

Exploratory hypothesis tests can be more compelling than confirmatory hypothesis tests

Mark Rubin & Chris Donkin

To cite this article: Mark Rubin & Chris Donkin (26 Aug 2022): Exploratory hypothesis tests can be more compelling than confirmatory hypothesis tests, *Philosophical Psychology*, DOI: 10.1080/09515089.2022.2113771

To link to this article: <https://doi.org/10.1080/09515089.2022.2113771>



© 2022 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.



Published online: 26 Aug 2022.



Submit your article to this journal [↗](#)



Article views: 13997



View related articles [↗](#)



View Crossmark data [↗](#)




Citing articles: 5 View citing articles [↗](#)

ARTICLE

 OPEN ACCESS

 Check for updates

Exploratory hypothesis tests can be more compelling than confirmatory hypothesis tests

Mark Rubin ^a and Chris Donkin ^b

^aDepartment of Psychology, Durham University, Durham, UK; ^bFaculty of Psychology and Educational Sciences, Ludwig Maximilian University of Munich, Munich, Germany

ABSTRACT

Preregistration has been proposed as a useful method for making a publicly verifiable distinction between confirmatory hypothesis tests, which involve planned tests of ante hoc hypotheses, and exploratory hypothesis tests, which involve unplanned tests of post hoc hypotheses. This distinction is thought to be important because it has been proposed that confirmatory hypothesis tests provide more compelling results (less uncertain, less tentative, less open to bias) than exploratory hypothesis tests. In this article, we challenge this proposition and argue that there are several advantages of exploratory hypothesis tests that can make their results *more* compelling than those of confirmatory hypothesis tests. We also consider some potential disadvantages of exploratory hypothesis tests and conclude that their advantages can outweigh the disadvantages. We conclude that exploratory hypothesis tests avoid researcher commitment and researcher prophecy biases, reduce the probability of data fraud, are more appropriate in the context of unplanned deviations, facilitate inference to the best explanation, and allow peer reviewers to make additional contributions at the data analysis stage. In contrast, confirmatory hypothesis tests may lead to an inappropriate level of confidence in research conclusions, less appropriate analyses in the context of unplanned deviations, and greater bias and errors in theoretical inferences.

ARTICLE HISTORY

Received 20 April 2022
Accepted 11 August 2022

KEYWORDS

Accommodation; exploratory analyses; confirmatory analyses; prediction; preregistration; hypothesis testing

The replication crisis may be partly explained by scientists' overconfidence in the replicability of their results. It has been argued that one source of this overconfidence is the false portrayal of exploratory hypothesis tests as confirmatory hypothesis tests (Nosek et al., 2018; Wagenmakers et al., 2012). *Exploratory* hypothesis tests involve unplanned tests of post hoc hypotheses that may be influenced by some of the research results.¹ In contrast, *confirmatory* hypothesis tests involve planned tests of ante hoc (a priori) hypotheses that are not influenced by any of the research results. It

CONTACT Mark Rubin  Mark-Rubin@outlook.com  Department of Psychology, Durham University, South Road, Durham DH1 3LE, UK

© 2022 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.
This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives License (<http://creativecommons.org/licenses/by-nc-nd/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited, and is not altered, transformed, or built upon in any way.

has been argued that the results of exploratory hypothesis tests tend to be more uncertain and tentative than those of confirmatory hypothesis tests, and so falsely portraying them as confirmatory results may lead to a false sense of confidence about their veracity and, consequently, their replicability (e.g., Nosek et al., 2018, p. 2600). The public preregistration of confirmatory hypothesis tests has been proposed as a method of better distinguishing them from exploratory hypothesis tests and preventing this false portrayal and overconfidence.

In the first part of this article, we criticize the usefulness of the confirmatory-exploratory distinction and argue that preregistration is not necessary to confirm the validity of hypothesis tests. Consequently, we argue that there is no fundamental reason why the result of an exploratory hypothesis test should be regarded as being more uncertain or tentative than the result of a confirmatory hypothesis test.

In the second part of the article, we argue that exploratory hypothesis tests have several *advantages* over confirmatory hypothesis tests and that, consequently, they have the potential to deliver *more* compelling research conclusions. We present six arguments to support this position.

Finally, in the third part of the article, we consider seven potential *disadvantages* of exploratory hypothesis tests that are related to issues such as falsification, overfitting, questionable research practices, researcher bias, and ethics. We conclude that these potential disadvantages may not outweigh the advantages of exploratory hypothesis tests. Consequently, exploratory hypothesis tests can be more compelling than confirmatory hypothesis tests.

To be clear, our claim is not that exploratory hypothesis tests are *always* more compelling than confirmatory hypothesis tests or even that they are *typically* more compelling. Our claim is only that specific exploratory tests *can be* more compelling than specific confirmatory tests in specific research situations. We begin by considering the argument for preregistering confirmatory hypothesis tests, which assumes that there is a fundamental reason to regard exploratory results as being more tentative and uncertain than confirmatory results.

Preregistration is not necessary to confirm the validity of hypothesis tests

The preregistration argument

Preregistration involves the formal, verifiable documentation of a study's planned hypotheses, methods, and analyses prior to data collection and analyses (e.g., Nosek et al., 2018). Preregistration increases research transparency (e.g., Paul et al., 2021, p. 54), but it does not automatically increase

research quality (Henderson, 2022; Nosek et al., 2018). Instead, preregistration's increased transparency helps to provide a clearer distinction between confirmatory and exploratory hypothesis tests. Preregistration advocates argue that it is important to clarify this distinction because confirmatory tests have several evidential advantages over exploratory tests. Confirmatory tests tend to be more *severe* (Mayo, 2018) when questionable research practices are prevalent (Lakens, 2019, p. 227). In addition, the results of confirmatory tests entail a 'lower risk of bias' (Hardwicke & Wagenmakers, 2021) and less 'uncertainty' (Nosek et al., 2018, p. 2601) and so tend to be less 'tentative' than the results of exploratory tests (Errington et al., 2021, p. 19; Nelson et al., 2018, p. 519; Nosek & Lakens, 2014, p. 138; Simmons et al., 2021, p. 154). Consequently, researchers' conclusions should be 'appropriately weighted in favor of the confirmatory outcomes' (Chambers & Tzavella, 2022, p. 36), and 'exploratory studies cannot be presented as strong evidence in favor of a particular claim' (Wagenmakers et al., 2012, p. 635).

A key reason for the proposed evidential advantage of confirmatory hypothesis tests relates to the fact that their results cannot be involved in their rationales. Confirmatory hypotheses are generated *before*, rather than *after*, their associated test results become known, and so their theoretical rationales cannot refer to those test results (Nosek et al., 2018). In other words, the rationale for an *ante hoc* (confirmatory) hypothesis cannot depend on the result of a test of that hypothesis. In contrast, the rationale for a *post hoc* (exploratory) hypothesis *may* refer to its own test result. For example, the hypothesis that $X > Y$ cannot have used the particular result that $x_1 > y_1$ in its rationale unless that result was known prior to the generation of the hypothesis (i.e., unless the hypothesis was generated *post hoc*).

The concern about whether a test result has been used as part of the rationale of a hypothesis has implications for the validity of the corresponding hypothesis test. According to statistical theory, it is invalid to use a result as part of the rationale for a hypothesis and to then argue that the same result provides a legitimate 'test' that increases or decreases support for that hypothesis (e.g., Kriegeskorte et al., 2009; Nosek et al., 2018, p. 2600; Wagenmakers et al., 2012, p. 633). Such circular reasoning invalidates the *use novelty* principle (e.g., Worrall, 2010, 2014), according to which a result cannot provide additional independent support for a hypothesis if it has already been 'used' in the rationale for that hypothesis. For example, a test result cannot be used to provide support for a hypothesis if it has already been used to determine the predicted effect size for the associated hypothesis test. Similarly, it is tautological to conclude that a nonsignificant result is due to low observed (achieved) power (O'Keefe, 2007). Relatedly, if a result is used to distinguish between

different parts of a data set (e.g., to distinguish between Group A and Group B), then it cannot be used again to provide further support for that hypothetical distinction without causing a selection bias (Kriegeskorte et al., 2009; see also André, 2022).

One of the proposed benefits of preregistration is that it helps to identify use novel results by providing a publicly verifiable distinction between (a) results that were known *before* hypotheses were generated, and so *may* have been used as part of the rationale for those hypotheses, and (b) results that were only known *after* hypotheses were tested, and so *cannot* have been used in the rationale for those hypotheses.

It is important to appreciate that preregistration does not prevent researchers from conducting unplanned tests of post hoc hypotheses (e.g., Simmons et al., 2021, p. 154). In other words, the argument for preregistering confirmatory hypothesis tests does not prohibit researchers from deviating from their planned research and undertaking exploratory hypothesis tests.

It is also important to appreciate that the preregistration argument does not imply that the rationales of *all* post hoc hypotheses depend on their own test results. Consequently, the preregistration argument does not assume that *all* exploratory results contravene the use novelty principle. The concern is only that exploratory results have the *potential* to contravene this principle, whereas confirmatory results do not. Consequently, the preregistration argument assumes that, all other things being equal, an exploratory result should be regarded as being more ‘uncertain’ and ‘tentative’ than a confirmatory result, because it is less clear whether the exploratory result has been used in the rationale for the associated post hoc hypothesis (e.g., Nosek et al., 2018, p. 2600). Hence, the preregistration argument delegitimizes the evidential value of exploratory results on the grounds that their use novelty is unclear and ambiguous.

Objections to the preregistration argument

Several critics have objected to the distinction between confirmatory and exploratory hypothesis tests and the associated assumptions in the preregistration argument (e.g., Devezer et al., 2021; Lewandowsky, 2019; Oberauer & Lewandowsky, 2019; Rubin, 2020, 2022; Szollosi & Donkin, 2021). Consistent with the use novelty principle, it is accepted that a hypothesis cannot receive additional support from a result that has been used in the rationale for that hypothesis (Rubin, 2020, 2022). However, there are two important nuances to the use novelty principle that allow a result to retain its use novelty in an unplanned, exploratory analysis that uses the same data to generate and test a hypothesis, even when the result has inspired that hypothesis.

First, a hypothesis whose rationale depends on one result may undergo an additional independent test in the same data analysis and receive legitimate support from a second independent test result. For example, as Worrall (2003, 2010) explained, researchers may use a test result to fix (specify) an otherwise free parameter in an ante hoc theory. They may then deduce a new post hoc hypothesis from this more specific, customized theory ('deduction from the phenomena'). This post hoc hypothesis can then be tested and receive independent support from another result in the same data analysis (see Path 1 in Figure 1). In the same vein, several experts agree that it is valid to use the result from one statistical test to help to create another statistical hypothesis/model as long as the test statistic value from the first test is independent from the test statistic value for the second test (Chow, 1998, p. 186; Devezer et al., 2021; Fisher, 1935, p. 194; Kriegeskorte et al., 2009, p. 535; Mayo, 2014, pp. 81–82; Spanos, 2010, p. 216; Worrall, 2010, p. 131). Here, 'independent' means that the first result has a different epistemic source (i.e., it is derived from a different test) to that of the second result. In this case, although the two tests may refer to the same raw data, the second test poses 'a different question' to this data (Fisher, 1935, p. 194), and the test statistic value of the second test will be use novel with respect to that question. Hence, contrary to the preregistration argument, the 'double use' of the same data to generate predictions and then test them is not necessarily problematic (Devezer et al., 2021; Hahn, 2011; Kriegeskorte et al., 2009; cf. Wagenmakers et al., 2012).

Second, a hypothesis may be deduced entirely from ante hoc theory, evidence, and background knowledge, even if that deduction occurs

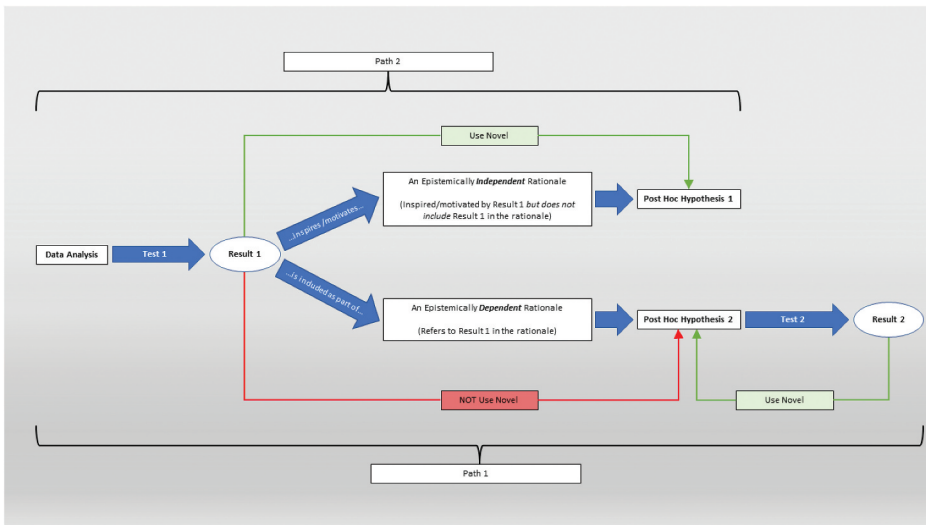


Figure 1. An illustration of two ways in which exploratory data analyses may provide legitimate support for post hoc hypotheses.

after the test result for the hypothesis has become known (Brush, 2015, p. 78; Keynes, 1921, pp. 305–306; Mill, 1843, p. 23; Oberauer & Lewandowsky, 2019; Scerri & Worrall, 2001; Worrall, 2010, 2014). This post hoc deduction provides an epistemically independent basis for the hypothesis that renders the associated test result use novel and, therefore, capable of increasing or decreasing evidential support for that hypothesis in a valid manner (Rubin, 2022; Scerri & Worrall, 2001, p. 424; Worrall, 2014).² Critically, a test result continues to be use novel for a hypothesis even if it has inspired or motivated the deduction of that hypothesis from independent theory and evidence (see Path 2 in Figure 1). As long as the test result is not formally included as part of the deduction (as represented in the published research report), it can be used to increase or decrease support for the associated hypothesis (Howson, 1984, 1985; Rubin, 2022; Worrall, 2014).

Given these two nuances, critics have argued that the distinction between confirmatory and exploratory hypothesis tests is too broad and imprecise because not all exploratory hypothesis tests violate the use novelty principle. In particular, a post hoc hypothesis whose rationale depends on one result from the current data analysis meets the use novelty principle if it is tested using a *different* (independent) result (e.g., Fisher, 1935, p. 194; Kriegeskorte et al., 2009, p. 535). In addition, a post hoc hypothesis that has been deduced from independent theory and evidence meets the use novelty principle if its test result is not included as part of that deduction (e.g., Rubin, 2022; Worrall, 2014).

Of course, both paths that are depicted in Figure 1 may be implemented in preregistered confirmatory analyses as well as in non-preregistered exploratory analyses. However, critics have also argued that preregistration is not necessary to identify use novelty (Rubin, 2020). Certainly, if a test is preregistered, then its result cannot violate the use novelty principle. However, as discussed earlier, the results of many non-preregistered, exploratory hypothesis tests may also be use novel, and preregistration is not required to distinguish between results that are use novel and results that are not use novel. Instead, a more direct and accurate method of identifying whether a result is use novel is to check whether it is included as part of the formal rationale for the associated hypothesis test (Rubin, 2022). If the result is *included* in the rationale, then it is *not* use novel. If the result is *excluded* from the rationale, then it *is* use novel. For example, preregistration is not required to check whether an observed effect size is used to determine the power of the test of that effect. For further examples of this content-based approach to assessing use novelty, please see Kriegeskorte et al.'s (2009, p. 536) analysis of 134 fMRI papers.

In summary, preregistration is a blunt and unnecessary tool for detecting violations of the use novelty principle. It is blunt because it does not distinguish between valid and invalid exploratory hypothesis tests. Consequently, it will often throw out the baby (use novel results) with the bathwater (nonuse novel results). Preregistration is also unnecessary because violations of use novelty can be identified by checking the *content* of the explanations for hypotheses (theoretical rationales) without knowing the *timing* of the construction of those explanations (i.e., ante hoc, post hoc). A test result is not use novel for a hypothesis if it is included in the formal explanation for that hypothesis.

Importantly, according to this critique of the preregistration argument, the use novelty principle does not provide a legitimate reason to argue that the result of an unplanned, post hoc, exploratory hypothesis test, which may have been inspired by or depend on other results in the data analysis, should be regarded as being more tentative or uncertain than that of a preregistered confirmatory test. Yes, circular reasoning may invalidate the use novelty of *some* exploratory results, and yes, this issue would cast doubt on the validity of *all* exploratory tests if there was no way of determining use novelty other than by preregistering confirmatory tests. However, there *is* another way of determining use novelty: Circular reasoning can be identified by checking the *contents* of that reasoning without knowing the *timing* of the reasoning. Consequently, the mere fact that exploratory results have the *potential* to violate the use novelty principle is not sufficient to view any particular exploratory result as being more tentative or uncertain than a confirmatory result. Instead, it is more appropriate to regard each specific exploratory hypothesis test as being either *invalid* or *valid* depending on whether its test result *has* or *has not* violated the use novelty principle, based on an assessment of the content of its test rationale (Kriegeskorte et al., 2009).

Although the use novelty principle does not provide a fundamental reason to regard specific exploratory results as being more tentative than confirmatory results, exploratory hypothesis tests may nonetheless suffer from several other disadvantages that impact on the credibility of their results. For example, they may be more open to bias than confirmatory hypothesis tests (e.g., Hardwicke & Wagenmakers, 2021). We consider some of these potential disadvantages in the third section of this article. However, a balanced assessment must also consider the potential *advantages* of exploratory hypothesis tests, and it is to these that we now turn. We argue that exploratory tests possess several advantages, which allow them to sometimes deliver *more* compelling evidence than confirmatory tests. In other words, we argue that exploratory results can sometimes be *less* uncertain and tentative than confirmatory results.

The advantages of exploratory hypothesis tests

Distinguishing predictions, hypotheses, and accommodation

Before discussing the advantages of exploratory hypothesis tests, it is useful to adopt more precise and clearly defined terminology. In particular, it is useful to distinguish study-specific *predictions* (e.g., ‘scores on this measure of self-esteem will be higher among male participants than among female participants in this study’) from more general and broader *hypotheses* (e.g., ‘men have higher self-esteem than women’) that are deduced from theories via theoretical rationales (e.g., ‘according to gender-status theory, men tend to have higher status in society than women, and higher status promotes higher self-esteem’). As Calder et al. (2021) pointed out, this distinction allows a clearer separation between expected associations between variables (i.e., study-specific predictions) and explanations for those associations in terms of constructs (i.e., theory-based hypotheses). This distinction is important in the current context because researchers preregister study-specific *predictions*, and they consider non-preregistered predictions to be ‘exploratory’ even if those predictions refer to the same broad *hypothesis*.

We also distinguish between *ante hoc* predictions and *post hoc* predictions. Ante hoc predictions are generated *before* the results of the associated data analysis are known, whereas post hoc predictions are generated *after* these results are known. As discussed earlier, post hoc predictions can be generated in two ways, both of which ensure epistemic independence from the result that they predict. First, if the rationale for a post hoc prediction depends on a result in the current data analysis, then it can be tested using a second independent result in that data analysis (i.e., Worrall, 2003, 2010). For example, after observing that $x_1 > y_1$, a researcher may create the new post hoc hypothesis that $X > Y$ and, based on this hypothesis, deduce the new prediction that $x_2 > y_2$, which is then tested and supported using the independent test result that $x_2 > y_2$. Second, a post hoc prediction may be deduced from ante hoc theory, evidence, and background knowledge after its test result is known (see also Scerri & Worrall, 2001, p. 424). For example, after observing that $x_1 > y_1$, a researcher may search the literature and find an existing theory that hypothesizes that $X > Y$.

Both ante hoc and post hoc prediction can be distinguished from *accommodation*. Accommodation can take two forms: *degenerative* and *progressive* (Worrall, 2010, pp. 143–144). A degenerative accommodation does not entail any novel predictions. It merely accommodates the current test result in an ad hoc manner. In contrast, a progressive accommodation not only accommodates the current test result (e.g., $x_1 > y_1$), but also allows the deduction of a post hoc prediction about an additional independent test result (e.g., $x_2 > y_2$). Hence, a progressive accommodation violates the use novelty principle with respect to the test result that it accommodates, but

not with respect to test results that it subsequently predicts (Worrall, 2010). It is worth noting that many of the past criticisms of exploratory hypothesis tests have tended to focus on accommodation (for discussions, see Hitchcock & Sober, 2004). Instead, in the current article, we consider exploratory hypothesis testing in relation to post hoc prediction.

Ante hoc and post hoc predictions refer to confirmatory and exploratory hypothesis tests respectively. Confirmatory hypothesis tests entail ante hoc predictions that have been deduced from independent theory and evidence before the data analysis. In contrast, exploratory hypothesis tests entail post hoc predictions that (a) are based on one result but tested by a different (epistemically independent) result or (b) deduced from ante hoc theory, evidence, and background knowledge. Below, we consider six advantages of testing post hoc predictions that increase the evidential value of exploratory hypothesis tests relative to confirmatory hypothesis tests.

(1) Post hoc predictions avoid researcher commitment biases

Publicly preregistering ante hoc predictions and hypotheses may trigger a number of *researcher commitment biases* that deter researchers from considering alternative post hoc predictions and hypotheses. In particular, an *automation bias* (Lyell & Coiera, 2017) and *plan continuation bias* (Winter et al., 2020) may lead researchers to make decision-making errors because they stick to their preregistered plan when a different approach is more appropriate. People are also less likely to change their attitude about an issue and more likely to implement their planned intentions when they have made an explicit commitment about the issue (e.g., Ajzen et al., 2009; Bettinghaus & Baseheart, 1969; Halverson & Pallak, 1978; Ronay et al., 2017). Hence, the act of preregistration may lead to an *attitudinal bias* in favor of the preregistered plan. Finally, the public preregistration of predictions may exacerbate the *first is best bias*, which inclines people to view their first ideas (e.g., ante hoc hypotheses) as being their best ideas (Carney & Banaji, 2012; Ihme & Wittwer, 2015).

In summary, the act of preregistering a research plan may reduce researchers' potential to deviate from that plan and increase their preference for and commitment to the plan. In turn, these commitment biases may deter researchers from testing alternative post hoc predictions and hypotheses that may provide better explanations of their observed results.

We should point out that our proposal that the preregistration of ante hoc predictions causes researcher commitment biases is only supported by general research, which has not focused on the scientific attitudes and behaviors of researchers. More specific, metascientific research is required to establish (a) whether these general findings apply to the more specific case of researchers and (b) to compare the impact of researcher commitment biases to other researcher biases, such as selective reporting.

(2) *Post hoc predictions reduce the probability of data fraud*

Dellsén (2021) recently argued that, compared to ante hoc prediction, accommodation reduces the probability of data fraud. Specifically, researchers may commit intentional or unintentional data fabrication and/or manipulation in order to produce results that fit their ante hoc predictions. However, they are less likely to commit data fraud when they engage in accommodation because, in this case, they are able to design their post hoc hypothesis to fit their results. Hence, the motive for data fraud is reduced in the case of accommodation, and greater confidence in the accuracy of the data should lead to greater trust in the research results and conclusions.

Based on economic modeling, Felgenhauer (2021) put forward a similar argument: Preregistration may discourage p -hacking but, in doing so, it may increase the rate of faked data. Following Dellsén (2021) and Felgenhauer (2021), we propose that post hoc prediction may also be regarded as yielding more trustworthy results than ante hoc prediction because it reduces the probability of data fraud.

(3) *Post hoc predictions avoid the researcher prophecy bias*

A researcher may include assumptions in the rationale for their prediction that are not entailed by the theory that they are testing (*holistic underdetermination*; Stanford, 2017).³ For example, in the absence of any clear theoretical rationale for a precise effect size, a researcher may make an educated guess about the predicted effect size, perhaps with reference to background knowledge about average effect sizes in the field. In addition, weak and vague theories allow the generation of multiple potential predictions, some of which may be consistent with other relevant theories (Kerr, 1998, p. 210; *contrastive underdetermination*; Stanford, 2017). Again, this theoretical underspecification requires researchers to make arbitrary choices about which specific prediction they will test (Del Giudice & Gangestad, 2021).

Importantly, the predictive success of a researcher's arbitrary decisions should not contribute to our evaluation of the predictive success of the associated theory (Keynes, 1921; Mill, 1843; Oberauer, 2019). As Keynes explained, if an ante hoc prediction is the result of a 'mere guess, [then] the lucky fact of its preceding some or all of the cases which verify it adds nothing whatever to its value' (p. 349). Nonetheless, contra Keynes, when evaluating a successful ante hoc prediction, people may not adequately distinguish between (a) a researcher's prophetic luck ('they were right') and (b) a theory's predictive power ('the theory was right'). Indeed, people may misattribute the success of the researcher's lucky guess to the verisimilitude of the theory, leading to an artificially inflated sense of confidence in the research conclusion (for an illustration in the case of manager and entrepreneurs, see Denrell & Fang, 2010; see also Cooper & Guest, 2014,

p. 43; Oberauer, 2019; Pham & Oh, 2020; Szollosi et al., 2020; Vancouver, 2018, p. 78). This misattribution of a researcher's prophetic luck to the verisimilitude of a theory may be termed a *researcher prophecy bias*.

Like the previously discussed researcher commitment biases, we are unable to estimate the size and impact of the researcher prophecy bias. Obviously, this bias is not sufficiently powerful to halt scientific progress. Nonetheless, it may represent a serious impediment to progress. Again, metascientific research is required to determine its impact relative to other biases, such as *p*-hacking.

What we can say for sure is that the practice of making transparently post hoc predictions will eliminate the researcher prophecy bias: No-one is impressed when you pick the winning horse *after* the race is run! The explicitly post hoc nature of the prediction process focuses attention on the predictive performance of theoretical deductions unconfounded from the atheoretical prophetic luck of the researcher (e.g., Brush, 2015, p. 78; Mill, 1843, p. 23). To be clear, researchers may still 'fill in the gaps' when making post hoc predictions by supplementing their theoretical rationales with additional information based on guesswork and/or background knowledge that is not entailed by the theory in question. Some of this information may turn out to be irrelevant to the prediction's success, and some may turn out to be critical and need to be incorporated into the theory. Only trial and error will tell. Whatever the case, in the post hoc situation, researchers' arbitrary decisions about these auxiliary assumptions will be clearly unconfounded from their personal prophetic luck. For example, readers cannot be impressed by a researcher's prophetic luck if the researcher does not make any ante hoc predictions and instead provides an explicitly post hoc argument that explains why the size of their observed effect is expected to be consistent with the average effect size in the field. Hence, transparent post hoc prediction allows a more accurate and diagnostic appraisal of the predictive power of theories independent from the prophetic luck of researchers.

(4) Post hoc predictions are more appropriate in the context of unplanned deviations

The methods that researchers end up using often differ from their planned approach (e.g., a smaller than planned sample size). In addition, some data analyses may reveal unexpected information that affects the validity of planned data analyses and calls for alternative, unplanned analyses. For example, preregistered preliminary tests may reveal that planned statistical models are inappropriate and that alternative but unplanned models are more suitable. In these cases, pragmatic researchers will deviate from their preregistered plans to adapt their predictions and analyses to the current context (Ansell & Samuels, 2016, p. 1811; Claesen et al., 2021). These

deviations produce post hoc predictions. Hence, post hoc predictions can be more appropriate than ante hoc predictions in the context of unplanned deviations (Devezer et al., 2021, p. 17).

To illustrate, consider a researcher who preregisters a study to test a gender difference in self-esteem using a previously validated 5-item self-esteem scale. During a preregistered test of the scale's internal reliability, the researcher finds that, unexpectedly, the scale has unacceptable internal consistency. An unplanned, exploratory analysis finds that Item 4 has caused the low internal reliability, most likely due to an error that the researcher made in the wording of that item. Further exploratory analyses find that removing Item 4 from the scale increases the scale's reliability to an acceptable level. Now, of course, with the benefit of hindsight, the researcher would have (a) undertaken sequential preregistration in which their first preregistered step would be to confirm the internal reliability of their self-esteem scale and/or (b) preregistered a decision tree in which they indicated what they would do in the event that the scale showed unacceptable internal reliability (Nosek et al., 2018, p. 2602). However, hindsight is not foresight, and the reality is that our researcher did not consider either of these options. Consequently, they find themselves in a similar position to many other researchers in the real world who identify an issue that compromises the appropriateness of their preregistered analysis plan (e.g., Abrams et al., 2020; Claesen et al., 2021; Heirene et al., 2021; see also Reinhart, 2015, p. 95).

Further imagine that our researcher proceeds to conduct both (a) their preregistered test, which uses the unreliable 5-item scale and (b) an unplanned test that uses a reliable version of the scale that excludes Item 4. However, the two tests yield contradictory results. Which result should have most influence on the final research conclusion (for a discussion of this type of problem, see Del Giudice & Gangestad, 2021, p. 4; Lakens, 2019, p. 226; Rubin, 2017a, pp. 326–327)? Note that the unplanned test refers to a post hoc prediction, because the decision to exclude an item from the self-esteem scale and the choice of which item to exclude both depend on a result from the current data analysis. However, it is clear that this dependency does not compromise the use novelty of the gender difference test, which refers to a different result to that of the scale reliability test. In other words, although the rationale for the post hoc prediction depends on the result of the scale reliability test, the test of the post hoc prediction refers to an independent test result (i.e., whether male participants have higher or lower self-esteem scores than female participants). Consequently, the unplanned test that uses the more reliable 4-item self-esteem scale provides a valid and, consequently, more compelling test than the preregistered test that uses the unreliable 5-item scale. Hence, in this case, the researcher should favor the result of the unplanned test of their post hoc prediction when reaching a final research conclusion.

Again, we are not suggesting that advocates of preregistration and confirmatory hypothesis tests are opposed to deviating from their planned tests. Instead, we are arguing that it is normal and justifiable for researchers to deviate from their preregistered plans and that, in this context, post hoc predictions can be more compelling than ante hoc predictions because they are more appropriate in the context of the changed circumstances of the study (see also Devezer et al., 2021, p. 17)

(5) Post hoc predictions facilitate inference to the best explanation

Inference to the best explanation involves a comparative assessment of the most plausible theoretical explanations of the observed results based on a variety of comparison dimensions (Calder et al., 2021; Greene, 2022; Haig, 2009; Mackonis, 2013). Compared to ante hoc predictions, post hoc predictions have a greater potential to facilitate this process of inference to the best explanation by allowing the consideration of previously unanticipated, but potentially superior, explanations. Put another way, post hoc predictions have a greater potential of avoiding incorrect theoretical inferences by making tests more severe (probative; Mayo, 2018; Morey, 2019).

For example, a researcher may undertake planned Test A of Theory A, find a positive result, and then claim support for Theory A. However, if they proceed to undertake several unplanned tests of post hoc predictions (e.g., Tests B, C, & D), then they may encounter additional results that lead them to believe that another, previously unanticipated theory – Theory B – provides a better explanation of their initial result. Hence, post hoc predictions can increase confidence in research conclusions by allowing a more thorough examination of alternative unanticipated explanations for the research results in order to achieve a more rigorous (severe) inference to the best explanation (Morey, 2019). Note that it is often impossible to plan thorough tests of alternative explanations in advance because the information that is required to do so may only become available during the data analyses. For example, Test D may only become relevant after Test C has yielded an unexpected result.

There is clear anecdotal evidence that post hoc predictions are used as a corrective to inferences based on ante hoc predictions. Readers are encouraged to search for the phrase ‘however, exploratory analyses’ in Google Scholar in order to view the numerous occasions in which the results of unplanned tests have had a substantial effect on researchers’ final theoretical inferences. For example, in his study of the emotion of awe as a predictor of interest in science, McPhetres (2019, Study 1) concluded that ‘although the preregistered analyses are consistent with the theoretical model, the exploratory results also lend to the possibility that awe has the effects it does because it is a positive emotion’ (p. 1605).

(6) Post hoc predictions allow peer reviewers to make additional contributions at the data analysis stage

Finally, post hoc predictions allow peer reviewers to make additional contributions at the data analysis stage by suggesting exploratory analyses in light of information provided by the current research results. As per the previous two advantages, the results of these additional post hoc analyses may help to make the final research conclusions more compelling.

Summary

In summary, compared to ante hoc predictions, post hoc predictions can provide more compelling results because they avoid researcher commitment and researcher prophecy biases, reduce the probability of data fraud, are more appropriate in the context of unplanned deviations, facilitate inference to the best explanation, and allow peer reviewers to make additional contributions at the data analysis stage. In particular, a greater willingness to consider post hoc explanations, a more appropriate level of confidence in the research conclusions, a greater probability of accurate data, more appropriate tests in the context of unplanned deviations, and a more rigorous investigation of alternative unanticipated explanations, all supported by advice from expert peer reviewers, can result in more convincing research conclusions relative to the more constrained approach of testing ante hoc predictions. Hence, a key conclusion here is that unplanned tests of post hoc predictions (i.e., exploratory hypothesis tests) can yield more compelling research conclusions than planned tests of ante hoc predictions (i.e., confirmatory hypothesis tests).

Addressing the disadvantages of exploratory hypothesis tests

We have established that the use novelty principle does not provide a sufficient reason to presume that specific exploratory results are *more* tentative and uncertain than confirmatory results. We have also established that there are some potential advantages of exploratory hypothesis tests that may make their results *more* compelling than those of confirmatory tests. However, to obtain a balanced view, we must also consider some potential *disadvantages* of exploratory hypothesis tests. Hence, in the third part of this article, we consider some common concerns that might be raised against the practice of testing post hoc predictions. We argue that none of these concerns prevent exploratory results from being more compelling than confirmatory results in specific situations. Again, we do not claim that exploratory results are *always* more compelling than confirmatory results or even that they are *typically* more compelling. We only claim that specific

exploratory results *can sometimes be* more compelling than specific confirmatory results.

(1) Post hoc predictions cannot be falsified

One objection to testing post hoc predictions is that they cannot be falsified (e.g., Kerr, 1998). However, there is evidence that this objection is incorrect. In particular, Kepes et al. (2022, p. 10) found that 41.7% of potentially HARKed (secretly post hoc) hypotheses were associated with nonsignificant results.

Like ante hoc predictions, post hoc predictions can be generated and disconfirmed by the current test results (Rubin, 2022). Indeed, researchers often follow this approach in the Discussion sections of their research reports, where they (and their peer reviewers) undertake a post hoc generation of alternative explanations for their results (i.e., post hoc predictions) and then provisionally rule out (disconfirm) these explanations through reasoned argument and additional post hoc tests. Again, this probative approach increases the severity of researchers' final inferences (Mayo, 2018).

(2) Post hoc predictions may cause overfitting

Another common concern is that post hoc predictions may overfit their test results (e.g., Hitchcock & Sober, 2004; Kriegeskorte et al., 2009). Again, however, this concern is unwarranted. Overfitting can only occur in the case of accommodation. By definition, it cannot occur in the case of prediction, even if that prediction is post hoc.

If a post hoc prediction is deduced from independent theory and evidence, then it cannot overfit its test result because its rationale is epistemically independent from that test result (Rubin, 2022). Of course, the test result may inspire the deduction of a prediction that perfectly fits the result. But it is more appropriate to describe this process as one of *targeted deduction* rather than accommodation and, if this deduction is based on unsound premises and/or an overly complex theoretical rationale, then the associated hypothesis will suffer an evaluative disadvantage during a process of inference to the best explanation (Rubin, 2022).

Overfitting is only possible when the current research results are formally incorporated into the rationale for a hypothesis or model (i.e., accommodation). Even in this case, a progressive accommodation may make a new post hoc prediction, and that new prediction cannot overfit an independent test result, because the independent test result does not form part of the rationale for that prediction. In short, the problem of overfitting only applies to cases of degenerative accommodation and not to cases of either ante hoc or post hoc prediction.

(3) *Post hoc predictions encourage questionable research practices*

Another concern is that post hoc predictions may encourage the use of questionable research practices (QRPs), such as failing to report all of a study's dependent measures or conditions (John et al., 2012). This concern implies that QRPs are unacceptable research practices. For example, Hartgerink and Wicherts (2016, p. 1) defined QRPs as 'practices that are detrimental to the research process ... [and that] harm the research process', and Chambers (2014) described QRPs as 'soft fraud'. However, contrary to this view, John et al. (2012, p. 531) noted that there is a great degree of variability in the acceptability of most QRPs, and Rubin (2020, 2022) pointed out that there are legitimate reasons for engaging in QRPs under certain conditions (see also Table 7, Moran et al., 2021; Sacco et al., 2019). Hence, QRPs are not always unacceptable.

Furthermore, there should be no 'unquestionable research practices' that are automatically accepted as being 'correct'. We should not allow demonstrations of planning or preregistration to provide a false sense of security about the validity of our research approaches. Research practices that have been preregistered may be biased and problematic (Devezer et al., 2021; Szollosi et al., 2020) and QRPs may be perfectly acceptable given a suitable context and verifiable justification (Fiedler & Schwarz, 2016; Moran et al., 2021; Rubin, 2020, 2022; Sacco et al., 2019). Ultimately, the assessment of any claim comes from an attempt to understand and criticize the contents of the scientific arguments that have been presented. Simplistic heuristics, such as 'QRPs tend to be problematic' or 'exploratory results tend to be more tentative' represent a form of *methodologism* (Chamberlain, 2000; Gao, 2014) that should not contribute to any such evaluation, since only the contents of the specific arguments themselves matter.

(4) *Post hoc predictions are unnecessary*

Another objection is that, in theory, it is possible for researchers to plan out all possible relevant analyses before they know their results, thereby making post hoc predictions redundant. According to this perspective, researchers are able to anticipate the entire *garden of forking paths* of potential sample-contingent data analyses that they might undertake (Gelman & Loken, 2013, 2014). However, contrary to this view, most researchers do not possess what (Navarro, 2020) described as 'godlike planning abilities' (see also Ansell & Samuels, 2016). Consequently, in practice, 'no analysis plan survives contact with the data' (Reinhart, 2015, p. 95, paraphrasing Helmuth von Moltke). Indeed, recent research shows that researchers often deviate from their preregistered plans (e.g., Abrams et al., 2020; Claesen et al., 2021; Heirene

et al., 2021). Hence, in practice, post hoc predictions are almost always necessary.

(5) Post hoc predictions allow researchers to predict anything

It might also be objected that the practice of generating post hoc predictions allows researchers to predict any result (e.g., Kerr, 1998, p. 210). This is true. A researcher can observe their current result and then generate a bespoke prediction that is confirmed by that result. However, a successful prediction is only a starting point for a scientific inference. During a process of inference to the best explanation, it is also necessary to compare the accompanying theoretical explanation with alternative explanations in terms of theoretical virtues such as parsimony, plausibility, specificity, and internal and external consistency (Greene, 2022; Kuhn, 1977; Mackonis, 2013; Rubin, 2022; Szollosi & Donkin, 2021). For example, a weak theory that can predict every potential result in a study should be preferred less than a stronger theory that can predict only a few potential results (Roberts & Pashler, 2000, p. 359). Hence, the question is not *whether* a researcher can make a successful post hoc prediction, because most researchers can cobble together a theoretical explanation that will predict their current result. Instead, the question is *how good* is the theoretical explanation for that prediction relative to other potential explanations. As scientists, we should be most impressed by the ‘best’ theoretical explanation of a post hoc prediction and least impressed by the ‘worst’ explanation.

(6) Post hoc predictions are susceptible to researcher bias

A further concern about post hoc predictions is that they are susceptible to various biases on the part of the researcher (e.g., Hardwicke & Wagenmakers, 2021). Four points mitigate this concern.

First, in some cases, the researcher bias that influences the reporting of results in unplanned exploratory research (*selective reporting*; e.g., *p*-hacking) may be *less* problematic than the researcher bias that influences the generation of the preregistered hypotheses, the design of preregistered methods and analyses to test these hypotheses, and the preregistered interpretation of the potential results (*selective questioning*; e.g., Clark & Tetlock, 2022; Dellsén, 2020; Jamieson et al., 2022; Landy et al., 2020; Silberzahn et al., 2018). Consequently, in some cases, the conclusion of a preregistered confirmatory test may be *more biased* than that of an exploratory test because the *selective questioning* in the confirmatory test is more influential than the *selective reporting* in the exploratory test.

Second, although preregistration may reduce some biases, such as the confirmation and hindsight biases (Hardwicke & Wagenmakers, 2021), it

may also increase other biases, such as the previously discussed automation bias, plan continuation bias, commitment bias, first-is-best bias, and researcher prophecy biases. Hence, in what we might term the *law of the conservation of bias*: Bias can neither be created nor destroyed; it may only be converted from one form to another! According to our ‘law’, it is naïve to assume that we can reduce one type of bias without increasing another. Instead, it is more realistic to try to expose biases and reflect on their potential influence (Field & Derksen, 2021), and this is why it is important for researchers to engage in open science practices such as making their research material and data available for scrutiny and providing robustness analyses (Rubin, 2020).

Third, preregistered confirmatory analyses may reduce a bias against the reporting of null results (e.g., Scheel et al., 2021). However, this bias is only potentially problematic in certain contexts. In particular, if researchers use a Neyman-Pearson (NP) test that is sufficiently powered to detect their smallest effect size of interest (SESOI), then they may regard a null result as informative.⁴ However, they must also acknowledge that an effect that is smaller than their SESOI may be present (Lakens et al., 2018, p. 260). Furthermore, researchers who acknowledge that their NP test has insufficient power to detect their SESOI should regard the probability of incorrectly accepting their null hypothesis as being unacceptably high. Consequently, failing to report such a null result should not be regarded as particularly problematic. A reporting bias is also less problematic when using a trichotomous NP approach, in which researchers either accept, reject, or ‘decide to remain in doubt’ about null hypotheses (Neyman & Pearson, 1933a, pp. 295–296; 1933b, p. 493). In this case, researchers who fail to report results that have led them to ‘remain in doubt’ will not bias their substantive conclusions. Similarly, a reporting bias against null results is not problematic in the Fisherian approach, in which null results only allow researchers to *fail to reject* null hypotheses rather than to *accept* them (Fisher, 1956, p. 45). In this case, null results represent the absence of evidence rather than evidence of an absence (Altman & Bland, 1995). Finally, during Bayesian hypothesis testing, evidence may be classed as ‘barely worth mentioning’ (Jeffreys, 1961, p. 432) when Bayes factors are close to 1.00, and so a bias against reporting this evidence is not particularly problematic. In summary, a reporting bias against null results is only problematic when using an NP test that is sufficiently powered to detect an SESOI. It is less problematic (a) when using an insufficiently powered NP test, (b) when an NP result falls into a region of doubt, (c) when using a Fisherian test, and (d) when a Bayes factor is close to 1.00.⁵

Finally, from a use novelty perspective, it does not matter whether a researcher (a) constructs a hypothesis because it fits a result or (b) reports a result because it fits a hypothesis. Instead, what matters is whether the

knowledge that is represented by the result is included in the researcher's theoretical rationale for the hypothesis. Hence, a result may inspire, motivate, or 'bias' a researcher to construct a theoretical rationale. However, if the result is not formally included as part of that rationale then it is epistemically independent from the associated hypothesis, and it may be used to either increase or decrease support for that hypothesis without contravening the use novelty principle.⁶ Hence, from a use novelty perspective, we should be more concerned about the *epistemic independence* between hypotheses and results than about the *operational independence* between researchers and either hypotheses or results.

(7) *The undisclosed timing of post hoc predictions is unethical*

Finally, post hoc predictions may be regarded as unethical when their post hoc timing is undisclosed. This research practice is sometimes referred to as *hypothesizing after the results are known* or HARKing (Kerr, 1998). There are three points to note here.

First, if a test result from the current data analysis is a crucial part of the rationale for a post hoc prediction, then it will not be possible to conceal this fact without rendering the rationale unacceptably unclear. Hence, researchers will find it difficult to deceive their readers about post hoc predictions whose rationales depend on their current results.

Second, although the word *prediction* is used to mean *foretell* or *prophecy* in everyday language, in scientific usage it is used to mean *deduce from a theory*, regardless of whether that deduction occurs before or after a relevant test result has become known (Brush, 2015, p. 78). Consequently, phrases such as 'we predicted that ...' and 'as hypothesized, ...' apply equally well to *both* ante hoc *and* post hoc predictions because theoretical deductions are timeless (Rubin, 2020, 2022; Worrall, 2014). In other words, in a scientific context, it is not deceptive to use the phrase 'as predicted' to refer to a post hoc prediction.

Finally, any active deception by a researcher about when a prediction has been deduced (i.e., before or after a test result was known) will not prevent readers from undertaking a valid evaluation of (a) the quality of the theory, (b) the quality of the deduction of the prediction from that theory (Rubin, 2022; see also Szollosi & Donkin, 2021, p. 5), and (c) the quality of the research methodology and the statistical analyses that are used to test the prediction. Consequently, HARKing does not conceal what Kerr (1998) described as 'a useful part of the "truth"' (p. 209).

In summary, researchers cannot deceive readers about a result-dependent post hoc prediction without fatally damaging the rationale for that prediction; they can use phrases such as 'as predicted' to refer to post hoc predictions; and they can deceive readers about the timing of the deduction

of post hoc predictions without obstructing a valid evaluation of their research. Hence, there are reasons to believe that the undisclosed timing of post hoc predictions is not problematic for valid scientific inference and, consequently, not unethical in this respect.

Summary

In summary, contrary to common objections, post hoc predictions (a) *can* be disconfirmed, (b) *do not* risk overfitting, (c) *do not* necessarily encourage unacceptable research practices, (d) *are* almost always necessary in practice, (e) can be used to predict anything but *may suffer an evaluative cost for doing so* in a process of inference to the best explanation, (f) are *not* necessarily more biased than ante hoc predictions, and (g) are *not* necessarily unethical when their post hoc timing is undisclosed. Consequently, it is possible for the advantages of exploratory hypothesis tests to outweigh their potential disadvantages and for specific exploratory results to be more compelling than confirmatory results.

General summary and conclusions

Preregistration advocates have put forward two key arguments to support the view that, all other things being equal, exploratory results are more uncertain and tentative than confirmatory results. The first, more fundamental argument is based on the use novelty principle. For example, Nosek et al. (2018, p. 2600) argued that the problem is ‘circular reasoning – generating a hypothesis based on observing data, and then evaluating the validity of the hypothesis based on the same data’. In the first part of this article, we discussed two important nuances to the use novelty principle that allow a result to retain its use novelty with respect to a hypothesis even when (a) the same data has been used to generate the result and hypothesis, and (b) the result has inspired the hypothesis. Based on these two nuances, we criticized the confirmatory-exploratory distinction as being too broad and imprecise, and we criticized preregistration as being a blunt and unnecessary tool for detecting violations of the use novelty principle. We argued that a more direct and accurate method of assessing use novelty is to check the contents of the formal rationale for a hypothesis. Based on these points, we concluded that the use novelty principle does not provide a sound basis for arguing that exploratory results are more tentative and uncertain than confirmatory results.

The second argument is that confirmatory hypothesis tests tend to have more advantages and fewer disadvantages than exploratory hypothesis tests. For example, they are supposed to be less open to bias than exploratory tests (e.g., Hardwicke & Wagenmakers, 2021). Contrary to this

argument, in the second part of this article, we highlighted three biases that are potentially problematic for confirmatory hypothesis tests (researcher commitment bias; higher probability of data fraud; researcher prophecy bias) and three advantages of exploratory hypothesis tests (more appropriate in the context of unplanned deviations; more rigorous inference to the best explanation; additional input from peer reviewers during data analysis). In addition, in the third part of the article, we addressed seven potential disadvantages of exploratory hypothesis tests. We concluded that exploratory hypothesis tests can have more advantages and fewer disadvantages than confirmatory tests and that, consequently, exploratory hypothesis tests can yield *more* compelling research conclusions than confirmatory tests. Again, however, we caution that metascientific research is required to identify situations in which the reverse may be true.

To be clear, we are dismissing neither the ubiquity nor the utility of confirmatory hypothesis tests. Ante hoc predictions are an important part of science, and they can have some advantages. Hence, we are not arguing that confirmatory results are *always* more tentative than exploratory results or even that they are *typically* more tentative. We are merely urging researchers to reconsider the assumption in the preregistration argument that, compared to exploratory results, confirmatory results are less open to bias, less uncertain, less tentative, represent stronger evidence, and deserve greater weighting (Chambers & Tzavella, 2022, p. 36; Errington et al., 2021, p. 19; Hardwicke & Wagenmakers, 2021; Nelson et al., 2018, p. 519; Nosek & Lakens, 2014, p. 138; Nosek et al., 2018, p. 2601; Simmons et al., 2021, p. 154; Wagenmakers et al., 2012, p. 635). To invert the reasoning of the preregistration argument, we think that confirmatory analyses should be allowed in exploratory research! However, we also think that researchers should concede that ante hoc, preregistered predictions may lead to an exaggerated level of confidence in their research conclusions, less appropriate tests in the context of unplanned deviations, and greater bias and errors in theoretical inferences.

How do our conclusions relate to the concern that the false portrayal of exploratory results as confirmatory results may have led to overconfidence about replicability and, consequently, to the replication crisis (e.g., Nosek et al., 2018, p. 2600)? In our view, the causes of the replication crisis may be entirely unrelated to the confirmatory-exploratory distinction. For example, the crisis may have been caused by a base rate fallacy, heterogeneous effects, poor validity, low power, and/or hidden moderators (e.g., Bird, 2020; De Boeck & Jeon, 2018; Fabrigar et al., 2020; Maxwell et al., 2015; Rubin, 2021a). A lack of attention to factors such as these may have led to overconfidence in replicability independent from the confirmatory-exploratory distinction.

Finally, at a broader level, the benefits of exploratory hypothesis tests that have been outlined here follow from the general principle that scientific progress comes from finding and correcting problems with our existing knowledge. According to this principle, scientific progress depends less on confirming our preexisting knowledge and more on specifying and exploring our ignorance (Firestein, 2012; Merton, 1987). In this context, it should not be surprising that a confirmatory hypothesis test can be inferior to its exploratory counterpart. A confirmatory analysis is only appropriate to the extent that its theoretical assumptions remain intact. However, since we conduct our experiments with a mind to destroy those exact foundations, we should rarely expect a confirmatory test to remain useful for long. On the other hand, exploratory tests must be the reason that we learn new things. So, while it is true that it is easy to build entire literatures on the basis of bad exploratory analyses, the solution is not to abandon an exploratory approach, but rather to rise to the challenge of understanding how we can use this approach to make good scientific arguments.

Notes

1. Exploratory *hypothesis tests* should not be confused with exploratory descriptive research in which researchers do not make any statistical claims about hypotheses (e.g., Tukey, 1977). In this paper, we consider exploratory *hypothesis tests*, in which statistical claims are made about hypotheses that have been generated after test results are known.
2. Note that, from this perspective, a test result may be use novel for a hypothesis when information that is similar to, or the same as, the test result is used in the theoretical rationale provided that the source of the information is shown to be epistemically independent from that of test result (e.g., the information can be traced to independent theory, evidence, or background knowledge).
3. Relatedly, commenting on the ability of researchers to obtain significant results with relatively small sample sizes, Baumeister (2016, p. 156) famously referred to ‘an intuitive flair for how to set up the most conducive situation and produce a highly impactful procedure’.
4. Some have argued that Neyman-Pearson null hypothesis testing is not valid in exploratory research due to an unknown inflation of the alpha level (e.g., Nosek & Lakens, 2014; Nosek et al., 2018). This argument is correct when the alpha level refers to a familywise error rate that is defined by all of the tests that are included in a study (Rubin, 2017b, 2021b). However, researchers are not usually interested in this *studywise error* rate, because they are not usually interested in testing the associated joint studywise null hypothesis, because this hypothesis does not usually possess any theoretical relevance. Instead, researchers are more interested in testing theoretically informative joint null hypotheses, and the relevant familywise error rates for these joint hypotheses can be computed on the fly (Rubin, 2021b, p. 10,992). See also Mayo (1996, Chapter 9) for the argument that the Neyman-Pearson approach is tenable in non-preregistered research.

5. In our view, the publication of null results is important when estimating the size of an effect that is already presumed to exist (i.e., effect size estimation) but not when determining *whether* an effect exist (i.e., basic hypothesis testing). If an effect exists (i.e., the null hypothesis is false), then selection for significance will inflate its reported size, because smaller, nonsignificant instances of the effect (i.e., false negatives) will not be reported. In this case, the conclusion that the effect exists will be valid (i.e., a valid decision during hypothesis testing), but the size of the effect will be overestimated (i.e., a biased estimation of the effect size). If an effect does not exist, then selection for significance will not make it appear as if it does exist, assuming that false positive confirmations and disconfirmations are reported with equal frequency.
6. Note that there is no free lunch to be had in using a result as part of a theoretical rationale for a hypothesis and then removing that result from the formal statement of the rationale in order to use it to claim support for the associated hypothesis. In this case, it is legitimate to use the result to add empirical support for the hypothesis. However, without the result in its formal theoretical rationale, the hypothesis now has a different and most likely less compelling theoretical basis, and readers will take this reduced plausibility into account when they make a final overall evaluation of the researcher's claims. Hence, if a researcher uses a result to make their formal rationale more convincing, then they cannot use it again to provide support for the associated hypothesis. And, if they remove the result from their formal rationale, then they can use it to provide support for the associated hypothesis, but the rationale for this hypothesis will now be less convincing.

Acknowledgement

We are grateful to E. J. Wagenmakers and Brian Haig for their generous feedback on previous versions of this article. Our acknowledgement does not imply their endorsement of any of the views that we have presented in this article.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Notes on contributors

Mark Rubin is a professor of social psychology at Durham University, UK. His recent work has focused on issues connected with the replication crisis in science, such as preregistration and HARKing. For more information, please visit: <http://bit.ly/rubinsync>

Chris Donkin is a professor of computational modeling at LMU Munich, Germany. He uses computational and mathematical models to help study various aspects of memory and decision-making.

ORCID

Mark Rubin  <http://orcid.org/0000-0002-6483-8561>

Chris Donkin  <http://orcid.org/0000-0002-4285-8537>

References

- Abrams, E., Libgober, J., & List, J. A. (2020). *Research registries: Facts, myths, and possible improvements* (No. w27250). National Bureau of Economic Research. <https://doi.org/10.3386/w27250>
- Ajzen, I., Czasch, C., & Flood, M. G. (2009). From intentions to behavior: Implementation intention, commitment, and conscientiousness. *Journal of Applied Social Psychology*, 39(6), 1356–1372. <https://doi.org/10.1111/j.1559-1816.2009.00485.x>
- Altman, D. G., & Bland, J. M. (1995). Absence of evidence is not evidence of absence. *BMJ*, 311(7003), 485–485. Article 485. <https://doi.org/10.1136/bmj.311.7003.485>
- André, Q. (2022). Outlier exclusion procedures must be blind to the researcher’s hypothesis. *Journal of Experimental Psychology General*, 151(1), 213–223. <https://doi.org/10.1037/xge0001069>
- Ansell, B., & Samuels, D. (2016). Journal editors and “results-free” research: A cautionary note. *Comparative Political Studies*, 49(13), 1809–1815. <https://doi.org/10.1177/00104140166669369>
- Baumeister, R. F. (2016). Charting the future of social psychology on stormy seas: Winners, losers, and recommendations. *Journal of Experimental Social Psychology*, 66, 153–158. <https://doi.org/10.1016/j.jesp.2016.02.003>
- Bettinghaus, E. P., & Baseheart, J. R. (1969). Some specific factors affecting attitude change. *The Journal of Communication*, 19(3), 227–238. <https://doi.org/10.1111/j.1460-2466.1969.tb00845.x>
- Bird, A. (2020). Understanding the replication crisis as a base rate fallacy. *The British Journal for the Philosophy of Science*, 72(4), 965–993. <https://doi.org/10.1093/bjps/axy051>
- Brush, S. G. (2015). *Making 20th Century science: How theories became knowledge*. Oxford University Press.
- Calder, B. J., Brendl, C. M., Tybout, A. M., & Sternthal, B. (2021). Distinguishing constructs from variables in designing research. *Journal of Consumer Psychology*, 31(1), 188–208. <https://doi.org/10.1002/jcpy.1204>
- Carney, D. R., & Banaji, M. R. (2012). First is best. *PloS One*, 7(6), e35088. <https://doi.org/10.1371/journal.pone.0035088>
- Chamberlain, K. (2000). Methodolatry and qualitative health research. *Journal of Health Psychology*, 5(3), 285–296. <https://doi.org/10.1177/135910530000500306>
- Chambers, C. (2014, June 10). Physics envy: Do ‘hard’ sciences hold the solution to the replication crisis in psychology?. *The Guardian*. <http://www.theguardian.com/science/head-quarters/2014/jun/10/physics-envy-do-hard-sciences-hold-the-solution-to-the-replication-crisis-in-psychology>
- Chambers, C. D., & Tzavella, L. (2022). The past, present and future of Registered Reports. *Nature Human Behavior*, 6(1), 29–42. <https://doi.org/10.1038/s41562-021-01193-7>
- Chow, S. L. (1998). Précis of statistical significance: Rationale, validity, and utility. *The Behavioral and Brain Sciences*, 21(2), 169–194. <https://doi.org/10.1017/S0140525X98001162>
- Claesen, A., Gomes, S., Tuerlinckx, F., Vanpaemel, W., & Leuven, K. U. (2021). Comparing dream to reality: An assessment of adherence of the first generation of preregistered studies. *Royal Society Open Science*, 8(10), 1–11. <https://doi.org/10.1098/rsos.211037>
- Clark, C. J., & Tetlock, P. E. (2022). Adversarial collaboration: The next science reform. In C. L. Frisby, R. E. Redding, W. T. O’Donohue, & S. O. Lilienfeld (Eds.), *Political bias in psychology: Nature, scope, and solutions*. Springer.

- Cooper, R. P., & Guest, O. (2014). Implementations are not specifications: Specification, replication and experimentation in computational cognitive modeling. *Cognitive Systems Research, 27*, 42–49. <https://doi.org/10.1016/j.cogsys.2013.05.001>
- De Boeck, P., & Jeon, M. (2018). Perceived crisis and reforms: Issues, explanations, and remedies. *Psychological Bulletin, 144*(7), 757–777. <https://doi.org/10.1037/bul0000154>
- Del Giudice, M., & Gangestad, S. W. (2021). A traveler's guide to the multiverse: Promises, pitfalls, and a framework for the evaluation of analytic decisions. *Advances in Methods and Practices in Psychological Science, 4*(1), 1–15. <https://doi.org/10.1177/2515245920954925>
- Dellsén, F. (2020). The epistemic impact of theorizing: Generation bias implies evaluation bias. *Philosophical Studies, 177*(12), 3661–3678. <https://doi.org/10.1007/s11098-019-01387-w>
- Dellsén, F. (2021). *An epistemic advantage of accommodation over prediction*. <http://philsci-archive.pitt.edu/19298/>
- Denrell, J., & Fang, C. (2010). Predicting the next big thing: Success as a signal of poor judgment. *Management Science, 56*(10), 1653–1667. <https://doi.org/10.1287/mnsc.1100.1220>
- Devezer, B., Navarro, D. J., Vandekerckhove, J., & Buzbas, E. O. (2021). The case for formal methodology in scientific reform. *Royal Society Open Science, 8*(3), 1–26. <https://doi.org/10.1098/rsos.200805>
- Errington, T. M., Denis, A., Perfito, N., Iorns, E., & Nosek, B. A. (2021). Reproducibility in cancer biology: Challenges for assessing replicability in preclinical cancer biology. *ELife, 10*, e67995. <https://doi.org/10.7554/eLife.67995>
- Fabrigar, L. R., Wegener, D. T., & Petty, R. E. (2020). A validity-based framework for understanding replication in psychology. *Personality and Social Psychology Review, 24* (4), 316–344. <https://doi.org/10.1177/1088868320931366>
- Felgenhauer, M. (2021). Experimentation and manipulation with preregistration. *Games and Economic Behavior, 130*, 400–408. <https://doi.org/10.1016/j.geb.2021.09.002>
- Fiedler, K., & Schwarz, N. (2016). Questionable research practices revisited. *Social Psychological and Personality Science, 7*(1), 45–52. <https://doi.org/10.1177/1948550615612150>
- Field, S. M., & Derksen, M. (2021). Experimenter as automaton; experimenter as human: Exploring the position of the researcher in scientific research. *European Journal for Philosophy of Science, 11*(1), 1–21. <https://doi.org/10.1007/s13194-020-00324-7>
- Firestein, S. (2012). *Ignorance: How it drives science*. Oxford University Press.
- Fisher, R. A. (1935). *The design of experiments*. Oliver & Boyd.
- Fisher, R. A. (1956). *Statistical methods and scientific inference*. Oliver & Boyd.
- Gao, Z. (2014). Methodologism—methodological imperative. In T. Teo (Ed.), *Encyclopedia of critical psychology* (pp. 1189–1193). Springer. https://doi.org/10.1007/978-1-4614-5583-7_614
- Gelman, A., & Loken, E. (2013). *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*. http://www.stat.columbia.edu/gelman/research/unpublished/p_hacking.pdf
- Gelman, A., & Loken, E. (2014). The statistical crisis in science. *American Scientist, 102*(6), 460. <https://doi.org/10.1511/2014.111.460>
- Greene, J. A. (2022). What can educational psychology learn from, and contribute to, theory development scholarship? *Educational Psychology Review*. <https://doi.org/10.1007/s10648-022-09682-5>

- Hahn, U. (2011). The problem of circularity in evidence, argument, and explanation. *Perspectives on Psychological Science*, 6(2), 172–182. <https://doi.org/10.1177/1745691611400240>
- Haig, B. D. (2009). Inference to the best explanation: A neglected approach to theory appraisal in psychology. *The American Journal of Psychology*, 122(2), 219–234. <https://doi.org/10.2307/27784393>
- Halverson, R. R., & Pallak, M. S. (1978). Commitment, ego-involvement, and resistance to attack. *Journal of Experimental Social Psychology*, 14(1), 1–12. [https://doi.org/10.1016/0022-1031\(78\)90056-2](https://doi.org/10.1016/0022-1031(78)90056-2)
- Hardwicke, T. E., & Wagenmakers, E. (2021, April 23). Preregistration: A pragmatic tool to reduce bias and calibrate confidence in scientific research. <https://doi.org/10.31222/osf.io/d7bcu>
- Hartgerink, C. H. J., & Wicherts, J. M. (2016). Research practices and assessment of research misconduct. *ScienceOpen Research*, 0(0), 1–10. <https://doi.org/10.14293/S2199-1006.1.SOR-SOCSCI.ARYSBL.v1>
- Heirene, R., LaPlante, D., Louderback, E. R., Keen, B., Bakker, M., Serafimovska, A., & Gainsbury, S. M. (2021). Preregistration specificity and adherence: A review of preregistered gambling studies & cross-disciplinary comparison. <https://doi.org/10.31234/osf.io/nj4es>
- Henderson, E. L. (2022, January 25). A guide to preregistration and Registered Reports. *MetaArxiv*. <https://doi.org/10.31222/osf.io/x7aqr>
- Hitchcock, C., & Sober, E. (2004). Prediction versus accommodation and the risk of overfitting. *The British Journal for the Philosophy of Science*, 55(1), 1–34. <http://dx.doi.org/10.1093/bjps/55.1.1>
- Howson, C. (1984). Bayesianism and support by novel facts. *The British Journal for the Philosophy of Science*, 35(3), 245–251. <https://www.jstor.org/stable/687475>
- Howson, C. (1985). Some recent objections to the Bayesian theory of support. *British Journal for the Philosophy of Science*, 36(3), 305–309. <https://www.jstor.org/stable/687574>
- Ihme, N., & Wittwer, J. (2015). The role of consistency, order, and structure in evaluating and comprehending competing scientific explanations. *Instructional Science*, 43(4), 507–526. <https://doi.org/10.1007/s11251-015-9349-6>
- Jamieson, M. K., Pownall, M., & Govaart, G. H. (2022). Reflexivity in quantitative research: A rationale and beginner's guide. *PsyArxiv*. <https://doi.org/10.31234/osf.io/xvrhm>
- Jeffreys, H. (1961). *The theory of probability* (3rd ed.). Oxford University.
- John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, 23(5), 524–532. <https://doi.org/10.1177/0956797611430953>
- Kepes, S., Keener, S. K., McDaniel, M. A., & Hartman, N. S. (2022). Questionable research practices among researchers in the most research-productive management programs. *Journal of Organizational Behavior*. <https://doi.org/10.1002/job.2623>
- Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review*, 2(3), 196–217. https://doi.org/10.1207/s15327957pspr0203_4
- Keynes, J. M. (1921). *A treatise on probability*. MacMillan.
- Kriegeskorte, N., Simmons, W. K., Bellgowan, P. S., & Baker, C. I. (2009). Circular analysis in systems neuroscience: The dangers of double dipping. *Nature Neuroscience*, 12(5), 535–540. <https://doi.org/10.1038/nn.2303>
- Kuhn, T. S. (1977). *The essential tension: Selected studies in the scientific tradition and change*. The University of Chicago.

- Lakens, D. (2019) The value of preregistration for psychological science: A conceptual analysis. *Japanese Psychological Review*, 62(3), 221–230 https://doi.org/10.24602/sjpr.62.3_221
- Lakens, D., Scheel, A. M., & Isager, P. M. (2018). Equivalence testing for psychological research: A tutorial. *Advances in Methods and Practices in Psychological Science*, 1(2), 259–269. <https://doi.org/10.1177/2515245918770963>
- Landy, J. F., Jia, M., Ding, I. L., Viganola, D., Tierney, W., Dreber, A., Johannesson, M., Pfeiffer, T., Ebersole, C. R., Gronau, Q. F., Ly, A., van den Bergh, D., Marsman, M., Derks, K., Wagenmakers, E.-J., Proctor, A., Bartels, D. M., Bauman, C. W., Brady, W. J. . . . Uhlmann, E. L. (2020). Crowdsourcing hypothesis tests: Making transparent how design choices shape research results. *Psychological Bulletin*, 146(5), 451–479. <https://doi.org/10.1037/bul0000220>
- Lewandowsky, S. (2019, January 22). Avoiding Nimitz Hill with more than a Little Red Book: Summing up #PSprereg. *Psychonomic Society*. <https://featuredcontent.psychonomic.org/avoiding-nimitz-hill-with-more-than-a-little-red-book-summing-up-psprereg/>
- Lyell, D., & Coiera, E. (2017). Automation bias and verification complexity: A systematic review. *Journal of the American Medical Informatics Association*, 24(2), 423–431. <https://doi.org/10.1093/jamia/ocw105>
- Mackonis, A. (2013). Inference to the best explanation, coherence and other explanatory virtues. *Synthese*, 190(6), 975–995. <https://doi.org/10.1007/s11229-011-0054-y>
- Maxwell, S. E., Lau, M. Y., & Howard, G. S. (2015). Is psychology suffering from a replication crisis? What does “failure to replicate” really mean? *The American Psychologist*, 70(6), 487–498. <https://doi.org/10.1037/a0039400>
- Mayo, D. G. (1996). *Error and the growth of experimental knowledge*. Chicago University Press.
- Mayo, D. G. (2014). Some surprising facts about (the problem of) surprising facts (from the Dusseldorf Conference, February 2011). *Studies in History and Philosophy of Science Part A*, 45, 79–86. <https://doi.org/10.1016/j.shpsa.2013.10.005>
- Mayo, D. G. (2018). *Statistical inference as severe testing*. Cambridge University Press.
- McPhetres, J. (2019). Oh, the things you don't know: Awe promotes awareness of knowledge gaps and science interest. *Cognition & Emotion*, 33(8), 1599–1615. <https://doi.org/10.1080/02699931.2019.1585331>
- Merton, R. K. (1987). Three fragments from a sociologist's notebooks: Establishing the phenomenon, specified ignorance, and strategic research materials. *Annual Review of Sociology*, 13(1), 1–29 <https://doi.org/10.1146/annurev.so.13.080187.000245>
- Mill, J. S. (1843). *A system of logic, ratiocinative and inductive: Being a connected view of the principles of evidence and the methods of scientific investigation (Vol. 2)*. Parker.
- Moran, C., Richard, A., Wilson, K., Twomey, R., & Coroiu, A. (2021, July 23). “I know it's bad but I have been pressured into it”: Questionable research practices among psychology students in Canada. <https://doi.org/10.31234/osf.io/kjby3>
- Morey, R. (2019). You must tug that thread: Why treating preregistration as a gold standard might incentivize poor behavior. *Psychonomic Society*. <https://featuredcontent.psychonomic.org/you-must-tug-that-thread-why-treating-preregistration-as-a-gold-standard-might-incentivize-poor-behavior/>
- Navarro, D. (2020). Paths in strange spaces: A comment on preregistration. <https://doi.org/10.31234/osf.io/wxn58>
- Nelson, L. D., Simmons, J., & Simonsohn, U. (2018). Psychology's renaissance. *Annual Review of Psychology*, 69(1), 511–534. <https://doi.org/10.1146/annurev-psych-122216-011836>

- Neyman, J., & Pearson, E. S. (1933a). IX. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society A*, 231(694–706), 289–337. <https://doi.org/10.1098/rsta.1933.0009>
- Neyman, J., & Pearson, E. S. (1933b). The testing of statistical hypotheses in relation to probabilities a priori. *Mathematical Proceedings of the Cambridge Philosophical Society*, 29(4), 492–510. <https://doi.org/10.1017/S030500410001152X>
- 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. <https://doi.org/10.1073/pnas.1708274114>
- Nosek, B. A., & Lakens, D. (2014). Registered reports. *Social Psychology*, 45(3), 137–141. <https://doi.org/10.1027/1864-9335/a000192>
- O’Keefe, D. J. (2007). Post hoc power, observed power, a priori power, retrospective power, prospective power, achieved power: Sorting out appropriate uses of statistical power analyses. *Communication Methods and Measures*, 1(4), 291–299. <https://doi.org/10.1080/19312450701641375>
- Oberauer, K. (2019, January 15). Preregistration of a forking path – What does it add to the garden of evidence?. *Psychonomic Society*. <https://featuredcontent.psychonomic.org/preregistration-of-a-forking-path-what-does-it-add-to-the-garden-of-evidence/>
- Oberauer, K., & Lewandowsky, S. (2019). Addressing the theory crisis in psychology. *Psychonomic Bulletin & Review*, 26(5), 1596–1618. <https://doi.org/10.3758/s13423-019-01645-2>
- Paul, M., Govaert, G. H., & Schettino, A. (2021). Making ERP research more transparent: Guidelines for preregistration. *International Journal of Psychophysiology*, 164, 52–63. <https://doi.org/10.1016/j.ijpsycho.2021.02.016>
- Pham, M. T., & Oh, T. T. (2020). Preregistration is neither sufficient nor necessary for good science. *Journal of Consumer Psychology*, 31(1), 163–176. <https://doi.org/10.1002/jcpy.1209>
- Reinhart, A. (2015). Statistics done wrong: The woefully complete guide. *No Starch Press*.
- Roberts, S., & Pashler, H. (2000). How persuasive is a good fit? a comment on theory testing. *Psychological Review*, 107(2), 358–367. <https://doi.org/10.1037/0033-295x.107.2.358>
- Ronay, R., Oostrom, J. K., Lehmann-Willenbrock, N., & Van Vugt, M. (2017). Pride before the fall: (Over) confidence predicts escalation of public commitment. *Journal of Experimental Social Psychology*, 69, 13–22. <https://doi.org/10.1016/j.jesp.2016.10.005>
- Rubin, M. (2017a). An evaluation of four solutions to the forking paths problem: Adjusted alpha, preregistration, sensitivity analyses, and abandoning the Neyman-Pearson approach. *Review of General Psychology*, 21(4), 321–329. <https://doi.org/10.1037/gpr0000135>
- Rubin, M. (2017b). Do p values lose their meaning in exploratory analyses? It depends how you define the familywise error rate. *Review of General Psychology*, 21(3), 269–275. <https://doi.org/10.1037/gpr0000123>
- Rubin, M. (2020). Does preregistration improve the credibility of research findings? *The Quantitative Methods for Psychology*, 16(4), 376–390. <https://doi.org/10.20982/tqmp.16.4.p376>
- Rubin, M. (2021a). What type of Type I error? Contrasting the Neyman-Pearson and Fisherian approaches in the context of exact and direct replications. *Synthese*, 198(6), 5809–5834. <https://doi.org/10.1007/s11229-019-02433-0>
- Rubin, M. (2021b). When to adjust alpha during multiple testing: A consideration of disjunction, conjunction, and individual testing. *Synthese*, 199(3–4), 10969–11000. <https://doi.org/10.1007/s11229-021-03276-4>

- Rubin, M. (2022). The costs of HARKing. *The British Journal for the Philosophy of Science*, 73(2), 535–560. <https://doi.org/10.1093/bjps/axz050>
- Sacco, D. F., Brown, M., & Bruton, S. V. (2019). Grounds for ambiguity: Justifiable bases for engaging in questionable research practices. *Science and Engineering Ethics*, 25(5), 1321–1337. <https://doi.org/10.1007/s11948-018-0065-x>
- Scerri, E., & Worrall, J. (2001). Prediction and the periodic table. *Studies in History and Philosophy of Science Part A*, 32(3), 407–452. [https://doi.org/10.1016/S0039-3681\(01\)00023-1](https://doi.org/10.1016/S0039-3681(01)00023-1)
- Scheel, A. M., Schijen, M. R., & Lakens, D. (2021). An excess of positive results: Comparing the standard Psychology literature with Registered Reports. *Advances in Methods and Practices in Psychological Science*, 4(2), 1–12. <https://doi.org/10.1177/25152459211007467>
- Silberzahn, R., Uhlmann, E. L., Martin, D. P., Anselmi, P., Aust, F., Awtrey, E., Bahník, Š., Bai, F., Bannard, C., Bonnier, E., Carlsson, R., Cheung, F., Christensen, G., Clay, R., Craig, M. A., Dalla Rosa, A., Dam, L., Evans, M. H., Flores Cervantes, I. . . Nosek, B. A. (2018). Many analysts, one data set: Making transparent how variations in analytic choices affect results. *Advances in Methods and Practices in Psychological Science*, 1(3), 337–356. <https://doi.org/10.1177/2515245917747646>
- Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2021). Pre-registration: Why and how. *Journal of Consumer Psychology*, 31(1), 151–162. <https://doi.org/10.1002/jcpy.1208>
- Spanos, A. (2010). Akaike-Type criteria and the reliability of inference: Model selection versus statistical model specification. *Journal of Econometrics*, 158(2), 204–220. <https://doi.org/10.1016/j.jeconom.2010.01.011>
- Stanford, K. (2017). Underdetermination of scientific theory. *Stanford Encyclopedia of Philosophy*. <https://plato.stanford.edu/entries/scientific-underdetermination/>
- Szollosi, A., & Donkin, C. (2021). Arrested theory development: The misguided distinction between exploratory and confirmatory research. *Perspectives on Psychological Science*, 16(4), 717–724. <https://doi.org/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 Science*, 24(2), 94–95. <https://doi.org/10.1016/j.tics.2019.11.009>
- Tukey, J. W. (1977). *Exploratory data analysis*. Pearson.
- Vancouver, J. B. (2018). In defense of HARKing. *Industrial and Organizational Psychology*, 11(1), 73–80. <https://doi.org/10.1017/iop.2017.89>
- Wagenmakers, E. J., Wetzels, R., Borsboom, D., van der Maas, H. L., & Kievit, R. A. (2012). An agenda for purely confirmatory research. *Perspectives on Psychological Science*, 7(6), 632–638. <https://doi.org/10.1177/1745691612463078>
- Winter, S. R., Rice, S., Capps, J., Trombley, J., Milner, M. N., Anania, E. C., Walters, N. W., & Baugh, B. S. (2020). An analysis of a pilot’s adherence to their personal weather minimums. *Safety Science*, 123(104576), 1–7. <https://doi.org/10.1016/j.ssci.2019.104576>
- Worrall, J. (2003). Normal science and dogmatism, paradigms and progress: Kuhn ‘versus Popper and Lakatos. In T. Nickles (Ed.), *Thomas Kuhn* (pp. 65–100). Cambridge University Press.
- Worrall, J. (2010). Theory confirmation and novel evidence. In D. G. Mayo, & A. Spanos (Eds.), *Error and inference: Recent exchanges on experimental reasoning, reliability, and the objectivity and rationality of science* (pp. 125–169). Cambridge University Press.
- Worrall, J. (2014). Prediction and accommodation revisited. *Studies in History and Philosophy of Science*, 45, 54–61. <https://doi.org/10.1016/j.shpsa.2013.10.001>