Means to valuable exploration:

I. The blending of confirmation and exploration and how to resolve it

Michael Höfler[1,2]

Stefan Scherbaum[1,3]

Philipp Kanske[1,2,4]

Brennan McDonald[1,2]

Robert Miller[1]

[1]: Faculty of Psychology, Technische Universität Dresden, Dresden, Germany. [2]: Clinical Psychology and Behavioural Neuroscience, Institute of Clinical Psychology and Psychotherapy, Technische Universität Dresden, [3]: Institute of General Psychology, Biopsychology, and Psychological Research Methods, [4]: Max Planck Institute for Human Cognitive and Brain Sciences, Leipzig, Germany,

Corresponding author: Michael Höfler, Chemnitzer Straße 46, Clinical Psychology and Behavioural Neuroscience, Institute of Clinical Psychology and Psychotherapy, Technische Universität Dresden, 01187 Dresden, Germany. michael.hoefler@tu-dresden.de, +49 351 463 36921

Submitted to *Meta-Psychology*, *2nd revision*

Abstract

Data exploration has enormous potential to modify and create hypotheses, models, and theories. Harnessing the potential of *transparent* exploration replaces the common, flawed purpose of intransparent exploration: to produce results that appear to confirm a claim by hiding steps of an analysis. For transparent exploration to succeed, however, methodological guidance, elaboration and implementation in the publication system is required. We present some basic conceptions to stimulate further development. In this first of two parts, we describe the current blending of confirmatory and exploratory research and propose how to separate the two via *severe testing*. A claim is confirmed if it passes a test that probably would have failed if the claim was false. Such a severe test makes a risky prediction. It adheres to an evidential norm with a threshold, usually $p < \alpha = .05$, but other norms are possible, for example, with Bayesian approaches. To this end, adherence requires control against questionable research practices like p-hacking and HARKing. At present, preregistration seems to be the most feasible mode of control. Analyses that do not adhere to a norm or where this cannot be controlled should be considered as exploratory. We propose that exploration serves to modify or create new claims that are *likely to pass severe testing with new data.* Confirmation and exploration, if sound and transparent, benefit from one another. The second part will provide suggestions for planning and conducting exploration and for implementing more transparent exploratory research.

Keywords

Exploration, confirmation, p-hacking, HARKing, preregistration, severity, replication, bias, Bayes

Introduction

Degrees of freedom in the specification of hypotheses and analyses are both a curse and a blessing. In improperly performed confirmatory analyses, they are misused to disguise incidental findings as evidence for a hypothesis; whereas in exploratory analyses they open ways to new insights (Thompson et al., 2020). Questionable research practices (QRPs) like p-hacking and "hypothesising after the results are known" (HARKing; Hollenbeck & Wright, 2017; Rubin, 2017) misuse degrees of freedom to produce nominally confirmatory results (p-value $< \alpha$). At the same time, results presented as confirmatory are more likely to be published (Francis, 2012; Gigerenzer & Marewski, 2015; Masicampo & Lalande, 2012; Rosenthal, 1979; Scargle, 2000). Both practices disrupt scientific communication by introducing findings that are not sufficiently substantiated by evidence to the literature. Moreover, results arising from QRPs are less likely to be replicated, such that p-hacking and HARKing are considered significant causes of the replication crisis: the failure to replicate many established experimental psychological findings (Head et al., 2015; but also see Lewandowsky & Oberauer, 2020; Ulrich & Miller, 2020).

Confirmation means that an evidential norm with a defined threshold (usually $p < \alpha$) is applied. It must be strictly adhered to without being affected by the data. In contrast, sloppy confirmation occurs when a scientist trawls through a dataset with various options on *how* to test a claim (Gelman & Loken, 2013) and cherry picks the data (e.g. the smallest p-value) while hiding the other results, thus missing the rigour required of confirmation (p-hacking). Or, with intransparent HARKing a new hypothesis is generated around one preferred result out of the many that were generated. It is presented as confirmed while suppressing the other results. By beginning with the flawed and improper intention of confirming hypotheses in this manner, one hiddenly engages in exploration.

Confirmation, on the other hand, is a constrained process by which established scientific concepts are held up to empirical scrutiny through precise prediction, study design, and analysis planning. In other words, the “researcher degrees of freedom” (Simonsohn et al., 2020) available during confirmation are intentionally restricted to provide evidence for (or against) a clearly stated hypothesis.

Basic concepts

In contrast to confirmation, transparent or “open exploration” (Thompson et al., 2020) embraces the degrees of freedom during the analysis to potentially reveal something of substantial interest within the data (Dirnagl, 2020). Exploration, when done transparently, thus allows a free inquiry into the behaviour of a dataset without preconceived delineations as to what patterns it shows (Thompson et al., 2020). Transparent exploration seems to be very rare (Gigerenzer & Marewski, 2015) and its potential to identify novel insights rarely utilised. In contrast, intransparent exploration for the purpose of confirmation appears to be worryingly common (Agnoli et al., 2017; Gopalakrishna et al., 2021a; John et al., 2012; Kerr, 1998). Our aim is to call upon transparent exploration’s potential and to revive this alternative approach to science. To this end, we present some basic concepts and methodical considerations to stimulate further in-depth and detailed elaborations.

We use the term "exploration" as referring to a toolbox of analytical methods to *generate and modify* hypotheses, models, and theories. With this purpose, HARKing may become transparent (Hollenbeck & Wright, 2017) and just describes the generating aspect. Likewise, even transparent p-hacking may serve the purpose to find novelty, as we shall explicite in part II of this first of two consecutive articles.

The more exploration succeeds in moving science forward through unpredicted findings, we suggest, the more *valuable* it is. Value may arise either directly, that is, through discoveries that provide ever more accurate depictions of reality. Indirect value might come from exploratively generated claims that are wrong but trigger alternative ideas and thus open other paths to novel insight (Nosek et al., 2018; Stebbins, 1992, 2001, 2006). Thus, we suggest, exploration not only possesses practical value to inform particular scientific domains, but also epistemic value as a systematic method to find the new. (Note that the general term "exploration" has a much broader meaning than used here including goals such as approaching a new area of research to begin with or "becoming familiar with something by testing it" (Stebbins, 2001).)

With the term “claim” we label statements that are “synthetic [either right or wrong at least in some occasions], testable, falsifiable, parsimonious, and (hopefully) fruitful” (Myers & Hansen, 2012; p. 167). A claim makes an assertion on a hypothesis, model, or theory. We follow the predictivistic tradition of the philosophy of science, where a claim is supported if it makes a correct prediction on new data (Barnes, 2008).

Confirmation and exploration are both imperative, but serve very different purposes. Confirmation is about rigidly testing a claim. If successful, a *new* claim becomes an *established* claim. We use the usual notation for statistical tests, where H_1 means that the assertion is true (operationalised as an alternative hypothesis in a statistical test) and H_0 that it is not true (null hypothesis in a statistical test). In exploration, the narrow focus of confirmation is replaced by the freedom to widen the scope with the inherent goal of identifying novelty. This could move existing claims in a wider range of directions or lead to new assertions about the world.

Confirmation describes the straightforward, iterative path of research: hypothesise – test – corroborate or discard. If confirmation fails, the straight path ends, opening up the opportunity for a divergent, less-travelled path toward insight. Exploration is the method of venturing from

(what should be) the well-trodden confirmatory path with *a quantitative quest*. It even seems necessary for discovery beyond the mainstream. However, for this alternative scientific track to succeed, researchers must be equipped with competencies on the conceptions, goals, and methods of both confirmation and exploration.

Structure of the two parts

Our two articles are intended to outline the required means for valuable confirmation and exploration in scientific research. We begin this first part by discussing how confirmation and exploration are often blended in today's research. We then describe the related pressure to produce nominally confirmatory results, which we believe can be reduced if one assigns exploration the value it deserves in scientific practice. This, however, requires an epistemically strict distinction of confirmation and exploration. We use the theory of *severe testing* for this and clarify the role of preregistration. Finally, we lay out how transparent exploration serves confirmation and vice versa. Part II will propose foundations on how to conceptualise and systematically do exploration and how to implement more exploration in scientific practice.

The blending of confirmation and exploration

Manifestations of blending

We define *blending* as the (mis)use of exploratory methods of analysis for confirmatory purposes. Conceptually, blending is not unfounded as studies may be meaningfully placed on a

continuum from purely exploratory ("where the hypothesis is found in the data") to purely confirmatory ("where the entire analysis plan has been explicated before the first participant is tested"; Wagenmakers et al., 2012). Such a continuum maps onto the experience of having unexpected difficulties with data. In many cases this leads scientists (either intentionally or unintentionally) to blend these approaches together, and exploratory results are reported as if they were confirmatory.

There are several indications that blending is all too common. Direct evidence comes from many researchers admitting to QRPs, such as excluding data, collecting more data, and making post-hoc claims about hypotheses (Agnoli et al., 2017; Gopalakrishna et al., 2021a; John et al., 2012; Kerr, 1998). Additionally, a disconcerting 9% (95% confidence interval = 6 - 11%) of researchers across scientific fields even concede to data fabrication and/or falsification (Gopalakrishna et al., 2021a), perhaps the worst practices used to trim results in a particular, presumptive confirmatory direction. Then there are multiple strands of indirect evidence. First, content analyses show that published studies are almost always framed as confirmatory (Banks et al., 2016; Gigerenzer & Marewski, 2015; Spector, 2015; Woo et al., 2017): Hypotheses with $p < \alpha$, even if they have only been established through the analysis of data, appear as already confirmed. Second, p-values just below the usual $\alpha = .05$ are found far more often than expected. This may be caused both by researchers doing QRPs when preparing a paper, as well as subsequent publication bias towards positive results (Francis, 2012; Masicampo & Lalande, 2012; Rosenthal, 1979; Scargle, 2000). Blending is also reflected in the fact that many more negative results are found in *registered reports*, the format that publishes a paper irrespective of whether the results confirm a claim (Allen & Mehler, 2019; Chambers & Tzavella, 2020; Scheel et al., 2021a). It is also evident in the one-sided focus on confirmation in the teaching of science and statistics, with statistical testing being misunderstood as "a universal method for scientific inference" (Gigerenzer & Marewski, 2015). Therefore not surprisingly, statistical

tests are also used for exploratory purposes. Finally, blending is carried further through harmed scientific communication, as the replication crisis shows (Aarts et al., 2015; Camerer et al., 2018; Open Science Collaboration, 2015): Recipients of a seemingly confirming result build their research on the false assumption of sound confirmation; or, if sensitive to the problem, do not know for sure whether a conclusion is based on confirmation or mere exploration. Moreover, empirical evidence suggests that the sharp increase in the number of publications exacerbates the problem by further favouring the already prevailing straight paths instead of opening up new ones (Chu & Evans, 2021).

The pressure to produce seemingly confirming results

Blending also seems to be caused by the *pressure to produce publications* in the current incentive system where the number of publications and citations dominates the evaluation of scientific performance and career opportunities (Gonzales & Cunningham, 2015; Kerr, 1998; McIntosh, 2017; Nosek et al., 2018; Nosek & Lindsay, 2018; Wagenmakers et al., 2012). Indeed, researchers have reported on such pressure (Gopalakrishna et al., 2021a, 2021b). Besides, pressure is related to more frequent participation in at least one “severe QRP” (Gopalakrishna et al., 2021a). Practices such as p-hacking and intransparent HARKing anticipate publication bias in favour of positive results. At a deeper level, researcher bias towards generating and publishing positive results seem to be influenced by the false belief that positive results were associated with more scientific novelty, the flawed “ideal of confirmation" (Kerr, 1998) and, as a consequence, a "positive testing strategy" (Klayman & Ha, 1987). This is, of course, in stark contrast to Popper's insight that the capacity to falsify hypotheses is actually more fundamental to advances in science (Glass & Hall, 2008; Kerr, 1998; Klayman & Ha, 1987; Locke, 2007; Mayo, 2018; Popper, 1959).

Preregistration of confirmatory analyses falls short of solving the problem

Whether something has been preregistered is not logically related to its quality (Szollosi et al., 2020). Kerr (1998) and others have been criticised for exaggerating the value of preregistration in this regard (Devezer et al., 2021; Rubin, 2019, 2020). Preregistration records an *a priori* plan of the claim and the analysis and thus creates *control* over the plan history (Heers, 2020; Rubin, 2019, 2020; Wagenmakers et al., 2012). It seems to be the best mode of control to date because it allows researchers to prove how they have planned a study (Wagenmakers & Dutilh, 2016). However, the following issues suggest that the pressure to publish positive results partially resists preregistration.

First, preregistration can be abused as an option pulled *only in the case of success* to sell a result as a more convincing *post-hoc*. A negative result would be kept secret (Bian et al., 2020; Claesen et al., 2019). Registered reports try to address the bias towards positive results through the guarantee of publication before the results are known, but in many cases a final report is not published (Claesen et al., 2019; Hardwicke & Ioannidis, 2018). Even registered reports can be abused to market positive and suppress negative results (Bian et al., 2020). Another issue is superficial preregistration: the underreporting of analyses and collected variables, which leaves room for intransparent exploration (Franco et al., 2016). Finally, despite preregistration becoming ever more widespread, only a minority of psychological researchers use any type of preregistration. Only around 30% of psychological researchers in French speaking countries mentioned such practice in a survey (Beffara-Bret & Beffara-Bret, 2019), and only 3% of 188 psychology articles published between 2014 and 2017 included a statement on preregistration (Hardwicke et al., 2020). This indicates remaining hindrances and unresolved issues.

Transparent exploration helps to reduce the pressure

Preregistration is sometimes misunderstood as eliminating flexibility in hypothesis formulation/modification (Hollenbeck & Wright, 2017) and data analysis (Goldin-Meadow, 2016; Scott, 2013). This is wrong, as preregistration only archives the *initial* plan for an analysis. There may be important reasons to deviate from a plan, with deviation allowed as long as a justification is provided and changes are clearly stated as soon as they occur. However, flexibility could also be indicative of deficits and gaps in theory formulation (Eronen & Bringmann, 2021; Fiedler, 2017; Gigerenzer, 2010; Szollosi & Donkin, 2021). This should be taken as a call to fill the gaps with explicit, transparent exploration (Woo et al., 2017). We will elaborate on this in part II. This opportunity, like any measure to implement more exploration through methods, teaching and publishing policies, would make it easier for researchers to dispense with confirmatory framing.

It has however been argued that, were preregistration to become mandatory, everything else might appear to be flawed confirmatory research (Goldin-Meadow, 2016). Besides, more open science practices would give rise to further QRP like "preregistering after the results are known" (Yamada, 2018). We believe that the best answer to these concerns is to promote transparent exploration to reduce these wrong and even "perverse incentives" (Chiacchia, 2017). To achieve this, however, we first require a clear distinction between confirmation and exploration.

Differentiating confirmation and exploration

It is crucial to resolve the blending in the reporting of scientific results. Otherwise, confirmation is damaged by intransparent exploration, and transparent exploration does not unfold as it could.

While replication (Zwaan et al., 2018) and multi-lab studies (Stroebe, 2019) try to address the consequences of blending, and preregistration has the mentioned insufficiencies, we address blending by clearing the path for explicit exploration as a viable alternative. However, differentiating confirmation and exploration is not as easy as it may first appear. Explorative results are also supported by a certain, albeit exaggerated, amount of evidence as seen in small p-values (Szollosi & Donkin, 2021).

Only confirmation uses an evidential norm

The principle difference between confirmation and exploration is that confirmation adheres to an *evidential norm* for the test of a hypothesis to pass. An evidential norm states that an H_1 hypothesis is confirmed, if the evidence for it at least exceeds a certain *threshold*. The usual norm requires that chosen 1 – p (p-value) must be greater than 1 – α. Specifying a threshold, like α = .05, makes researchers "accountable" for what they would report as confirming (Mayo, 2018). In short: *The choice of the threshold norm and its strict application must not be influenced by the data.* Adherence is violated if variation in p according to the available analytical options (Gelman & Loken, 2013) is misused to fish for a particular p that happens to be smaller than α, thus with p-hacking. Adherence is also violated by HARKing when multiple relations are tested and an assertion is made around the one that happens to yield $p < \alpha$. In this case, 1 - α is not adhered to in the context of *all* that has been tested (see the discussion on global vs. local claims at the end of the chapter).

We suggest, however, that deviations from an analytical plan are unproblematic if they are not chosen in order to obtain a smaller p, but, for example, to account for deviations from otherwise violated model assumptions (Field & Wilcox, 2017). This keeps an analysis conceptionally within the bounds of confirmation. It is in accordance with the logical possibility of changing

analytical decisions after seeing the data without reducing the rigour in testing (Szollosi et al., 2020). Likewise, it is in principle possible to create or modify a claim after looking at the data without being influenced by what is seen in them. However, these possibilities are difficult to control unless they can be anticipated and incorporated into a preregistered plan (e.g. run model with option A if the parameter estimation converges, run model with option B otherwise). Such instances might constitute confirmation, but it is difficult to assess whether they truly do.

We propose that confirmation is *testing with a high risk to fail if a claim is wrong* (“*severe testing”, see below), and that this high risk must not be reduced by analytical decisions*. With adherence to an evidential norm, confirming a claim is supported with true evidence. However, because of blending and the experiences of the replication crisis, adherence requires *control*. Reliable control should be the prerequisite for an analysis to be accepted as confirmation. Preregistration seems to be the most feasible and effective mode of control, wherefore we agree with others that *only preregistered analyses should be accepted as confirmatory* (Lakens, 2019; Yamada, 2018). This should apply *from now on* and until perhaps a better mode of control is found. Note that other modes have been proposed. The retrospective “21 word solution” demands a post-hoc statement with which scientists declare that they have worked properly (Simmons et al., 2012). Open analysis may be very effective, but is pretty effortful (see part II). Both alternatives, however, do not offer transparency on the plan history. (Deciding whether to accept old analyses as confirmations, especially in the period before preregistration, is in itself a difficult question.)

Control measures like requiring preregistration place the burden of proof on scientists with the price of false negative assessments. Assuming that researchers have not worked properly, although they have, is rigid but seems necessary in psychology, which has been severely harmed by exaggerated evidence. Accordingly, new analyses that have not been or been improperly

preregistered, or where the preregistered analytical plan contradicts the report, should be considered as exploratory.

A norm must be used, but which norm is disputable

The usual norm of $1 - p > 1 - \alpha$ is disputable and subject to intense debate. P-values and statistical tests have several interpretational pitfalls and fundamental drawbacks (Greenland, 2017a; Greenland et al., 2016; Wagenmakers, 2007). Besides, they are based on many assumptions on the path from a substantive claim via study design, the produced data and the model that describes them. This relates to "Duhem's problem" (Ivanova, 2021; Mayo, 2018; Rakover, 2003), which states that a hypothesis cannot be tested without making assumptions beyond the data. Such assumptions often refer to *bias* and remain intransparent. If true, they would mean that issues like selection, measuring, non-compliance, and unconsidered shared causes between factor and outcome (Maclure & Schneeweiss, 2001) do not introduce any bias (in the Bayesian framework with a probability of 100%, Greenland, 2005). We propose that a well-chosen norm does well in staking out the boundary between the new and the established, considering major sources of bias.

Evidential norms and Mayo's theory of severe testing

We use Mayo's (2018) much debated (Gelman et al., 2019) philosophy of "severe testing" to discuss the choice of a norm. For Mayo (2018) *severity* is the *probability with which a given test with given data would have found a hypothesis to be wrong if it was truly wrong.* A test might have yielded a positive result, but the test might have been hardly capable of giving a negative result if the claim was wrong. In short: "A test is severe when it is highly capable of

demonstrating a claim is false" (Lakens, 2019). Importantly, this concept is not bound to any specific statistical theory or school (e.g. Frequentist vs. Bayesian), rather it is a conceptual framework by which to judge the appropriateness of whatever evidential norm. Grounded in Popper (1959), Lakatos (1977) and others, a severe test is difficult to pass and, in case of success, provides evidential support because it could have easily failed. With a non-severe test, a hypothesis is not sufficiently *probed*, that is, the test was not capable of finding the "flaws or discrepancies of a hypothesis" (Mayo, 2018). Severity calls for study designs that produce data capable of separating the truth or falsity of a hypothesis from all alternative explanations and thus closely link a hypothesis with the associated empirical observations used in testing it. Such awareness should recall the insight related to Duhem's problem that any single study is incapable of ruling out all alternative assumptions (e.g. Greenland, 2005; Milde, 2019). It also links to the fundamental question to what extent truth can be approached (e.g. via "truthlikeness", Cevolani & Festa, 2018; Niiniluoto, 2020).

Once a study has been designed and data have been collected, a claim can only be statistically probed. Any statistical method then is limited by the study's ability to produce a certain data result (e.g., a high average error rate in a cognitive test) that exceeds an evidential norm if the claim is indeed true (e.g. cognitive impairment is present), but would not be expected to do so under alternative assumptions (e.g., lack of compliance *or* misunderstanding the instructions). Likewise, Mayo's (2018) elaborations on *calculating* severity relies on this and thus involves only probing against chance. This, however, is the subject of a controversial discussion (Gelman et al., 2019). Anyway, a general understanding of severity might encourage researchers to reflect on substantive reasons for a claim to be wrong rather than falling prey to self-delusion and hiding behind statistical rituals (Gigerenzer, 2018).

In regards to the replication debate, the severity framework makes it transparent that QRPs like HARKing and p-hacking create the illusion of *greater test severity*. Thus, a result is sold by hiddenly exceeding the evidential norm. However, "preregistration makes it possible to evaluate the severity of a test" (Lakens, 2019). The framework also sheds light on the limitations of replication. A study design and analytical model might poorly map the phenomenon of interest and barely probe why a hypothesis may be wrong, in which case a wrong result could replicate (Devezer et al., 2021; Mayo, 2018; Steiner et al., 2019). In addition, a finding may indeed replicate (e.g. in a very large sample), but not translate into practical use like an intervention (Yarkoni, 2020).

Bayesian severity

We propose that evidential norms should be reconsidered along severity considerations *to set the right boundaries beyond which exploration should take over*. This should involve the capacity of both a study design and an analytical model to probe a substantive hypothesis against alternative non-causal explanations, especially bias. Whereas Mayo's frequentist-oriented elaborations on calculating severity are not capable of incorporating assumptions beyond the data, elaboration on *Bayesian severity* assessment opens the door for formalising this. Although controversial for epistemic reasons (Gelman et al., 2019 and papers cited therein; Mayo, 2018), severity can be handled in the Bayesian framework through a new interpretation. Such "falsificationist Bayesianism" (Gelman et al., 2019; Gelman & Shalizi, 2013) makes "risky and specific predictions" that could easily turn out to be wrong (van Dongen et al., 2020). A prediction might be made on the "posterior probability" for a claim to be true (given a prior distribution for an effect and, as in frequentist statistics, the data and the model that describes the data). Then, the norm requires that this posterior probability must exceed $1 - \alpha$ for a test to

pass. However, to achieve severity it has been argued that one needs to consider how likely a claim was already *before* seeing the data (Mayo, 2018). This shifts the focus to the *increment* in this probability through data observation (Held et al., 2021; Wagenmakers et al., 2018). Whatever norm is chosen, it must be preregistered to counteract that its choice or how it was evaluated was affected by data inspection. The same is true for the prior distribution since the posterior distribution may heavily depend on it (Gelman et al., 2013).

Bayesian approaches open up the possibility of addressing two further issues. First, one may probe against scepticism with a "sceptical prior". This expresses the belief in values around 0 before seeing the data (in terms of a normal distribution) and serves the purpose of convincing a sceptic (Good, 1950, p. 80 ff.; Held, 2020; Held et al., 2021). Second, Bayesian norms could take advantage of the - in psychology little-known - ability of Bayesian methods to probe causal hypotheses against pure associations via a *causal model* with explicit assumptions on *bias*. Bias may arise, for instance, from misclassification, selection probabilities and effects of a common cause on factor and outcome (Greenland, 2005, 2009; Höfler et al., 2007; Lash et al., 2009). Uncertainty in certain assumptions on bias themselves can be expressed, and such uncertainty carries over to more cautious conclusions about causal effects (Greenland, 2005). In addition, the assumptions might be varied, and thus sensitivity of the result against them is evaluated. This informs the readership of how *probing against a particular bias scenario* (Lash et al., 2014; Smith et al., 2021; VanderWeele & Mathur, 2020) relates to meeting a norm. Although seldom used, such Bayesian bias models make assumptions on bias transparent. At the very least, they encourage reflection on such mechanisms and generate awareness which of them require better understanding.

Because severity is a fairly new concept, there is much room for the development of rigorous norms beyond frequentist statistical tests. In particular, methodological progress should lead to norms that are based on calculating severity under more defendable assumptions, especially on bias. Another way to account for bias is using different studies that probe against different sources of bias through a range of studies. This embraces the methodical diversity that addresses the requirements of various causal quests in different domains (Gigerenzer & Marewski, 2015; Greenland, 2017b). Although it is very difficult to integrate diverse and differently biased evidence (Greenland, 2017b), scientists could find ways to approach multi-method norms that advance science through *multidimensional confirmation*. Single studies might nevertheless model unaddressed sources of bias or, at least, be explicit in that they have only been probed against certain bias and thus communicate a better understood piece of evidence.

At the cost of falling short of a norm, exploration opens the door for novelty

The common epistemic price of doing exploration is the possibility of falling short of adherence to a norm. For example, the usual frequentist α may be exceeded. P may be smaller than the nominal $\alpha = .05$, but through exploration it was only compared with, say, $\alpha = .20$. This happens with intransparent HARKing if one explores several outcomes, selects a particular outcome with $p < .05$ and presents only that outcome (Altman et al., 2017). And, as mentioned, α may be exceeded by p-hacking, since the multiple chances of passing a norm through different analytical options are not taken into account. Then, to meet the norm, new data are required, the more data, the more exploration has been used. The second aspect of this price is that the extent of the exceedance quickly becomes incalculable when multiple explorative steps are used. In case of an entire lack of transparency, this may call for a new study that meets the norm on its own. The effortful replication initiatives take this stance (Schimmack, 2018).

We suggest that exploration should consider "turning all the knobs" (Hofstadter & Dennett, 1981) around a given hypothesis or model, or when generating new hypotheses or models. This enables the core benefit of exploration: the potential of finding the new, wherever it might be hidden, whatever it might look like. An explorative quest might include the functional shapes of an effect, factor and outcome categorization and different effects in subpopulations. We propose that quests should be guided by the following key ideas building on severity:

- The stronger and more specific a claim, the more available it is to severe testing through confirmation (with new data). Thus, the greater is its ability to advance science if it were true (Lakens, 2019).
- Claims should be searched for that are *likely to pass severe testing* with new data.

Global versus local claims

Another idea to be elaborated in part II concerns the distinction between *global* and *local* claims. As we shall see, this distinction is important in planning and conducting explorations. It deserves to be mentioned here already because it sheds further light on how HARKing may practically violate the boundary between affirmation and exploration. Assume that a set of k factors and a set of l outcomes is explored with regard to factor-outcome associations. Each possible association is analysed with a frequentist level α test. In this case, the probability that at least one p-value is smaller than α is high for a large number of tests (k * l). Then the *global* existence of any (at least one) such relation is tested with poor severity. This would not adhere to the norm because multiple testing would be ignored (Bender & Lange, 2001). The test result for a *particular* factor-outcome association with $p < \alpha$, however, could have well been negative. Now one may ask whether a local claim on the existence of this association is confirmed. The

answer depends on whether the global context can be ignored in substantive terms. The problem is that a claim may be shifted by data analysis from local to global by overgeneralising a single factor and a single outcome as if they would represent two latent variables with a relation between them. This is another instance where adherence to a norm is difficult to control without preregistration.

Part II will discuss a couple of ambiguities, such as whether to aim at global versus local assertions and how to address them with background knowledge.

Transparent exploration serves confirmation and scientific communication

With transparency, explorative research practices are no longer questionable

If scientists become committed to conducting and publishing transparent exploration, there will be less pressure and incentive to blend exploration and confirmation. With transparency regarding exploratory results, communication is no longer harmed by hidden information, and evidence is no longer overstated.

Transparency also provides the answer to the question whether the double-use of data for confirmation and exploration is problematic. For example, one could misunderstand Wagenmakers and colleagues (2012) this way: “The interpretation of common statistical tests in terms of Type I and Type II error rates are valid only if the data were used once and if the statistical test was not chosen on the basis of suggestive patterns in the data”. Actually, while exploration must not *affect* what and how to confirm (Barnes, 2008), it may well be used later to *modify* a hypothesis through exploration (Devezer et al., 2021). Double-use, when done in this temporal order, is established practice in medical research. For instance, the consort data

from the UK Biobank (2020) are made available for exploration by everyone after the confirmatory results have been published. Remarkably, having a hypothesis versus not having one appears to severely impair the ability to detect striking data patterns (Yanai & Lercher, 2020), and it would be interesting to assess whether teaching exploration could reduce this effect.

We suspect that researchers fear their confirmation trials failing ("All my work will be in vain if I do not confirm my hypothesis!") or, at least, that data are only able to analyse what has been pre-specified. In the case of non-confirmation of, say, an intervention effect, a researcher might proceed with common practices like subgroup analysis ("Is the intervention effective in females and males, respectively?"). This is not problematic as long as a then found data pattern does not lead to a confirming assertion ("We confirmed that intervention is effective in females"), but as a *modified hypothesis*, yet to be confirmed or not with new data ("We propose the modified hypothesis that the intervention is effective in females.").

Concatenated exploration

Science has been argued to be most productive if confirmation and exploration co-exist in a "good balance", back and forth from theories to derived claims, study design and data (Bogen & Woodward, 1988; Box, 1980; Scheel et al., 2021b). But intransparent HARKing and *p*-hacking hinder researchers from recognizing the two (Woo et al., 2017). If their difference becomes transparent, exploratively generated or modified hypotheses, models and theories openly invite confirmatory studies. With "concatenated exploration", Stebbins (1992) denotes a cooperative strategy that pays off for all who participate in a "longitudinal research process". Such a process may start with exploration (according to Popper, new scientific claims may start from anywhere; Klayman & Ha, 1987; Popper, 1959; in Lakatos' conception of science new

additions from exploration contribute to the further development of theory; Lakatos, 1977). Explorative results may give rise to a new claim, a confirmation trial (with perhaps explorative refinement), subsequent studies with confirmation and extensions (maybe using different populations), further adjustment, confirmation and so on. For similar proposals, see Behrens (1997), Nosek and colleagues (2018) and Thompson and colleagues (2020). Such a chain process may provide the impetus for the evolution of hypotheses (e.g. excess screen time causes a heightened stress response in adolescents), the development of models (different causes of the stress response and how they interact together) or an entire theory (the evolution and development of the stress response).

Stebbins (2001) describes a range of sociological examples, where such a chain of research has advanced science including postpartum changes in women and the development of women's occupational aspirations across the lifespan. In psychology, concatenated exploration appears to describe the idea behind the very common trial-and-error proceeding along the development of interventions (e.g. the history of origins of dialectic behavior therapy, Linehan & Wilks, 2015). In addition, an interplay between exploration and confirmation has long been established practice in factor analysis, where the construction of a scale is a chain of setting up a model, testing it, modifying, again testing and so on (Hurley et al., 1997).

Researchers who participate in such a chain of research are able to publish at least once and can expect to be cited several times in the currently prevailing quantitative incentive system of publications and citations. In qualitative terms this commitment to concatenated exploration may be favoured in upcoming new criteria of sustainable scientific achievement (Pavlovskaia, 2014; Spangenberg, 2011). Additionally, given the iterative nature of a research chain scientists can hope for more confirmed findings, cooperatively generated insights and the ability to look back on research of enduring validity later in life (McKiernan et al., 2016; Nosek et al., 2012;

Pavlovskaia, 2014). Yet such an outlook on incentives for conducting science well beyond just publishing a paper might help to counteract behaviour geared towards short-term benefits.

Conclusion

Several researchers had already called for a major up-valuing of exploration as a complement to confirmation (Gonzales & Cunningham, 2015; McIntosh, 2017; Nosek et al., 2018; Scheel et al., 2021b). However, without elaborations on the conceptions, methods, good examples, and teaching and implementation practices in the publication system uncertainty may prevent researchers from abandoning the ritualised (Gigerenzer, 2018), almost obsessive, restriction to blended confirmatory research. Additional obstacles may hinder more transparent exploration such as the social barriers that have been described for change in general (Nosek & Bar-Anan, 2012; Zwaan et al., 2018).

We believe that transparent exploration is fundamental to the advance of science. A starting point for transparent exploration is an understanding that, to date, the blending of confirmation and exploration has been all too common and that distinguishing these two concepts is vital to the health of science. A sound norm is severe. Adherence to and control over such a norm establish a sharp boundary for the transition of a new assertion into an established one. This promotes scientific communication by requiring that future research be only built on sufficient evidence.

In the second part we shall outline how to plan and conduct transparent exploration in practice, setting the goals of “comprehensive exploration” and “effective exploration” with some ideas on filtering and smoothing data patterns to separate the signals from the noise. We will discuss

the roles of preregistration, open data and open analysis. Part II will end with the key points of a research agenda on how to explore in a specific domain and a checklist with recommendations to stakeholders who have the means to establish more transparent exploration in the publication system.

Author Contributions:

Michael Höfler worked out most of the content and had the lead in writing. Robert Miller and Stephan Scherbaum have contributed to the elaboration of the basic idea and contributed to details. Brennan McDonald joined in this version, refined epistemic details and was involved in the writing and wording of the entire manuscript. Philipp Kanske commented on and edited the manuscript.

Conflicts of Interest:

The authors declare that there were no conflicts of interest with respect to the authorship or the publication of this article.

Funding:

Stefan Scherbaum and Philipp Kanske are supported by the German Research Foundation (CRC940/A08 and KA4412/2-1, KA4412/4-1, KA4412/5-1, CRC940/C07, respectively).

Acknowledgements

We wish to thank the three reviewers on our first submission for their detailed comments, for instance, Matt Williams for his epistemical suggestions. Several new arguments are based on the reviewers' suggestions. We also thank Annekathrin Rätsch for aid with the references.

References

Aarts, A., Anderson, J., Anderson, C., Attridge, P., Attwood, A., Axt, J., Babel, M., Bahník, Š., Baranski, E., Barnett-Cowan, M., Bartmess, E., Beer, J., Bell, R., Bentley, H., Beyan, L., Binion, G., Borsboom, D., Bosch, A., Bosco, F., & Penuliar, M. (2015). Estimating the reproducibility of psychological science. *Science*, 349. doi: 10.1126/science.aac4716

Agnoli, F., Wicherts, J. M., Veldkamp, C. L. S., Albiero, P., & Cubelli, R. (2017). Questionable research practices among italian research psychologists. *PLoS ONE*, *12*(3): e0172792. doi: 10.1371/journal.pone.0172792

Allen, C., & Mehler, D. M. A. (2019). Open science challenges, benefits and tips in early career and beyond. *PLOS Biol*, *17*, e300024. doi: 10.1371/journal.pbio.3000246

Altman, D. G., Moher, D., & Schulz, K. F. (2017). Harms of outcome switching in reports of randomised trials: CONSORT perspective *BMJ 356*: j396 doi: 10.1136/bmj.j396

Banks, G. C., Rogelberg, S. G., Woznyj, H. M., Landis, R. S., & Rupp, D. E. (2016). Editorial: Evidence on Questionable Research Practices: The Good, the Bad, and the Ugly. *Journal of Business and Psychology*, *31*(3), 323–338. doi: 10.1007/s10869-016-9456-7

Barnes, E. (2008). *The Paradox of Predictivism.* Cambridge: Cambridge University Press. doi:10.1017/CBO9780511487330

Beffara-Bret, A., & Beffara-Bret, B. (2019). Accessed on Sep 29th, 2021. Open Science in European French Speaking Countries. Brice-Beffara

Behrens, J. T. (1997). Principles and procedures of exploratory data analysis. *Psychological Methods, 2*(2), 131–160. doi: 10.1037/1082-989X.2.2.131

Bender, R., & Lange, S. (2001). Adjusting for multiple testing--when and how? *Journal of Clinical Epidemiology*, *54*(4), 343-349. doi: 10.1016/s0895-4356(00)00314-0.

Bian, J., Min, J. S., Prosperi, M., & Wang, M. (2020). Are preregistration and registered reports vulnerable to hacking? *Epidemiology*, *31*(3), e32. doi: 10.1097/EDE.0000000000001162

Bogen, J., & Woodward, J. (1988). Saving the phenomena. *Philosophical Review*, *97*(3) 303–352. doi: 10.2307/2185445

Box, G. E. P. (1980). Sampling and Bayes inference in scientific modelling and robustness (with discussion and rejoinder). *Journal of the Royal Statistical Society A, 143*(4), 383–430.

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

Cevolani, G., & Festa, R. (2018). A partial consequence account of truthlikeness. *Synthese*, *197*, 1627-1646. doi: 10.1007/s11229-018-01947-3

Chambers, C., & Tzavella, L. (2020). Registered reports: Past, present and future. Preprint at MetaArXiv doi: 10.31222/osf.io/43298

Chiacchia, K. (2017, July 12). Perverse Incentives? How Economics (Mis-)shaped Academic Science. HPC Wire. Retrieved October 26, 2021 from https://www.hpcwire.com/2017/07/12/perverse-incentives-economics-mis-shaped-academic-science/

Chu, J. S. G., & Evans, J. A. (2021). Slowed canonical progress in large fields of science *Proceedings of the National Academy of Sciences, 118* (41). e2021636118; doi: 10.1073/pnas.2021636118

Claesen, A., Gomes, S. L. B. T., Tuerlinckx, F., & Vanpaemel, W. (2019, May 9). *Preregistration: comparing dream to reality.* Retrieved October 14, 2020 from https://psyarxiv.com/d8wex/

Devezer, B., Navarro, D. J., Vandekerckhove, J., & Ozge Buzbas, E. (2021). The case for formal methodology in scientific reform. *Royal Society Open Science, 31*; 8(3), 200805. doi: 10.1098/rsos.200805

Dirnagl, U. (2020). Preregistration of exploratory research: learning from the golden age of discovery. *PLOS Biol, 18*(3), e3000690. doi: 10.1371/journal.pbio.3000690

Eronen, M. I., & Bringmann, L. F. (2021). The Theory Crisis in Psychology: How to Move Forward. *Perspectives on Psychological Science, 16*(4), 779–788. doi: 10.1177/1745691620970586

Fiedler, K. (2017). What constitutes strong psychological science? The (neglected) role of diagnosticity and a priori theorizing. *Perspectives on Psychological Science*, 12(1), 46–61. doi: 10.1177/1745691616654458

Field, A. P., & Wilcox, R. R. (2017). Robust statistical methods: A primer for clinical psychology and experimental psychopathology researchers. *Behaviour Research and Therapy, 98,* 19-38. doi: 10.1016/j.brat.2017.05.013

Francis, G. (2012). Publication bias and the failure of replication in experimental psychology. *Psychonomic Bulletin & Review, 19*, 975–991. doi: 10.3758/s13423-012-0322-y

Franco, A., Malhotra, N., & Simonovits, G. (2016). Underreporting in Psychology Experiments: Evidence From a Study Registry. *Social Psychological and Personality Science*, *7*(1), 8-12. doi: 10.1177/1948550615598377

Gelman, A., & Loken, E. (2013, November 14). The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time. Retrieved October 14, 2020 from http://www.stat.columbia.edu/~gelman/research/unpublished/p_hacking.pdf

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian Data Analysis* (3nd ed.). Chapman and Hall/CRC. doi: 10.1201/b16018

Gelman, A., Haig, B., Hennig, C., Owen, A., Cousins, R., Young, S., Robert, C., Yanofsky, C., Wagenmakers, E. J., Kenett, R., & Lakeland, D. (2019). Many perspectives on Deborah Mayo's "Statistical Inference as Severe Testing: How to Get Beyond the Statistics Wars" Retrieved November 2, 2021 from http://www.stat.columbia.edu/~gelman/research/unpublished/mayo_reviews_2.pdf

Gelman, A., & Shalizi, C. R. (2013). Philosophy and the practice of Bayesian statistics. *British Journal of Mathematical and Statistical Psychology, 66*, 8-38. doi: 10.1111/j.2044-8317.2011.02037.x

Gigerenzer, G. (2010). Personal reflections on theory and psychology. *Theory & Psychology*, *20*(6), 733–743. doi: 10.1177/0959354310378184

Gigerenzer, G. (2018). Statistical rituals: The replication delusion and how we got there. *Advances in Methods and Practices in Psychological Science, 1*(2), 198–218. doi: 10.1177/0959354310378184

Gigerenzer, G., & Marewski, J. N. (2015). Surrogate science: The idol of a universal method for scientific inference. *Journal of Management, 41*(2), 421–440. doi: 10.1177/0149206314547522

Glass, D. J., & Hall, N. (2008). *A brief history of the hypothesis. Cell, 134*(3): 378–381. doi: 10.1016/j.cell.2008.07.033

Goldin-Meadow, S. (2016, August 31). *Why preregistration makes me nervous.* Retrieved October 14, 2020, from http://www.psychologicalscience.org/observer/why-preregistration-makes-me-nervous

Gonzales, J. E., & Cunningham, C. A. (2015, August). *The promise of preregistration in psychological research. Psychological Science Agenda.* Retrieved October 14, 2020 from https://www.apa.org/science/about/psa/2015/08/preregistration

Good, I. J. (1950). *Probability and the Weighing of Evidence*. Griffin.

Gopalakrishna, G., Riet, G. t., Cruyff, M. J., Vink, G., Stoop, I., Wicherts, J. M., & Bouter, L. (2021a, July 6). Prevalence of questionable research practices, research misconduct and their potential explanatory factors: a survey among academic researchers in The Netherlands. doi:10.31222/osf.io/vk9yt

Gopalakrishna, G., Wicherts, J. M., Vink, G., Stoop, I., Van den Akker, O., Riet, G. t., & Bouter, L. (2021b, July 6). Prevalence of responsible research practices and their potential explanatory factors: a survey among academic researchers in The Netherlands. doi: 10.31222/osf.io/xsn94

Greenland, S. (2005). *Multiple-bias modeling* for analysis of observational data (with discussion). *Journal of the Royal Statistical Society, Series A, 168*, 267-306. doi: 10.1111/j.1467-985X.2004.00349.x

Greenland, S. (2009). Bayesian perspectives for epidemiologic research: III. Bias analysis via missing-data methods. *International Journal of Epidemiology, 38*(6), 1662-73. doi: 10.1093/ije/dyp278

Greenland, S. (2017a). Invited Commentary: The Need for Cognitive Science in Methodology, *American Journal of Epidemiology*, *186*(6), 639–645. Doi. 10.1111/j.1467-985X.2004.00349.x

Greenland, S. (2017b). For and Against Methodologies: Some Perspectives on Recent Causal and Statistical Inference Debates. *European Journal of Epidemiology*, 32(1), 3-20. doi: 10.1007/s10654-017-0230-6

Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., & Altman, D. G. (2016). Statistical tests, P values, confidence intervals, and

power: A guide to misinterpretations. *European Journal of Epidemiology*, *31*(4), 337–350. doi: 10.1007/s10654-016-0149-3

Hardwicke, T. E., & Ioannidis, J. P. A. (2018). Mapping the universe of registered reports. *Nature Human Behaviour, 2*, 793–796. doi: 10/gf9db

Hardwicke, T. E., Thibault, R. T., Kosie, J., Wallach, J. D., Kidwell, M. C., & Ioannidis, J. (2020). Estimating the prevalence of transparency and reproducibility-related research practices in psychology (2014-2017). *MetaArXiv*. doi: 10.31222/osf.io/9sz2y

Head, M. L., Holman, L., Lanfear, R., Kahn, A. T., & Jennions, M. D. (2015). The extent and consequences of p-hacking in science. *PLOS Biol, 13*(3), e1002106. doi: 10.1371/journal.pbio

Heers, M. (2020). *Preregistration and registered reports.* FORS Guide No. 09, Version 1.0. Lausanne: Swiss Centre of Expertise in the Social Sciences FORS. doi: 10.24449/FG-2020-00009

Held, L. (2020). A new standard for the analysis and design of replication studies. *Journal of the Royal Statistical Society*, *183*(2), 431–448. doi: 10.1111/rssa.12493#

Held, L., Matthews, R., Ott, M., & Pawel, S. (2021). Reverse-Bayes methods for evidence assessment and research synthesis. Research Synthesis Methods. doi: 10.1002/jrsm.1538

Höfler, M., Lieb, R., & Wittchen, H. U. (2007). Estimating causal effects from observational data with a model for multiple bias. *International Journal of Methods in Psychiatric Research, 16*(2), 77–87. doi: 10.1002/mpr.205

Hofstadter, D. R., & Dennett, D. C. (1981). The mind's I: Fantasies and reflections on self and soul. New York: Basic Books.

Hollenbeck, J. R., & Wright, P. M. (2017). Harking, Sharking, and Tharking: Making the case for post hoc analysis of scientific data. *Journal of Management, 43*(1), 5–18. doi: 10.1177/0149206316679487

Hurley, A. E., Scandura, T. A., Schriesheim, C. A., Brannick, M. T., Seers, A., Vandenberg, R. J., & Williams, L. J. (1997). Exploratory and confirmatory factor analysis: guidelines, issues, and alternatives. *Journal of Organizational Behavior, 18*(6), 667-683. doi: 10.1002/(SICI)1099-1379(199711)18:6<667::AID-JOB874>3.0.CO;2-T

Ivanova, M. (2021). Duhem and Holism. *Elements in the Philosophy of Science*. doi: 10.1017/9781009004657

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science, 23*, 524–532.

Kerr, N. L. (1998). "HARKing: Hypothesizing after the results are known". *Personality and Social Psychology Review, 2*(3), 196–217. doi: 10.1207/s15327957pspr0203_4

Klayman, J., & Ha, Y.-W. (1987). Confirmation, disconfirmation, and information in hypothesis testing. *Psychological Review, 94*(2), 211–228. doi: 10.1037/0033-295X.94.2.211

Lakatos, I. (1977). *The Methodology of Scientific Research Programmes: Philosophical Papers Volume 1*. Cambridge University Press, Cambridge.

Lakens, D. (2019). The Value of Preregistration for Psychological Science: A Conceptual Analysis. doi: 10.31234/osf.io/jbh4w

Lash, T. L., Fox, M. P., MacLehose, R. F., Maldonado, G., McCandless, L. C., & Greenland, S. (2014). Good practices for quantitative bias analysis. *International Journal of Epidemiology, 43*(6), 1969–1985. doi: 10.1093/ije/dyu149

Lash, T. L., Fox, M. P., & Fink, A. K. (2009). *Applying Quantitative Bias Analysis to Epidemiologic Data.* New York, NY: Springer.

Lewandowsky, S., & Oberauer, K. (2020). Low replicability can support robust and efficient science. *Nature Communications, 11*(1), 1-12. doi: 10.1038/s41467-019-14203-0

Linehan, M. M., & Wilks, C. R. (2015). The course and evolution of dialectical behavior therapy. *American Journal of Psychotherapy, 69*(2), 97-110. doi: 10.1176/appi.psychotherapy.2015.69.2.97

Locke, E. A. (2007). The case for inductive theory building. *Journal of Management, 33*, 867-890. doi: 10.1177/0149206307307636

Maclure, M., & Schneeweiss, S. (2001). Causation of bias: the episcope. *Epidemiology 12*(1),114-22. doi: 10.1097/00001648-200101000-00019.

Masicampo, E. J., & Lalande, D. R. (2012). A peculiar prevalence of p values just below .05, *The Quarterly Journal of Experimental Psychology*, *65*(11), 2271-2279. doi: 10.1080/17470218.2012.711335

Mayo, D. G. (2018). *Statistical Inference as Severe Testing: How to Get Beyond the Statistics Wars*. Cambridge: Cambridge University Press. doi: 10.1017/9781107286184

McIntosh, R. D. (2017). Exploratory reports: A new article type for Cortex. *Cortex*, 96, A1–A44. doi: 10.1016/j.cor-tex.2017.07.014

McKiernan, E. C., Bourne, P. E., Brown, C. T., Buck, S., Kenall, A., Lin, J., McDougall, D., Nosek, B. A., Ram, K., Soderberg, C. K., Spies, J. R., Thaney, K., Updegrove, A., Woo, K. H., & Yarkoni, T. (2016). How open science helps researchers succeed. *eLife*, *5*, e16800. doi: 10.7554/eLife.16800

Milde, C. (2019). What Can Be Concluded from Statistical Significance? Severe Testing as an Appealing Extension to Our Standard Toolkit (SSRN Scholarly Paper ID 3413808). *Social Science Research Network*. doi: 10.2139/ssrn.3413808

Myers, A., & Hansen, C. H. (2012). *Experimental psychology*, 7th edition. Pacific Grove, CA: Wadsworth/Thomson Learning.

Niiniluoto, I. (2020). Truthlikeness: Old and new debates. *Synthese*, *197*(4), 1581–1599. doi: 10.1007/s11229-018-01975-z

Nosek, B. A., & Bar-Anan, Y. (2012). Scientific utopia: I. Opening scientific communication. *Psychological Inquiry*, *23*(3), 217–243. doi: 10.1080/1047840X.2012.692215

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *PNAS Proceedings of the National Academy of Sciences of the United States of America, 115*(11), 2600–2606. doi: 10.1073

Nosek, B. A., & Lindsay, D. S. (2018, February 2). Preregistration becoming the norm in psychological. science. *APS Observer, 31*(3). Retrieved October 14, 2020 from https://www.psychologicalscience.org/observer/preregistration-becoming-the-norm-in-psychological-science

Nosek, B. A., Spies, J. R., & Motyl, M. (2012). Scientific Utopia: II. Restructuring Incentives and Practices to Promote Truth Over Publishability. *Perspectives On Psychological Science, 7,* 615-31. doi: 10.1177/1745691612459058

Open Science Collaboration. (2015). "Estimating the reproducibility of psychological science" (PDF). *Science, 349* (6251), aac4716. doi:10.1126/science.aac4716

Pavlovskaia, E. (2014). Sustainability criteria: their indicators, control, and monitoring (with examples from the biofuel sector). *Environmental Sciences in Europe, 26***,** 17. doi: 10.1186/s12302-014-0017-2

Popper, K. (1959). The logic of scientific discovery. Basic Books.

Rakover, S. S. (2003). Experimental Psychology and Duhem's Problem. *Journal for the Theory of Social Behaviour*, *33*(1), 45–66. doi: 10.1111/1468-5914.00205

Rosenthal, R. (1979). The "file drawer problem" and tolerance for null results, *Psychological Bulletin, 86*(3), 838-641.

Rubin, M. (2017). When Does HARKing Hurt? Identifying When Different Types of Undisclosed Post Hoc Hypothesizing Harm Scientific Progress. *Review of General Psychology*, *21*(4), 308-320. doi:10.1037/gpr0000128

Rubin, M. (2019). The costs of HARKing. *British Journal for the Philosophy of Science*. doi: 10.1093/bjps/axz050

Rubin, M. (2020). Does preregistration improve the credibility of research findings? *The Quantitative Methods for Psychology*, *16*(4), 376–390. doi: 10.23668/psycharchives.4839

Scargle, J. (2000). Publication bias: The "file-drawer" problem in scientific inference, *Journal of Scientific Exploration, 14*(1), 91-106.

Scheel, A. M., Schijen, M. R. M. J., & Lakens, D. (2021a). An excess of positive results: Comparing the standard psychology literature with registered reports. *Advances in Methods and Practices in Psychological Science*, 4(2), article 25152459211007467. doi: 10.1177/25152459211007467

Scheel, A. M., Tiokhin, L., Isager, P. M., & Lakens, D. (2021b). Why Hypothesis Testers Should Spend Less Time Testing Hypotheses. *Perspectives on Psychological Science*, *16*(4), 744–755. doi: 10.1177/1745691620966795

Schimmack, U. (2018). The replicability revolution. *Behavioral and Brain Sciences*, *41*, e147. doi: 10.1017/S0140525X18000833

Smith, L. H., Mathur, M. B., & VanderWeele, T. J. (2021). Multiple-bias Sensitivity Analysis Using Bounds. *Epidemiology*, *32*(5), 625–634. doi: 10.1097/EDE.0000000000001380

Scott, S. K. (2013). Preregistration would put science in chains. Retrieved October 14, 2020 from https://www.timeshighereducation.com/comment/opinion/preregistration-would-put-science-in-chains/2005954.article

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2012). A 21 Word Solution (October 14, 2012). Retrieved February 11, 2021 from https://ssrn.com/abstract=2160588 doi: 10.2139/ssrn.2160588

Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2020). Specification curve analysis. *Nature Human Behavior*. doi: 10.1038/s41562-020-0912-z

Spangenberg, J. (2011). Sustainability science: A review, an analysis and some empirical lessons. *Environmental Conservation, 38*(3), 275-287. doi: 10.1017/S0376892911000270

Spector, P. E. (2015). Induction, deduction, abduction: Three legitimate approaches to organizational research. Video lecture for consortium for advancement of research methods and analysis. University of North Dakota (https://razor.med.und.edu/carma/video).

Stebbins, R. A. (1992). Concatenated exploration: notes on a neglected type of longitudinal research. *Quality & Quantity, 26*, 435-442. doi: 10.1007/BF00170454

Stebbins, R. A. (2001). *Exploratory research in the social sciences*. Thousand Oaks, Calif: Sage Publications. doi: 10.4135/9781412984249

Stebbins, R. A. (2006). Concatenated exploration: aiding theoretic memory by planning well for the future. *Journal of Contemporary Ethnography*, *35*(5), 483-494. doi: 10.1177/0891241606286989

Steiner, P. M., Wong, V. C., & Anglin, K. (2019). A causal replication framework for designing and assessing replication efforts. Zeitschrift für Psychologie, *227*(4), 280-292. doi: 10.1027/2151-2604/a000385

Stroebe, W. (2019). What Can We Learn from Many Labs Replications? *Basic and Applied Social Psychology*, *41*(2), 91–103. doi: 10.1080/01973533.2019.1577736

Szollosi, A., & Donkin, C. (2021). Arrested Theory Development: The Misguided Distinction Between Exploratory and Confirmatory Research. *Perspectives on Psychological Science, 16*, 717 - 724. doi:10.1177/1745691620966796

Szollosi, A., Kellen, D., Navarro, D. J., Shiffrin, R., van Rooij, I., Van Zandt, T., & Donkin, C. (2020). Is Preregistration Worthwhile? *Trends in Cognitive Sciences*, *24*(2), 94–95. doi: 10.4135/9781412984249

Thompson, W. H., Wright, J., & Bissett, P. G. (2020). Point of view: open exploration. *eLife*, *9*, *e52157*. doi: 10.7554/eLife.52157

Ulrich, R., & Miller, J. (2020). Questionable research practices may have little effect on replicability. *eLife*, *9*, e58237. doi: 10.7554/eLife.58237

UK Biobank. (2020). Retrieved October 14, 2020 from https://www.ukbiobank.ac.uk/

van Dongen, N. N. N., Wagenmakers, E., & Sprenger, J. (2020, December 16). A Bayesian Perspective on Severity: Risky Predictions and Specific Hypotheses. doi: 10.31234/osf.io/4et65

VanderWeele, T. J., & Mathur, M. B. (2020). Commentary: Developing best-practice guidelines for the reporting of E-values. *International Journal of Epidemiology, 49* (5), 1495 - 1497. doi: 10.1093/ije/dyaa094

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review*, *14*(5), 779–804. doi: 10.3758/BF03194105

Wagenmakers, E. J., & Dutilh, G. (2016). Seven selfish reasons for preregistration. *APS Observer*, *29*(9). https://www.psychologicalscience.org/observer/seven-selfish-reasons-for-preregistration

Wagenmakers, E. J., Wetzels, R., Borsboom, D., van der Maas, H. J. L., & Kievit, R. A. (2012). An agenda for purely confirmatory research. *Perspectives on Psychological Science, 7*, 632–638. doi: 10.1177/1745691612463078

Wagenmakers, E.-J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Love, J., Selker, R., Gronau, Q. F., Šmíra, M., Epskamp, S., Matzke, D., Rouder, J. N., & Morey, R. D. (2018). Bayesian inference for psychology. Part 1: Theoretical advantages and practical ramifications. *Psychonomic Bulletin & Review*, *25*(1), 35–57. doi: 10.3758/s13423-017-1343-3

Woo, S. E., O'Boyle, E. H., & Spector, P. E. (2017). Best practices in developing, conducting, and evaluating inductive research [editorial]. *Human Resource Management Review*, *27*(2), 255–264. doi: 10.1016/j.hrmr.2016.08.004

Yamada, Y. (2018). How to Crack Preregistration: Toward Transparent and Open Science. *Frontiers in Psychology, 9*, 1831. doi: 10.3389/fpsyg.2018.0183

Yanai, I., & Lercher, M. A. (2020). A hypothesis is a liability. *Genome Biology, 21***,** 23. doi: 10.1186/s13059-020-02133-w

Yarkoni, T. (2020). Implicit Realism Impedes Progress in Psychology: Comment on Fried (2020). *Psychological Inquiry* 31, 326-333. doi: 10.1080/1047840X.2020.1853478.

Zwaan, R. A., Etz, A., Lucas, R. E., & Donnellan, M. B. (2018). Making replication mainstream. *Behavioral and Brain Sciences*, *41*. doi: 10.1017/S0140525X17001972