Replicability: Terminology, Measuring Success, and Strategy

Werner A. Stahel[1*]

1 Seminar for Statistics, ETH, Zurich, Switzerland * stahel@stat.math.ethz.ch

Abstract

Empirical science needs to be based on facts and claims that can be reproduced. This calls for replicating the studies that pronounce the claims, but practice in many fields does not implement this idea. When such studies emerged in the past decade, the results were generally disappointing. There have since been an overwhelming number of papers addressing the "reproducibility crisis" in the last 17 years. Nevertheless, terminology is not yet settled, and there is no consensus about when a replication should be called successful. This paper intends to clarify such issues.

A fundamental problem in empirical science is that usual claims only state that effects are non-zero, and such statements are scientifically void. An effect must have a *relevant* size to become a reasonable item of knowledge. Therefore, estimation of an effect, with an indication of precision, forms a substantial scientific problem, whereas testing it against zero does not. A relevant effect is one that is shown to exceed a relevance threshold. This paradigm has implications for the judgement on replication success.

A further issue is the unavoidable variability between studies, called heterogeneity in meta-analysis. Therefore, it is of little value, again, to test for zero difference between an original effect and its replication, but exceedance of a corresponding relevance threshold should be tested. In order to estimate the degree of this variability, more than one replication is needed, and an appropriate indication of the precision of an estimated effect requires such an estimate.

These insights show the complexity of obtaining solid scientific results, implying the need for a strategy to make replication happen.

1 Introduction

Science is supposed to be based on facts that are reproducible: If the same phenomenon is studied again with the same methods, the results should be the same. Even though such validation is considered essential for the establishment of sound knowledge, it is not commonly performed in many fields of science. The term *reproducibility* in a wide sense encompasses all kinds and steps of validation (Section 2). In empirical science, claims are based on data subject to random variation. Performing a whole study again is called a *replication*, and such projects have been rare in most fields. (For psychology, see [1].)

The literature on replicability focusses on studies assessing a quantitative or qualitative "effect," and this will also be the topic here.

1.1 The replication crisis

In the last seventeen years, there have been ample studies that found replications to fail in too many instances, and this experience has lead to such a vast literature that we limit ourselves to citing the book by the National Academies of Sciences, Engineering and Medicine [2] for a general introduction.

A wealth of papers have argued and speculated about the reasons for the widespread failure of replicability. Considering empirical research, this author is convinced that the dominant cause in many fields is the so-called *selective reporting bias* (or, more simply, selection bias, [3, p.1328]) at various levels, such as:

- Data dredging, P hacking and HARK-ing, that is, Hypothesizing After the Results are Known (Acad. Med. Sci., 2015, and others): searching for patterns without a pre-conceived question or hypothesis, picking the most salient ones, and then formulating a statistical hypothesis for the most salient one(s), followed by formal statistical testing;
- *garden of forking paths* [5] or *researcher degrees of freedom* [6]: selecting, among different possibilities of analyzing the data, the one that gives the largest effect;
- *reporting bias* [7], including *confirmation bias* [8]: focussing on the effects that appear most significant and plausible; and
- *publication bias*: journals and their readers are interested in significant effects and therefore, these get published and achieve attention, even if they were the result of chance and correspond to the testing error of the first kind.

They all reflect a way of selecting, among several potential results, those that are statistically most signifikant, and therefore, effects that are estimated, by chance, stronger than their true values are, will have a larger likelihood to be documented in publications. Thus, effects in publications of "original studies" tend to be larger (in absolute value) than the potential true effects they claim to estimate, and consequently, their nominal statistical significance is (much) more pronounced than it should be. Clearly, this bias is enhanced for studies with low statistical power [9].

The high need for new topics for publications, including Ph.D. projects, has lead to an inflation of studies that do not examine clearly relevant and plausible scientific questions. The tactic is to collect some data and search for patterns in many ways in order to find any that appear statistically significant. While such exploratory analyses may play a positive role in the advancement of science and lead to surprising discoveries, the effects found in this manner suffer from selective reporting bias and therefore need independent confirmation.

Comments on the crisis of reproducibility often complain about mediocre quality, lack of education leading to inadequate experimental and statistical procedures, and low statistical power, see [10] for a review. These aspects may lead to even more blind exploration entailing higher selective reporting biases, but are otherwise not related to replicability.

In the famous paper on "Estimating the reproducibility of psychological science" [11], abbreviated as OSC15 in the sequel, 223 scientists were involved in replicating 100 statements about an effect of some sort that had been published in high ranking psychological journals. Among the 97 effects that had been statistically significant in the original study, 35 reached significance in the same direction in the replication. This result was received as a shock, documenting the crisis in empirical science. [12] (and papers cited there) put the results into perspective. In the meantime, there have been several similar attempts to obtain rates of successful replication, see [13], [14], [15], [16], [17], [18].

We do not discuss the important aspects of Good Practice in empirical research in general here, see [19] and their summary (their Table 1). For broad discussions about reproducibility, mostly in the sense of successful replication,
see [19], [20], [21], [22], [23], [3], [24], as well as the "consensus report" of the U.S. National Academies of Sciences, Engineering, and Medicine [2] with its broad view on Best Practices in the scientific process.

1.2 The generic case

A generic problem consists of assessing the difference between two groups of values of a continuous target variable, the groups referring to different situations or treatments. The difference is interpreted as the *effect* of the situation or treatment. This problem will be used as the basic example when introducing concepts and criteria below.

There are, of course, plenty of other ways in which data can lead to or confirm knowledge. The majority of well-posed scientific problems leads to measuring or observing a target variable (or sometimes more than one) in different situations determined by "explanatory" or "input" variables. The interest is again in *effects* of the latter variables on the target. Concepts and criteria should easily generalize to such situations.

While estimation of one or several effects is the central paradigm for which reproducibility is widely discussed, other types of statistical problems, like prediction, model development, or search for potential causal relationships in big data, are neglected, see [25].

1.3 Focus of this paper

A first topic of this paper is based on the observation that the discussions have suffered from a lack of common language. Even the term "reproducibility" has been used to name specific aspects of validation. And what does it mean to say "This claim has not been reproduced"? Was the analysis of the data misguided in the first place? Has replication of the study not been possible or not been tried? Have the results of a replication shown ambiguous or even contradictory results? Here, we try to establish clear terminology, building on the different earlier proposals from different fields of research and models of the scientific process.

A second point is the assessment of success of a replication study. Clearly, when a *replication study* is undertaken to validate a claim found in an empirical study, one must expect the new data to lead to non-identical results due to random variability. When is such a replication successful? In the literature, a popular but criticized criterion consists of finding a statistically significant effect again. Otherwise, the answer is often left to vague formulations such as "consistent effect sizes," "consistent measures of statistical significance," or "the results [should be] within the range of values predicted by estimates from the original study" (see, e.g, [26]). Anderson and Maxwell [27] distinguish between different goals of replication and define the respective criteria for success. We will discuss and introduce precise notions and a classification of results that goes beyond a simple binary answer in Section 4. The lack of criteria has lead to disputes about the interpretation of replication studies, diagnosed as a "war" by Ioannidis [28].

A fundamental issue is the perversion of scientific reasoning mentioned above, consisting of a search for patterns in data and subsequently treating the most salient ones by the statistical inference tools that are adequate for testng pre-conceived hypotheses. The road that leads to scientific knowledge starts from a clear scientific question or hypothesis and then chooses the experiment or observation study and the statistical tools to find the answer. Now, a suitable question asks if there is a certain effect of interest. We argue that since there is almost always at least a tiny effect, the question needs to be enhanced by specifying a threshold of relevance. This leads to shifting the focus away from testing to estimation (Section 3). This issue concerns empirical studies in general, but also affects the interpretation of replication results specifically.

Another fundamental issue in empirical science emerges from the experience that there is generally a variability of results that goes beyond the exptected statistical variation stemming from the randomness of the single observations. No replication study exactly mimiks the original, and assuming that the new observations are independent realizations from precisely the same distribution as the original ones is an over-simplification. A reasonable model postulates a "between study variance component" in addition to the variance of observations within the same study. It is called the random effects model in

meta-analysis (Section 4.5). As a consequence, more than one replication is needed to assess the precision of an effect estimate adequately.

These considerations show the complexity of the basic and seemingly simple task of assessing an effect. The final sections 5 and 6 add thoughts about ways to overcome the difficulties.

2 Terminology

As mentioned above, terms used in discussions about reproducibility are sometimes unclear or ambiguous. Here, we collect them in best agreement with the literature as far as there is a widespread common understanding and mention alternative meanings. Words in *slanted font* are meant to be used as terms with the given meaning.

Reproducibility. The term *reproducibility* should not be used in a specific sense but rather left as a *name for the theme*, that is, it should encompass all aspects of examination of the reliability and relevance of a *scientific claim* or *statement*.

As mentioned in the Introduction, we focus on the situation where a scientific claim resulting from an *original study* is examined by conducting a *validation study*. An empirical study typically consists of the following steps, which may or may not be copied as far as possible in the validation study (see also the Supplement of [23]).

1. Specification of the scientific hypothesis or *claim* with a supposed domain of validity (like population, conditions, ranges of specified variables). Often, a replication only picks up a part of the original study.
2. Design of an experiment or specification of observation units, like subjects, animals, plots.
3. Tools: Auxiliary material, measurement devices, experimenter or observer. These may show batch or calibration effects, temporal variations, and environmental influences.
4. Generation of the data.
5. Data cleaning.
6. Statistical model and procedure of analysis.
7. Selection and presentation of results.
8. Interpretation.

Transparency, Re-assessment. All these steps should be documented well enough to allow for copying them as far as possible. This includes public availability of the data and the code to repeat the formal results, typically the output of statistical methods. We call this aspect the *transparency* of the steps. (Nosek *et al.* [29] label it *process reproducibility*, and Seibold *et al.* [30], *computational reproducibility.)* The steps should be verified in peer *re-assessments* of publications, and the result should be positive—a *confirming re-assessment*. The actual *re-computation* might also be called a *verification* [10] of the analysis.

In computer science and related applications, the term "reproducibility" has been used for this aspect, leading to "reproducible research" [31], [32]. Alternatively, we may call it *transparent research*.

Remark 1 *In fact, this restricted meaning has lead to detrimental confusion, when the "reproducibility crisis" was interpreted as a problem of mistakes or ambiguities in applying statistical software. Therefore, the term reproducibility should definitely be used in the wide—and vague—sense encompassing the whole theme of validation of scientific results.*

Re-analysis. Some steps from "data" to "interpretation" may be subject to criticism, or alternatives may be available. Then, a *re-analysis* may be appropriate as a validation of the claim of the original study. A *new analysis* may use the data for reaching new claims in the spirit of an exploratory study.

Replication. When the experiment or the observation campaign is done again to obtain new data—that is, steps "design" to "data generation" are again executed in the same way—the study is called a *repetition* if it is done by the same team with the same set of tools. Such a repetition should usually be published as part of an original study. If a different team sets up the same experiment or observes according to the same plan, it leads to a ("independent") *replication*, more precisely, a *direct* or *close replication*.

The term *replicability* is generally used in an unfortunate because misleading way. The literal meaning clearly is the *feasibility of replication*, without any specification of the result. A statement like "This paper (or claim) is not replicable" might (and should) mean that there is not enough documentation or that the nature of the phenomenon does not allow for a repetition (like the Big Bang), but it usually means that there has been a replication study that failed to find the same result. One should therefore state if a replication of a claim is feasible and whether the result is a *confirmation*, a *failed confirmation*, or even a *contradiction*. We come back to this assessment of the result in Section 4 and in the conclusions.

Robustness, generalization, extension. It may be informative to examine whether scientific conclusions remain unchanged when experimental methods or schemes of observation are varied, or alternative data cleaning and statistical methods are used. This desired stability has been called *robustness* by Goodman *et al.* [33]. Since this word has many meanings throughout science, it is recommendable to specify the steps (2, 3, 5, 6) towards which the validatoin is tuned.

More broadly, the degree to which a scientific claim remains valid if conditions and populations different from those in the original study is of interest, Corresponding studies modify steps "design" to "data generation" and are called *generalization* studies.

Similarly, a *conceptual replication* is a study to assess the validity of a claim under different circumstances, for other populations, or using alternative methods of measurement, thus varying the first three steps. The "conceptual replication" has been practiced in psychology in order to derive psychological "constructs" (like measures of dimensions of intelligence) and to confirm relations between these using different questionnaires or tests.

An *extension* study would also extend the scientific claim itself.

Further literature about terminology. Patil *et al.* [23] propose terms that are mostly similar to ours. They distinguish between a "replicable study" which means that the estimates of parameters are compatible, and a "replicable claim," for which the scientific conclusion is confirmed by the replication.

Goodman *et al.* [33] introduce a distinction between

- "methods reproducibility," which corresponds to our "transparency" and to reproducibility as used in computational sciences,
- "results reproducibility," which corresponds to confirming replication, and
- "inferential reproducibility," defined as "making knowledge claims of similar strength from a study replication or reanalysis."

For further elaborations on terminology, see [23], [24], [33], [10].

3 Estimation, not "zero hypothesis" testing

It is common practice in most fields of empirical research to report effects if and only if they are statistically significantly different from zero. This habit of "Null Hypothesis Statistical Testing (NHST)" has been critized since it has emerged in the literature, but has lead to more intense controversy in the last decades. Since the p-value is usually given as a summary of the test's outcome, the discussion is also named the "p-value debate." In a preceding paper [34], this author has treated the issue in depth and suggested a simple way to deal with it. Here is a summary.

To fix ideas, consider a paired sample study, in which two treatments are applied to each of n observation units. The effect is measured by the average of the n differences D_i of a target variable for the two treatments. The parameter of interest, θ, is the expected value of the D_i's. The classical test for the null hypothesis of zero difference is the paired samples t-test, that is, the one-sample t-test on the differences D_i. The confidence interval is the t-interval corresponding to this test.

Remark 2 *Note that in practice, it is preferable to use the Wilcoxon signed rank test and the corresponding confidence interval. The t test and interval are nevertheless still more often used, and they generalize to other situations easily.*

In a general case, there is a model for n observations Y_i, containing a parameter θ of interest. In most cases, a suitable estimator $\widehat{\theta}$, like the maximum likelihood estimator, follows aprroximately a normal distribution, $\widehat{\theta} \sim \mathcal{N}(\theta, V/n)$, where V is the asymptotic variance.

The essential argument against NHST runs as follows (see also [35]).

The Zero Hypothesis Testing Paradox. Testing an effect against zero does not answer a scientifically meaningful question. When a study is undertaken to find some difference between groups or some influence between variables, the *true* effect θ will never be precisely zero. Therefore, the strawman null hypothesis of zero true effect (the "zero hypothesis") could in almost all reasonable applications be rejected if one had the patience and resources to obtain enough observations. Consequently, the question that is answered mutates to: "Did we produce sufficiently many observations to prove the (alternative) hypothesis that was true on an apriori basis?" This does not seem to be a fascinating task. This paradox has been stated prominently as a problem in the philosophy of science over fifty years ago in a highly cited long paper by Meehl [36].

Remark 3 *Researchers have taken the paradox into account by refraining from "too large" samples, thereby avoiding that tiny effects become significant. This pragmatic behavior nevertheless appears difficult to justify rationally.*

Parametric models. The scientifically justified question is therefore: "How large is the effect?" The question makes sense only if the parameter is part of a model that describes the phenomenon under study. It can be asked independently of a design of an experiment or obbervation scheme that is used to provide an answer.

Estimation! The straightforward answer to the question is given by an estimate with a confidence interval, based on data related to the model. Many authors and teachers have propageted the routine use of confidence intervals for statistical inference, but the magic of expressing a result in just a single, scaleless number—the p-value—has won in practice. Here, we need to note a different limitation: If effects are reported just when their confidence intervals do not include zero, the selection bias still operates, and this problem of the zero hypothesis testing "culture" is not avoided.

Relevance threshold. In view of the Zero Hypothesis Testing Paradox, the sensible question is: "Is the effect relevant?" This question asks for a *threshold of relevance,* to be set by informed judgement. The threshold has been labelled "Smallest Effect Size Of Interest (SESOI)" [37], "Minimum Practically Significant Distance (MPSD)" [38]. or the limit of the "Region of Practical Equivalence (ROPE)" [39].

Effect scale. The choice of a relevance threshold is often eased by expressing the effect of interest on a natural scale, resulting from a transformation of the original effect parameter. For example, distinctions are often expressed naturally as percentages. Then, inference should be based on parameters and data transformed to logarithmic scale. This translates percentage differences and multiplicative effects into linear differences and effects, which are simpler to interpret and treat mathematically. Specifically, a chosen relevance threshold on the log scale corresponds to a threshold on a multiplicative effect on the original scale. For proportions or probabilities, the log-odds or logistic scale turns effects on odds into linear differences. Whereas a difference of propabilities of 0.1 has a very different importance depending on the two values—changing 0.5 to 0.6 is much less severe than changing 0.88 to 0.98—equal differences in log-odds can be interpreted as being equally relevant. A relevance threshold for log-odds relates to a threshold on multiplicative changes of odds and identifies changes in probabilities that are (arguably) intuitively comparable regardless of their numerical values—a change from 0.5 to 0.6 appears equivalent to a move from 0.88 to 0.92.

An effect scale is suitable if equal differences on it correspond naturally to equally important effects on the original scale, and a constant relevance threshold therefore applies to the whole range of possible values. The transformed parameter will be called the effect ϑ.

For many quantitative target variables, equal differences on their original scale correspond to equally important effects, and no transformation is called for. However, it makes sense to compare an effect to the variable's random variability between observations. In the generic case of a paired sample, this standardization amounts to dividing the expected mean θ by the standard deviation σ of the distribution of the D_i's to get the standardized difference $\delta = \theta/\sigma$. The standardized effect shall be $\vartheta = \delta/2$ for the sake of consistency with the standardized coefficient in regression, see [34].

Choice of a relevance threshold. The choice of a threshold may appear like an undesirable burden for the researcher and a source of arbitrariness. The Zero Hypothesis Testing Paradox suggests that avoiding it is the source of irrelevant or misleading research—certainly a worse option.

In order to alleviate the burden, Stahel [34] gives advice for a "default" choice for the most commonly used statistical models. If the logarithmic scale is appropriate, a threshold of 0.1, corresponding to a descrepancy of approximately 10% on the original scale, may be a plausible choice. An analogous choice applies to the logistic scale, which is suitable for proportions. For a standardized effect, the recommendation is again 0.1.

Remark 4 *Effect scales thus also make effects on the various corresponding types of target variables comparable. In many replicability studies, effects have been aligned by transforming them into correlations, following OSC15. However, effects on this scale are comparable only since they are centered such that the null effect turns into zero correlation, and transformed effects are bound by -1 and 1. According to the arguments above, the correlation scale is not a suitable effect scale.*

Usually, the effect is supposed to be in one of the two possible directions. In order to simplify the wording, we assume it to be positive in the rest of the section.

Relevance measure. Relevance can be expressed as the effect ϑ, divided by the relevance threshold ζ, $\mathrm{RI} = \vartheta/\zeta$. Then, a value of RI larger than 1 indicates a relevant result. It is a parameter of the model, and the point and interval estimates for the original model parameter that determines the effect leads to an estimate RIe and a confidence interval for the relevance. The lower and upper ends of the interval are called the "secured" relevance RIs and the "potential" relevance RIp, respectively. If the secured relevance is larger than 1, the effect is statistically proven to be relevant in a clearly defined sense. Thus, RIs can be used as a new single number to summarize the most important aspect of inference—what the p-value was meant to accomplish.

Classification of results. Based on relevance, a differential answer to the research question can be given according to the following distinction.

- **RIv** The effect is clearly relevant if the whole confidence interval is larger than the threshold, $\mathrm{RIs} \geq 1$.
- **Ngl** The effect is clearly irrelevant or negligible if the whole confidence interval lies on the low side of the threshold, $\mathrm{RIp} < 1$—whether or not zero is covered, that is, the null hypothesis is rejected.
- **Ctr** The assumed direction of the effect proves wrong if the whole confidence interval lies on the negative side, $\mathrm{RIp} < 0$, a clear contradiction.
- **Amb** The result is ambiguous if relevance 1 is contained in the confidence interval, $\mathrm{RIs} < 1 < \mathrm{RIp}$.
- **Amb.Sig** It may be worthwile to label the sub-case of **Amb** in which the result is at least significantly larger than 0, $\mathrm{RIs} > 0$, as **Amb.Sig**.
- **Ngl.Sig** Similarly, the sub-case of **Ngl** with a significant effect is **Ngl.Sig**. It will be very rare unless the sample size or the relevance threshold is large.

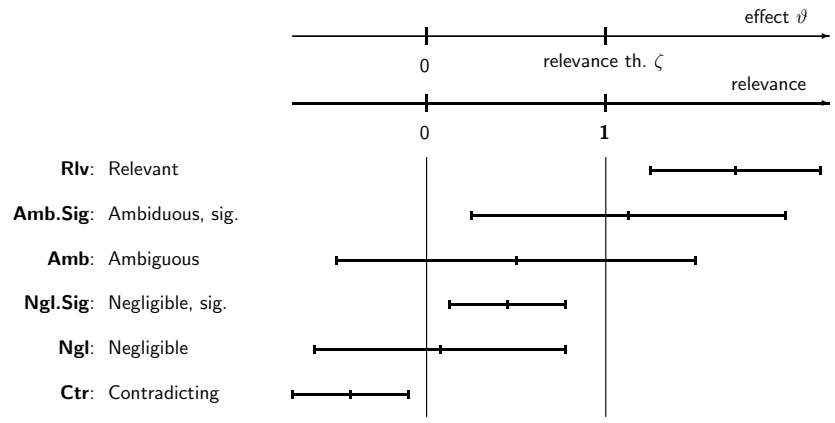

Fig 1. Classification of cases based on a confidence interval and a relevance threshold

The problem of Zero Hypothesis Testing has often been mentioned as an important cause of the reproducibility crisis. In fact, the relation between the two themes is only indirect: The testing routine leads to efficient screening among multiple possible effects for the most appearently significant ones and thereby entails the selective reporting bias.

Remark 5 *Bayesian inference has been advocated as the fruitful alternative to NHST. An important argument is the possibility to assign and develop a probability for the null hypothesis to be true. If the null hypothesis consists of a zero effect, our argument says that its probability is zero, which contradicts this idea. For a "fat" null hypothesis, Bayesian analysis makes sense, but also needs a relevance threshold, in addition to the specification of a prior distribution on the effect, see [39].*

4 Assessment of Success

Let us come back to the theme of replicating an "original" study that examined an effect. How should we assess the success of the replication?

There are two main aspects of success or failure:

- Does the replication lead to the same conclusion as the original? See 4.1.
- Are the two studies consistent in the sense that their quantitative results are similar? See 4.2 to 4.5.

4.1 "Significant again"

The hypothesis testing mode of reasoning suggests that the replication of a significant effect is successful if it turns out significant again and the estimate has the same sign in both studies. Let us call this criterion "significant again," "sigag."

In the rare case when an insignificant original effect is examined by replication, success would be to reach insignificance again, but a study with this goal would be, one could say, of "null merit," c.f. the Zero Hypothesis Testing Paradox. A meaningful question in this case may ask if the effect can be shown to be negligible, see Section 3.

In the light of the preceding section, "sigag" is not a sensible criterion. Note that the probability of achieving it depends on the p-value of the original result. Assume for a moment that the original result was just significant (p=0.05) and the original estimate happened to be the true value of the effect. Then (neglecting that the scale parameter needs to be re-estimated), obtaining a larger estimated effect in the replication with the same number of observations amounts to "sigag" and has a probability of about 0.5 [40], [41].

In OSC15, the proportion of studies achieving "sigag" was 36%.

One can ask for the same conclusion again also using relevance: A relevant result in the original study (case Rlv) is successfully replicated if the case Rlv shows up again, leading to "relevant again." Clearly, in the respective borderline case, this has again a probability of 0.5. We will come back to this requirement and call it a "confirmation" when we distinguish more cases than just "success" from "failure" in Section 4.6.

4.2 Consistency

In empirical studies, the data is subject to random variation. This applies to the original as well as to the replication study. A reasonable question to ask is whether the data in the two studies could be described as coming from the same statistical population. In the generic case, this can be checked by testing if the average D_i in the original study and in the replication show a statistically significant difference. The question is answered by a two-sample t-test. (Note that, again, a nonparametric rank sum test would be more appropriate.) If the test shows no significance, one can say that *the two samples are consistent*.

A closely related approach calculates a prediction interval for the estimate of the replication from the estimate of the original study, its precision, and the sample size of the replication, and checks if the actual estimate of the replication falls into this interval,

see [23]. Applying this procedure to OSC15, they find that 75% of the replications produced consistent data.

Remark 6 *Note that this approach does not compare the data in all aspects, but concentrates on the effects. In our generic case, the response values in the replication could be quite divergent from those in the original study. The test only checks the differences D_i between the responses for the two treatments. If the response shows different values, the new study indeed examines a generalization of the original results.*

It is tempting to say that success of the replication is achieved if the confidence interval of the replication overlaps with the interval of the original study. A second thought shows that this criterion accepts compatibility more often than is likely assumed by most readers: Under the null hypothesis of equal effects and sample sizes in the two studies, the probability of such an overlap is about 99.7% (based on the normal instead of a t distribution, $\Phi(2 \cdot 1.96/\sqrt{2})$) instead of 95%.

Remark 7 *The OCS15 study used confidence intervals in an inappropriate way: The replication was labelled as successful when the confidence interval of the replication covered the estimated effect of the original. This criterion does not consider the randomness of the original result. It is easy to see that if the power of the replication study was increased sufficiently, the criterion woud almost certainly fail, regardless of the quality of the original study. A symmetrized version checks "whether the estimates are within each other's confidence intervals" [42] and suffers from the same flaw. In spite of these undesirable properties, these criteria are still in use, see [18].*

4.3 Relevant Effect difference

Since we do not want to fall back on testing a hypothesis, we now re-formulate the problem. In the generic case, the quantity to be estimated is *half the difference between the true treatment effects*—or more generally of a parameter in a given model—in the two studies, $\mathrm{ED} = (\theta^{(r)} - \theta^{(o)})/2$. (The reason for choosing *half* the difference is mentioned above.)

In order to ease interpretation and avoid cumbersome details, assume that the effect in the original study is positive, $\theta^{(o)} > 0$. The typical case of an attenuated effect then leads to a negative ED.

The (approximate) confidence interval for ED is determined by the standard error $\mathrm{se}_{\mathrm{ED}}$ obtained from the standard errors $\mathrm{se}^{(s)}$ of the effect estimates in the two studies,

$$\widehat{\mathrm{ED}} \pm q\, \mathrm{se}_{\mathrm{ED}} \,, \quad \mathrm{se}^2_{\mathrm{ED}} = \big((\mathrm{se}^{(o)})^2 + (\mathrm{se}^{(r)})^2\big)/4 \,,$$

where q is the appropriate quantile of a t distribution.

As discussed in the previous section, the result should be interpreted with reference to a threshold of relevance. Since the plausible and relevant direction of the difference is to the negative side (a smaller effect in the replication than in the original), the threshold is applied with a minus sign. Then, the case "relevant (Rlv)" occurs if the confidence interval for ED lies on the low side of this threshold, and analogously for the other cases of the classification in Section 3.

Standardized Effect Difference, EDS. Apart from this, the considerations on selecting an effect scale apply to the comparison of the replication with the original again. The possibly transformed or standardized effect difference is called EDS. The relevance threshold for the comparison may be chosen differently from the threshold used for expressing the relevance of the effects in the two studies.

Note that EDS is a *parameter of the model*. It is estimated by plugging in estimates of the parameters. In our generic case, it is plausible to use the standardization $\mathsf{EDS} = (\theta^{(r)} - \theta^{(o)})/(2\sigma) = \vartheta^{(r)} - \vartheta^{(o)}$, where σ is the standard deviation of the D_is and is assumed to be the same in the two studies.

Remark 8 *In the generic case, $2\,\widehat{EDS}$ equals an index that is well-known in the social sciences, called Cohen's d and is sometimes called Standardized Mean Difference in other sciences. Note, however, that the index here refers to the difference between studies, not as usual to the difference between groups within study.*

In fact, in many replication studies, Cohen's d between groups has been used as the effect size and calculated for both the original and the replication study. It is misleading to compare the "d's" between the studies. A difference, $d^{(o)} - d^{(r)} = \overline{D}^{(o)}/\widehat{\sigma}^{(o)} - \overline{D}^{(r)}/\widehat{\sigma}^{(r)}$, could easily occur if the unstandardized effects $\overline{D}^{(s)}$ were equal in the two studies, but the variabilities $\widehat{\sigma}^{(o)}$ and $\widehat{\sigma}^{(r)}$ of the observations were different. Such a difference between variabilities could be due to a true difference in precision of the measurements or to chance. Thus, the effect θ itself must not be standardized when compared between studies, but the standaridization of their difference by a common scaling parameter σ is appropriate. The same argument applies to some other transformations of effects, like the transformation to a correlation coefficient applied in OSC15 and other replication campaigns.

In the generic case, EDS is therefore estimated by

$$\widehat{\mathsf{EDS}} = (\widehat{\vartheta}^{(r)} - \widehat{\vartheta}^{(o)})/\widehat{\sigma} = \widehat{\vartheta}_p^{(r)} - \widehat{\vartheta}_p^{(o)} \,,$$

where $\widehat{\sigma}$ is the pooled estimate of σ, and thus $\widehat{\vartheta}_p^{(o)} = \widehat{\theta}^{(o)}/\widehat{\sigma}$ (subscript p for "pooled") differs from the standardized effect in the original study, $\widehat{\vartheta}^{(o)} = \widehat{\theta}^{(o)}/\widehat{\sigma}^{(o)}$, and analogously for the estimates in the replication. $\widehat{\mathsf{EDS}}$ is then proportional to the t test statistic T for comparison of two independent samples, $T = 2\,\widehat{\mathsf{EDS}}/c_n$ with $c_n = \sqrt{1/n^{(o)} + 1/n^{(r)}}$. The confidence interval for $2\,\mathsf{EDS}$ is thus given by $2\,\widehat{\mathsf{EDS}} \pm q \cdot c_n$, where q is the quantile of the t distribution with $n^{(o)} + n^{(r)} - 2$ degrees of freedom.

A threshold of 0.1 for EDS equals the threshold 0.2 for "small" values for Cohen's standardized difference $d = 2\,\mathsf{EDS}$ that is popular when interpreting d.

Standardization in the general case. In a general setup, the "effect" θ is any parameter of interest in a given model describing the observations. An estimator $\widehat{\theta}$ has a given distribution, derived from the model. Usually, this distribution approximately equals a Gaussian with a variance that is inversely proportional to the sample size, $\mathrm{var}(\widehat{\theta}) = V/n$, $\widehat{\theta} \sim \mathcal{N}(\vartheta, V/n)$. Then, the estimator of $2\,\mathsf{ED} = \theta^{(r)} - \theta^{(o)}$ entails the confidence interval

$$\left(\widehat{\theta}^{(r)} - \widehat{\theta}^{(o)}\right) \pm q\,\sqrt{\widehat{V}^{(o)}/n^{(o)} + \widehat{V}^{(r)}/n^{(r)}}\,.$$

Often, V does not depend on the value of θ, or this can be achieved by a transformation of θ. Then, the standardized effect difference is

$$\mathsf{EDS} = \left(\theta^{(r)} - \theta^{(o)}\right) \Big/ \left(2\sqrt{V}\right)\,.$$

Note that standardization is not needed nor recommended if the effect scale is logarithmic or logistic.

4.4 Desirable properties of discrepancy measures.

The quantity EDS has the first three of the following properties that we consider essential for any index RD of *Replication Discrepancy*.

(P1) RD should be a function of parameters of the model for the original and the replication(s) and thus should not depend on the number of observations used in the studies. It is then *estimated* on the basis of the data. (See "Parametric models" in Section 3.)

(P2) RD should measure the discrepancy between the quantities of interest in the original and the replication study (or studies) and be rather insensitive to differences in other aspects.

(P3) It is desirable that RD can be generalized to multivariate effects, as for instance to analysis of variance, where several contrasts are of interest.

(P4) RD should generalize to the situation of more than one replication study.

The desired property (P4) leads us to extending our model as follows.

4.5 Heterogeneity of studies

Following the preceding arguments, one might expect that the test for zero difference $\mathrm{ED} = 0$ or $\mathrm{EDS} = 0$ should fail only in about 5 percent of replications of original studies that can be assumed to be free of selective reporting bias. General experimental-statistical experience dampens this hope. The hypothesis of exactly equal expected effects, $\vartheta^{(r)} = \vartheta^{(o)}$, is not realistic in practice. It is clear from experience of any types of measurements or observations that their random variation within the same study will be smaller than the variation of measurements from different studies. In technical terms, there is a variance component reflecting the differences between studies, the *between studies variance*. The concept is usually called "heterogeneity" and forms the basis of the random effects model in meta-analysis. In the generic case, the effects $\vartheta^{(s)}$ in different studies s are modelled as realizations of a random variable. The quantity of interest would be the expected value Θ of this random variable, and its variance is the "between study variance component" σ^2_ϑ. Estimation of Θ is best achieved by an average (possibly a weighted one) of the $\widehat{\vartheta}^{(s)}$ that are available, and the width of a confidence interval would need to contain an estimate of σ_ϑ or of the (relative) *Between Study Variability* $\mathrm{BSV} = \sigma_\vartheta/\sigma$.

Heterogeneity has received increasing attention in recent years, see, e.g, [43], [3], [44].

Remark 9 *In meta-analysis, an index (H in [46]) compares a version of the between-study variability σ^2_ϑ with the average of precisions $1/\mathrm{var}\big(\widehat{\vartheta}^{(s)}\big) = n^{(s)}/V^{(s)}$ of the effect estimates for the individual studies. In contrast, BSV uses the average of the $V^{(s)}$ quantities, which describe the information contained in individual observations rather than the variability of the estimates. This makes BSV a parameter of the model for the observations that is independent of the numbers $n^{(s)}$ of observations in the studies, thus fulfilling property (P1), whereas the random effects model of meta-analysis starts from the effect estimates in the studies and their precisions and therefore fails to characterize the basic phenomenon generating the data.*

Remark 10 *Clemens [10] states that "In expectation, these tests [the tests for 0 difference of effect sizes] are supposed to yield estimates identical to the original study. If they do not, then either the original or the replication contains a fluke, a mistake, or fraud." In the light of the concept of a between study variance component, such a conclusion is not warranted. Several authors suggest potential reasons for heterogeneity and urge researchers to*

investigate and eliminate them. Experience from interlaboratory studies in chemistry and metrology in general shows that often, no reasons for a variance component between batches can be identified, but it remains relevant anyway.

If only the original and one replication study are available, an estimate of the between studies variance component is only possible if the selective reporting bias of the original is assumed to be zero, and it then relies on one degree of freedom and would therefore be ill-determined and useless. (In fact, $\widehat{\text{BSV}}^2 = \big(\widehat{\text{EDS}}^2 - (1/n_0 + 1/n_1)\big)/2$—or 0, if this is negative—can be interpreted as a point estimate of the between study variability BSV.)

Several replications! Remembering that a valid confidence interval for the true effect Θ needs a reliable value for the between study variance σ_ϑ^2, a reasonable number of replications is needed for its estimation, as pointed out and justified by Hedges and Schauer [44]. Due to the selective reporting bias, the original study should not be used in such an evaluation [47].

Alternatively, if a number of replication studies for different original claims in a field of application should lead to similar values of BSV, such a value may be used to calculate a rough factor by which the confidence interval for $\vartheta^{(s)}$ of a replication study should be widened in order to be used as a confidence interval for the "global" true effect Θ.

4.6 A classification of outcomes

The goal of replication is to validate a scientific claim. Here, we deal with the case of an "effect" that has been found relevant or at least significant in the original study. On the basis of the confidence interval "IEff$^{(r)}$" for the effect ϑ obtained in the replication and on the confidence interval "IEDS" for EDS, the result may be characterized, using the scheme of Section 3. Besides a threshold of relevance for the effect ϑ, a threshold for relevant values of the standardized effect difference EDS is needed. EDS is relevant if it is lower than the negative relevance threshold. Then, the result is a

(Cnf) *Confirmation,* if IEff$^{(r)}$ only contains relevant values (case Rlv), and the negative standardized effect difference EDS is small (cases Ngl or Amb); if IEff$^{(r)}$ is only significant (Amb.Sig) and the estimate $\widehat{\vartheta}_1$ is larger than the relevance threshold, we call it a *weak confirmation* (CnfW),

(Att) *Attenuation,* if IEff$^{(r)}$ lies on the same side of 0 as in the original study (Rlv or Amb.Sig) and IEDS is relevant (Rlv),

(Enh) *Enhancement,* if the replication suggests a clearly stronger effect, that is, case (Rlv) for IEff$^{(r)}$ and significantly positive EDS (Ctr); this will be rare,

(Amb) *Ambiguous,* if IEff$^{(r)}$ covers the relevance threshold and it also covers zero (Amb) or the estimate $\widehat{\vartheta}_1$ is below the reference threshold,

(Anh) *Annihilation,* if IEff$^{(r)}$ covers only irrelevant values (Ngl),

(Ctr) *Contradiction,* if all values of IEff$^{(r)}$ have the opposite sign (Ctr),

(Drp) *Dropout,* if the replication failed to mimik the experimental or observational setup.

The classification is summarized in Table 1 and displayed in Figure 2. The first three cases are identified by the "significant again" criterion as a *successful replication.* Nevertheless, conclusions might be rather different between them, see Section 6.

Effect estimate $\text{IEff}^{(r)}$ in replication	Effect Difference (standardized), IEDS relevant, Rlv	Amb or Ngl	contradicting, Ctr
relevant, Rlv	attenuation, Att	confirmation, Cnf	enhancement, Enh
significant, Sig	attenuation, Att	weak conf., CnfW*	—
ambiguous, Amb	ambiguous, Amb	ambiguous, Amb	—
negligible, Ngl	annihilation, Anh	annihilation, Anh**	—
contradicting, Ctr	contradiction, Ctr	—	—

Table 1. Classification of results of a replication of a relevant effect, based on the classificaion of the confidence interval $\text{IEff}^{(r)}$ for the effect in the replication and the confidence interval IEDS of the EDS. It is assumed that the original effect was relevant or at least significant. Then, the cases marked — cannot occur. * This conclusion also requires Rle ≥ 1; otherwise, it counts as ambiguous. ** This cannot occur if the original effect was relevant.

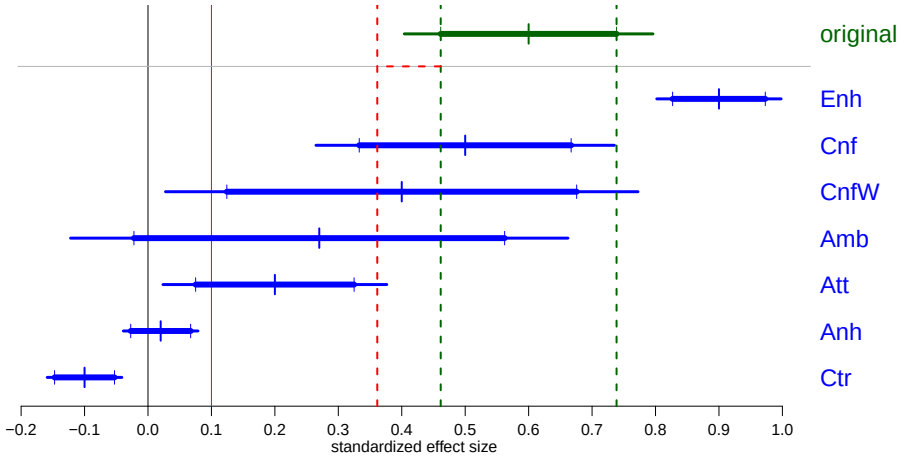

Fig 2. Classification of replication results for a relevant original effect. Confidence intervals for the original effect (top, green) and for the effect obtained in the replication in the different possible cases (blue). The additional ticks and thicker lines on the bars allow for an assessment of EDS. If the intervals bounded by them do not overlap, EDS is significantly different from 0. In order for EDS to be also relevant, the gap must be longer than $\rho_D = 0.1$, say. Thus, for "attenuation", Att, the right hand end point of the bar's thicker part must be to the left of the red dashed vertical line.

Remark 11 *Bonett [48] also uses* $\text{IEff}^{(r)}$ *and IEDS to classify replication results. His classes are defined by one of these intervals or the other. Therefore, they overlap. See supplementary material for more detail.*

As an illustration, Figure 3 shows confidence intervals for standardized effects in ten "simple cases" studied in OSC15, together with their classification. The data is displayed in the Supplement. We suggest that showing the confidence intervals for the original study and the replication(s), including the relevance threshold (cf. end of Section 3) should become a standard display of the information in replication studies. (Unfortunately, we had to compare standardized effects here, disregarding Remark 8, since unstandardized effects are not given in the data provided by OSC15.)

The figures display additional ticks on the interval bars that allow for checking the significance of the effect difference ED. If the shortened intervals do not overlap, the

difference is significant, and the relative width of the gap or overlap visualizes how significant the difference is indeed. The position of the ticks is given, for the original study, by $\widehat{\vartheta}^{(o)} \pm q\nu^{(o)}\mathrm{se}^{(o)}$, where $\nu^{(o)} = 2\ \mathrm{se}_{\mathsf{ED}}/(\mathrm{se}^{(o)} + \mathrm{se}^{(r)})$, and analogously for the replication results. Then, the gap is, if $\mathsf{ED} > 0$,

$$\widehat{\vartheta}^{(r)} - q\nu^{(r)}\mathrm{se}^{(r)} - (\widehat{\vartheta}^{(o)} + q\nu^{(o)}\mathrm{se}^{(o)}) = 2\,\mathsf{ED} - (q\nu^{(o)}\mathrm{se}^{(o)} + q\nu^{(r)}\mathrm{se}^{(r)}) = 2\,(\mathsf{ED} - q\ \mathrm{se}_{\mathsf{ED}})\ ,$$

which is positive if and only if the difference is significant. (This enhancement corresponds to the idea of "notched box plots" [49]) and makes it exact for the comparison of two samples.)

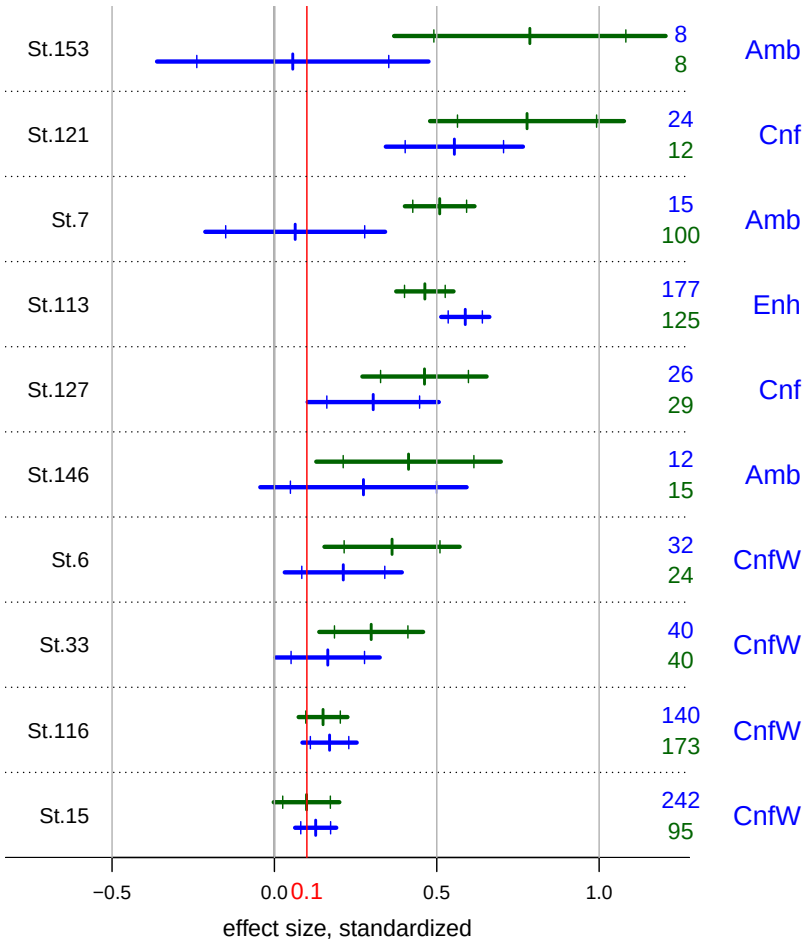

Fig 3. Confidence intervals for effects in ten items studied in OSC15, all based on paired or simple t-tests. Original and replication results are shown in green and blue, respectively. The number of observations and the classification according to Table 1 are shown in the right margin.

5 Replication!

Let us briefly come back to the different sources of selective reporting bias. It emerges in settings where data may show several or many effects or patterns. By a formal or informal search, the most prominent ones are selected, and statistical methods are applied that ignore

this selection process. Note that if a study is designed in the first place to find "real" effects among a clearly defined set of possible ones in a structured way, there often are statistical tools to avoid selective reporting bias, notably the well-known Bonferroni rule, but also more refined methods like those proposed in [50], [51], [52]. But in general, such methods to correct for selection are not available or not used.

Pre-registration! Publication bias is avoided by *pre-registration*: The scientific claim(s) to be examined and the plan to perform a respective study are published in an accessible and visible registry, before the data is collected. Such *pre-registration reports*—usually called "registered reports"—should be peer-reviewed [53]. The results of the study must then be published regardless of whether they confirm or contradict the claim. Whereas such practice can be useful even for original work, it must be standard for replication studies, since otherwise, selection bias may easily play again and even a joint meta-analysis of several replications could not be reliably interpreted. *A replication that is not pre-registered should not obtain the label "replication," and should be avoided.*

Several replications are needed! If a trustworthy confidence interval for the size of an effect is desired, the between studies variance component must be estimated. As argued above, several replications—at least 5, say—should be planned and pre-registered.

Remark 12 *Experience in the series of replications in cancer biology shows that in several studies, some of the preliminary steps needed before acquiring the data for validating the original claims have failed and a section on "Deviations from registered report" became necessary. When essential deviations occur, the replication should be counted as a "dropout."*

Remark 13 *A confidence interval obtained in a pre-registered study would have the correct probability of covering the true parameter value if there was no between studies variance component, see above. Note, however, that such a study may in fact be a step in a stepwise strategy (Section 6): If the replication fails to be significant again, there is an incentive to try it again by another pre-registered replication. If this was a clear strategy, a formal treatment as a sequential procedure would be appropriate to derive correct coverage probabilities.*

Remark 14 *Extending this consideration further makes it clear that the selective reporting bias of scientific claims can be greatly reduced by replication studies and strategies, but not eliminated, because claims that are not confirmed will be forgotten, and those that are will be retained. This again constitutes a selection and leads to a "secondary selective reporting bias." The flaw is reduced if replications are restricted to the most relevant claims. A blind routine of replicating as much as possible would increase the secondary selective reporting bias.*

Remark 15 *When planning a single replication, it is essential to consider the power of the new study—as a low powered single replicaton will enhance the problems just mentioned. Using the estimated effect from the original study as the true effect for power analysis would be inappropriate because of the selective reporting bias that is to be assumed for the original work. Samantha et al. [54] and Bonnett [48] consider the uncertainty of the original estimate in addition to the bias, discuss the implications and provide a method to calculate more appropriate sample sizes for a replication. Implications of statistical power are discussed extensively by Morey and Lakens [55] under the title "Why most of psychology is statistically unfalsifiable."*

Performing replications can be made attractive. Even though the benefits of replication studies are widely recognized, many authors seem rather pessimistic since they judge such studies to be unattractive. They state that researchers will not get the necessary recognition and funds if they invest their time and resources into such activities.

However, we see two ways that lead to the desired studies:

- Beginning PhD students need to learn and practice the methods of scientific studies in their field. They often work in directions that start from an existing publication. If they are asked to perform a replication of such a study at the start, this guarantees a first publication, which is counted towards number needed to complete their thesis [56], [57].

- More generally, research often aims at a generalization or extention in the sense of Section 2. Such projects should contain a (pre-registered!) replication as a first part [58], [59].

Chambers [60] gives good arguments for a change of culture rewarding replications.

Let's establish the rules! The following rules should be suitable for establishing the replication paradigm:

- **The Pottery Barn Rule** [61]. Journals should adopt the policy of accepting pre-registrations for replication studies of the "original" papers that they have published. They entertain a publicly available list of these pre-registered projects, preferably integrated with the lists of other journals. They guarantee that the results of the replication will be published: A short version must appear in their main mode of publication, possibly leaving the documentation to an online "supplementary" part. This should allow journals and their readers to keep their enthusiasm for novel findings any yet promote the establishment of reliable knowledge.

  Adequate power for deciding about the relevance of the effect need *not* be required in each replication, as several independent "underpowered" replications are more easily obtained and more useful than a large replication study, since such studies eventually allow for estimation of the inter-study variability. Note, however, that a conclusive joint evaluation is only warranted for a pre-planned series, since otherwise, the results of the first studies might influence the likelihood that later studies will be undertaken.

- Supervisors and funding agencies of beginning PhD students ask that they start with a replication study, preferably of an original study from another research group. This would even enhance cooperation among groups. (Clearly, this principle cannot be followed in fields where experiments take long—more than a year, say.)

- In addition, the principles of open science, i.e., complete transparency, data availability and re-computability of analysis help to improve reproducibility. Here, the platform of the Open Science Framework [62, 63] is a very useful resource, and badges or medals [64] can help acknowledge the efforts.

An extensive discussion on "making replications mainstream" is provided by Zwaan *et al.* [65] and the 36 evoked comments.

6 Conclusions

The basic paradigm in science states that facts should only be recognized as such if they can be and have been reproduced. In many empirical science fields, this is not often practiced, and it is difficult to judge which statements should be regarded as reliable. Even worse,

when replication studies were undertaken, their results have shown a disappointing rate of confirmation. This insight has lead to the "reproducibility crisis" in large parts of science.

An important trigger is the urge to find new "facts" by a kind of raster screening. The content of research is often not guided by interesting relevant questions but by exploring many potential effects of minor importance with the hope to find statistically significant ones in some niche. The trap of selection bias or "p-hacking" snaps [66] [67].

Awareness and concern about the problem have increased and lead to a flood of editorials and articles—and even books—that have dealt with diagnoses, interpretations, reviews, and proposals for procedures and policies. In this contribution, we have focussed on some basic issues:

- **Estimation, not testing.** Solid empirical research concerns important relations and aims at estimation of relevant effects, intending not only to prove that they are different from zero—which they are "almost surely" in any case. An estimation problem asks for a relevance threshold, and the adequate and straightforward way to present the result is by a confidence interval. If a conclusion is needed, it should consist of checking whether the confidence interval covers the threshold. The small but important step from the misguided use of p values to providing confidence intervals with their simple and direct relation to the scientific problem of assessing an effect must finally be consistently implemented.

- **Between study variation.** In any two studies, we should expect a variation between effects that is not restricted to the statistical variability of their estimates within each study as quantified by the formal standard error. This insight is best described by the random effects model of meta-analysis. Consequently, large replication studies should not be expected to yield effect estimates that are compatible with the original in the sense of statistically insignificant difference, due both to this heterogeneity and to the selective reporting bias.

  The other important consequence is that the confidence interval obtained from a single study does not cover the true effect with the probability expressed by the nominal confidence level. It should be widened by a factor reflecting the between-study variance. A whole set of studies is needed to estimate this variance, or an informed value must be assumed. Due to the selective reporting bias lurking in "original studies," sincere estimation is only achieved from pre-registered replications, thereby devaluating the quantitative results of the original study in favor of unbiasedness.

  This paradigm entails a fundamental change in planning replications. It is not really useful to conduct a single replication with a desired power, calculated on the basis of the assumption of equal true effects between original and replication. Instead, a series of replications should be planned (possibly using a sequential design) and the sample sizes should be determined by power calculations respecting the heterogeneity, see [47].

- **"Success" of a replication.** Statements about reproducibility should be careful in their use of terminology and differentiate possible outcomes. A binary answer is not helpful. We have suggested a classification with six different outcomes.

Usually, a replication is designed as "close" as possible to the original, using, in the applicable sense, the same "population" and the same methods. This principle is meant to lead to a high probability of getting consistent results—which, as we just argued, will be "successful" ($p < 0.05$) only if the statistical power is kept low enough to avoid detection of the interstudy variability. For important scientific problems, however, an essential criterion is the generalizability of the result. Therefore, an adequate compromise between confirmation and extension of results is needed (see also conclusions in [22]).

In summary, then, a strategy is needed in order to obtain reliable scientific facts. An attempt to draft such a standard is the following.

- If a claim is of basic interest for the field, multiple replications should be planned. A judgement is needed on the extent to which these replications should generalize the context and thus extend the domain of validity, and on the relevance threshold. These decisions may be a topic for professional societies.
- For findings stemming from exploratory studies, a first close (pre-registered) replication should be conducted, and depending on the result, more replications should follow: In the case of a confirmation (Cnf according to Section 4.6), generalizations can be studied, but in case of an attenuation or ambiguous result, more close replication is suggested. When replications are conducted without a pre-planned strategy, meta-analyses need to take the sequential aspects into account.
- If a study serves to validate a theoretical proposition, it should be pre-registered in the first place.
- In other cases, claims should be interpreted as working hypotheses ([68] and others). Such exploratory results should still play an important role in science and, if done with enough care, get published as the potential source of replication or, more generally, as indications for generating theoretical hypotheses to be examined by pre-registered studies.

Such a strategy aims at structuring the process of knowledge generation. They should avoid rules that restrict creativity, but rather help distinguish the degrees of reliability of empirically based claims and thereby save resources.

Our recommendations ask for substantial changes in the practice of empirical research. We are convinced that the crisis of empirical science is even deeper than recognized by the current discussion, and it is time to ask for the changes needed to overcome it even though they sound overly challenging at present. In the long run, solid establishment of scientific facts will prove sustainable, whereas past and present practices dilute the credibility of science and threaten to erode the support from society it still enjoys.

Acknowledgement. Useful comments on an earlier version of the typescript have been provided by Markus Kalisch. Samuel Pawel gave very valualble hints to relevant literature and helped strengthening arguments.

References

1. Makel MC, Plucker JA, Hegarty B. Replications in psychology research: how often do they really occur? Perspect Psychol Sci. 2012;7(6):537–542.
2. National Academies of Sciences, Engineering and Medicine. Reproducibility and replicability in science. Washington, D.C.: National Academies Press; 2019.
3. Stanley TD, Carter EC, Doucouliagos H. What Meta-Analyses Reveal About the Replicability of Psychological Research. Psychological Bulletin. 2018;144(12):1325–1346.
4. Academy of Medical Sciences. Reproducibility and reliability of biomedical research: improving research practice. Academy of Medical Sciences; 2015.
5. Gelman A, Loken E. The statistical crisis in science. American Scientist. 2014;10/20/14.
6. Simmons JP, Nelson LD, Simonsohn U. False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. Psychological Science Online. 2011; p. doi:10.1177/0956797611417632.

7. Dwan K, Gamble C, Williamson PR, Kirkham JJ. Systematic Review of the Empirical Evidence of Study Publication Bias and Outcome Reporting Bias–An Updated Review. PLOS ONE. 2013;8(e66844).

8. Munafò MR, Nosek BA, Bishop DVM, Button KS, Chambers CD, du Sert NP, et al. A manifesto for reproducible science. Nature human behaviour. 2017;1(0021):1–9.

9. van Zwet1 EW, Cator EA. The significance filter, the winner's curse andthe need to shrink. Statistica Neerlandica. 2021;75:437–452.

10. Clemens MA. The meaning of failed replications: a review and proposal. J Economic Surveys. 2015;.

11. Open Science Collaboration. Estimating the reproducibility of psychological science. Science. 2015;349:943–952.

12. Patil P, Peng RD, Leek JT. A statistical definition for reproducibility and replicability. bioRxiv preprint; 2016.

13. Klein RA, Ratliff KA, Vianello M, et al . Investigating variation in replicability: A "many labs" replication project. Social Psychology. 2014;45(3):142–152.

14. Klein RA, Vianello M, Hasselman F, Adams BG, Adams RB, Alper S, et al. Many Labs 2: Investigating Variation in Replicability Across Samples and Settings. Advances in Methods and Practices in Psychological Science. 2018;1(4):443–490.

15. Ebersole CR, et al . Many Labs 5: Testing Pre-Data-Collection Peer Review as an Intervention to Increase Replicability. Advances in Methods and Practices in Psychological Science. 2020;3(3):309–331. doi:10.1177/2515245920958687.

16. Camerer CF, Dreber A, Holzmeister F, Ho TH, Huber J, Johannesson M, et al. Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. Nature Human Behaviour. 2018;2:637–644. doi:10.1038/s41562-018-0399-z.

17. Cova F, Strickland B, Abatista A, Allard A, Estimating the reproducibility of experimental philosophy. RevPhilPsych. 2018;.

18. Errington TM, Mathur M, Soderberg CK, Denis A, Perfito N, Iorns E, et al. Investigating the replicability of preclinical cancer biology. eLife. 2021;10:e71601. doi:10.7554/eLife.71601.

19. Shrout PE, Rodgers JL. Psychology, Science, and Knowledge Construction: Broadening Perspectives from the Replication Crisis. Annu Rev Psychol. 2018;69:487–510.

20. Begley CG, Ioannidis JP. Reproducibility in Science: Improving the Standard for Basic and Preclinical Research. Circulation Research. 2015;116(1):116–126.

21. Leek JT, Jager LR. Is most published research really false? A Rev Statist Appl. 2017;4:109–122.

22. Maxwell SE, Lau MY, Howard GS. Is psychology suffering from a replication crisis? What does "failure to replicate" really mean? Am Psychol. 2015;70(6):487–498.

23. Patil P, Peng RD, Leek JT. What should researchers expect when they replicate studies? A statistical view of replicability in psychological science. Perspect Psychol Sci. 2016;11:539–544.

24. Plesser HE. Reproducibility vs. Replicability: A Brief History of a Confused Terminology. Frontiers in Neuroinformatics. 2018;11:76.

25. Stahel WA. Statistical Issues in Reproducibility. In: Atmanspacher H, Maasen S, editors. Reproducibility: Principles, Problems, Practices, and Prospects. Wiley; 2016. p. 87–114.

26. Nosek BA, Errington TM. Making sense of replications. eLife. 2017;6:e23383.

27. Anderson SF, Maxwell SE. There's More Than One Way to Conduct a Replication Study: Beyond Statistical Significance. Psychological Methods. 2016;21:1–12.

28. Ioannidis JPA. The reproducibility wars: Successful, unsuccessful, uninterpretable, exact, conceptual, triangulated, contested replication. Clinical Chemistry. 2017;63(5):943–945.

29. Nosek BA, Hardwicke TE, Moshontz H, Allard A, Corker KS, Dreber A, et al. Replicability, Robustness, and Reproducibility in Psychological Science. Annual Review of Psychology. 2021;73:114157.

30. Seibold H, Czerny S, Decke S, Dieterle R, Eder T, Fohr S, et al. A computational reproducibility study of PLOS ONE articles featuring longitudinal data analyses. PLOS ONE. 2021;16:e0251194.

31. Peng RD. Reproducible research in computational science. Science. 2011;334:1226–1228.

32. Peng RD. Reproducible research and Biostatistics. Biostatistics. 2009;10(3):405–408.

33. Goodman SN, Fanelli D, Ioannidis JPA. What does research reproducibility mean? Science Translational Medicine. 2016;8(341):341ps12–341ps12.

34. Stahel WA. New relevance and significance measures to replace p-values. PLOS ONE. 2021;16:e0252991.

35. Gelman A, Hill J, Yajima M. Why we (usually) don't have to worry about multiple comparisons. arXiv:09072478 [statAP]. 2009;.

36. Meehl PE. Theory-testing in psychology and physics: A methodological paradox. Philosophy of Science. 1967;34:103–115.

37. Heirene RM. A call for replications of addiction research: Which studies should we replicate and what constitutes a "successful" replication?; 2019.

38. Goodman WM, Spruill SE, Komaroff E. A proposed hybrid effect size plus p-value criterion: Empirical evidence supporting its use. The American Statistician. 2019;73(1, suppl.):168–185.

39. Kruschke JK. Rejecting or accepting parameter values in Bayesian estimation. Advances in Methods and Practices in Psychological Science. 2018;1(2):270–280.

40. Goodman SN. A comment on replication, P-values and evidence. Statistics in Medicine. 1992;11:875–879.

41. Piper SK, Grittner U, Rex A, Riedel N, Fischer F, Nadon R, et al. Exact replication: Foundation of science or game of chance? Plos Biology. 2019;17(4).

42. Errington TM, Iorns E, Gunn W, Tan FE, Lomax J, Nosek BA. An open investigation of the reproducibility of cancer biology research. eLife. 2014;3:e04333.

43. Kenny DA, Judd CM. The Unappreciated Heterogeneity of Effect Sizes: Implications for Power, Precision, Planning of Research, and Replication. Psychological Methods. 2019;.

44. Hedges LV, Schauer JM. More Than One Replication Study Is Needed for Unambiguous Tests of Replication. Journal of Educational and Behavioral Statistics. 2019;44(5):543–570.

45. Hedges LV, Schauer JM. Statistical analyses for studying replication: Meta-analytic perspectives. Psychological Methods. 2019;24(5):557–570.

46. Higgins JPT, Thompson SG. Quantifying heterogeneity in a meta-analysis. Statistics in Medicine. 2002;21:1539–1558.

47. Pawel S, Held L. Probabilistic forecasting of replication studies. PLOS ONE. 2020;15(4):1–23. doi:10.1371/journal.pone.0231416.

48. Bonett DG. Design and Analysis of Replication Studies. Organizational Research Methods. 2020;24(3).

49. McGill R, Tukey JW, Larsen WA. Variations of box plots. The American Statistician. 1978;32:12–16.

50. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple -Testing. J R Statist Soc B. 1995;57(1):289–300.

51. Meinshausen N, Meier L, Bühlmann P. P-values for high-dimensional regression. J Am Statist Assoc. 2009;104:1671–1681.

52. Berk R, Brown L, Buja A, Zhang K, Zhao L. Valid post-selection inference. Ann Statist. 2013;41(2):802–837.

53. Chambers CD, Tzavella L. The past, present and future of Registered Reports. Nature Human Behaviour. 2022;6:29–42.

54. Anderson SF, Kelley K, Maxwell SE. Sample-Size Planning for More Accurate Statistical Power: A Method Adjusting Sample Effect Sizes for Publication Bias and Uncertainty. Psychological Science. 2017;28(11):1547–1562.

55. Morey RD, Lakens D. Why most of psychology is statistically unfalsifiable; 2016.

56. Everett JAC, Earp BD. A tragedy of the (academic) commons: interpreting the replication crisis in psychology as a social dilemma for early-career researchers. Frontiers in Psychology. 2015;6:1152.

57. Kochari AR, Ostarek M. Introducing a replication-first rule for Ph.D. projects. Comment on Zwan et al, "Making replication mainstream". Behavioral and Brain Sciences. 2018;41, E138.

58. Morey RD, Chambers CD, Etchells PJ, Harris CR, Hoekstra R, Lakens D, et al. The Peer Reviewers' Openness Initiative: incentivizing open research practices through peer review. R Soc Open Sci. 2016;3:150547.

59. Bonett DG. Replication-extension studies. Current Directions in Psychological Science. 2012;21:409–412.

60. Chambers C. The registered reports revolution. Lessons in cultural reform. significancemagazinecom. 2019;.

61. Srivastava S. A Pottery Barn rule for scientific journals; 2012. thehardestscience.com/2012/09/27/a-pottery-barn-rule-for-scientific-journals/.

62. Nosek BA, Alter G, Banks GC, Borsboom D, Bowman SD, Breckler SJ, et al. Promoting an open research culture. Science. 2015;348:1422–1425.

63. Foster ED, Deardorff A. Open Science Framework (OSF). J Med Libr Assoc. 2017;105(2):203–206.

64. Kidwell M, Lazarević L, Baranski E, Hardwicke T, S P, Falkenberg L, et al. Badges to acknowledge open practices: a simple, low-cost, effective method for increasing transparency. PLOS Biol. 2016;14:e1002456.

65. Zwaan RA, Etz A, Lucas RE, Donnellan MB. Making replication mainstream. Behavioral and Brain Sciences. 2018;41:E120.

66. Amrhein V, Greenland S, McShane B. Retire statistical significance. Nature. 2019;567:305–307.

67. Wasserstein RL, Schirm AL, Lazar NA. Moving to a World Beyond "p ¡ 0.05". The American Statistician. 2019;73:sup1:1–19.

68. Sorić B. Statistical "discoveries" and effect size estimation. J Am Statist Assoc. 1989;84(406):608–610.