytic estimate, we only expect theoretically close effects to be included (LeBel et al. 2018), which should imply low effect size heterogeneity. This means that, in practice, the difference between RV estimates based on fixed-effects and random-effects models should be low whenever close replication results are combined.

where RV_{rand} is the estimate of replication value, C is the citation count of the original article reporting on the target claim, Y is the number of years since the original article was published, $SE_M{}^*$ is the standard error of the summary effect for the mixed effect meta-analysis (see Borenstein et al. 2009, equation 12.8 and 12.9), W is the inverse variance weight of each study including τ^2 (see Borenstein et al. 2009, 2013, equation 12.6, 12.7), and i is a given study in the set k included in the RV_{rand} estimate.

While the random effects model is theoretically straightforward to calculate for a set of studies, there are two practical obstacles to using random effects variance in the estimate of RV_{Cn}:

1. In addition to variance estimates, which can be derived using only the sample size, we need to determine the effect sizes of interest in order to calculate the between-study heterogeneity estimate τ^2 (see Borenstein et al. 2009, equation 12.2 and 12.3).
2. We need a sufficient sample of studies in order to reliably estimate τ^2 (Borenstein et al. 2009, 84).

In addition to the practical difficulties of deriving random effects precision estimates, it can also be difficult to determine which among a set of findings should be combined in a meta-analysis (Sharpe 1997, @Esteves2017). Because closely related findings are rarely linked to each other in meta-data, identifying such findings will currently require manual inspection by the replicating researcher. However, platforms like CurateScience could perhaps make automatic identification of replications possible in the future (LeBel et al. 2018).

One might reasonably ask whether the citation count of all replications should also be combined in the meta-analytic replication value estimate. On the one hand, more studies entail a larger literature, which in theory could increase the overall impact and visibility of the claims studies, and perhaps citation count would reflect such increases. However we regard it as likely that replication and original studies are usually cited together, or at least for similar reasons, which means that each replication's citation count provides highly overlapping information about the underlying value of the replication target. We therefore only include the citation count of the original study in the definitions of RV_{fixed}/RV_{rand}, though we recognize the appropriateness of this choice is a largely unresolved empirical question.

Example: Applying RV_{fixed} to studies on the Stroop effect

The R script containing the data material and exact calculations used to produce the numbers reported in this section can be found on OSF (https://osf.io/e35pu/).

Suppose we would like to calculate RV_{fixed} for the classic Stroop effect (Stroop 1935). The Stroop effect is an extremely impactful finding, and one of the most cited publications in psychology. On the other hand, the original results have been consistently replicated in many research efforts [e.g., MacLeod (1991); Ebersole et al. (2016); Verhaeghen and De Meersman (1998), not to mention psychology classrooms around the world. Considering whether to, at this point, commit further resources to replicating the Stroop effect, we need to consider our uncertainty about the Stroop effect given the total weight of evidence from both the original study as well as from replications.

As of 2021-08-14, the citation count of the original Stroop effect (Stroop 1935) was `citations` according to Crossref), and the age of the publication at that time was 83 years. There are three separate studies reported in Stroop (1935). Study 2 directly tests the well-known interference effect of word meaning on color naming that most later replications have been based on (MacLeod 1991; Ebersole et al. 2016).

Study 2 includes data from 100 participants, but we should adjust this sample size for the fact that Stroop (Stroop 1935, Study 2) is a within-subject design (see supplementary material 2). Unfortunately, like many repeated measures experiments, Stroop does not report the correlation between dependent measures, which is necessary to accurately calculate the standard error and effect size of a repeated measures experiment (Dunlap et al. 1996). However, we can estimate the within-subject correlation from data generated by a similar Stroop paradigm. For example, a close replication of the original Stroop paradigm was performed by Burns et al. (Burns et al. 2019). For the conditions relevant for the replication of Stroop (1935), Study 2, the within-subject correlation in this study is 0.932, 95%CI[0.901, 0.954]. With this correlation estimate we can convert the within-subject effect size to a corresponding between-subject effect size that would have the same amount of precision. The adjusted sample size is (100*2)/(1-0.932) = 2958.737 (see supplementary material 2, equation SM2-1). The replication value for Stroop (Stroop 1935, Study 2) thus becomes:

$$\frac{w(C_S)}{Y+1} \times \frac{1}{\sqrt{n}} = \frac{9423}{86+1} \times \frac{1}{\sqrt{2958.737}} = 1.991 \tag{SM1-3}$$

Suppose we would like to update this replication value estimate after replications of the Stroop effect are performed. A collection of close replications of the original Stroop paradigm can be found in Verhaeghen and De Meersman (1998). Study designs and sample characteristics (within the young group) were similar to Stroop (1935), Study 2 in all but two of the studies reported in this meta-analysis (in two cases, subjects were told to read the words, not name the colors; Dulaney and Rogers 1994; Park et al. 1996).

We can track change in replication value as replications accumulate by recalculating equation SM1-1 after every successive replication attempt, including in each calculation all replication studies published up until that point. Assuming equal standard deviations, the only parameter changing between successive replications is the sample size. Therefore, replication value will always decrease monotonically under these assumptions[2]. Figure SM1-1 displays the reduction in replication value with every replication reported in Verhaeghen and De Meersman (1998), in the order by which these replications were published.

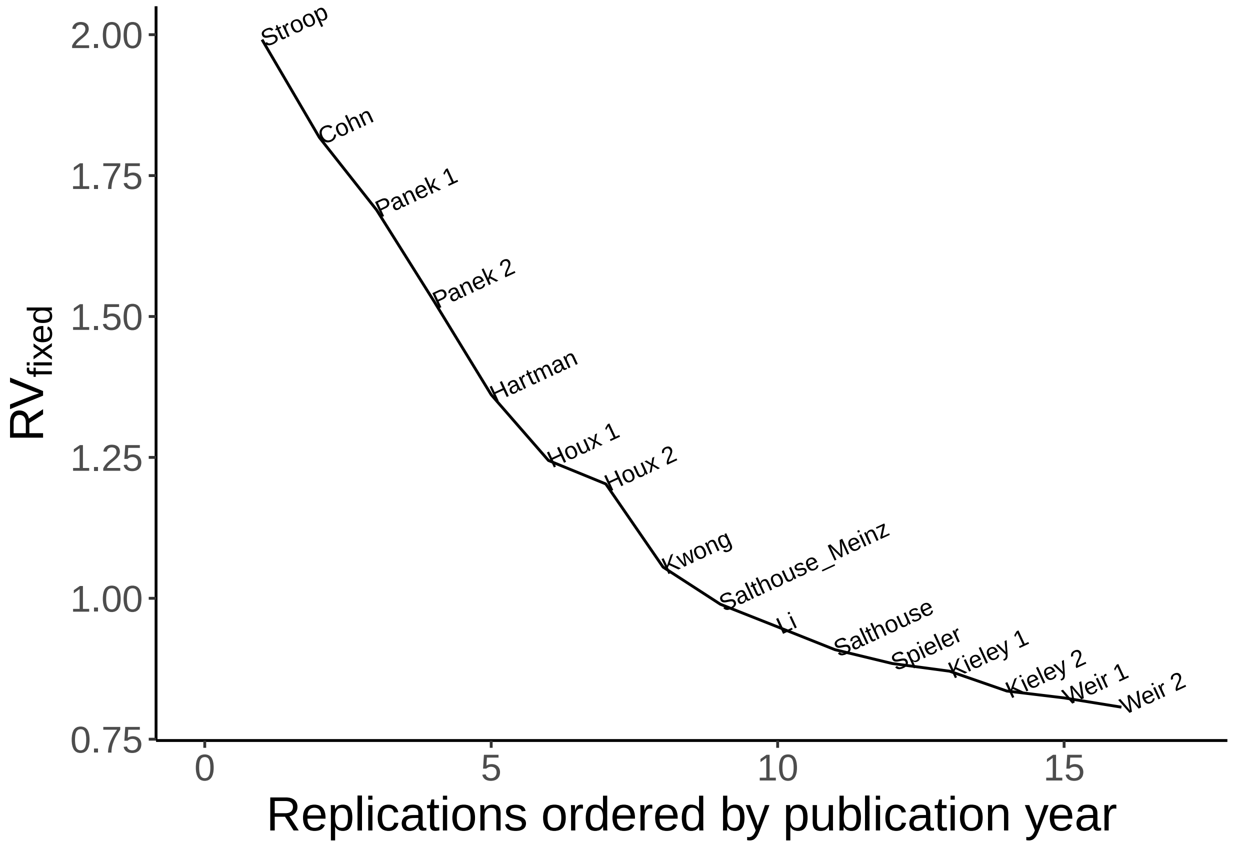

Figure SM1-1: Cumulative replication value of the Stroop effect over time, derived by recalculating equation SM1-1 after every successive replication attempt. Studies included are reported in Verhaeghen and De Meersman (1998), table 1, with the exception of Dulaney and Rogers (1994), and Park et al. (1996).

[2] Monotonic decrease may not hold under different assumptions. For example, if we instead use equation SM1-2 to update replication value, replication value could in theory increase after a replication if the effect size heterogeneity τ^2 increases substantially.

References

Borenstein, Michael, Larry V. Hedges, Julian P. T. Higgins, and Hannah R. Rothstein. 2009. *Introduction to Meta-Analysis*. Chichester, UK: John Wiley & Sons, Ltd. https://doi.org/10.1002/9780470743386.

Burns, Devin M., Elizabeth (Betsy) Fox, Michael Greenstein, and Demaris A. Montgomery. 2019. "An Old Task in New Clothes: A Preregistered Direct Replication Attempt of Enclothed Cognition Effects on Stroop Performance," March. https://doi.org/10.31234/osf.io/cj6kv.

Dulaney, C. L., and W. A. Rogers. 1994. "Mechanisms Underlying Reduction in Stroop Interference with Practice for Young and Old Adults." *Journal of Experimental Psychology. Learning, Memory, and Cognition* 20 (2): 470–84. https://doi.org/10.1037//0278-7393.20.2.470.

Dunlap, William P., Jose M. Cortina, Joel B. Vaslow, and Michael J. Burke. 1996. "Meta-Analysis of Experiments with Matched Groups or Repeated Measures Designs." *Psychological Methods* 1 (2): 170–77. https://doi.org/10.1037//1082-989X.1.2.170.

Ebersole, Charles R., Olivia E. Atherton, Aimee L. Belanger, Hayley M. Skulborstad, Jill M. Allen, Jonathan B. Banks, Erica Baranski, et al. 2016. "Many Labs 3: Evaluating Participant Pool Quality Across the Academic Semester via Replication." *Journal of Experimental Social Psychology* 67 (November): 68–82. https://doi.org/10.1016/j.jesp.2015.10.012.

Esteves, Sandro C., Ahmad Majzoub, and Ashok Agarwal. 2017. "The Problem of Mixing 'Apples and Oranges' in Meta-Analytic Studies." *Translational Andrology and Urology* 6 (S4): S412–S413. https://doi.org/10.21037/tau.2017.03.23.

LeBel, Etienne P., Randy J. McCarthy, Brian D. Earp, Malte Elson, and Wolf Vanpaemel. 2018. "A Unified Framework to Quantify the Credibility of Scientific Findings." *Advances in Methods and Practices in Psychological Science* 1 (3): 389–402. https://doi.org/10.1177/2515245918787489.

MacLeod, Colin M. 1991. "Half a Century of Research on the Stroop Effect: An Integrative Review." *Psychological Bulletin* 109 (2): 163–203. https://doi.org/10.1037/0033-2909.109.2.163.

Park, D. C., A. D. Smith, G. Lautenschlager, J. L. Earles, D. Frieske, M. Zwahr, and C. L. Gaines. 1996. "Mediators of Long-Term Memory Performance Across the Life Span." *Psychology and Aging* 11 (4): 621–37. https://doi.org/10.1037//0882-7974.11.4.621.

Sharpe, Donald. 1997. "Of Apples and Oranges, File Drawers and Garbage: Why Validity Issues in Meta-Analysis Will Not Go Away." *Clinical Psychology Review* 17 (8): 881–901. https://doi.org/10.1016/S0272-7358(97)00056-1.

Stroop, J. R. 1935. "Studies of Interference in Serial Verbal Reactions." *Journal of Experimental Psychology* 18 (6): 643–62. https://doi.org/10.1037/h0054651.

Verhaeghen, Paul, and Lieve De Meersman. 1998. "Aging and the Stroop Effect: A Meta-Analysis." *Psychology and Aging* 13 (1): 120–26. https://doi.org/10.1037/0882-7974.13.1.120.