Primer: Analysis of Replication Studies

Charlotte Micheloud, Leonhard Held
doi: 10.5281/zenodo.10137483

What is a replication study?

Replication studies are conducted to examine whether an original finding can be confirmed in an independent study where new data are collected. We focus here on direct replications, where the experimental procedures of the replication study must closely match those of the original study.
In recent years, large-scale replication projects have been conducted in various fields, such as the Reproducibility Project: Psychology (*RPP*, Open Science Collaboration, 2015), the Experimental Economics Replication Project (*EERP*, Camerer et al., 2016), the Social Sciences Replication Project (*SSRP*, Camerer et al., 2018) and the Reproducibility Project: Cancer Biology (*RPCB*, Errington et al., 2021), and several criteria have been used to assess the success or failure of the replication attempts. In this primer, we introduce the most frequently used criteria as presented in the introduction of Micheloud (2023), discuss their properties and point to alternative approaches in the discussion.

Effect size type

Depending on the outcome of interest, different effect size types are used to measure the impact of an intervention (such as a new treatment or therapy): standardized mean differences (SMD) or correlations (Pearson's r) for continuous outcomes, odds ratios (OR) for binary outcomes, and hazard ratios for survival outcomes. In both the meta-analysis and the replication settings, these effect sizes are frequently transformed to a common scale. Figure 1 shows standard conversions between effect sizes (Cooper et al., 2019, Chapters 11.3 and 11.6).

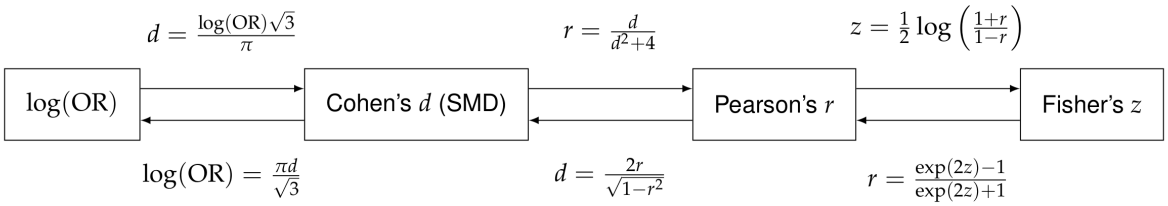

Figure 1: Conversion between commonly used effect size types (taken from Micheloud, 2023, Figure 1).

Effect sizes were transformed to the SMD scale in the *RPCB*, and Pearson's correlation coefficient r was chosen in the *RPP*, *EERP*, and *SSRP*. The attractiveness of using Pearson's correlation coefficient is that a normal distribution is approximately achieved after applying Fisher z-transformation, with standard errors being a function of the sample size only.

Replication success criteria

There is currently no universally agreed-upon standard for evaluating the success or failure of replication efforts. We will present the most commonly used criteria, and it is essential to differentiate between criteria for original 'positive' and 'null' results.
Original results which are selected for replication are usually 'positive', that is, they are significant or show at least a trend to significance at the standard two-sided significance level 0.05. The criteria for original positive results are summarized in Figure 2 and classified into four categories depending on whether they are based on p-*values*, *interval estimates*, *effect size* or *peer belief*. However, it might

also happen that original 'null' (that is, non-significant) results are chosen for replication. This was for example the case for 3 studies in the *RPP* and 15 numerical effects in the *RPCB*. Not all criteria presented in Figure 2 can be used in this situation and modified criteria are necessary. We do not discuss those in this primer, see Pawel et al. (2023) for a critical view.

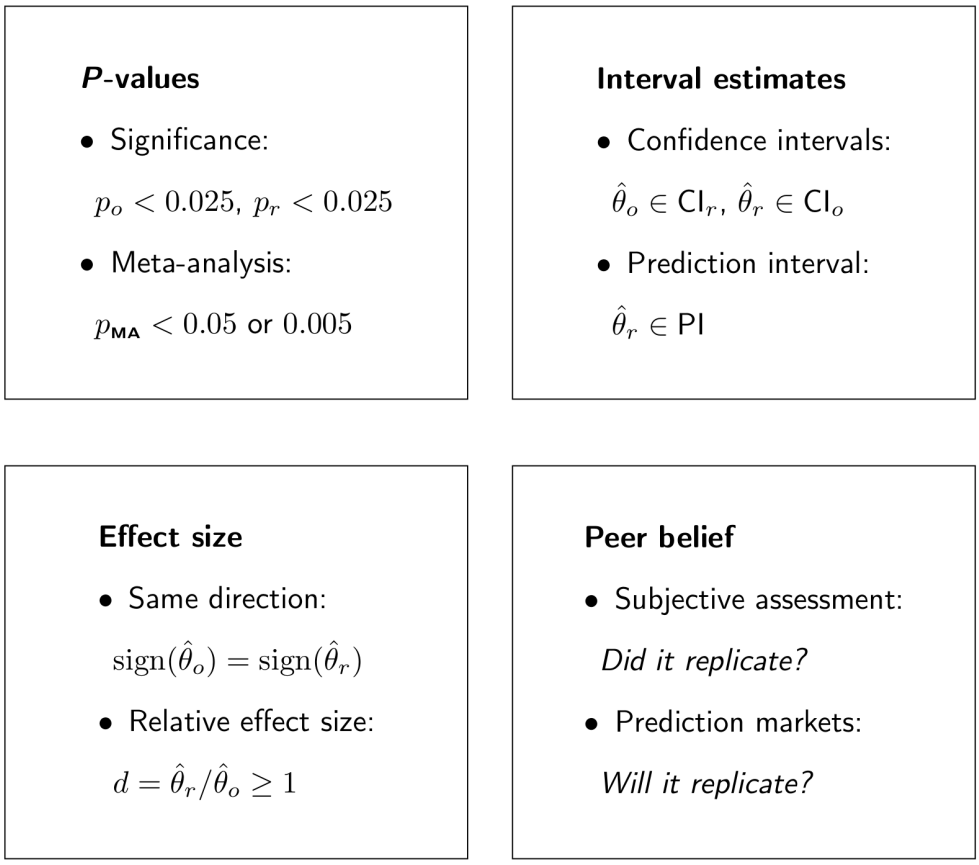

Figure 2: Common criteria to assess replication success. Adapted from Micheloud (2023, Figure 2).

Four study pairs are used to illustrate the properties of the different criteria: the original study by Dodson et al. (2008) and its replication by the *RPP*, the original study by de Clippel et al. (2014) and its replication by the *EERP*, the original study by Pyc and Rawson (2010) and its replication by the *SSRP* and the original study by DeNicola et al. (2011) and its replication by the *RPCB*, see Figure 3.

Criteria based on p-values

Significance Significance of both the original and the replication study, with an effect estimate in the same direction, is the most commonly used criterion. In practice, this means that both the original one-sided p-value p_o and the replication one-sided p-value p_r must be smaller than the one-sided significance level $\alpha = 0.025$, which is half of the standard two-sided 0.05 threshold. This criterion is analogous to the 'two-trials rule' in drug development (Senn, 2021).
Replication success with this criterion is achieved in the examples from de Clippel et al. (2014) and Pyc and Rawson (2010), despite the fact that the replication effect estimate $\hat{\theta}_r$ is a lot smaller than the original effect estimate $\hat{\theta}_o$ in the latter. This is a well-known weakness of p-values: they ignore the magnitude of the effect estimate (Sullivan and Feinn, 2012). Any positive effect, regardless of how small, can reach statistical significance if the sample size is large enough. Moreover, the example from DeNicola et al. (2011) illustrates well the importance of considering the effect direction: both studies have a significant effect estimate (as the confidence intervals do not overlap 0) but in opposite directions.

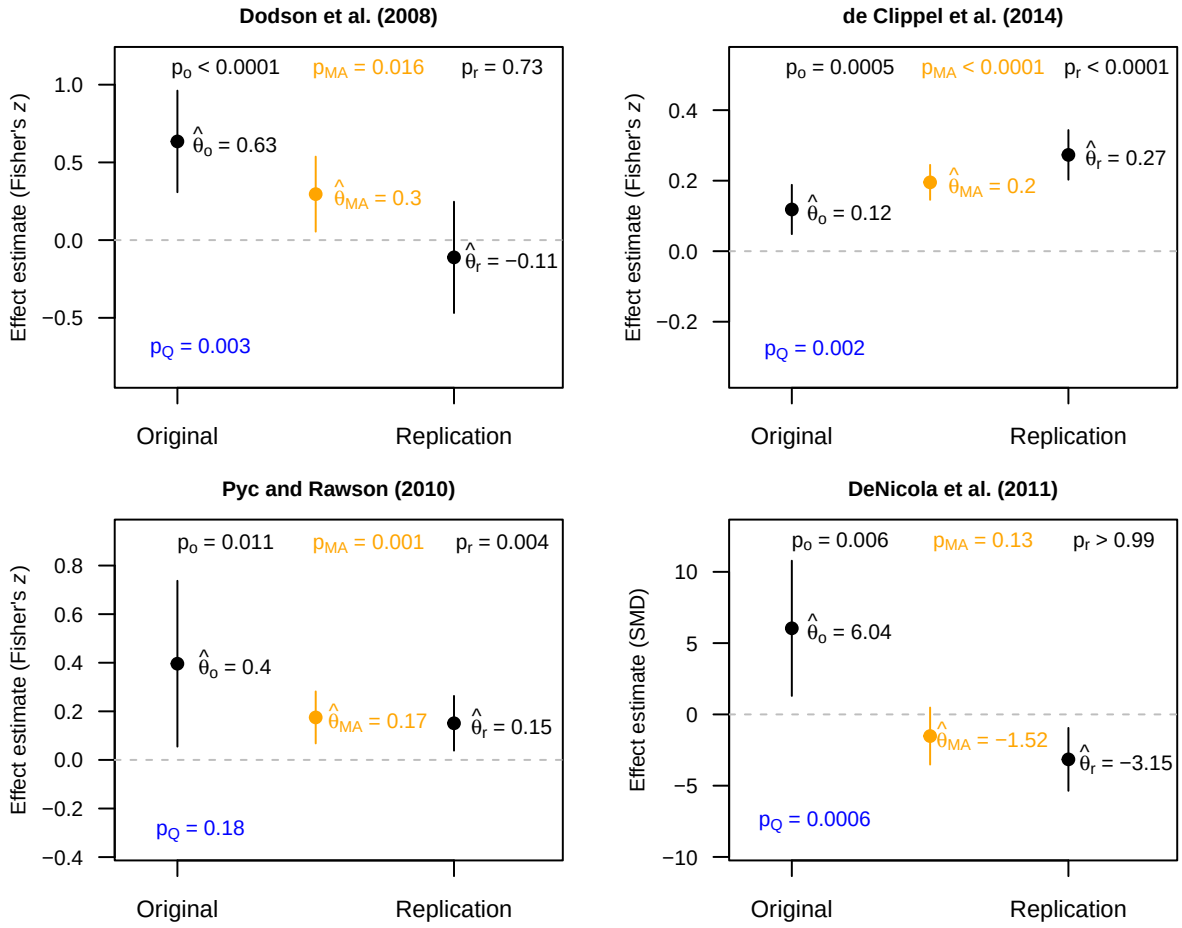

Figure 3: Original $\hat{\theta}_o$, replication $\hat{\theta}_r$ and meta-analytic $\hat{\theta}_{\text{MA}}$ effect estimates of the four examples with the corresponding 95%-CIs. The original (p_o) and replication (p_r) p-values are one-sided, while the meta-analysis (p_{MA}) p-value as well as p-value (p_Q) from the Q-test are two-sided. Adapted from Micheloud (2023, Figure 3).

Meta-analysis Another widely used approach is to pool the effect estimates from the original and replication studies into a meta-analytic combined effect estimate $\hat{\theta}_{\text{MA}}$, using a fixed-effect meta-analysis (Deeks et al., 2019). Replication success is declared if the associated meta-analytic p-value p_{MA} is significant at a certain (two-sided) threshold. There is not established convention regarding this threshold, both 0.05 and 0.005 are used in practice (Camerer et al., 2018; Errington et al., 2021). However, some authors recommend a two-sided significance level of $2 \times 0.025 \times 0.025 = 0.00125$ to ensure the same strength of evidence as the two-trials rule (Fisher, 1999; Shun et al., 2005). The Dodson et al. (2008) example highlights a drawback of the meta-analysis criterion: it does not take the direction of the effect estimates into account and can flag success (in this case, at level 0.05) even if the estimates go in opposite directions.

Criteria based on interval estimates

This category contains three different criteria based on either the replication 95% confidence interval (CI_r), the original 95% confidence interval (CI_o), or the 95% prediction interval (PI) for the replication effect estimate.

Original effect estimate in replication CI Replicability is declared with this criterion if the original effect estimate $\hat{\theta}_o$ is contained within the 95% CI_r of the replication effect estimate. Note that this

method ignores the uncertainty of the original effect estimate and is therefore miscalibrated. This means that even if the true underlying effect is the same in both the original and replication study, this criterion will not be fulfilled 95% of the time as it should.

Replication effect estimate in original CI This criterion is complementary to the previous one and requires the replication effect estimate $\hat{\theta}_r$ to be contained within the 95% CI_o of the original effect estimate. The uncertainty of the replication effect estimate is ignored in this method and so this criterion is also miscalibrated.

Replication effect estimate in PI A 95% PI contains the range of values a future observation will have with 95% probability. With this criterion, replicability is declared if the replication effect estimate $\hat{\theta}_r$ is within the 95% PI of the original effect estimate. This criterion is equivalent to $p_Q > 0.05$, where p_Q is the p-value from the Q-test for heterogeneity in meta-analysis (Higgins, 2003). If $p_Q \leq 0.05$, there is evidence for incompatibility of the effect estimates and replication success is therefore not achieved. This criterion is well calibrated as it takes into account the uncertainty of both the original and the replication effect estimates (Patil et al., 2016). This means that it will be fulfilled 95% of the time if the true underlying effect is the same in both studies.
The confidence and prediction interval criteria assess the incompatibility of the estimates and do not consider their significance. For example, in the de Clippel et al. (2014), both the significance and the meta-analytic criteria are fulfilled and the replication effect estimate is increased as compared to the original one. However, all three criteria based on interval estimate are not fulfilled as the estimates are in conflict. Furthermore, replication success is always declared with these three criteria if the two estimates are the same, even is they are null.

Criteria based on effect size

The two criteria in this category only consider the estimates and ignore their uncertainty. As a result, they are too simplistic to be used on their own and are rarely used as primary replicability indicator but rather in combination with other criteria. Moreover, both criteria do not consider significance of the estimates, so can both be fulfilled even in cases where the evidence for an effect is very low in both the original and the replication study.

Same direction The same direction criterion is fulfilled if both the original and replication effect estimates are in the same direction. However, if the true underlying effect in both the original and the replication study is null, one would still expect this criterion to be fulfilled in 50% of the cases. As mentioned by Errington et al. (2021), this is a weakness of the criterion which is considered a 'low-bar' for assessing replication success.

Relative effect size The relative effect size criterion is fulfilled if the replication effect estimate is at least as large as the original one, which is only the case in the de Clippel et al. (2014) example. One would expect this criterion to be fulfilled 50% of the time if the underlying effect is the same in both studies and is accurately estimated.

Criteria based on peer belief

Unlike the methods described above, the criteria in this category are not based on quantitative methods, but rather on peer belief.
Before the replication study is conducted, peers who are familiar with the subject can bet on whether or not the replication study will reach a significant result in the same direction as the original one. The probability of replication success produced by the **prediction market** is then interpreted as a replicability indicator (Dreber et al., 2015). Furthermore, the **subjective assessment** criterion is fulfilled

if, based on the available replication results, the replication team considers that the original finding was replicated.

Application

Table 1 shows the replication rates which were reported in the respective publications of each of the four projects, as well as the average shrinkage of the replication effect estimate as compared to the original one. The replication rates greatly differ depending on the replication success criterion. For example, 79% of study pairs in the *RPCB* fulfill the 'same direction' criterion, while only 3% achieve replication success with the 'relative effect size' criterion. The authors of the replication projects therefore recommend to assess the criteria collectively instead of interpreting them individually.
A common feature of these four replication projects is that the shrinkage of the replication effect estimate as compared to the original one is considerable. This effect estimate shrinkage has been attributed to low power, publication bias and other types of biases which inflate the original effect estimates (Ioannidis, 2008). In contrast, the replication studies are usually conducted with higher standards and therefore do not suffer from an effect estimate inflation.

	RPP	**EERP**	**SSRP**	**RPCB**
Significance	36% (35/97)	61% (11/18)	62% (13/21)	43% (42/97)
Meta-analysis ($p_{\text{MA}} < 0.05$)	68% (51/75)	78% (14/18)	76% (16/21)	62% (60/97)
$\hat{\theta}_o \in \text{CI}_r$	41% (30/73)	67% (12/18)	–	18% (17/97)
$\hat{\theta}_r \in \text{CI}_o$	–	–	–	43% (41/97)
$\hat{\theta}_r \in \text{PI}$	–	83% (15/18)	67% (14/21)	58% (56/97)
Same direction	–	–	–	79% (80/101)
Relative effect size	18% (18/99)*	–	–	3% (3/97)
Subjective assessment	39% (39/100)*	–	–	–
Prediction markets	–	75%	63%	–
Average shrinkage	50%	33%	50%	85%

Table 1: Replication rates in the four projects with different replication success criteria. The numbers in brackets indicate how many study pairs fulfill the criterion out of the total number of study pairs for which the criterion could be calculated. Results are shown at the effect level for the *RPCB*, see Errington et al. (2021, Table 1) for more details. *Replication rate reported for original positive and null results.

Discussion

The quantitative criteria presented in this primer are used as binary indicators of replication success. However, instead of being dichotomized, they could also be considered quantitatively. One could for example look at the distribution of the original and replication p-values, and at criteria based on interval estimates at levels different than 95%.
Furthermore, they all focus on either statistical significance or effect size, but none of them combines both aspects. Held (2020) recently proposed a new criterion which simultaneously takes into account effect size and significance: the sceptical p-value. This criterion penalizes shrinkage of the replication effect estimate as compared to the original one while ensuring that both studies are statistically significant to some extent. Moreover, the sceptical p-value can be interpreted as a quantitative measure of replication success, with smaller values indicating a higher replication success degree. Several extensions of this method have since been published (Held et al., 2022; Micheloud et al., 2023).
Another important aspect of replication studies, besides their analysis, is their design. The sample size of the replication study is usually calculated such that the power to reach statistical significance is 80% or 90%. It is however recommended to use the same method in sample size planning and analysis (Anderson and Kelley, 2022). The sceptical p-value mentioned in the previous section can be used

to design replication studies accordingly. A primer focusing on the design of replication studies will be published in the future.

Note

This primer is inspired by the introduction of the PhD thesis of the first author 'Advances in Statistical Methods for the Design and Analysis of Replication Studies' (Micheloud, 2023).

References

Anderson, S. F. and Kelley, K. (2022). Sample size planning for replication studies: The devil is in the design. *Psychological Methods*. doi:10.1037/met0000520.

Camerer, C. F., Dreber, A., Forsell, E., et al. (2016). Evaluating replicability of laboratory experiments in economics. *Science*, 351(6280):1433–1436. doi:10.1126/science.aaf0918.

Camerer, C. F., Dreber, A., Holzmeister, F., et al. (2018). Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behaviour*, 2(9):637–644. doi:10.1038/s41562-018-0399-z.

Cooper, H., Hedges, L. V., and Valentine, J. C. (2019). *The Handbook of Research Synthesis and Meta-Analysis*. Russell Sage Foundation. doi:10.7758/9781610448864.

de Clippel, G., Eliaz, K., and Knight, B. (2014). On the selection of arbitrators. *American Economic Review*, 104(11):3434–58. doi:10.1257/aer.104.11.3434.

Deeks, J. J., Higgins, J. P., and Altman, D. G. (2019). Analysing data and undertaking meta-analyses. In *Cochrane Handbook for Systematic Reviews of Interventions*, chapter 10, pages 241–284. John Wiley & Sons, Ltd, Chichester.

DeNicola, G. M., Karreth, F. A., Humpton, T. J., et al. (2011). Oncogene-induced Nrf2 transcription promotes ROS detoxification and tumorigenesis. *Nature*, 475(7354):106–109. doi:10.1038/nature10189.

Dodson, C. S., Darragh, J., and Williams, A. (2008). Stereotypes and retrieval-provoked illusory source recollections. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 34(3):460–477. doi:10.1037/0278-7393.34.3.460.

Dreber, A., Pfeiffer, T., Almenberg, J., Isaksson, S., Wilson, B., Chen, Y., Nosek, B. A., and Johannesson, M. (2015). Using prediction markets to estimate the reproducibility of scientific research. *Proceedings of the National Academy of Sciences*, 112(50):15343–15347. doi:10.1073/pnas.1516179112.

Errington, T. M., Mathur, M., Soderberg, C. K., Denis, A., Perfito, N., Iorns, E., and Nosek, B. A. (2021). Investigating the replicability of preclinical cancer biology. *eLife*, 10:e71601. doi:10.7554/eLife.71601.

Fisher, L. D. (1999). One large, well-designed, multicenter study as an alternative to the usual fda paradigm. *Drug Information Journal*, 33(1):265–271. doi:10.1177/009286159903300130.

Held, L. (2020). A new standard for the analysis and design of replication studies (with discussion). *Journal of the Royal Statistical Society, Series A*, 183:431–469. doi:10.1111/rssa.12493.

Held, L., Micheloud, C., and Pawel, S. (2022). The assessment of replication success based on relative effect size. *The Annals of Applied Statistics*, 16:706–720. doi:10.1214/21-AOAS1502.

Higgins, J. P. T. (2003). Measuring inconsistency in meta-analyses. *BMJ*, 327(7414):557–560. doi:10.1136/bmj.327.7414.557.

Ioannidis, J. P. A. (2008). Why most discovered true associations are inflated. *Epidemiology*, 19(5):640–648. doi:10.1097/ede.0b013e31818131e7.

Micheloud, C. (2023). *Advances in Statistical Methods for the Design and Analysis of Replication Studies*. PhD thesis, University of Zurich.

Micheloud, C., Balabdaoui, F., and Held, L. (2023). Assessing replicability with the sceptical p-value: Type-I error control and sample size planning. *Statistica Neerlandica*, 77:573–591. https://doi.org/10.1111/stan.12312.

Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science*, 349(517):aac4716. doi:10.1126/science.aac4716.

Patil, P., Peng, R. D., and Leek, J. T. (2016). What should researchers expect when they replicate studies? a statistical view of replicability in psychological science. *Perspectives on Psychological Science*, 11(4):539–544. doi:10.1177/1745691616646366.

Pawel, S., Heyard, R., Micheloud, C., and Held, L. (2023). Replication of "null results" – absence of evidence or evidence of absence? *eLife*. doi:10.48550/arXiv.2305.04587. to appear.

Pyc, M. A. and Rawson, K. A. (2010). Why testing improves memory: Mediator effectiveness hypothesis. *Science*, 330(6002):335–335. URL https://doi.org/10.1126/science.1191465.

Senn, S. (2021). *Statistical Issues in Drug Development*. Wiley. doi:10.1002/9781119238614.

Shun, Z., Chi, E., Durrleman, S., and Fisher, L. (2005). Statistical consideration of the strategy for demonstrating clinical evidence of effectiveness—one larger vs two smaller pivotal studies. *Statistics in Medicine*, 24(11):1619–1637. doi:10.1002/sim.2015.

Sullivan, G. M. and Feinn, R. (2012). Using effect size – or why the p-value is not enough. *Journal of Graduate Medical Education*, 4(3):279–282. doi:10.4300/jgme-d-12-00156.1.

Version from November 15, 2023.

Edited by Charlotte Micheloud (0000-0002-4995-4505) and Leonhard Held (0000-0002-8686-5325).