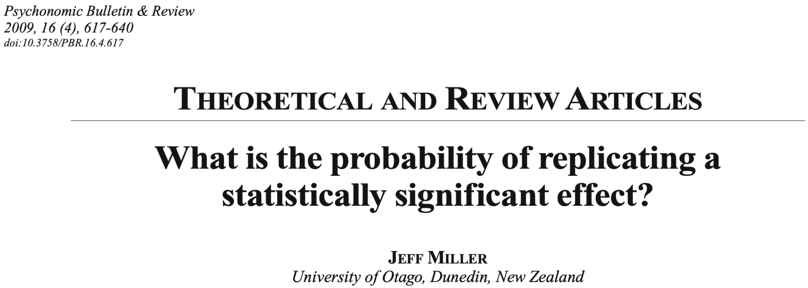
Replicability in Science:
Day 1, Paper 2:
Probability of Replication

giovanni_parmigiani@dfci.harvard.edu

Padova, July 8, 2024



Psychonomic Bulletin & Review
2009, 16 (4), 617-640
doi:10.3758/PBR.16.4.617

THEORETICAL AND REVIEW ARTICLES

What is the probability of replicating a statistically significant effect?

JEFF MILLER
University of Otago, Dunedin, New Zealand

glossary

nas 2019 report

National Academies of Science of the U.S.A. (NAS)

nas replicability

We define ... replicability to mean

obtaining **consistent** results
across **studies**
aimed at **answering** the same scientific question,
each of which has obtained its own **data**.

nas reproducibility

NAS: We define **reproducibility** to mean ... obtaining consistent computational results using the same input data, computational steps, methods, and code, and conditions of analysis

nas generalizability

NAS: A third concept, **generalizability**, refers to the extent that results of a study apply in other contexts or populations that differ from the original one.

repeatability

A repeatable prediction approach produces predictions without variation across independent tests carried out by repeating the entire process, including data collection, on the same individual or sampling unit.

Miller’s definitions

aggregate versus individual

the **aggregate** *replication probability is the probability that researchers who obtain significant results in their initial experiments will also obtain significant effects in identical follow-up experiments.*

Probability is defined as the frequency *across a large pool of researchers working within a common experimental or theoretical context but testing different null hypotheses*

the **individual** *replication probability, is the long-run proportion of significant results that would be obtained by a particular researcher in exact replications of that researcher's own initial study*

miller versus nas

— **Likely:** probability

— **Result:** positive significance test

— **Same:** significant both times in the same direction

topics for breakout discussion

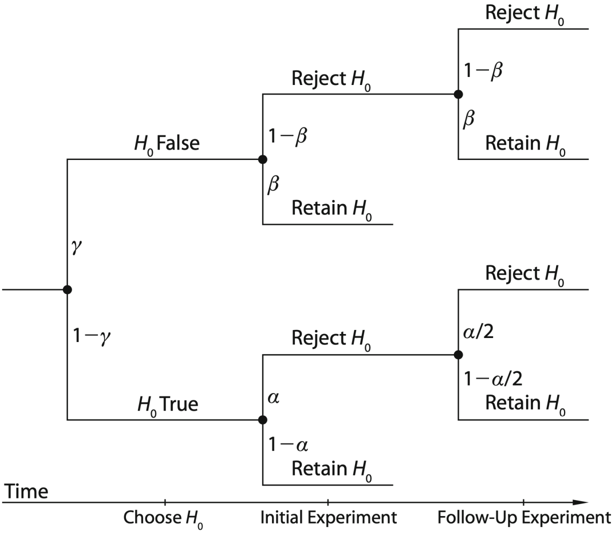
Miller's reasons for assessing replicability:

— First, this probability is relevant in assessing the implications of discrepant results (Is this a real effect that by chance was not replicated, or was the initial finding spurious?).

— Second, it is also relevant when researchers want to show that an effect obtained in one circumstance disappears in some other situation (e.g., a control experiment); the absence of the effect in the new situation is only diagnostic if the experiment had a high probability of replicating a true effect.

— Third, replication probability is relevant when planning a series of experiments (What are the chances that I will obtain this effect again in future experiments like this one?).

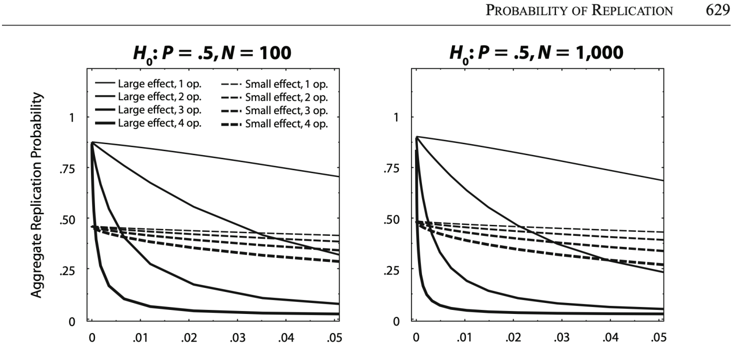
replicating statistical significance



probability of replication

$$
\begin{aligned}
p_{\mathrm{ra}} &= \Pr(S_2 | S_1) \\
&= \Pr(S_2 \cap S_1) \,/\, \Pr(S_1) \\
&= \frac{\Pr(H_1) \cdot \Pr(S_2 \cap S_1 | H_1) + \Pr(H_0) \cdot \Pr(S_2 \cap S_1 | H_0)}{\Pr(H_1) \cdot \Pr(S_1 | H_1) + \Pr(H_0) \cdot \Pr(S_1 | H_0)} \\
&= \frac{\gamma \cdot (1 - \beta)^2 + (1 - \gamma) \cdot \alpha \cdot \alpha/2}{\gamma \cdot (1 - \beta) + (1 - \gamma) \cdot \alpha}. \qquad (1)
\end{aligned}
$$

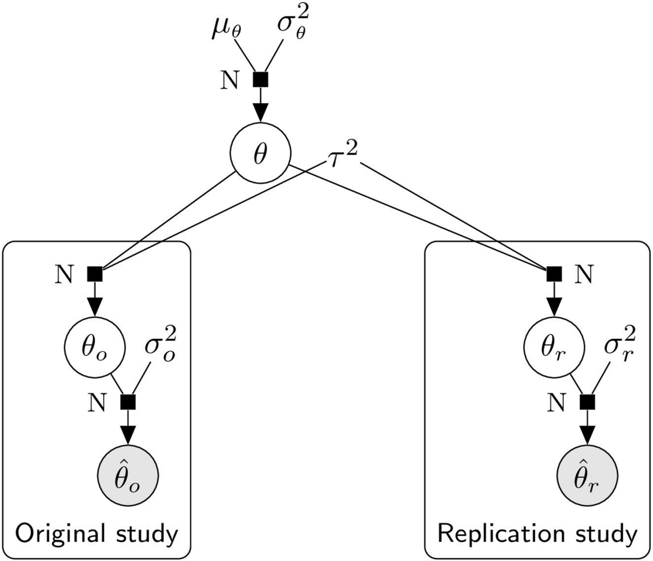
p-values and probability of replication



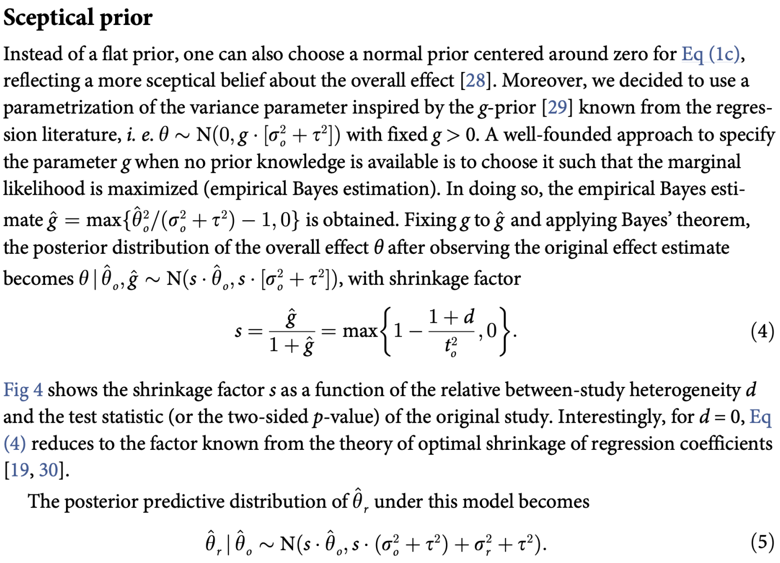
PROBABILITY OF REPLICATION 629

Aggregate replication probability as a function of the p value of the initial experiment, the number of opportunities for significant results (op.), the size of the true effect when it is present, and the sample size of the experiment. In all cases, real effects were assumed to be present for 50% of the null hypotheses tested. Solid lines represent theories for which the real effects are larger, whereas dashed lines represent theories for which these effects are smaller. Probability of rejecting the null hypothesis P=.5 using a binomial test with the indicated sample size of N=100 or 1,000.

probability of replication: Pawel & Held model



posterior predictive distribution



Sceptical prior

Instead of a flat prior, one can also choose a normal prior centered around zero for Eq (1c), reflecting a more sceptical belief about the overall effect [28]. Moreover, we decided to use a parametrization of the variance parameter inspired by the *g*-prior [29] known from the regression literature, *i. e.* $\theta \sim \mathrm{N}(0, g \cdot [\sigma_o^2 + \tau^2])$ with fixed $g > 0$. A well-founded approach to specify the parameter *g* when no prior knowledge is available is to choose it such that the marginal likelihood is maximized (empirical Bayes estimation). In doing so, the empirical Bayes estimate $\hat{g} = \max\{\hat{\theta}_o^2/(\sigma_o^2 + \tau^2) - 1, 0\}$ is obtained. Fixing *g* to $\hat{g}$ and applying Bayes' theorem, the posterior distribution of the overall effect θ after observing the original effect estimate becomes $\theta \,|\, \hat{\theta}_o, \hat{g} \sim \mathrm{N}(s \cdot \hat{\theta}_o, s \cdot [\sigma_o^2 + \tau^2])$, with shrinkage factor

$$s = \frac{\hat{g}}{1 + \hat{g}} = \max\left\{1 - \frac{1 + d}{t_o^2}, 0\right\}. \tag{4}$$

Fig 4 shows the shrinkage factor *s* as a function of the relative between-study heterogeneity *d* and the test statistic (or the two-sided *p*-value) of the original study. Interestingly, for $d = 0$, Eq (4) reduces to the factor known from the theory of optimal shrinkage of regression coefficients [19, 30].

The posterior predictive distribution of $\hat{\theta}_r$ under this model becomes

$$\hat{\theta}_r \,|\, \hat{\theta}_o \sim \mathrm{N}(s \cdot \hat{\theta}_o, s \cdot (\sigma_o^2 + \tau^2) + \sigma_r^2 + \tau^2). \tag{5}$$

shrinkage

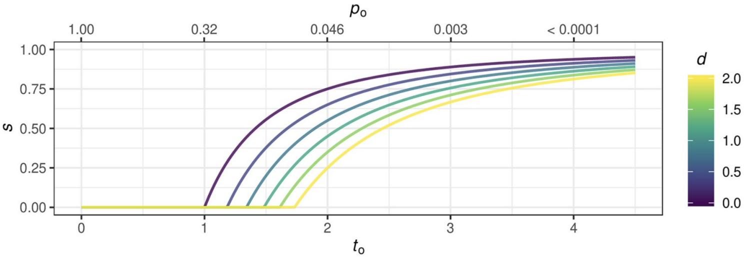


Fig 4. Evidence-based shrinkage. Shrinkage factor s as function of the test statistic t_o (bottom axis) and the two-sided p-value p_o (top axis) of the original study and the relative between-study heterogeneity $d = \tau^2/\sigma_o^2$.

shrinkage

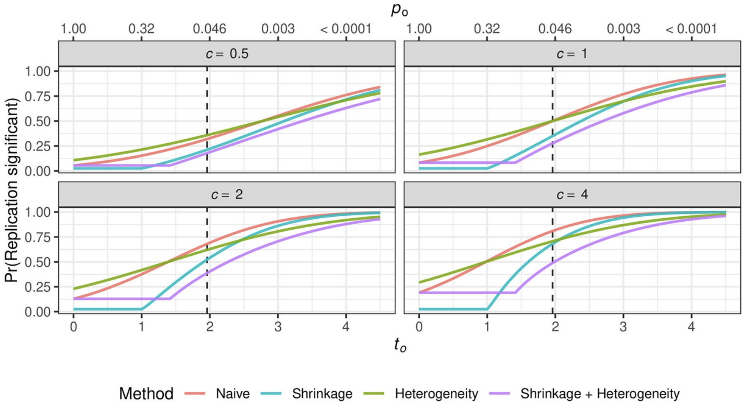


Fig 5. Replication probability. Probability of a significant replication outcome in the same direction as in the original study at (two-sided) $\alpha = 0.05$ level as a function of the test statistic t_o (bottom axis) and p-value p_o (top axis) of the original study and variance ratio $c = \sigma_o^2/\sigma_r^2$. The dashed line indicates $z_{0.025} \approx 1.96$. In the case of heterogeneity, $d = \tau^2/\sigma_o^2$ is set to one, otherwise to zero.

https://doi.org/10.1371/journal.pone.0231416.g005

topic for breakout discussion

Miller’s skepticism:
“answer is, in practice, virtually unknowable under either interpretation”
(i.e. the aggregate and individual interpretation)

topic for breakout discussion

Would it make more sense to consider replicability of *both* positive and negative results?

Think about this question in the context of testing a point null (as does Miller) and then more generally.

topics for breakout discussion

some nuggets from the NAS report

- "the assessment of replicability may not result in a binary pass/fail answer"
- "it is restrictive and unreliable to accept replication only when when the p-values in both studies have exceeded a selected threshold"
- "replicability of individual studies is an inefficient way to assure the reliability of scientific knowledge. Rather, reviews of cumulative evidence on a subject, to assess both the overall effect size and generalizability, is often a more useful"
- "Non-replicability occurs for a number of reasons that do not necessarily reflect that something is wrong. Some occurrences of non-replicability may be helpful to science, e.g. discovering previously unknown effects or sources of variability"

meta-analysis and replicability

Ganzfried etal 2007

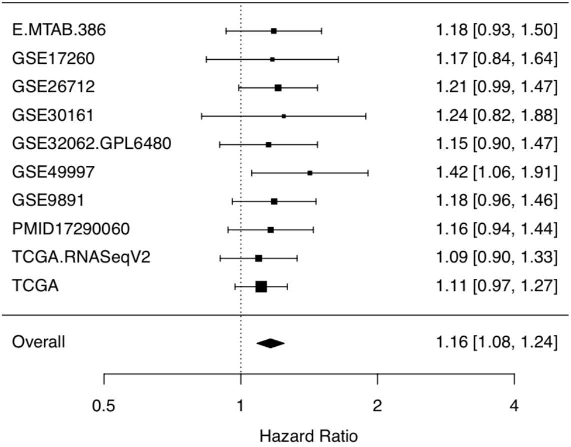
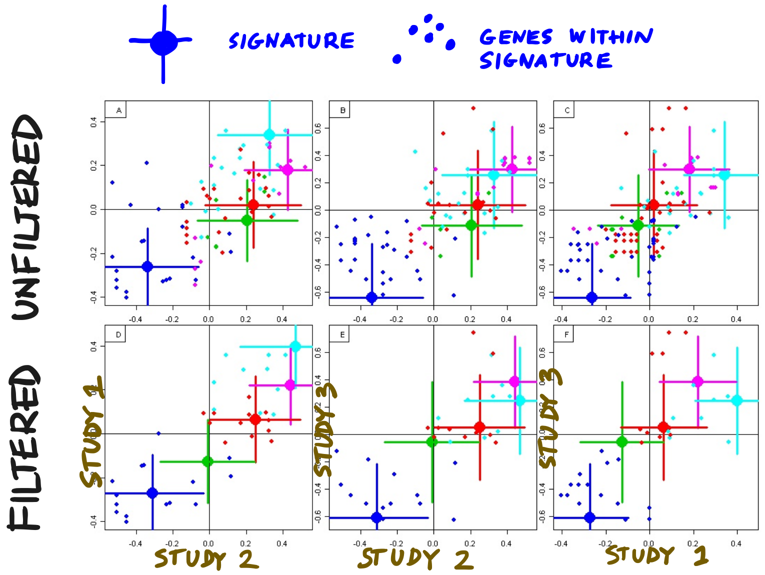


Figure 2: **Validation of CXCL12 as an independent predictor of survival**
This figure shows a forest plot as in Figure 1, but the CXCL12 expression levels were adjusted for debulking status (optimal versus suboptimal) and tumor stage. The p-value for the overall HR, found in res$pval, is 1.8e-05.

aggregate replicability

Zhong etal 2007



SCALE : PREDICTION STRENGTH (COX EFFECT SIZE)

urls

— Miller

— Pawel and Held

— NAS Report

— Zhong

— Ganzfried

"OF COURSE YOU CAN'T REPLICATE MY
EXPERIMENTS. THAT'S THE BEAUTY OF THEM."