Making our “meta-hypotheses” clear: heterogeneity and the role of direct replications in science

Eirik Strømland¹

Received: 30 January 2020 / Accepted: 22 January 2021 / Published online: 24 March 2021
© The Author(s) 2021

Abstract

This paper argues that some of the discussion around meta-scientific issues can be viewed as an argument over different “meta-hypotheses” – assumptions made about how different hypotheses in a scientific literature relate to each other. I argue that, currently, such meta-hypotheses are typically left unstated except in methodological papers and that the consequence of this practice is that it is hard to determine what can be learned from a direct replication study. I argue in favor of a procedure dubbed the “limited homogeneity assumption” – assuming very little heterogeneity of effect sizes when a literature is initiated but switching to an assumption of heterogeneity once an initial finding has been successfully replicated in a direct replication study. Until that has happened, we do not allow the literature to proceed to a mature stage. This procedure will elevate the scientific status of direct replication studies in science. Following this procedure, a well-designed direct replication study is a means of falsifying an overall claim in an early phase of a literature and thus sets up a hurdle against the canonization of false facts in the behavioral sciences.

Keywords Replication crisis · Heterogeneity · Truthlikeness · Direct replications

1 Introduction

In the last decade, a crisis of confidence has characterized the behavioral sciences. Failures to replicate central findings are pervasive across fields such as psychology

I thank two anonymous reviewers for many helpful comments that substantially improved the quality of this paper.

This article belongs to the Topical Collection: *Philosophical Perspectives on the Replicability Crisis*

Guest Editors: Mattia Andreoletti, Jan Sprenger

✉ Eirik Strømland
eias@hvl.no

¹ Department of Business Administration, Western Norway University of Applied Sciences, P.O. 7802, 5020 Bergen, Norway

(Open Science Collaboration, 2015), economics (Camerer et al., 2016) and the social sciences in general (Camerer et al., 2018). Perhaps the most worrying part of this credibility crisis is that certain literatures seem to have lived on for a very long time with seemingly stable and robust effects, yet central findings later fail in highly powered direct replication studies. Such empirical patterns raise doubts about whether science can be said to always be self-correcting (Ioannidis, 2012).

Theoretical studies have suggested that the current publication system, where negative results are rarely published, promotes the “canonization of false facts” (Nissen et al., 2016) – a false positive finding may initiate a literature, and a series of subsequent false positive findings eventually leads the field to accept the overall claim as established as the missing studies are typically not published. Questionable research practices such as “p-hacking” (Simmons et al., 2011) exacerbate this tendency to canonize false facts. One process by which canonization of a false fact can occur in practice is that “conceptual replications” – typically studies with a minor change in the original study design intending to test the robustness of the same concept – will be taken as evidence corroborating the overall claim, whereas failed conceptual replications will be interpreted as evidence that the study design deviated too much from the original study to be considered a fair test of the overall claim. One likely example of this process of “canonization of false facts” is the “ego depletion effect” – the idea that willpower relies on a limited pool of available mental resources that can be “depleted” through cognitive effort (Baumeister et al., 1998). The idea enjoyed substantial empirical support for almost two decades across many experimental psychology studies, but in 2016 a many-laboratory direct replication failed to obtain evidence for an ego depletion effect (Hagger et al., 2016). Another example is the “facial feedback effect” (Strack et al., 1988) – the idea that people’s facial expression can influence their affective responses. This idea was supported by several independent related studies and was commonly discussed in psychology textbooks but failed to replicate in a preregistered and highly powered direct replication study (Wagenmakers et al., 2016). Further, analyses of citation data from experimental psychology indicate that there is no difference in citation rates for studies that passed or failed a replication attempt – suggesting that statistical “flukes” are just as likely to influence future research as papers with reproducible results (Yang et al., 2020).

The ongoing credibility crisis has sparked a widespread methodological discussion about our methods and means of interpreting data. Proposed reforms include a suggestion of preregistration of studies (Nosek et al., 2018; Strömmland, 2019), greater focus on replications (Coffman & Niederle, 2015), increased reliance on meta-analysis (Braver et al., 2014; Cumming, 2014), making statistical inferences more stringent (Benjamin et al., 2018), abandoning statistical significance altogether (Amrhein et al., 2019; McShane et al., 2019) and a shift towards Bayesian statistics (e.g. Etz & Vandekerckhove, 2016).

This paper suggests that some of the methodological disagreement among scholars discussing the credibility crisis can be viewed as fundamental disagreements about the appropriate “meta-hypothesis” – a hypothesis about the relationship among different effects in a literature. Currently, such meta-hypotheses are rarely made explicit by authors of empirical studies, but sometimes are in methodological papers. As a meta-hypothesis has implications for how to interpret the outcome of a replication study, this practice makes it hard to interpret replications clearly and makes it hard to prevent the canonization of false facts.

I suggest that the behavioral sciences should adopt a two-step procedure I dub the “limited homogeneity assumption”. Intuitively, the procedure could be thought of as a way of elevating the scientific status of direct replication studies. This two-step procedure uses a different meta-hypothesis in the early phase of a literature than in a more mature stage of the literature. In the early phase of a scientific literature – when an original claim is first formulated and demonstrated in an experimental study – a direct replication study needs to succeed before the literature can proceed to a more mature stage. In this early stage, one assumes no “population heterogeneity” (variation in the effect size due to variation in the study population), and no design variation if the original author has agreed that the planned design constitutes a direct replication study. Thus, following the limited homogeneity assumption, a replication intended as a direct replication is considered a proper “copy” of the original study and is therefore a potential falsifier of the overall claim in the early phase of the literature. When a replication succeeds, the field can adopt a meta-hypothesis assuming more heterogeneity and proceed to try to understand and revise the meta-hypothesis by understanding the form of heterogeneity. If the replication fails, the field assumes that there is nothing to be found in this literature.

The advantage of the limited homogeneity assumption lies in setting up a hurdle against the “canonization of false facts”, while still allowing us to discover real and heterogeneous effects when they exist. Thus, the limited homogeneity assumption can be defended from a “verisimilitude” or truthlikeness criterion (Popper, 1976; Oddie, 2016), which evaluates different scientific theories or hypotheses according to their relative closeness to the truth. For literatures with no or very few true effects, the limited homogeneity assumption procedure would lead to a conclusion that there is nothing of importance to be discovered. For literatures with a nontrivial number of genuine effects, it will typically lead to a conclusion that there is a genuine effect. For the alternative of assuming more heterogeneity even initially, one will discover genuine effects but there is no clear hurdle against the “canonization of false facts”. This alternative meta-hypothesis, which has its proponents (e.g., Yarkoni, 2020; McShane et al., 2019), therefore seems harder to defend.

This paper proceeds as follows. First, I explain the concept of a “meta-hypothesis”. Second, I discuss the extent to which one’s position on central meta-scientific issues can be traced to a meta-hypothesis. Then, I argue in favor of a “limited homogeneity assumption”. Finally, I discuss some potential objections to this proposal and conclude the paper.

2 The role of meta-hypotheses in research

We will define a scientific literature as a class of studies assumed to be related in some form, each class defined by a list of effect sizes or parameters each representing a true effect $\Theta_l = \{\theta_{1l}, \theta_{2l}, \dots, \theta_{kl}\}$, and assume these classes are clearly defined in the sense that there is agreement on what belongs in the same literature. We can distinguish between the effect size in a population, θ_l and the effect size from a random sample drawn from that population, $\hat{\theta}_i$. The population effect size is the target parameter in a single study – and we want to use methods that allow us to estimate this parameter without systematic error. We will generally think of observed effect sizes as estimates from randomized experiments, so

that $\hat{\theta}_i$ is an estimated effect of an experimental treatment and θ_i is the true effect in study i . In the following discussion, “effect size” will generally refer to the population effect of interest and not a single estimate from a particular experiment. Observed effect sizes in different experiments can differ both because of sampling error and systematic or random “heterogeneity” (variation in the effect resulting from changes in the population, study design or other, unpredictable factors). If there is no such heterogeneity, two studies with an infinite sample size would arrive at the same value of the effect size. If there is heterogeneity, then even with infinite sample sizes two studies would arrive at somewhat different values of the effect size.

One example of a literature constituting one class of effects studying the same phenomenon is the literature on the “watching eye effect”, on how visual cues suggesting someone is watching primes people to behave prosocially (Haley & Fessler, 2005). Different experimental designs and study populations used to test this overall hypothesis (e.g., Rigdon et al., 2009; Ekström, 2012; Sparks & Barclay, 2013) are members of the same class as they study the same phenomenon, but whether these different experiments measure the same underlying parameter is so far undefined.

The extent to which a single study has implications for other studies in the same class is determined by a “meta-hypothesis” concerning how the studies in the class are related to each other. A meta-hypothesis can be explicit, or it can be implicit, but the important feature is that one’s meta-hypothesis determines how we interpret the implications of the results of a study for other studies in the same class. Defining θ_i as the true effect in a random study in a literature indexed by i , μ as the mean effect in the literature, $f(X_i)$ a function of observable and unobservable characteristics which captures “systematic heterogeneity” (predictable sources of variation) and η_i a continuous random variable capturing residual or “random” heterogeneity (other, unpredictable, sources of variation in the effect), a meta-hypothesis is a model

$$\theta_i = \mu + f(X_i) + \eta_i \quad (1)$$

Notably, for each literature, there is an unknown true form of the Eq. (1) and the accuracy of a given meta-hypothesis may be judged according to how close it is to this true form.¹ A meta-hypothesis A can be said to have greater truthlikeness than a meta-hypothesis B if it is closer to the true form of Eq. 1 than B. There is a very large number of possible ways to specify a meta-hypothesis. On the extreme points, one finds an assumption of extreme homogeneity versus extreme heterogeneity. When both $f(X_i) = 0$ and $\eta_i = 0$ the meta-hypothesis is that there is one single effect in the literature, and all tests are tests concerning that parameter alone. This amounts to a very strict homogeneity assumption. Under this assumption, any test has broad and general implications about all others in the class. There would be no distinction between a direct replication and a conceptual replication. For the example of “watching eyes”, all studies using eye-like cues to study the effect on prosocial behavior, despite substantial variation in populations and experimental designs, are regarded as exact copies of each other – with

¹ While equation (1) is specified at the study level with “i” indexing the study, one could also specify a meta-hypothesis at a more detailed level, indexing “i” as the individual subject.

infinite sample sizes, they would be expected to arrive at the same effect size. Any study, supposing the sample is large enough, would have implications for all other studies in the class.

The problem with using this extreme homogeneity assumption as a meta-hypothesis is that it can quickly lead to contradictions, which subsequently makes it difficult to interpret studies in relation to each other. For instance, suppose we study a literature with a positive mean effect with some heterogeneity in the effect across different studies. In the example above from the literature on the “watching eyes” effect, under the extreme homogeneity assumption a success or failure of one of the different experiments would have the same implication for the underlying substantive claim as the meta-hypothesis suggests they are both measuring the same parameter. So, one could end in a situation where the same hypothesis is simultaneously both true and untrue, which implies a contradiction that would only be resolved by assuming that the effect under study is heterogeneous.

The other extreme is a strict “heterogeneity assumption”, where the effect is very context-sensitive, and depends on the function $f(X_i)$ as well as the random heterogeneity captured by the term η_i . Under this assumption, the null hypothesis is never true and could at best be an approximation. To see this, suppose that there is no mean effect ($\mu = 0$) and that $f(X_i) = 0$ for all values of X_i so that the heterogeneity across studies would be randomly determined from study to study. As η_i is a continuously distributed random variable the probability of η_i (and thus θ_i) taking a value of exactly zero is zero.² Thus, rejecting the null hypothesis is simply an issue of having a large enough sample size – with a large enough sample the null hypothesis would necessarily be rejected.

Between the extremes of an extreme homogeneity assumption and an extreme heterogeneity assumption lies a continuum where one can assume either a relatively high degree of homogeneity or a relatively high degree of heterogeneity. One’s placement on this scale will matter for one’s position on some important meta-scientific issues.

3 Meta-hypotheses and the “statistics wars”

A lot of the discussion in the wake of the “replication crisis” about what tools or methods to use to analyze data – what Deborah Mayo (2018) broadly refers to as the “statistics wars” – can be viewed as a discussion concerning what meta-hypothesis is typically most appropriate.

For instance, there has been a lot of debate concerning the appropriateness of null hypothesis significance testing in the social and behavioral sciences, with some arguing that we should lower our significance level (Benjamin et al., 2018), and others arguing that the use of “statistical significance” should be abandoned (Amrhein et al., 2019; McShane et al., 2019). One’s position on this issue logically follows from an underlying meta-hypothesis, though the meta-hypothesis itself may not be explicitly defended. If one believes there is little to no heterogeneity, null hypothesis significance

² if $f(X_i)$ took another value than zero one could also make a similar argument noting that the probability of $f(X_i)$ and η_i taking exactly the same value in opposite directions would be zero. This is because of the assumption that η_i is a continuous random variable, which has a probability of zero of taking a particular value.

testing makes sense, but if one believes extensive heterogeneity is the norm then significance testing can never be more than a strawman.

This link between meta-hypotheses and one's position on the role of null hypothesis significance testing is sometimes made explicit. For instance, proponents of abandoning null hypothesis testing often argue that testing a point null hypothesis is meaningless as it is highly unlikely. The reason why they state it is unlikely is that effect sizes in the biomedical and social sciences are "generally small and variable" (i.e., heterogeneous) (e.g., McShane et al., 2019). A recent paper by Wilson et al. (2020) also makes this connection. They argue that science is not a signal detection problem, in the sense that a more appropriate model of the world is to assume that true effect sizes follow a continuous distribution and not a binary "null versus alternative". They also cite Cohen (1990) on a statement that the null hypothesis is always false in the strict sense, as some variation across experiments will mean that the true effect under study is always non-zero. It is clear from this that they hold a meta-hypothesis where heterogeneity is very important, and under this meta-hypothesis we saw from the discussion concerning Eq. (1) that it immediately follows from an assumption of sufficient unpredictable heterogeneity that null hypothesis testing is meaningless and that the null hypothesis is always a strawman. Under this assumption, with an infinite sample size two studies are always expected to obtain different values of the same intended parameter.

In contrast, defenders of significance testing typically argue that null hypothesis significance testing is a useful approximation even in cases where the null is not literally true, but the effect is so small that it cannot be practically distinguished from zero. E.J. Wagenmakers recently made this point (Wagenmakers, 2017), and the argument dates back at least 50 years (e.g., Cornfield, 1966). This argument implies that its proponent places himself closer to "homogeneity" on the "meta-hypothesis scale". If heterogeneity is relatively small and negligible, the random and unpredictable error term in Eq. (1) can be effectively ignored as the world is "as if" the term was exactly zero. Thus, the meaningfulness of null hypothesis significance testing is conditional on a meta-hypothesis of a relatively large degree of homogeneity of effect sizes.

One's placement on the meta-hypothesis scale also matters for the view of what we can learn from a well-designed direct replication study. Starting from an assumption that effect sizes in a literature are highly heterogeneous, even with an infinite sample size, a highly powered and preregistered original study will give different results. This can happen if the replicators fail to implement the same systematic factors (X_i) as those present in the original study, but it can also happen if we succeed in copying all the design conditions of the original experiment, as some part of the heterogeneity in the effect cannot be directly controlled by the researcher (the "random heterogeneity" captured by the term η_i in Eq. 1). Thus, under this meta-hypothesis, a "direct replication" should necessarily "fail" according to conventional definitions of a successful replication (e.g., a statistically significant effect in the direction of the original study). One example of this view, and the connection between a meta-hypothesis with extensive heterogeneity and one's view of direct replication studies, is a paper by Stroebe and Strack (2014), who write that the "[...] alleged 'crisis of replicability' is primarily due to an epistemological misunderstanding that emphasizes the phenomenon instead of its underlying mechanisms. As a consequence, a replicated phenomenon may

not serve as a rigorous test of a theoretical hypothesis because identical operationalizations of variables in studies conducted at different times and with different subject populations might test different theoretical constructs.” These authors are here implicitly invoking a meta-hypothesis of extensive heterogeneity to downplay the value of direct replications. Under such an assumption, a “direct replication” is a meaningless term as one can never consider two studies “copies” of each other. Amrhein et al. (2019) make a related point, indicating that any systematic factor (e.g., failing to implement the same measurement in the replication and the original study) may give rise to variations in the effect size, thus implying that replication success is not necessarily to be expected. If replication success is never expected ahead of time, then one can never conclude from a failed replication that the original finding is wrong.

In contrast, under a meta-hypothesis assuming more homogeneity, a replication measures the same parameter as the original study, so the implication of a replication failure is clearly that the original statement is wrong – under a homogeneity assumption a direct replication can and will be considered a “copy” of the original study.

4 What meta-hypothesis should we adopt? The case for a limited homogeneity assumption

As meta-hypotheses are central to disagreements around the replication crisis, it would be very valuable if we could evaluate which meta-hypothesis is typically most appropriate. In the current research climate, meta-hypotheses are rarely made explicit except in purely methodological papers (e.g., Wilson et al., 2020). The disadvantage of keeping meta-hypotheses implicit or not stating them in original papers lies in what happens when direct replication attempts fail. When this happens and the meta-hypothesis for the literature is unclear, it is also unclear what a replication failure implies for the overall literature on the topic.

In many of the responses to recent pre-registered replication efforts, the original researchers explain the replication failure as resulting from some form of heterogeneity in the underlying effect size or methodological deviation in the replication that allegedly caused variation in the effect size (e.g., Amir et al., 2018; Lee & Schwarz, 2018; Rand, 2018; Strack, 2016). These arguments are consistent with the observed results, but problematic in the sense that in the papers subject to the replication attempt, the original authors often make general statements that go beyond their specific experimental design (Yarkoni, 2020). Thus, the formulations in the original paper may suggest an implicit meta-hypothesis of something quite close to homogeneity. Yet, when invoking heterogeneous effects as a defense against a replication failure, we cannot consider the original statements falsified, as the meta-hypothesis in the original paper is not stated. Thus, the problem is that currently a meta-hypothesis of heterogeneity works as a rescue hypothesis when replications fail. As meta-hypotheses are left unstated in original studies, we cannot hold original authors accountable for making this argument.

What meta-hypothesis is typically reasonable to adopt? Ideally, we want to pick an initial meta-hypothesis that will lead to long-run convergence to the true meta-hypothesis over time. However, all scientific hypotheses (meta-hypotheses or

ordinary hypotheses) are probably wrong, and it is doubtful that we will ever arrive at the exact true form of Eq. (1). Comparing different initial meta-hypotheses can therefore better be assessed from a truthlikeness or “verisimilitude” criterion (Popper, 1976; Oddie, 2016). We want to start with a meta-hypothesis that is likely to get us closer to the true meta-hypothesis than an alternative meta-hypothesis. “Closer” is hard to define, but one possibility is to think about “closeness” in terms of a Pareto principle (Oddie, 2016). An unknown fraction p of possible literatures will have (essentially) nothing to discover (no to practically no true effects in the list $\Theta_l = \{\theta_{1l}, \theta_{2l}, \dots, \theta_{kl}\}$), and another fraction of literatures $(1-p)$ will contain a nontrivial number of true effects. Suppose a meta-hypothesis A will lead to just as many true discoveries as meta-hypothesis B in the proportion $(1-p)$ of literatures, but that meta-hypothesis A will lead to less false discoveries for the other proportion p . In that case, from a truthlikeness perspective meta-hypothesis A will clearly be preferable as adopting it moves us closer to the truth in the proportion of literatures with nothing to discover but leaves us no further away from the truth for the rest of the $(1-p)$ literatures. In practice, there may be some tradeoffs between the ability to make true discoveries and the ability to avoid false discoveries. In that case, we could modify the truthlikeness criterion to be somewhat less strict: A meta-hypothesis A is considered more “truthlike” than meta-hypothesis B if it leads to a net decrease in the fraction of false conclusions in the two types of literatures. It is, however, immediately clear why this less strict criterion is more problematic; researchers will inevitably vary in how much they weight the two types of false conclusions.

4.1 Heterogeneity is problematic in an initial phase of a literature

From the point of view of the above truthlikeness criterion, it is problematic to assume a lot of heterogeneity in an early phase of a literature. This is because a hypothesis of extensive heterogeneity makes overall claims harder to falsify – in a setting with no effect in any setting (the null is generally true for the parameters in the same literature), starting from a heterogeneity assumption would not prevent the canonization of false facts. If another meta-hypothesis could be picked to prevent this from happening, while still allowing us to generally discover effects that are real, the meta-hypothesis could be said to be preferable from a truthlikeness perspective.

Consider again a scientific literature classed by a list of parameters measuring the same substantive effect, $\Theta_l = \{\theta_{1l}, \theta_{2l}, \dots, \theta_{kl}\}$. In the most extreme case, the effect is purely unpredictable due to random heterogeneity (in the model in Eq. 1 captured by the continuous random variable η_i), so it is expected that from study to study two teams testing the same hypothesis would arrive at different results even with infinite sample sizes. Moreover, under the assumption of a lot of heterogeneity, we do not necessarily expect a replicator to be able to pick the correct combination of systematic factors X_i necessary for a replication attempt to succeed. Therefore, even a study intended to be a direct replication cannot really be considered a direct replication in a strict sense.

When proceeding from an initial meta-hypothesis assuming a lot of heterogeneity it is hard to see how one can prevent the “canonization of false facts” – the process by which a false positive finding initiates a literature and the overall claim is eventually canonized as a “fact” (Nissen et al., 2016). A failure of a direct replication can under this assumption be explained away due to variation in design factors or population

factors that the replication researchers fail to consider, or some source of random (unpredictable) heterogeneity. Thus, even when a replication fails the overall claim cannot be considered falsified. Failed conceptual replications can be interpreted in a similar way, whereas successful conceptual replications can also just be interpreted as further inquiry into the form of Eq. (1). Notably, under such a heterogeneity assumption the distinction between a direct and conceptual replication is also somewhat blurred. The only way of making original claims falsifiable under an initial meta-hypothesis of extensive heterogeneity would be if original authors strictly specify the exact conditions under which the effect is expected to reliably appear. But the larger this list, the higher the probability is that at least one experimental design specified in the list (i.e., at least one “conceptual replication”) will come out as a false positive. If there really is no effect across conditions and populations, starting out with a meta-hypothesis assuming extensive heterogeneity will make it very difficult to reach a conclusion that there is no effect, with little to no variation across contexts.

For instance, the literature on “intuitive cooperation”, initiated by a study by Rand et al. (2012), was quickly followed by a high-powered failed replication study by Tinghög et al. (2013). However, the original authors had earlier suggested that the effect was expected to vary across individuals and contexts – an explicit meta-hypothesis of heterogeneity. This discussion was then followed by studies arguing for specific moderators of the effect such as subject experience with economic games (Rand et al., 2014) and trust in everyday institutions (Rand & Kraft-Todd, 2014). Subsequently, two pre-registered replication studies have failed to replicate two of the experiments in the original study by Rand et al. (2012) (Bouwmeester et al., 2017; Camerer et al., 2018), but David Rand (Rand, 2016; Rand, 2017, 2018) still argues that heterogeneity in the underlying effect size may account for replication failures and the literature keeps growing. Under the stated theoretical expectation that the effect will depend on individual-specific factors, the immediate replication failure (Tinghög et al., 2013) could not be considered as a falsifier of the overall claim. Even though the direct replications had very clear results, the explicit heterogeneity assumption early formulated by the original authors of the claim has clearly prevented this claim from being considered falsified.

One possible way to test for an overall lack of an effect when starting from a meta-hypothesis of extensive heterogeneity is to use a meta-analysis – a quantitative review of the literature – to summarize the evidence for the overall claim in the literature. One would then meta-analyze studies estimating effect sizes in some class Θ_i , imposing that the parameters in the class follow some probability distribution $f(\theta_{1i}, \theta_{2i}, \dots, \theta_{ki} | \mu, \tau^2)$, where μ is the mean of the distribution of true effects and τ^2 is the systematic variance (Higgins et al., 2009). A failure to reject the null hypothesis that $\mu = 0$ could be taken to falsify the claim of an overall positive but heterogeneous effect, as the statistical power would be very high for a meta-analysis, implying that a meta-analysis is very likely to discover a true underlying mean effect if present. A failed rejection of $\mu = 0$ in such a high-powered setting could be viewed as showing that on average, across different implementations of the hypothesis, there is no clear evidence that the treatment influences behavior. However, simulation studies indicate that conventional meta-analyses methods are very likely to produce false positive findings under realistic research conditions. Carter et al. (2019) show through different simulations assuming different degrees of questionable research practices and publication bias that the

conventional random effects meta-analysis method can produce a false positive rate of 98% or higher under certain conditions (high publication bias and no heterogeneity in effect sizes). They also find that across a broad set of simulation conditions and evaluation criteria, no meta-analytic method consistently outperforms the others. Recent empirical evidence (Kvarven et al., 2019) points in the same direction; comparing meta-analysis results to the results of pre-registered, many laboratory replication studies, they find that meta-analyses typically find much larger effect sizes than the replications. Interestingly, the meta-analysis typically finds a statistically significant effect in the same direction as the original study even in cases where the many-lab replication systematically fails. Kvarven et al. also compare different meta-analytic methods and similar to the simulation study by Carter et al. conclude that no meta-analytic method considered consistently outperforms the others.

Even if we ignore the problem of a potentially very high false positive rate of existing meta-analyses, if the null fails to be rejected and the meta-analysis is extremely well-powered, researchers could still maintain a belief that $\mu = 0$ and $\tau^2 > 0$ (which amounts to a specific form of the heterogeneity assumption) so that for some θ_{il} the effect would be nonzero, and that it is worth continuing to explore “moderators” of the hypothesized effect. When the heterogeneity assumption is maintained, the underlying substantive claim may thus never be refuted even under ideal conditions for conducting a meta-analysis. It is worth noting that an estimated $\tau^2 > 0$ is generally to be expected as the conventional estimators of this variance parameter are truncated (Ruhkin, 2013).

4.2 The “Limited homogeneity assumption”

The above discussion suggests that the behavioral sciences need a procedure that allows us to discover real (and possibly heterogeneous) effects when they exist, yet allow us not to prematurely reject the possibility that there is nothing of importance to be discovered in the literature on the topic considered. Our reason for this is that we want to prevent the scenario where an initially false positive finding may lead to the development of a full literature and the overall fact is over time eventually “canonized” as true (Nissen et al., 2016).

A particularly interesting candidate for such a procedure is a two-step procedure that I here dub the “limited homogeneity” assumption, which adopts a different meta-hypothesis in the initial phase of the literature than in the subsequent stage of the literature. In the initial phase of the literature, direct replication studies are regarded as falsifiers of the overall claim. This procedure assumes that when a high-powered direct replication is carried out, there is no “population heterogeneity” (variation in the effect size due to demographic factors uncontrolled by the researchers). We allow for “design heterogeneity” (variation in the effect size due to changes in the experimental design), but as long as the original author has agreed on the experimental design before the replication is carried out, one assumes a replication failure cannot be explained away as resulting from such design heterogeneity. This type of replication study has been carried out previously (e.g., Hagger et al., 2016; Bouwmeester et al., 2017), and if the limited homogeneity assumption is adopted, such “registered replication” studies become the gold standard for evaluating whether a research literature is allowed to proceed to a more mature stage. The limited homogeneity assumption is “limited” because in the initial phase of the literature we remain agnostic about whether or not to

adopt a homogeneity assumption or a heterogeneity assumption in a subsequent stage, and it is a homogeneity assumption because we assume that a direct replication study is a proper “copy” of the original study.

Following the above procedure, if a direct replication of the agreed upon replication design subsequently fails, the original claim is considered falsified, and one proceeds as if there is no effect to be detected in this literature – i.e., switching to an extreme homogeneity assumption. But if a direct replication succeeds, we change our meta-hypothesis in the literature to an assumption that there is a real and potentially heterogeneous effect (potentially both systematic and random heterogeneity) as captured in the general model in Eq. (1) and proceed to try to understand exactly what factors explain variation in the effect. Thus, we use the replication study as the means of evaluating whether a literature should proceed to a heterogeneity assumption or not.

From a “truthlikeness” perspective, the “limited homogeneity assumption” is attractive compared to an initial meta-hypothesis of extensive heterogeneity, as it will guard against the canonization of false facts while still allowing us to make discoveries in literatures with true effects. If literatures with no effect, universally (or approximately universally) across contexts, are sufficiently common, the limited homogeneity assumption may be argued to be preferable from a truthlikeness criterion. If there really is no effect, then under the limited homogeneity assumption a replication attempt will typically fail and the field will adopt an extreme homogeneity assumption, assuming there is no effect in this literature. Specifically, if there really is no effect, so that the class of effect sizes can be written $\Theta_l = \{\theta_{1l}, \theta_{2l}, \dots, \theta_{kl}\} = \{\theta_l\} = \{0\}$, and an original study claims that there is an effect, for the conventional criterion of “replication success” (a statistically significant effect in the same direction as the original study), a direct replication attempt is only expected to succeed 2.5% of the time. Thus, the literature would rarely be able to proceed to a more mature stage.

However, if there really is an underlying stable (and possibly heterogeneous effect), a high-powered replication is expected to succeed, and the literature will subsequently quickly proceed to a more mature stage – assuming effect size heterogeneity and trying to learn about the form of heterogeneity. This contrasts with an approach that starts out assuming a lot of effect size heterogeneity – as we saw in section 4.1., if there really is no effect to be discovered, it will be hard to reach this conclusion when starting out from a meta-hypothesis of extensive heterogeneity.

Adopting a limited homogeneity assumption amounts to giving replications a very clear and important role in the behavioral sciences: a replication is a hurdle to pass before a literature can proceed to a mature stage. Perhaps more importantly, the limited homogeneity assumption works against the “canonization of false facts” (Nissen et al., 2016) – the process by which an initially false positive finding starts a literature and the substantive claim never gets falsified or takes decades to falsify. If there really is no effect, this hurdle will make it harder for dubious claims to get accepted and for a literature to get going on flawed foundations. This is a likely explanation for what happened in the “ego depletion” literature, where after almost two decades of reported positive findings a many-laboratory replication study found no ego depletion effect across different labs (Hagger et al., 2016). Had a limited homogeneity assumption been adopted from the start, it is unlikely that the literature would have proceeded like this, as a direct replication of the original ego depletion design would most likely not have replicated in the first place. While we cannot go back in time and prevent this from

occurring in the first place, we can adopt procedures to prevent it from happening again in the future. It could be that the limited homogeneity assumption brings with it the costs of fewer true discoveries. In that case, the truthlikeness criterion used in this paper cannot be used to argue in favor of it as there would be a clear cost to adopt the limited homogeneity assumption compared to an alternative initial meta-hypothesis assuming more heterogeneity. Theoretically, it seems reasonable that “true findings” will be canonized as true – for instance, Nissen et al. (2016) show that true findings are almost universally canonized as true under a wide range of simulation conditions based on a mathematical model. However, theoretical support is not sufficient to defend the limited homogeneity assumption from a “truthlikeness” perspective – we also need to know whether there is an empirical basis for adopting it. Specifically, can we expect that direct replications will typically succeed in literatures with mostly true effects? If this is the case, it would support the idea that there would be few costs involved in adopting the limited homogeneity assumption procedure.

Importantly, there is growing evidence that effect size heterogeneity, at least for direct replication studies, is relatively unimportant. This is particularly true for failed replications. Several papers have used the “many lab” studies, where many different laboratories across study population replicate some psychological effect (Klein et al., 2014; Klein et al., 2018; Ebersole et al., 2016) to argue that heterogeneity is universal in the behavioral science literature. The argument is that even in these replication studies where the experimental design is kept constant, there is nontrivial heterogeneity, suggesting that heterogeneity is likely to be much bigger once we let design factors vary. Some have argued based on the many labs studies that these studies show that a non-negligible “lower bound” on heterogeneity is substantial. This is based on inspection of the so-called I^2 which measures the fraction of variation attributable to factors other than chance (Higgins & Thompson, 2002). The Many Labs project found statistically significant heterogeneity in more than half of the replication studies, and many of the I^2 measures suggest moderate or large levels of heterogeneity (particularly for the “anchoring” replications). However, Simonsohn (2017) shows through permutation tests that the observed I^2 measures for the numerical anchoring replication in the many labs study are in line with what one would expect from chance alone. Moreover, it is worth noting that the I^2 measure does not measure the magnitude of heterogeneity and depends on the sample size. Even very small heterogeneity in the effect size will lead the I^2 measure to approach 100% as the sample size increases without bounds, and a study by Rücker et al. (2008) finds that as precision increases the I^2 measure rapidly approaches 100%. Finally, in the many labs study the large I^2 measures were typically found for cases where the replication succeeded, and not for the cases with smaller or no replication effect (Klein et al., 2014), supporting a homogeneity assumption as a good approximation for studies with no apparent effect.

In the “Many Labs 2” (Klein et al., 2018) study, the average between-study standard deviation (τ) is reported, which unlike the I^2 measure measures the magnitude of heterogeneity in the effect size. The between-study standard deviation was 0.04 on average and 19 out of 28 studies had an estimated τ of zero. Further, only one of the replication studies with no evidence for an aggregate effect found significant heterogeneity, and there was even little systematic variation in the effect size between

WEIRD (Western, Educated, Industrialized, Rich and Democratic) countries and less WEIRD countries. A particularly interesting finding in the Many Labs 2 study is that the estimated between-study variation seems to be more relevant in cases where a replication succeeded, and not in cases where the replication failed. This is in line with the limited homogeneity assumption advocated in this paper when starting a scientific literature; from the Many Labs 2 evidence, it seems that if an effect is not present in one study population, this quite well predicts null findings in other study populations, and heterogeneity generally does a poor job in explaining replication failures. It is also noteworthy that the between-study standard deviation measure is a truncated estimator so by construction it has a built-in positive bias (Rukhin, 2013), indicating that if anything the between-study standard deviation may be overestimated in the Many Labs 2 study. This bias will be negligible in settings where there is large heterogeneity but will be more relevant in cases where the effect size is homogeneous.³

The main takeaway from the above empirical studies seems to be that population-based sources of variation in effect sizes seems limited exactly in the way that would favor the limited homogeneity assumption. When effect sizes are real, they are typically real but somewhat variable across contexts. However, the direct replication typically succeeds across laboratories (suggesting that a true effect will typically be corroborated in a well-powered direct replication study). When there is no effect, there typically is no effect across contexts. Thus, empirical evidence suggests that adopting the limited homogeneity assumption would not involve substantial costs from a truthlikeness perspective when compared against an alternative meta-hypothesis assuming more heterogeneity even in an early stage of a literature – true effects are likely to be corroborated under either of these initial meta-hypotheses. When there is no effect, adopting the limited homogeneity assumption initially will lead us to stay where we are (i.e., not allowing the literature to get going based on a false positive, and the overall claim will be considered falsified). When there is a real effect, a replication will typically succeed, and we will revise our meta-hypothesis to one of a real and possibly heterogeneous effect and proceed to try to model this variation better over time.

5 Possible objections

5.1 The “generalizability crisis”

A recent paper by Yarkoni (2020) argues that the replication crisis in the behavioral sciences is a consequence of a larger “generalizability crisis”. This crisis consists in the fact that original authors tend to make sweeping and bold claims about the generalizability of their findings. However, Yarkoni holds that effect sizes are highly variable so that such generalizations are unwarranted. The fact that replications tend to fail is an expression of this intrinsic variation in effect sizes.

The model Yarkoni fits to different data sets to illustrate his points includes a subject-level error term that is interacted with a treatment indicator to allow for

³ Note that this truncation would affect the I^2 measure as well, so with a large dataset and homogeneous effects one would expect this measure to be higher than 0%.

subject-specific slopes. This model is consistent with the general model in Eq. (1) of this paper where there are both deterministic and unpredictable sources of variation in effect sizes – Yarkoni’s model amounts to positing a specific meta-hypothesis where there is subject-specific treatment effects and no other modelling sources of heterogeneity. Thus, his arguments can be interpreted in light of the concept of a meta-hypothesis as defined in the present paper.

In terms of this terminology, Yarkoni is essentially positing a meta-hypothesis close to the extreme heterogeneity assumption as his model does not feature any predictable sources of variation in effects beyond individual-specific random subject effects. If he is correct, it is meaningless to adopt a limited homogeneity assumption in an initial phase of a literature, as it would be wasting time on testing an essentially straw man hypothesis. There are no true null effects – effect sizes are inherently variable and should be modeled accordingly from the start. Thus, carrying out replications and regarding them as falsifiers of an overall claim would also be a waste of time as a claim can never be general in the first place – and direct replication studies are not to be regarded as “copies” of the original study.

Yarkoni’s paper serves to illustrate the point made in Section 4 of this paper about the difficulty in falsifying meta-hypotheses assuming extensive heterogeneity. The problem with Yarkoni’s approach is that he assumes without any empirical support that his meta-hypothesis is true. However, if there really is no effect (and it does not vary over individuals), starting from Yarkoni’s modelling approach will never allow us to confirm it. Fitting the model that Yarkoni fits to various data sets (interacting a treatment variable with a normally distributed subject-level random effect) will mechanically increase noise in the estimation as adding these random effects subtract some subjects from the estimation of the overall treatment effects. The addition of these terms to his model will also probably fit a given data set better than a model that assumes no heterogeneity. But we do not know the true form of Eq. (1) and we can either assume *a priori* that we know whether effects in a literature are heterogeneous or not, or we can adopt a procedure that allows us to discover both possibilities. As we have seen in section 4, much evidence already suggests that systematic heterogeneity in effect sizes is lower than many have suggested, and that heterogeneity is particularly limited precisely in settings where replications typically fail. It therefore seems useful to adopt a limited homogeneity assumption in the initial phase of a literature as it guards against the possibility of letting a literature with no true effects proceed to a mature stage and allow errors to cumulate over time.

Note that proponents of a heterogeneity assumption would have no problem with the limited homogeneity assumption in the sense that if there really is a real and variable effect, one could easily pass the hurdle by conducting a successful replication study. One would then freely be able to adopt a meta-hypothesis of heterogeneity and proceed to model its sources in any way one might prefer. However, people who believe that there are literatures where there is nothing to be discovered would never be able to strictly refute Yarkoni’s heterogeneity assumption were it to be false – replication failures could never work against this assumption as the meta-hypothesis denies the possibility of a “direct replication” in the first place. For practical purposes, the limited homogeneity assumption therefore seems like a reasonable solution.

5.2 Potential problem of false negatives

One further issue with adopting the limited homogeneity assumption as a default is that it could lead to false negatives. For instance, for a literature with mostly real effects, original authors initiating the literature could pick a design unlikely to replicate (a true null). This could be the case if the distribution of true effects in a literature is wide and overlaps with zero, and original authors are unlucky or unskilled in picking their experimental design. A replication study will then fail, but it should not have failed, as most designs in this literature would be replicable. However, it seems unlikely that this objection would often hold in practice as one would expect original authors to aim for designs that are more likely to succeed, for instance by going for “low-hanging fruits” when making an initial discovery.

Ultimately, one could judge how much heterogeneity to allow in the initial phase of the literature – that is, how many different original designs that can be subjected to a direct replication before deciding whether the literature should be allowed to proceed to a more mature stage or not – by the relative importance of the topic studied. For medical studies, for instance, one could argue that the cost of failing to discover some context-dependent effects is very large, even when the distribution of true effects in the literature is wide and overlaps with zero. The reason would be that some of these effects could potentially be lifesaving. However, the relative cost of missing some positive discoveries in the behavioral sciences seems less obvious, and for robust and important phenomena, the effect will most likely consistently replicate across contexts. An example in this regard is a recent replication of prospect theory (Kahneman & Tversky, 1979) where the effect robustly appeared in almost all replication teams (Ruggeri et al., 2020).

5.3 The meaning of a “replication success” is controversial

Perhaps the most important objection to my proposal of following a limited homogeneity assumption is that the proposed two-step procedure assumes that there is agreement on what constitutes a replication success. There is considerable controversy on this topic, and many of the most well-known replication studies use several different criteria of “replication success” and report results for each of these criteria. One possible way to circumvent this issue lies in the requirement that the original author must agree to the planned replication design for the replication to be considered a proper copy of the original study. We could use the same procedure for the criterion used to define a replication success – requiring the original author to specify in advance what the criterion would be for considering the replication successful or not. If this is unclear in advance, the outcome of a replication would spark disagreement even among researchers committed to the idea of following the limited homogeneity assumption to decide on which meta-hypothesis to proceed to.

6 Conclusion

I have argued that much of the discussion around the replication crisis can be seen as a discussion over overarching “meta-hypotheses” – hypotheses about how different effects in a literature are related to each other. Currently, such hypotheses are rarely made explicit except in methodological papers. This means that currently, the scientific status of direct

replication studies is somewhat unclear – a replication failure typically cannot be interpreted very clearly as we cannot hold original authors accountable for stating that the replication failure was due to some form of systematic variation in the effect size unaccounted for in the replication. It also means that it is hard to prevent the “canonization of false facts” (Nissen et al., 2016), as original claims are hard to falsify when the meta-hypothesis for the field is unclear to begin with. I suggest that the behavioral sciences should adopt a two-step procedure dubbed the “limited homogeneity assumption” to circumvent these issues. This procedure would consider direct replication studies as falsifiers of overall claims in the initial phase of a scientific literature. Once a claim has been put forth, a direct replication needs to corroborate the finding for the literature to proceed to a more mature stage and study the source of possible effect size heterogeneity. If the direct replication fails, the field proceeds to an assumption that there is no effect in this literature. If the direct replication succeeds, the field proceeds to an assumption of a possibly heterogeneous and real effect and try to understand the sources of variation in the effect.

The advantage of this two-step approach is that it will allow us to discover real and possibly heterogeneous effects when present but will also guard against the possible problem that entire literatures may get going under an assumption of heterogeneous effect sizes when the null hypothesis is universally (or approximately universally) true. This two-step procedure seems reasonable based on observed heterogeneity in the recent Many Labs 2 study (Klein et al., 2018) which finds that heterogeneity is typically prevalent in cases where the replication succeeds but less so in settings where the replication fails. In this respect, the limited homogeneity assumption seems attractive when considered from a “truthlikeness” (Popper, 1976) criterion. It seems realistic that the “limited homogeneity assumption” procedure will imply greater truthlikeness, at least on average, than if starting from an initial meta-hypothesis allowing for more heterogeneity already from the start.

The limited homogeneity assumption approach takes a “Popperian” stance and elevates a pre-registered, high-powered direct replication study to the role of a falsifier of an overall claim in an initial phase of a scientific literature. A recent paper by Nosek and Errington (2020), which also deals with how to treat replications in science, takes a more “Lakatosian” stance. By their account, one can distinguish between a progressive and degenerative research programme by judging whether replication studies keep succeeding in replicating findings across conditions not deemed relevant by theory (progressive), or persistently fail to replicate findings and thereby narrow the conditions under which the theory in question applies (degenerative). Their procedure therefore offers an alternative to the limited homogeneity assumption as one could imagine researchers adhering to this Lakatosian approach would eventually learn what literatures are “degenerative” and what literatures are not. However, a problematic feature of Lakatos’ philosophy of science, as pointed out by among others Paul Feyerabend, is that it does not allow us to make any normative prescriptions about what to do when faced with a degenerative research programme (e.g., Musgrave & Pigden, 2016). If researchers admit to the degenerative state, they can still cling to their current procedures and claim, validly, that to do so is rational. The “modified” Lakatosian approach applied to replication studies has the risks of falling into the same trap. One could imagine that replications are treated as suggested by Nosek & Errington, but that the supporters of a meta-hypothesis that the effect is real but heterogeneous, could just keep acknowledging new replication failures, yet form hypotheses about novel conditions under which an effect may still appear. To do this, potentially indefinitely, would not be

irrational; no clear normative prescription to abandon a degenerative programme follows from the Lakatosian approach alone.

While adopting a limited homogeneity assumption would not eliminate problems, it would help us guard against the problem of cumulating errors in the behavioral sciences (the “canonization of false facts”) by setting up a hurdle for researchers to pass for a literature to proceed to a more mature stage. It would also give direct replications a clear and well-defined role in the development of a scientific literature. Following the limited homogeneity assumption, the format of a “registered replication report” (Simons et al., 2014) becomes the gold standard for determining whether a scientific field should proceed to a more mature stage or abandoned.

The current paper implicitly assumes that the “limited homogeneity assumption” procedure could be easily adopted and coordinated upon across researchers. However, in practice it may be unlikely to get researchers to coordinate on this procedure as there would be no clear incentives to adopt it. A theoretical investigation of how a research community could coordinate on such a norm is an interesting topic for future research.

Funding Open access funding provided by Western Norway University Of Applied Sciences.

Code availability Not applicable.

Data availability Not applicable.

Declarations

Ethical approval Not applicable (as no data was collected, no ethical approval was necessary by a local or national ethics board).

Informed consent Not applicable.

Conflict of interest The author declares no conflict of interest.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Amir, O., Mazar, N., & Ariely, D. (2018). Replicating the effect of the accessibility of moral standards on dishonesty: Authors' response to the replication attempt. *Advances in Methods and Practices in Psychological Science*, 1(3), 318–320.
- Amrhein, V., Greenland, S., & McShane, B. (2019). Scientists rise up against statistical significance. *Nature*, 567, 305–307.
- Baumeister, R. F., Bratslavsky, E., Muraven, M., & Tice, D. M. (1998). Ego depletion: Is the active self a limited resource? *Journal of Personality and Social Psychology*, 74, 1252–1265. <https://doi.org/10.1037/0022-3514.74.5.1252>.

- Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E. J., Berk, R., et al. (2018). Redefine statistical significance. *Nature Human Behaviour*, 2(1), 6–10.
- Bouwmeester, S., Verkoeijen, P. P., Aczel, B., Barbosa, F., Bègue, L., Brañas-Garza, P., et al. (2017). Registered replication report: Rand, Greene, and Nowak (2012). *Perspectives on Psychological Science*, 12(3), 527–542.
- Braver, S. L., Thoenes, F. J., & Rosenthal, R. (2014). Continuously cumulating meta-analysis and replicability. *Perspectives on Psychological Science*, 9(3), 333–342.
- Camerer, C. F., Dreber, A., Forsell, E., Ho, T. H., Huber, J., Johannesson, M., et al. (2016). Evaluating replicability of laboratory experiments in economics. *Science*, 351(6280), 1433–1436.
- Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T. H., Huber, J., Johannesson, M., 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.
- Carter, E. C., Schönbrodt, F. D., Gervais, W. M., & Hilgard, J. (2019). Correcting for bias in psychology: A comparison of meta-analytic methods. *Advances in Methods and Practices in Psychological Science*, 2(2), 115–144.
- Coffman, L. C., & Niederle, M. (2015). Pre-analysis plans have limited upside, especially where replications are feasible. *Journal of Economic Perspectives*, 29(3), 81–98.
- Cohen, J. (1990). Things I have learned (thus far). *American Psychologist*, 45, 1304–1312.
- Cornfield, J. (1966). A Bayesian test of some classical hypotheses—with applications to sequential clinical trials. *Journal of the American Statistical Association*, 61(315), 577–594.
- Cumming, G. (2014). The new statistics: Why and how. *Psychological Science*, 25(1), 7–29.
- Ebersole, C. R., Atherton, O. E., Belanger, A. L., Skulborstad, H. M., Allen, J. M., Banks, J. B., et al. (2016). Many labs 3: Evaluating participant pool quality across the academic semester via replication. *Journal of Experimental Social Psychology*, 67, 68–82.
- Ekström, M. (2012). Do watching eyes affect charitable giving? Evidence from a field experiment. *Experimental Economics*, 15(3), 530–546.
- Etz, A., & Vandekerckhove, J. (2016). A Bayesian perspective on the reproducibility project: Psychology. *PLoS One*, 11(2), e0149794.
- Hagger, M. S., Chatzisarantis, N. L., Alberts, H., Anggono, C. O., Batailler, C., Birt, A. R., et al. (2016). A multilab preregistered replication of the ego-depletion effect. *Perspectives on Psychological Science*, 11(4), 546–573.
- Haley, K. J., & Fessler, D. M. (2005). Nobody's watching?: Subtle cues affect generosity in an anonymous economic game. *Evolution and Human Behavior*, 26(3), 245–256.
- Higgins, J. P., & Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine*, 21(11), 1539–1558.
- Higgins, J. P., Thompson, S. G., & Spiegelhalter, D. J. (2009). A re-evaluation of random-effects meta-analysis. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 172(1), 137–159.
- Ioannidis, J. P. (2012). Why science is not necessarily self-correcting. *Perspectives on Psychological Science*, 7(6), 645–654.
- Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decisions under risk. *Econometrica*, 47, 278.
- Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Jr., Bahník, Š., Bernstein, M. J., & Nosek, B. A. (2014). Investigating variation in replicability: A “many labs” replication project. *Social Psychology*, 45(3), 142–152.
- Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Adams, R. B., Jr., Alper, S., et al. (2018). Many labs 2: Investigating variation in replicability across samples and settings. *Advances in Methods and Practices in Psychological Science*, 1(4), 443–490.
- Kvarven, A., Strömberg, E., & Johannesson, M. (2019). Comparing meta-analyses and preregistered multiple-laboratory replication projects. *Nature Human Behaviour*, 4, 423–434. <https://doi.org/10.1038/s41562-019-0787-z>.
- Lee, S. W., & Schwarz, N. (2018). Methodological deviation from the original experiment. *Nature Human Behaviour*, 2(9), 605–605.
- Mayo, D. G. (2018). *Statistical inference as severe testing*. Cambridge: Cambridge University Press.
- McShane, B. B., Tackett, J. L., Böckenholt, U., & Gelman, A. (2019). Large-scale replication projects in contemporary psychological research. *The American Statistician*, 73(sup1), 99–105.
- Musgrave, Alan and Charles Pigden (2016). Imre Lakatos. Edward N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (winter 2016 edition), URL = <<https://plato.stanford.edu/archives/win2016/entries/lakatos/>>.
- Nissen, S. B., Magidson, T., Gross, K., & Bergstrom, C. T. (2016). Publication bias and the canonization of false facts. *Elife*, 5, e21451.
- Nosek, B. A., & Errington, T. M. (2020). What is replication? *PLoS Biology*, 18(3), e3000691.
- Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences*, 115(11), 2600–2606.

- Oddie, Graham (2016). Truthlikeness. Edward N. Zalta (ed.), *The Stanford encyclopedia of philosophy* (winter 2016 edition). URL = <<https://plato.stanford.edu/archives/win2016/entries/truthlikeness/>>
- Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, 349(6251), aac4716.
- Popper, K. (1976). A note on verisimilitude. *The British Journal for the Philosophy of Science*, 27(2), 147–159.
- Rand, D. G. (2016). Cooperation, fast and slow: Meta-analytic evidence for a theory of social heuristics and self-interested deliberation. *Psychological Science*, 27(9), 1192–1206.
- Rand, D. G. (2017). Reflections on the time-pressure cooperation registered replication report. *Perspectives on Psychological Science*, 12(3), 543–547.
- Rand, D. G. (2018). Non-naïvety may reduce the effect of intuition manipulations. *Nature Human Behaviour*, 2(9), 602.
- Rand, D. G., & Kraft-Todd, G. T. (2014). Reflection does not undermine self-interested prosociality. *Frontiers in Behavioral Neuroscience*, 8, 300.
- Rand, D. G., Greene, J. D., & Nowak, M. A. (2012). Spontaneous giving and calculated greed. *Nature*, 489(7416), 427–430.
- Rand, D. G., Peysakhovich, A., Kraft-Todd, G. T., Newman, G. E., Wurzbacher, O., Nowak, M. A., & Greene, J. D. (2014). Social heuristics shape intuitive cooperation. *Nature Communications*, 5(1), 1–12.
- Rigdon, M., Ishii, K., Watabe, M., & Kitayama, S. (2009). Minimal social cues in the dictator game. *Journal of Economic Psychology*, 30(3), 358–367.
- Rücker, G., Schwarzer, G., Carpenter, J. R., & Schumacher, M. (2008). Undue reliance on I 2 in assessing heterogeneity may mislead. *BMC Medical Research Methodology*, 8(1), 79.
- Rukhin, A. L. (2013). Estimating heterogeneity variance in meta-analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(3), 451–469.
- Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, 22(11), 1359–1366.
- Ruggeri, K., Ali, S., Berge, M. L., Bertoldo, G., Bjørndal, L. D., Cortijos-Bernabeu, A., et al. (2020). Replicating patterns of prospect theory for decision under risk. *Nature Human Behaviour*, 4, 622–633. <https://doi.org/10.1038/s41562-020-0886-x>
- Simons, D. J., Holcombe, A. O., & Spellman, B. A. (2014). An introduction to registered replication reports at perspectives on psychological science. *Perspectives on Psychological Science*, 9(5), 552–555.
- Simonsohn (2017). “Many Labs” overestimated the importance of hidden moderators. Retrieved from: <http://datacolada.org/63>
- Sparks, A., & Barclay, P. (2013). Eye images increase generosity, but not for long: The limited effect of a false cue. *Evolution and Human Behavior*, 34(5), 317–322.
- Strack, F. (2016). Reflection on the smiling registered replication report. *Perspectives on Psychological Science*, 11(6), 929–930.
- Strack, F., Martin, L. L., & Stepper, S. (1988). Inhibiting and facilitating conditions of the human smile: A nonobtrusive test of the facial feedback hypothesis. *Journal of Personality and Social Psychology*, 54(5), 768–777.
- Stroebe, W., & Strack, F. (2014). The alleged crisis and the illusion of exact replication. *Perspectives on Psychological Science*, 9(1), 59–71.
- Strömmland, E. (2019). Preregistration and reproducibility. *Journal of Economic Psychology*, 75, 102143.
- Tinghög, G., Andersson, D., Bonn, C., Böttiger, H., Josephson, C., Lundgren, G., Västfjäll, D., Kirchler, M., & Johannesson, M. (2013). Intuition and cooperation reconsidered. *Nature*, 498(7452), E1–E2.
- Wagenmakers, E.J. (2017). Redefine statistical significance part IX: Gelman and Robert join the fray, but are quickly chased by two kangaroos. Retrieved from: <https://www.bayesianspectacles.org/redefine-statistical-significance-part-ix-gelman-and-robert-join-the-fray-but-are-quickly-chased-by-two-kangaroos/>
- Wagenmakers, E. J., Beek, T., Dijkhoff, L., Gronau, Q. F., Acosta, A., Adams, R. B., Jr., et al. (2016). Registered replication report: Strack, Martin, & Stepper (1988). *Perspectives on Psychological Science*, 11(6), 917–928.
- Wilson, B. M., Harris, C. R., & Wixted, J. T. (2020). Science is not a signal detection problem. *Proceedings of the National Academy of Sciences*, 117(11), 5559–5567.
- Yang, Y., Youyou, W., & Uzzi, B. (2020). Estimating the deep replicability of scientific findings using human and artificial intelligence. *Proceedings of the National Academy of Sciences*, 117(20), 10762–10768.
- Yarkoni, T. (2020). The generalizability crisis. *Behavioral and Brain Sciences* Advance online publication. <https://doi.org/10.1017/S0140525X20001685>.