

# When Does HARKing Hurt? Identifying When Different Types of Undisclosed Post Hoc Hypothesizing Harm Scientific Progress

Mark Rubin

The University of Newcastle, Australia

Hypothesizing after the results are known, or HARKing, occurs when researchers check their research results and then add or remove hypotheses on the basis of those results without acknowledging this process in their research report (Kerr, 1998). In the present article, I discuss 3 forms of HARKing: (a) using current results to construct post hoc hypotheses that are then reported as if they were a priori hypotheses; (b) retrieving hypotheses from a post hoc literature search and reporting them as a priori hypotheses; and (c) failing to report a priori hypotheses that are unsupported by the current results. These 3 types of HARKing are often characterized as being bad for science and a potential cause of the current replication crisis. In the present article, I use insights from the philosophy of science to present a more nuanced view. Specifically, I identify the conditions under which each of these 3 types of HARKing is most and least likely to be bad for science. I conclude with a brief discussion about the ethics of each type of HARKing.

**Keywords:** accommodation, falsification, HARKing, prediction, replication crisis

In his seminal article on the subject, Kerr (1998) described the research practice of *hypothesizing after the results are known* (HARKing). HARKing occurs when researchers check their research results and then add and/or remove hypotheses from their research report on the basis of those results. This process can be disclosed or undisclosed to the readers of research reports (Hollenbeck & Wright, 2017; Schwab & Starbuck, 2017). Following Kerr (1998), the present article is mainly concerned with undisclosed HARKing.

Kerr (1998) distinguished between several different types of HARKing, and these can be grouped into two broad categories. In the first category, researchers *include* one or more post hoc hypotheses in their research report as if they were a priori hypotheses. In the second category, researchers *exclude* one or more a priori hypotheses from their research report. Due to implicit pressures from the wider research community (Fanelli, 2010; Motyl et al., 2017; Nosek, Ebersole, DeHaven, & Mellor, 2017; O'Boyle, Banks, & Gonzalez-Mulé, 2017), researchers usually include post hoc hypotheses that are *confirmed* by their research results and exclude a priori hypotheses that are *disconfirmed* by their results. Hence, from the readers' perspective, a larger proportion of the researchers' "a priori" hypotheses are supported than is actually the case.

To illustrate these two categories of HARKing, consider a researcher who aims to test an a priori hypothesis—Hypothesis A—that expressions of prejudice increase self-esteem. To test this hypothesis, the researcher randomly assigns a sample of partici-

pants to either describe their negative feelings about immigrants or to describe their positive feelings about immigrants. The researcher then measures participants' state self-esteem. Contrary to Hypothesis A, she finds that participants in the negative feelings condition have significantly *lower* self-esteem than those in the positive feelings condition. In an effort to accommodate this unexpected finding, the researcher engages in the two categories of HARKing. First, she constructs a new post hoc hypothesis—Hypothesis B—that predicts the unexpected result. Specifically, Hypothesis B predicts that expressions of prejudice *reduce* self-esteem. The researcher then includes this post hoc hypothesis in her research report as if it was an a priori hypothesis. Second, she removes any mention of Hypothesis A from her research report. Crucially, she does not reveal any of these post hoc changes in her research report. Hence, from the readers' perspective, the researcher predicted and found that the prejudice reduces self-esteem.

Kerr (1998) considered how cases such as the one above may be detrimental to scientific progress. He also differentiated between different types of HARKing and argued that they are unlikely to be equivalent in terms of their potential costs to scientific progress. However, neither he nor subsequent discussants have explored this issue any further (Hollenbeck & Wright, 2017; Kerr, 1998; Leung, 2011; Schwab & Starbuck, 2017). In particular, it remains unclear *when* different types of HARKing will be most likely and least likely to harm science. It is important to address this issue in order to better understand the relationship between HARKing and the replication crisis. In particular, under what conditions do different types of HARKing contribute to the publication of spurious effects that may represent Type I errors?

The current article addresses this question in order to provide a more articulated and sophisticated understanding of HARKing's threat to scientific progress and add some nuance to the common view that "all HARKing is bad." I begin by considering the

This article was published Online First October 26, 2017.

Correspondence concerning this article should be addressed to Mark Rubin, School of Psychology, The University of Newcastle, Behavioural Sciences Building, Callaghan, New South Wales 2308, Australia. E-mail: [mark.rubin@newcastle.edu.au](mailto:mark.rubin@newcastle.edu.au)

assumptions that underpin this common view, focusing in particular on HARKing's preclusion of falsification and its presumed implication in science's replication crisis via the promulgation of false positive results.

### How HARKing Harms Science

HARKing is considered to be problematic for scientific progress because it results in hypotheses that are always confirmed and never falsified by the results. Falsification is an essential part of the scientific process because it allows researchers to distinguish hypotheses that are confirmed (i.e., clearly supported by the evidence) from those that are disconfirmed (Ferguson & Heene, 2012; Kerr, 1998; Leung, 2011). However, as illustrated in the example of the prejudice research study, HARKing can preclude reports of falsification by (a) generating post hoc hypotheses that always confirm the observed results and (b) suppressing a priori hypotheses that have been disconfirmed by those results. Hence, HARKing can harm science by preventing the research community from accurately assessing which hypotheses are true and which are false.

To illustrate HARKing's preclusion of falsification, it is useful to consider the Texas sharpshooter analogy that is often used to describe this problem (e.g., de Groot, 2014; Wagenmakers, Wetzel, Borsboom, Kievit, & van der Maas, 2015). In this analogy, a Texas sharpshooter aims and fires his gun at target on a barn wall but misses. He then walks up to the wall, rubs out the initial target, and draws a second target around his bullet hole in order to make it appear as if he is a good shot. In this analogy, the sharpshooter represents a researcher, the bullet hole is his evidence, the first target is an excluded a priori hypothesis, and the second target is an included post hoc hypothesis. The analogy illustrates the impossibility of reported falsification when HARKing takes place: It does not matter where the sharpshooter's shot hits the barn wall; he will always make it look as if he has hit his target.

HARKing's preclusion of falsification would not be particularly problematic for scientific progress if HARKing was practiced by only a few researchers. However, this is not the case. As shown in Table 1, recent surveys have found that self-admission rates for HARKing are quite high, with close to half of researchers (43%) having HARKed at least once.

Why do researchers HARK? The evidence seems to suggest that they do so in order to increase their chances of publishing their work in a system that places a high value on the hypothetico-deductive approach to science (Fanelli, 2010; Kerr, 1998; Mazzola & Deuling, 2013; Motyl et al., 2017; O'Boyle et al., 2017). For example, Mazzola and Deuling (2013) analyzed 215 published journal articles and 127 unpublished PhD dissertations that were produced in the area of industrial-organizational psychology during 2010–2012. They found that, compared with the unpublished dissertations, the published journal articles contained a significantly higher percentage of supported hypotheses and a significantly lower percentage of unsupported hypotheses. This evidence suggests that researchers HARKed when writing their journal articles by including confirmed post hoc hypotheses and excluding disconfirmed a priori hypotheses, and that these actions facilitated the publication of their articles (for similar results, see O'Boyle et al., 2017).

Despite (or perhaps because of) its relatively widespread occurrence, the research community does not appear to have been particularly concerned about HARKing. Indeed, Kerr's (1998) seminal article on HARKing was discussed in a collection of the most "underappreciated" and "unloved" work by social psychologists (Kerr, 2011). Commenting on his 1998 article, Kerr (2011) lamented that, "if there has been lively discussion and debate on this issue [HARKing] in (or outside of) social psychology since the paper's appearance, it has escaped my notice" (p. 129). However, things have changed markedly since 2011, and the issue of HARKing has now become a hot topic. This increased interest is illustrated in Figure 1, which shows the relative frequency of

Table 1  
Self-Admission Rates of HARKing in Self-Report Surveys

Survey	Population	Survey Item	N	Self-Admission Rate
John et al. (2012)	U.S. psychologists	"In a paper, reporting an unexpected finding as having been predicted from the start."	2,155	27.0%
Agnoli et al. (2017)	Italian psychologists	"In a paper, reporting an unexpected finding as having been predicted from the start."	277	37.4%
Bosco et al. (2016; Study 1)	Researchers who published in <i>Personnel Psychology</i> and the <i>Journal of Applied Psychology</i> during 2005 to 2010	"Whether any changes in hypotheses had occurred between the completion of data collection and subsequent publication"	53	38%
Fiedler and Schwarz (2016)	German psychologists	"Reporting an unexpected finding as having been predicted from the start."	1,138	47%
Banks et al. (2016; Studies 1 & 2)	Management researchers	"Selectively reported hypotheses on the basis of statistical significance . . . and presented a post hoc hypothesis as if it were developed a priori."	749	50%
Motyl et al. (2017; Study 1)	Personality and social psychologists from Australian, European, and the U.S.	"Report that unexpected findings were expected."	1,166	58%
			Mean	43%

Note. Self-admission rates are for undertaking the stated behavior "at least once." Self-admission rates are likely to be underestimates because researchers tend to underreport practices that they perceive to be undesirable (Agnoli et al., 2017).

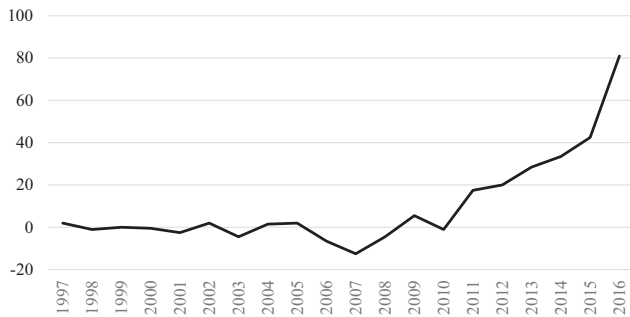


Figure 1. Difference in number citations per year to Kerr's (1998) seminal article on HARKing compared with the average number of citations to two other articles that were published in the same issue of the same journal. The difference in citations remained relatively low during the period 1997–2010 and then increased dramatically after 2011. Data is sourced from Google Scholar.

citations to Kerr's (1998) article over the 20-year period from 1997 to 2016. In order to control for potential changes in citation rates over this period, Figure 1 shows the *differences* between the number of citations to Kerr's (1998) article and the average number of citations to two other articles that were published in the same issue of the same journal (Glaser & Salovey, 1998; Helgeson & Fritz, 1998). As can be seen in Figure 1, Kerr's article had a similar number of citations to the other two articles until 2011 when the number of citations to Kerr's article increased substantially from year to year. In fact, Kerr's article received more citations in 2016 ( $n = 88$ ) than it did in the previous 15-year period from 1997 to 2011 ( $n = 82$ ).

The dramatic rise in citations to Kerr's (1998) article after 2011 may be attributed to the research community's increased concern about research practices in general. This concern was triggered by a number of key events during 2011. In that year, social psychologist Diederik Stapel was suspended from his university on the grounds of fabricating scientific data. (To date, 55 articles have been implicated.) In the same year, social psychologist Daryl Bem (an advocate of certain types of HARKing; Bem, 1987) published questionable work that purported to provide evidence of precognition and premonition (Bem, 2011). Finally, 2011 also saw the publication of an influential paper that demonstrated how researchers can use undisclosed flexibility in their data analyses to produce statistically significant results—an approach that has come to be known as *p-hacking* (Simmons, Nelson, & Simonsohn, 2011). For these reasons, 2011 is usually taken to mark the beginning of serious concerns about the validity of commonly used research practices in psychology and science in general. More recently, these concerns have intensified following an attempt to replicate the findings of 100 psychology studies (Open Science Collaboration, 2015) that found that replicated effect sizes were only half the size of original effect sizes, and that only “39% of effects were subjectively rated to have replicated the original result” (Open Science Collaboration, 2015, p. 943). The reasons for these replication results are the subject of much debate. Certainly, it is difficult to draw strong conclusions about the replicability of an effect based on a single failed replication attempt (e.g., Maxwell, Lau, & Howard, 2015). Nonetheless, several researchers have concluded that the use of *questionable research practices* (John,

Loewenstein, & Prelec, 2012) has caused an unexpectedly low rate of replication in psychology and other fields (e.g., Baker, 2016; Motyl et al., 2017; Munafò et al., 2017; Świątkowski & Dompnier, 2017).

The increased number of citations to Kerr's (1998) article after 2011 suggests that researchers perceive a connection between HARKing and the replication crisis. Consistent with this view, a recent survey of over 1,500 researchers found that “selective reporting of results” was regarded as the most important factor contributing to irreproducible research (Baker, 2016). In addition, several commentators have argued that HARKing is one of the questionable research practices that reduces the replicability of published effects (e.g., Aguinis, Cascio, & Ramani, 2017; Hollenbeck & Wright, 2017; John et al., 2012; Kerr, 1998; Mazzola & Deuling, 2013; Munafò et al., 2017; Schwab & Starbuck, 2017; Świątkowski & Dompnier, 2017; Unkelbach, 2016; Wagenmakers, Wetzels, Borsboom, & van der Maas, 2011). According to this argument, although many researchers claim to conduct confirmatory tests of *a priori* hypotheses, they actually conduct multiple exploratory tests that are uncorrected for having a greater chance of detecting false positive results. They also use *researchers' degrees of freedom* (Simmons et al., 2011) to “massage” their data into yielding statistically significant results (i.e., *p-hacking*). After checking the results of these exploratory tests, researchers then omit any disconfirmed *a priori* hypotheses and construct new post hoc hypotheses that have no potential to be falsified by their results. Finally, they misrepresent their post hoc hypotheses as *a priori* hypotheses in order to give the impression of confirmatory tests. As a consequence, the research literature reports many effects that are supposedly predicted by *a priori* hypotheses but that are actually unanticipated artifacts of multiple uncorrected exploratory tests and/or *p-hacking*. These spurious effects are more likely to be limited to the specific samples in which they were discovered and, consequently, less likely to replicate in other samples.

The above viewpoint has led to the conclusions that (a) HARKing is bad for science, and (b) greater openness and transparency is required in the research process. Two complementary approaches have been put forward to address these issues.

The first approach is for researchers to *preregister* their *a priori* hypotheses, materials, and analysis plans (e.g., Bosco, Aguinis, Field, Pierce, & Dalton, 2016; Lindsay, Simons, & Lilienfeld, 2016; Nosek et al., 2017; Richards, 2016). Preregistration does not prevent researchers from conducting unplanned statistical tests and/or putting forward new post hoc hypotheses. However, it does ensure that these post hoc activities are undertaken openly and transparently, and this transparency allows readers to adjust their expectations regarding the potential replicability of associated results.

The second approach is to *change the culture* of the scientific community so that it is more accepting of exploratory research (e.g., Hollenbeck & Wright, 2017; Schwab & Starbuck, 2017). Such a change is intended to reduce researchers' motivation to engage in undisclosed HARKing. Hence, like preregistration, this cultural change is intended to replace *undisclosed* HARKing with *transparent* HARKing.

Both preregistration and greater cultural acceptance of exploratory research represent important approaches towards pre-

venting undisclosed HARKing. However, both approaches are predicated on the assumption that undisclosed HARKing is bad for science, and it is here that I believe that a more nuanced view may be warranted. HARKing represents an umbrella term for a collection of several different research practices, and each practice can have different implications for scientific progress under different conditions. Consequently, the unqualified assumption that “HARKing is bad for science” is likely to be inaccurate. Instead, it is more likely to be the case that only *some* types of HARKing are bad for science under *some* conditions. In the present paper, I aim to unpack this more articulated view of HARKing in order to allow researchers to make more informed decisions about this questionable research practice.

In the following sections, I leverage several concepts from the philosophy of science in order to identify the specific conditions under which different types of HARKing pose the greatest threat to scientific progress. I argue that treating post hoc hypotheses as if they are a priori hypotheses only has the potential to harm science when those hypotheses lack independence from the observed evidence, and even in this case the potential for harm is reduced when tests are rigorous and hypothesis construction follows certain key principles. I also argue that failing to report a priori hypotheses is only harmful when those hypotheses are related to final research conclusions and/or the subject of rigorous tests.

### Lessons From the Philosophy of Science: Use Novelty and Test Severity

Discussions of HARKing have tended to lag behind related discussions of hypothesis testing in the philosophy of science. Discussions of HARKing often refer to Popperian falsification and the distinction between prediction and accommodation (e.g., Bosco et al., 2016; Hollenbeck & Wright, 2017; Kerr, 1998; Leung, 2011). However, there has been not been any in-depth consideration of more modern insights about the extent to which evidence provides *novel* information in relation to hypotheses (Worrall, 1985, 2010, 2014) or the extent to which *severe* tests of those hypotheses provide diagnostic information (Mayo, 1991, 1996, 2008, 2010, 2014; Mayo & Spanos, 2006). These philosophical insights are important because they provide a basis for determining when different types of HARKing are problematic for science and when they are not. I introduce the concept of use novelty first and then consider test severity.

#### Use Novelty

Many philosophers of science believe that for evidence to confirm a hypothesis the evidence must be not only *consistent* with the hypothesis but also *novel* in some way (e.g., Mayo, 1996; Musgrave, 1974; Worrall, 1985; Zahar, 1973, based on Lakatos, 1970). The general argument is that evidence that is not novel in relation to a hypothesis may be biased towards either confirming or disconfirming that hypothesis because it may have been used as the basis for constructing that hypothesis.

According to the *temporal* conceptualization of novelty, evidence must be observed *after* a hypothesis has been constructed in order for it to represent novel evidence for that hypothesis. From

this perspective, evidence can only be novel for *a priori* hypotheses and not for *post hoc* hypotheses because a priori hypotheses are constructed prior to data collection whereas post hoc hypotheses are constructed after data collection. Consequently, only a priori hypotheses have the potential to *predict* and be *falsified* by evidence (Musgrave, 1974; Worrall, 1985).

However, several philosophers of science have pointed out that temporal novelty is really only a heuristic for distinguishing between hypotheses that are constructed *independent* from the current evidence and hypotheses that are constructed *on the basis of* the current evidence (e.g., Hitchcock & Sober, 2004; Mayo, 1991, 1996; Musgrave, 1974; Worrall, 1985, 2014; Zahar, 1973; see also Dienes, 2016; Kerr, 1998). Hence, a more accurate conceptualization of novelty refers to a hypothesis' independence from the evidence rather than whether or not the hypothesis was constructed before the evidence was known. There are several different conceptualizations of independence (Musgrave, 1974; Worrall, 1985; Zahar, 1973). Here, I refer to Worrall's (1985, 2010, 2014) concept of *use novelty* because it is the most clearly articulated approach (Mayo, 1991). According to Worrall, evidence is only novel for a hypothesis if information about that evidence has not been “used” in the construction of the hypothesis.

It is important to note that the use novelty approach assumes that both a priori *and* post hoc hypotheses have the potential to predict and be falsified by previously-observed evidence (e.g., Worrall, 2014). In particular, post hoc hypotheses can be falsified by previously-observed evidence as long as they are constructed independent from that evidence. This independence ensures that the contents of the hypotheses are not biased towards being either confirmed or disconfirmed by the evidence. Similar reasoning allows post hoc hypotheses to predict a researcher's current results (Worrall, 1985, 2014). If information from a set of evidence is not used in the construction of a hypothesis, then that evidence is use novel for the hypothesis, and the hypothesis may be used to predict that evidence *even if the evidence was observed by the researcher prior to the construction of the hypothesis*.

To provide a practical illustration, imagine that our prejudice researcher conducted her study and obtained the opposite result to that predicted by her original a priori Hypothesis A. In other words, she found that the prejudice *reduced*, rather than *increased*, self-esteem. Lacking any theoretical explanation for this result, the researcher does not try to publish it. Instead, she files it away. Two years later, she reads a newly published paper in which an independent group of researchers has put forward a different hypothesis—Hypothesis B—that prejudice *reduces* self-esteem. From a temporal novelty perspective, Hypothesis B represents a post hoc hypothesis because it was constructed *after* our researcher observed her evidence. Nonetheless, it was constructed independent from this evidence by an independent group of researchers. Consequently, from a use novelty perspective, the researcher's evidence provides an informative test of Hypothesis B. Note that, in this particular scenario, the researcher's evidence supports Hypothesis B. However, this confirmation does not alter the fact the test procedure is unbiased and has the potential to produce disconfirmations. This point can be demonstrated if we imagine that the prejudice researcher read about *two* independent hypotheses after observ-



ing her results. The first hypothesis—Hypothesis B—predicts that prejudice *reduces* self-esteem, whereas the second hypothesis—Hypothesis C—predicts that prejudice has *no effect* on self-esteem. Although our researcher's observed evidence confirms Hypothesis B, it disconfirms Hypothesis C. Hence, the post hoc nature of a hypothesis does not necessarily make it unfalsifiable.

Most researchers tend to employ temporal novelty rather than use novelty when judging hypothesis tests. However, temporal novelty is a rather blunt and sometimes inaccurate criterion with which to gauge the independence between a hypothesis and its associated evidence. Certainly, temporal novelty is *sufficient* to demonstrate independence because it is impossible to construct a hypothesis on the basis of evidence that has yet to be known. However, temporal novelty is not always *necessary* to demonstrate independence because it is possible to construct an independent hypothesis after knowing the evidence that is to be tested by that hypothesis (Mayo, 1991; Worrall, 1985; Zahar, 1973). Consequently, use novelty is more accurate than temporal novelty because it does not incorrectly imply that all post hoc hypotheses lack independence from the known results.

A few previous discussions of HARKing have made indirect reference to the concept of use novelty and its associated prohibition against the double use of evidence in constructing and testing the same hypothesis. For example, Kerr (1998) argued that “it is asking too much of one set of data both to suggest a new hypothesis to an investigator and simultaneously to provide an ‘independent’ empirical confirmation of that hypothesis” (p. 206). More recently, Wagenmakers, Wetzels, Borsboom, van der Maas, and Kievit (2012) made a similar point with regards to exploratory research: “A hypothesis that is developed on the basis of exploration of a data set is unlikely to be refuted by that same data” (p. 633). But these brief nods towards use novelty do not highlight the alternative possibility: Post hoc hypotheses that *are* developed independently from the current results *do* have the potential to predict and be falsified by those results. Kerr (1998) came closest to acknowledging this possibility when he asked, “is it HARKing if one advances a new and useful theory that comes to one's attention independently of but after one's knowledge of the results?” (p. 199). He did not provide a definite answer to this question but instead noted that some types of HARKing fall into a “gray region.” In the present article, I argue that this “gray region” is larger than might be assumed, and that concepts from the philosophy of science can help to clarify the contents of this region and their implications for scientific progress. In particular, based on the distinction between temporal novelty and use novelty, I draw a parallel distinction between hypothesizing *after* the results are known and hypothesizing *on the basis of* the known results. I argue that in cases in which researchers hypothesize *on the basis of* the known results, evidence is not use novel and prediction and falsification are not possible. In contrast, in cases in which researchers hypothesize *after* the results are known but *not on the basis of* those results, use novelty is preserved, prediction and falsification are possible, and there is no detriment to scientific progress. However, before considering these issues in greater depth, it is necessary to briefly introduce a second philosophical concept that has been brought to bear on the issue of hypothesis testing, and that is test severity.

## Test Severity

Worrall's (1985, 2010, 2014) concept of use novelty is restricted to the relation between hypotheses and evidence. However, there are actually three key aspects of any hypothesis test: the hypothesis, the evidence, and the *test* of the hypothesis. Based on Popper (1979), Mayo (1991, 1996, 2008, 2014; Mayo & Spanos, 2006) argued that it is important to consider the rigor of the test of the hypothesis, or *test severity*, when reaching conclusions about the confirmation or disconfirmation of hypotheses. According to Mayo, a test is severe if it has a low probability of confirming a hypothesis that is false. Test severity relates to the statistical and methodological reliability and validity of hypothesis tests.<sup>1</sup> Hence, severe tests are those that use reliable, sensitive measures, high statistical power, and stringent Type I error control (Mayo, 1991; Parker, 2015). Severe tests also avoid *p*-hacking and refer to sensitivity analyses in order to check the robustness of results to different assumptions and procedures (Mayo, 1991). In addition, severe tests use research designs and methods that have a high degree of internal, external, and construct validity (Mayo, 1996). Finally, severe tests employ both direct and conceptual replications within and across studies in order to reduce the probability of incorrect confirming a false hypothesis.

Severe tests are necessary for meaningful confirmation and falsification (Parker, 2015). A nonsevere test provides unreliable and/or invalid test results that cannot be taken to provide stringent evidence either for or against a hypothesis. In particular, confirmation based on a nonsevere test may represent a false positive error: The confirmed hypothesis may actually be false. Similarly, falsification based on a nonsevere test may represent a false negative error: The falsified hypothesis may actually be true. Returning to our Texas sharpshooter analogy, if it is a rainy, windy day and the sharpshooters' rifle sights are not correctly adjusted, then the test of his marksmanship is nonsevere, and we cannot draw any strong conclusions based on any of the targets that he hits or misses.

## Summary

In summary, use novelty refers to the independence between a hypothesis' construction and the observed evidence, and test severity refers to the reliability and validity of tests of the hypothesis. Put another way, use novelty refers to the way in which the hypothesis is produced, whereas test severity refers to the way in which the evidence is produced. Both use novelty and test severity are necessary for hypotheses to meaningfully

<sup>1</sup> Mayo (1991, 1996) proposed that her concept of severity encompasses and supersedes Worrall's (1985) concept of use novelty. Specifically, she argued that although use novelty can contribute to severe tests, it is neither necessary nor sufficient for severity (Mayo, 2014). Instead, Mayo (2008) argued that severity depends on the way in which *both* evidence and hypotheses are generated (i.e., test severity and stringent use-construction rules; Mayo, 2008, 2010, 2014). In response, Worrall (2010) argued that it is helpful to distinguish use novelty from test severity because using evidence to *construct* a hypothesis is different from using evidence to *test* a hypothesis. In the present article, I adopt Worrall's approach and distinguish between use novelty and test severity because this distinction provides greater clarity when discussing HARKing. Nonetheless, I refer to Mayo's (2008, 2010, 2014) arguments regarding stringent hypothesis construction.

predict and be falsified by evidence. Critically, if both use novelty and test severity are achieved, then even post hoc hypotheses can meaningfully predict and be falsified by previously-observed evidence.

In the following sections, I leverage the concepts of use novelty and test severity in order to identify when three different types of HARKing are most and least likely to be detrimental to scientific progress. I begin by considering two types of HARKing in which researchers add new hypotheses to their research report. The first type occurs when researchers construct new hypotheses on the basis of their results and then report those hypotheses as if they are a priori hypotheses. Kerr (1998) described this type of HARKing as “pure HARKing.” In the present article, I use the phrase *constructing hypotheses after the results are known*, or CHARKing, in order to emphasize that researchers are *constructing* new hypotheses.

The second type of HARKing occurs when researchers undertake a post hoc literature search in order to retrieve previously-published hypotheses that are then presented as a priori hypotheses in their research reports. Kerr (1998) described this type of HARKing as “empirically-inspired scholarship.” In the present article, I use the phrase *retrieving hypotheses after the results are known*, or RHARKing, in order to emphasize that researchers are *retrieving* hypotheses from a post hoc literature search.

I then move on to consider a third type of HARKing in which researchers remove disconfirmed a priori hypotheses from their research reports. Kerr (1998) describes this type of HARKing as “suppressing loser hypotheses.” In the present article, I use the phrase *suppressing hypotheses after the results are known*, or SHARKing.

### CHARKing

CHARKing entails the construction of hypotheses that are specifically designed to account for, or *accommodate*, the observed results. Hence, CHARKing involves hypothesizing *on the basis* of the known results and, consequently, the results are not use novel with respect to the hypotheses that are developed.

CHARKing suffers from two key problems. First, it precludes prediction and falsification because it produces hypotheses that are always confirmed by the results (Collins, 1994; Hitchcock & Sober, 2004; Kerr, 1998). Second, it can produce complex hypotheses that contain many caveats and qualifications in order to account for complex and irregular patterns of results (e.g., prejudice only increases self-esteem among black women who are aged 50 years or more). Hitchcock and Sober (2004) described this process as *overfitting* because the resulting hypotheses accommodate not only general effects that exist in the population, but also nonreplicable idiosyncratic effects that are limited to the specific sample (see also Bosco et al., 2016; Gigerenzer, 2004; Schwab & Starbuck, 2017). Overfitting is problematic because it increases the probability that hypotheses will be falsified in future samples.

Despite these problems, several philosophers of science have argued that accommodation (CHARKing) can be as valid and useful as prediction when certain conducive conditions are met (e.g., Collins, 1994; Howson, 1988; Harker, 2006; Hitchcock & Sober, 2004; Lange, 2001; Mayo, 1996, 2008, 2010, 2014;

Schlesinger, 1987). Below, I consider some of these conducive conditions, drawing in particular on Mayo’s work.

Hypothesis that are based on accommodation (i.e., ad hoc hypotheses) are always confirmed by the observed evidence. However, this does not mean that they will always be false. Indeed, some ad hoc hypotheses may be true (Mayo, 2008). According to Mayo (1996, 2008, 2014), two factors can increase the probability that ad hoc hypotheses will be true: (a) severe tests and (b) stringent hypothesis construction rules.

Researchers who use severe tests (e.g., tests based on reliable and valid research designs and methods) are more likely to identify genuine effects than spurious error-based effects. Furthermore, hypotheses that are constructed in order to accommodate genuine effects are more likely to be true than hypotheses that are based on spurious effects. Consequently, although ad hoc hypotheses will always be confirmed by the evidence on which they are based, they will have a higher probability of being true when they accommodate results that are based on severe tests rather than nonsevere tests (Mayo, 1996, 2008, 2014).

To illustrate, imagine that a coin is biased such that it has a greater chance of landing on heads when tossed. Further imagine that two researchers are informed that the coin is biased but they are not informed about the direction of the bias. The researchers are then asked to conduct tests on the coin and generate ad hoc hypotheses about the direction of the bias. Researcher A conducts a nonsevere test in which he only tosses the coin three times. He observes that the coin lands on heads once and on tails twice. Based on this evidence, he constructs the (incorrect) ad hoc hypothesis that the coin is biased towards tails. Researcher B conducts a more severe test in which she tosses the coin 100 times. She observes that the coin lands on heads 66 times and on tails 34 times. Based on this more reliable evidence, she constructs the (correct) ad hoc hypothesis that the coin is biased towards heads. Note that both researchers’ ad hoc hypotheses are supported by the evidence that they observed. However, Researcher B’s more severe test provides more reliable data, and it is for this reason that her ad hoc hypothesis is more likely to be correct.

Severe tests ensure that ad hoc hypotheses are based on high quality evidence. However, high quality evidence is not sufficient to produce high quality hypotheses. Stringent hypothesis construction rules are also required (Mayo, 2008, 2010, 2014). In particular, ad hoc hypotheses should be (a) parsimonious and (b) consistent with prior theory and evidence (Hitchcock & Sober, 2004; Hollenbeck & Wright, 2017; Kerr, 1998; Murayama, Pekrun, & Fiedler, 2014; Stroebe, 2016). I consider each of these stringent hypothesis construction rules in turn.

First, ad hoc hypotheses should be *parsimonious*. For example, the hypothesis that “prejudice reduces self-esteem” is more parsimonious than the convoluted hypothesis that “prejudice reduces self-esteem, but not among men, unless they are white men in which case prejudice increases self-esteem, unless they are young white men, in which case prejudice has no effect on self-esteem.” Parsimony helps to prevent ad hoc hypotheses from overfitting the observed results.

Second, ad hoc hypotheses should also have a certain degree of *consistency* with prior theory and evidence. All other things being equal, ad hoc hypotheses that have greater consistency with prior theory and evidence are less likely to overfit the results. From a

Bayesian perspective, such hypotheses also have a higher prior probability of being true (Dienes, 2016; Kerr, 1998; Murayama et al., 2014).<sup>2</sup> To illustrate, consider a researcher who finds that prejudice increases self-esteem but only in relation to a state-based measure of body image (e.g., “At the moment, I feel good about my weight”) and not in relation to a trait-based measure of global self-esteem (e.g., “In general, I feel good about myself overall”). The researcher considers three potential ad hoc hypotheses to explain this unexpected effect: (a) prejudice only increases state self-esteem, (b) prejudice only increases body-image self-esteem, and (c) prejudice only increases state-based body-image self-esteem. Upon surveying the literature, the researcher finds that some prior theory and evidence is consistent with the hypothesis that prejudice increases state self-esteem, but that no prior theory or evidence is consistent with the hypothesis that prejudice increases body-image self-esteem. Following stringent hypothesis construction rules, the researcher then advances the ad hoc hypothesis that prejudice only increases state self-esteem as an explanation for their results. In doing so, they avoid the less plausible hypothesis that prejudice only increases body-image self-esteem. In addition, they avoid overfitting their results by proposing that prejudice only increases state-based body-image self-esteem.

The stringent hypothesis construction rules of parsimony and consistency with prior theory and evidence increase the prior probability that the resulting ad hoc hypotheses will be true (Dienes, 2016; Kerr, 1998; Murayama et al., 2014). These rules also constrain the structure and content of ad hoc hypotheses and so limits the potential of the hypotheses to overfit the observed results. In this sense, stringent hypothesis construction rules can be said to limit *researchers’ degrees of freedom* (Simmons et al., 2011) during the construction of ad hoc hypotheses so that there is less potential for the hypotheses to be biased towards the evidence. Of course, researchers may ignore the rules of parsimony and consistency and instead construct convoluted and/or entirely unprecedented ad hoc hypotheses in order to accommodate results that would be otherwise inexplicable. However, this unconstrained approach to ad hoc hypothesis construction is more likely to lead to overfitting and inferential error and, consequently, it is more likely to harm scientific progress.

To summarize, researchers can engage in low quality accommodation based on nonsevere tests and unconstrained hypothesis construction or they can engage in high quality accommodation based on severe tests and stringent hypothesis construction rules. This distinction between low and high quality accommodation is important when considering the potential harm that CHARKing may do to scientific progress and its potential role in the replication crisis. HARKing is thought to contribute to a lack of replicable results as part of a multistage process (e.g., Munafò et al., 2017, Figure 1). In this process, low statistical power, *p*-hacking, and uncorrected multiple testing provide nonsevere tests. These nonsevere tests produce spurious results that feed into a process of unconstrained accommodation. And this undisclosed accommodation then produces ostensibly confirmed a priori hypotheses that are actually false ad hoc hypotheses. However, this process is only tenable in the case of *low quality* accommodation. In the case of *high quality* accommodation, tests are more severe (i.e., high power, no *p*-hacking, no uncorrected multiple testing), the observed effects are more likely to be genuine, and the ad hoc hypotheses that accommodate those effects are more likely to be

true. In addition, hypothesis construction is constrained by stringent construction rules, making hypotheses more likely to be true and less likely to overfit spurious effects. Hence, although low quality accommodation may result in false hypotheses that predict nonreplicable effects, high quality accommodation is more likely to result in true hypotheses that predict replicable effects. In other words, not all accommodation (CHARKing) harms scientific progress, and only low quality accommodation is likely to be implicated in the replication crisis.

Critics might argue that it is difficult to distinguish low quality accommodation from high quality accommodation in research reports. However, the signs are quite obvious. Low quality accommodation suffers from exactly the same problems as low-quality research in general: poor-quality methodology (e.g., low statistical power and unreliable, invalid, and/or insensitive measures, manipulations, and designs) and poor-quality hypothesizing (unparsimonious hypotheses that are inconsistent with prior theory and evidence). Peer reviewers and end-users are able to identify these problems and reduce their confidence in the reported research regardless of whether or not they suspect that researchers have engaged in HARKing.

In summary, CHARKed hypotheses cannot be used to predict or falsify the results on which they are based. Nonetheless, CHARKed hypotheses that are derived through a process of high quality accommodation are more likely to be true than those that are derived through a process of low quality accommodation. And, if the ultimate aim of science is to uncover the truth, then high quality accommodation may be regarded as being more helpful to scientific progress than low quality accommodation. It is important to note that this conclusion does not dispute the fact that it is deceptive for researchers to misrepresent accommodation as prediction. Hence, CHARKing should never be concealed; researchers should always indicate in their research reports any hypotheses that are ad hoc and any analyses that are exploratory.

## RHARKing

CHARKing is not the only way in which researchers can engage in post hoc hypothesizing. RHARKING involves retrieving hypotheses after the results are known. RHARKing occurs when empirical disconfirmation of a priori hypotheses inspires researchers to engage in a post hoc search for other relevant hypotheses in the literature—a process that Kerr (1998) described as “empirically-inspired scholarship.” Researchers then present their retrieved hypotheses as a priori hypotheses in their research reports. For example, if our prejudice researcher found that prejudice

<sup>2</sup> Although Kerr (1998) did not advocate CHARKing, he did suggest a Bayesian approach towards improving the usefulness of accommodation. Specifically, he proposed that after researchers had constructed a new hypothesis based on their results, they could “counterfactually estimate the prior probability of that hypothesis being true given knowledge of all evidence available except for those new results, and then use Bayes’s theorem to estimate how much belief now to place in the new hypothesis in light of the new results (i.e., the posterior probability)” (p. 206). Kerr doubted that this approach would work in practice because researchers’ knowledge of their results would be likely to bias their selection of other relevant evidence for estimating the prior probability of the hypothesis. However, he may have dismissed this approach too quickly, because it is possible to remove this bias by asking independent experts to estimate the prior probability of the hypothesis.



reduced self-esteem rather than increased it, then she might proceed to search the literature and find Researcher B's (1989) Hypothesis B that the prejudice reduces self-esteem. She might then present Hypothesis B as a confirmed *a priori* hypothesis in her research report despite its post hoc inspiration.

RHARKing is markedly different from CHARKing for three reasons. First, unlike hypotheses based on CHARKing, hypotheses based on RHARKing are constructed *independent* from researchers' known results. Specifically, the fact that such hypotheses are proposed in research articles that were published *prior to* researchers' analysis of their data guarantees the use novelty of that evidence. It is important to note that this use novelty is not undermined by the facts that (a) researchers come to know about the hypotheses *after* they know about their results or (b) those results include the disconfirmation of one or more of the researchers' *a priori* hypotheses. Hence, unlike CHARKing, RHARKing does not involve hypothesizing *on the basis of* the known results, and the resulting hypotheses have the potential to both predict and be falsified by evidence that is already known by the researcher. To illustrate, consider the Texas sharpshooter analogy again. Imagine that the sharpshooter misses the original target but then dusts off the barn wall around his bullet hole to reveal that it landed inside a second target that was drawn on the wall several years ago by someone else. This "dusting off" represents RHARKing, and the second target that the sharpshooter uncovers represents a previously-constructed hypothesis. Note that the hypothesis and evidence remain independent in this case because the researcher did not draw the target around his own bullet hole. Furthermore, the second target was able to *predict* where the sharpshooter's shot would land.

Second, compared with CHARKing, RHARKing leaves a different scholarly footprint in research reports, and that footprint enables an objective verification of use novelty. In the case of undisclosed CHARKing, readers have no way of objectively verifying whether or not hypotheses were constructed independent from the observed evidence. In contrast, in the case of undisclosed RHARKing, readers are able to confirm the independence of hypotheses by checking the cited articles from which those hypotheses are claimed to have been retrieved. For example, our prejudice researcher might state: "In the current study, I tested Researcher B's (1989) Hypothesis B that prejudice reduces self-esteem." In this case, readers are able to confirm the use novelty of Hypotheses B by consulting Researcher B's (1989) article. Note that the researcher could also disclose that she read about Hypothesis B *after* observing her results (Schwab & Starbuck, 2017, p. 131). However, it is unclear how this additional information would be helpful because the time at which the researcher became aware of the hypothesis has no bearing on the extent to which the researcher's evidence was used in the construction of that hypothesis.

Finally, compared with CHARKing, RHARKing makes a more modest, incremental contribution to scientific progress. Researchers who engage in CHARKing are able to construct *new* hypotheses. In contrast, researchers who engage in RHARKing are limited to testing *old* hypotheses that have been previously advanced in the extant literature. Hence, hypotheses that result from RHARKing will be less innovative than those that result from CHARKing. Nonetheless, researchers who RHARK may still make an important contribution to science by testing the replica-

bility and generalizability of old hypotheses. In the absence of any qualifying information, most hypotheses contain an implicit assumption that the effects that they predict are replicable and generalizable to some extent (Simons, Shoda, & Lindsay, 2017). Hence, researchers can test the replicability and generalizability of previously-hypothesized effects to previously-untested populations, measures, methods, contexts, and cultures. For example, after engaging in RHARKing, our prejudice researcher might find that her evidence that prejudice reduced self-esteem is predicted by Researcher B's (1989) Hypothesis B, but that Hypothesis B has only been tested using *global* measures of self-esteem. Our researcher may then make a significant contribution by demonstrating that Hypothesis B is also confirmed using *multidimensional* measures of specific self-esteem that assess several different aspects of self-esteem. Note that if researchers discover boundary conditions to preexisting hypotheses (e.g., Hypothesis B is only confirmed on some dimensions of self-esteem and not on others), then those findings may either be predicted via further RHARKing or transparently accommodated via high quality CHARKing.

In summary, RHARKing does not pose any threat to science. Instead, it makes a modest contribution to scientific progress by testing the replicability and generalizability of old hypotheses. Critics may raise two objections to this conclusion.

First, it might be argued that the research literature contains an abundance of unfalsified hypotheses that may be retrieved via RHARKing (Ferguson & Heene, 2012), and that researchers often experience implicit pressures to select hypotheses that are confirmed by their results in order to increase their chances of publishing their work (e.g., Fanelli, 2010; Mazzola & Deuling, 2013; Motyl et al., 2017; Nosek et al., 2017; O'Boyle et al., 2017). Hence, RHARKing may preclude reported falsification because researchers are biased towards retrieving confirmed hypotheses rather than disconfirmed hypotheses. However, although this selection bias may occur, it represents a separate type of HARKing called *suppressing hypotheses after the results are known* or SHARKing, and I address it in depth in the next section. At this stage, it should suffice to point out that if researchers report *both* the confirmed *and* the disconfirmed hypotheses that they discover during their RHARKing, then that post hoc scholarship may operate in an unbiased manner and without any harm to scientific progress. It should also be noted that erudite peer review teams are able to reduce this type of selection bias. As Kerr (1998) explained, "editors and reviewers who are sufficiently knowledgeable should be able to distinguish a biased and selective appeal to the literature from a balanced and comprehensive one" (p. 212).

A second argument against RHARKing is that researchers can find flexible theories or hypotheses that "predict nearly any pattern of results in nearly any context" (Kerr, 1998, p. 210; see also van 't Veer & Giner-Sorolla, 2016, p. 4). However, this argument is more applicable to *theories* than it is to *hypotheses*. Certainly, researchers may interpret preexisting *theories* in many different ways in order to predict numerous patterns of results, and this unconstrained ad hoc hypothesis construction leads to the low quality accommodation that was addressed in relation to CHARKing. However, researchers have less freedom to predict numerous patterns of results from preexisting *hypotheses* because hypothesis are more specific than theories in the predictions that they make. For example, returning to the prejudice study example, recall that Researcher B's (1989) Hypothesis B predicts a single unidirec-



tional effect: that prejudice reduces self-esteem. In the absence of any post hoc modification, Hypothesis B cannot be interpreted as predicting either an increase in self-esteem or no change in self-esteem. A similar constraint is likely to apply to most hypotheses. Hence, researchers who engage in RHARKing per se are unlikely to find hypotheses that are sufficiently flexible to predict multiple patterns of results. However, even if they are able to identify such vague and flexible hypotheses, then they still need to convince their audience about the usefulness of these hypothesis, and, in most cases, this will not be an easy task because research is usually judged on the quality and strength of not only the evidence but also the hypotheses. Hence, hypotheses that are sufficiently vague and flexible to predict numerous mutually exclusive patterns results will tend to be dismissed as suffering from “predictive impotence” (Hitchcock & Sober, 2004, p. 7). Returning to the Texas sharpshooter example, if the sharpshooter fires multiple shots at the barn wall at random and with his eyes closed (i.e., in the absence of any genuine effects), then his bullet holes will be scattered all over the side of the barn, and the size of the target that he needs to find after dusting off the barn wall (RHARKing) will need to be the size of the barn wall if he wants to make it look as if he hit the target. Such a target is likely to be entirely unimpressive to his onlookers!

In summary, RHARKing involves the retrieval of old hypotheses that are independent from results that are already known by the researcher, and these hypotheses can therefore predict and be falsified by those results. RHARKing can make a modest contribution to scientific progress by testing the replicability and generalizability of previously-hypothesized effects. Threats to this approach include researchers restricting their selection of hypotheses to those that confirm their results, including relatively flexible and undiagnostic hypotheses. However, these potential selection biases do not compromise the use novelty of the observed evidence, and they can be identified and addressed during the peer review process.

### SHARKing

The third type of HARKing involves researchers failing to report a priori hypotheses, most likely because those hypotheses are unsupported by their results (Kerr, 1998). *Suppressing hypotheses after the results are known*, or SHARKing, can be harmful to science because it precludes the reported falsification of hypotheses. When associated results are also suppressed, SHARKing can bias information about the size and replicability of effects in meta-analyses (Bosco et al., 2016; Schwab & Starbuck, 2017). Nonetheless, there are some conditions under which SHARKing may not impede scientific progress.

Leung (2011) suggested that hypotheses and evidence can be suppressed without any harm to science when they are *unrelated* to the final research conclusions. So, for example, if the theory and evidence for Hypothesis X is unrelated to the theory and evidence for Hypothesis Y, then failing to report Hypothesis X and its evidence will have no impact on the conclusions that are drawn regarding Hypothesis Y and its evidence. However, if Hypotheses X and Y are related to one another, then it will be necessary to report the results of tests of both hypotheses, because failing to report the results for Hypothesis X may affect the perceived veracity of the research conclusions regarding Hypothesis Y.

Leung's (2011) proposal makes sense in relation to the scientific integrity of individual research articles. However, it overlooks the impact of SHARKing on the broader scientific process. Although a

priori hypotheses and their associated results may be unrelated to the final conclusions of a research article, they may be nonetheless important to the conclusions of other research articles, including meta-analyses (Bosco et al., 2016; Ferguson & Heene, 2012; Kerr, 1998, p. 208). From this broader perspective, the results of *all* a priori hypotheses should be reported regardless of whether or not they are related to a specific research report's final conclusions.

Having said this, concerns about the disclosure of all hypothesis tests need to be balanced with concerns about the quality of those tests. After all, low quality tests and methodology are purported to be one of the reasons for the current replication crisis (e.g., John et al., 2012; Munafò et al., 2017; Świątkowski & Dompnier, 2017). Hence, it is also necessary to take account of test severity when considering whether or not to report tests of hypotheses that are unrelated to an article's final conclusions (Mayo, 1991, 1996; Parker, 2015). Nonsevere tests are more likely to lead researchers to incorrectly accept or reject hypotheses and, consequently, these tests and hypotheses may be omitted from the broader scientific record without adversely affecting scientific progress. In contrast, severe tests provide valuable information about hypotheses regardless of the relation between those hypotheses and the final conclusions of a specific research article. Consequently, such tests and hypotheses should be retained in the scientific record.

Putting the arguments about relatedness (Leung, 2011) and test severity (Mayo, 1991, 1996) together, it can be concluded that a priori hypotheses may be omitted from a research article without any harm to science when they are (a) unrelated to the final conclusions of that article *and* (b) the subject of nonsevere tests. In contrast, if a priori hypotheses are *either* related to the article's final conclusions *or* the subject of severe tests, then they should be reported.

Three objections might be raised against the above approach. First, it might be argued that we need greater transparency in the research process, not less, and that there is no reason to exclude information about *any* a priori hypotheses or *any* research results. If tests are nonsevere, then information about severity should be provided, and readers should be empowered to form their own opinions about the value of the evidence. However, this argument ignores a key limitation regarding the communication of science: Consumers of research have limited time and ability to comprehend long and/or complex sets of results. Consequently, it is necessary for researchers to balance concerns about providing comprehensive information with concerns about clarity and concision (Kerr, 1998; Leung, 2011; Vazire, 2014). As Kerr (1998) pointed out, “research reports neither can nor should be detailed laboratory diaries. Research report writing must necessarily be selective” (p. 203). Given this practical constraint, it seems sensible to omit information from research reports that has the least value, and nonsevere tests of hypotheses that are unrelated to the final research conclusions fit this criterion.

Second, it might be argued that excluding hypotheses from research reports reduces the apparent number of null hypothesis significance tests that have been undertaken and therefore lowers readers' expectations about encountering Type I errors as a result of multiple testing (Szucs, 2016). For example, if a researcher Tests 20 hypotheses with an alpha level of .05, then he has a 64.15% chance of making at least one Type I error. However, if his results confirm only one of these hypotheses, and he decides to suppress the other 19 disconfirmed hypotheses, then he will give the incorrect impression that he only conducted a single hypothesis test and that, consequently,

he only had a 5.00% chance of making a Type I error. Hence, SHARKing may artificially inflate readers' confidence in the probity of the reported results. However, this error rate problem is only valid for cases in which hypotheses are *related* to the final research conclusions. If hypotheses are *unrelated* to the final research conclusions (which is the approach being advocated here), then they do not constitute part of the same family of hypotheses that are contingent on the same universal null hypothesis. Consequently, they should not be counted in the calculation for familywise error control (for related discussions, see Matsunaga, 2007; Rubin, 2017). Hence, in the example above, if the 19 suppressed hypotheses are unrelated to the final research conclusions, then tests of those hypotheses will not inflate the Type I error rate for the single hypothesis that is reported, and SHARKing will not bias readers' expectations regarding Type I errors.

Finally, researchers' judgments about whether hypotheses are "related" or "unrelated" to research conclusions and whether tests are "severe" or "nonsevere" are relatively arbitrary and subjective. Consequently, researchers should always attempt to err on the side of caution when making judgment about these matters and, if necessary, seek independent advice from their peer review team.

In summary, it is not appropriate to suppress *a priori* hypotheses or their associated evidence when *either* (a) the hypotheses are related to a research article's final conclusions *or* (b) the tests of those hypotheses are severe. However, it is appropriate to suppress nonsevere tests of *a priori* hypotheses that are unrelated to an article's final conclusions. Suppressing such hypotheses does not artificially increase readers' confidence in the results, and it helps rather than harms science by allowing clearer and more concise communications of results. In cases in which it is unclear whether or not hypotheses are related to final conclusions and/or tests are severe, independent judgments should be sought from editors and peer reviewers.

### When Is HARKing Unethical?

So far, I have identified the conditions under which three different types of HARKing are likely to be more or less harmful to scientific progress. However, progressing science is not the only determinant of research practice. Scientists are also obligated to undertake research in an ethically responsible manner. Hence, it is also important to consider the extent to which different types of HARKing are more or less ethical.

Kerr (1998) was against moralizing about the ethics of HARKing because "too many complex arguments exist on each side to make 'the evils of HARKing' the theme of a compelling sermon" (p. 197). Despite this view, there is a growing trend to regard undisclosed HARKing as unethical (e.g., Świątkowski & Dompnier, 2017) because it contradicts the general principles of openness and transparency in the research process. Again, however, I think that it is useful to adopt an articulated and contextual approach to this issue that considers when different types of undisclosed HARKing may be unethical and when they may be ethical (for a related discussion, see Leung, 2011).

Failing to disclose CHARKing (accommodation) may be considered to be unethical because it conceals two important pieces of information. First, it conceals the fact that the reported hypotheses are unable to predict or be falsified by the evidence. Second, it conceals the fact that overfitting is possible during the hypothesis construction process.

In contrast, failing to disclose RHARKing may be considered to be ethical because disclosing that one read about a preexisting hypothesis before or after analyzing one's data does not provide any useful information to readers. In particular, reading about a preexisting hypothesis after checking one's results does not alter the use novelty of those results for that hypothesis. Furthermore, statements such as "in the present research, I tested Researcher B's (1989) Hypothesis B that prejudice reduces self-esteem" are true regardless of whether researchers read about Hypothesis B before or after knowing their results. Similarly, follow-up statements such as "as predicted, prejudice caused a reduction in participants' self-esteem" are also accurate when they refer back to a RHARKed hypothesis (e.g., Researcher B's, 1989, Hypothesis B).

SHARKing may be either ethical or unethical depending on whether the hypotheses are the subject of severe or nonsevere tests and whether they are related or unrelated to the final research conclusions. The suppression of hypotheses that are *either* severely tested *or* related to the final research conclusions may be considered to be unethical because it conceals important information from readers and the broader scientific record. In contrast, the suppression of hypotheses that are *both* nonseverely tested *and* unrelated to the final research conclusions may be considered to be ethical because it conceals unimportant information from readers.

Finally, it is helpful to distinguish between *active* HARKing and *passive* HARKing when considering the ethics of HARKing. Active HARKing is undertaken by researchers *prior* to the submission of their research report to the peer review team. In contrast, passive HARKing is undertaken by researchers *in response* to requests by editors and peer reviewers to change hypotheses, add new hypotheses, and/or suppress loser hypotheses (Bedeian, Taylor, & Miller, 2010; Bosco et al., 2016; Giner-Sorolla, 2012; Kerr & Harris, 1998, as cited in Kerr, 1998; Kepes & McDaniel, 2013; Hollenbeck & Wright, 2017; Leung, 2011; Motyl et al., 2017; O'Boyle et al., 2017; Schwab & Starbuck, 2017). For example, passive HARKing may occur when researchers comply with the requests of editors and/or peer reviewers to suppress null findings and their associated hypotheses, test new hypotheses, and/or "re-frame" or "refocus" the narrative of articles. Again, the circumstances behind these requests need to be understood before judgments can be made about their potential harm to science and their ethical status. However, all other things being equal, passive HARKing is likely to be more ethical than active HARKing because it is not used by researchers to try to conceal information that might otherwise influence the publication decision.

In summary, not all types of HARKing are unethical under all conditions. Instead, HARKing falls into a "gray zone" of ethical practice (Butler, Delaney, & Spoelstra, 2017; O'Boyle et al., 2017), with only some types of HARKing being ethically unacceptable under some conditions. Consequently, judgments about the ethics of HARKing need be made on a case-by-case basis that takes into account the context of specific research situations.

### Conclusions

HARKing has become a hot topic in the wake of the replication crisis. However, the discussion seems to have polarized towards the view that all HARKing is bad for science. In the present article, I provided a more nuanced view by identifying when different

types of HARKing are most likely and least likely to be harmful to science. I arrived at the following conclusions.

CHARKing (accommodation) occurs when researchers construct hypotheses after the results are known. CHARKed hypotheses lack independence from the observed evidence and, consequently, these hypotheses cannot be used to predict or be falsified by the evidence. Accordingly, CHARKing should always be disclosed in research reports. CHARKing is less harmful to scientific progress (a) when it accommodates results that have been obtained using severe tests (e.g., tests that are based on reliable and valid research designs and methodology) and (b) when it is based on stringent hypothesis construction rules (that produce ad hoc hypotheses that have satisfactory parsimony and consistency with prior theory and evidence).

RHARKing occurs when researchers retrieve hypotheses from the extant literature after the results are known. RHARKed hypotheses have been constructed independent from known results, and so they can predict and be falsified by the known results. RHARKing allows researchers to make a modest contribution to scientific progress by testing the replicability and generalizability of previously-hypothesized effects in relation to new populations, measures, methods, contexts, and cultures. To enable independent verification of RHARKing, researchers should provide citations to the articles that propound the hypotheses being tested. To prevent a selection bias, researchers should report hypotheses in the literature that are both confirmed and disconfirmed by their evidence.

Finally, SHARKing involves the suppression of a priori hypotheses after the results are known. The suppression of hypotheses that are related to a research article's final conclusions can artificially inflate the perceived veracity of those conclusions. In addition, the suppression of a priori hypotheses that have undergone severe testing represents the omission of important information. However, the suppression of a priori hypotheses that are (a) unrelated to an article's final conclusions and (b) that have undergone nonsevere tests represent the omission of unimportant information. Consequently, this type of SHARKing may be undertaken without detriment to either specific or broad scientific progress. It may also help in the communication of science by increasing the clarity and concision of research reports.

The replication crisis may be related to (a) CHARKing based on nonsevere tests and nonstringent hypothesis construction rules and (b) SHARKing of severe tests of a priori hypotheses and nonsevere tests that are related to final research conclusions. However, the replication crisis is less likely to be related to (a) CHARKing based on severe tests and stringent hypothesis construction rules, (b) RHARKing, and/or (c) SHARKing of nonsevere tests of a priori hypotheses that are unrelated to the final research conclusions.

It is difficult to arrive at generic ethical principles about HARKing given the diversity of the research practices and conditions that are involved. However, the concealment of (a) CHARKing and (b) hypothesis that are either severely tested or related to the final research conclusions represent the concealment of important information and, consequently, these practices can usually be considered to be unethical.

## References

Agnoli, F., Wicherts, J. M., Veldkamp, C. L., Albiero, P., & Cubelli, R. (2017). Questionable research practices among Italian research psychol-

ogists. *PLoS ONE*, 12, e0172792. <http://dx.doi.org/10.1371/journal.pone.0172792>

Aguinis, H., Cascio, W. F., & Ramani, R. S. (2017). Science's reproducibility and replicability crisis: International business is not immune. *Journal of International Business Studies*, 48, 653–663. <http://dx.doi.org/10.1057/s41267-017-0081-0>

Baker, M. (2016, May 25). Is there a reproducibility crisis? *Nature*, 533, 452–454. <http://dx.doi.org/10.1038/533452a>

Banks, G. C., O'Boyle, E. H., Jr., Pollack, J. M., White, C. D., Batchelor, J. H., Whelpley, C. E., . . . Adkins, C. L. (2016). Questions about questionable research practices in the field of management: A guest commentary. *Journal of Management*, 42, 5–20. <http://dx.doi.org/10.1177/0149206315619011>

Bedeian, A. G., Taylor, S. G., & Miller, A. N. (2010). Management science on the credibility bubble: Cardinal sins and various misdemeanors. *Academy of Management Learning & Education*, 9, 715–725. <http://dx.doi.org/10.5465/AMLE.2010.56659889>

Bem, D. J. (1987). Writing the empirical journal. In M. P. Zanna & J. M. Darley (Eds.), *The compleat academic: A practical guide for the beginning social scientist* (pp. 171–201). Mahwah, NJ: Erlbaum.

Bem, D. J. (2011). Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology*, 100, 407–425. <http://dx.doi.org/10.1037/a0021524>

Bosco, F. A., Aguinis, H., Field, J. G., Pierce, C. A., & Dalton, D. R. (2016). HARKing's Threat to organizational research: Evidence from primary and meta-analytic sources. *Personnel Psychology*, 69, 709–750. <http://dx.doi.org/10.1111/peps.12111>

Butler, N., Delaney, H., & Spoelstra, S. (2017). The gray zone: Questionable research practices in the business school. *Academy of Management Learning & Education*, 16, 94–109. <http://dx.doi.org/10.5465/amle.2015.0201>

Collins, R. (1994). Against the epistemic value of prediction over accommodation. *Noûs*, 28, 210–224. <http://dx.doi.org/10.2307/2216049>

de Groot, A. D. (2014). The meaning of “significance” for different types of research (E. J. Wagenmakers, D. Borsboom, J. Verhagen, R. Kievit, M. Bakker, A. Cramer, . . . H. L. J. van der Maas). *Acta Psychologica*, 148, 188–194. <http://dx.doi.org/10.1016/j.actpsy.2014.02.001>

Dienes, Z. (2016). How Bayes factors change scientific practice. *Journal of Mathematical Psychology*, 72, 78–89. <http://dx.doi.org/10.1016/j.jmp.2015.10.003>

Fanelli, D. (2010). Do pressures to publish increase scientists' bias? An empirical support from U.S. states' data. *PLoS ONE*, 5, e10271. <http://dx.doi.org/10.1371/journal.pone.0010271>

Ferguson, C. J., & Heene, M. (2012). A vast graveyard of undead theories: Publication bias and psychological science's aversion to the null. *Perspectives on Psychological Science*, 7, 555–561. <http://dx.doi.org/10.1177/1745691612459059>

Fiedler, K., & Schwarz, N. (2016). Questionable research practices revisited. *Social Psychological and Personality Science*, 7, 45–52. <http://dx.doi.org/10.1177/1948550615612150>

Gigerenzer, G. (2004). Mindless statistics. *Journal of Socio-Economics*, 33, 587–606. <http://dx.doi.org/10.1016/j.socrec.2004.09.033>

Giner-Sorolla, R. (2012). Science or art? How aesthetic standards grease the way through the publication bottleneck but undermine science. *Perspectives on Psychological Science*, 7, 562–571. <http://dx.doi.org/10.1177/1745691612457576>

Glaser, J., & Salovey, P. (1998). Affect in electoral politics. *Personality and Social Psychology Review*, 2, 156–172. [http://dx.doi.org/10.1207/s15327957pspr0203\\_1](http://dx.doi.org/10.1207/s15327957pspr0203_1)

Harker, D. (2006). Accommodation and prediction: The case of the persistent head. *The British Journal for the Philosophy of Science*, 57, 309–321. <http://dx.doi.org/10.1093/bjps/axl004>



- Helgeson, V. S., & Fritz, H. L. (1998). A theory of unmitigated communion. *Personality and Social Psychology Review*, 2, 173–183. [http://dx.doi.org/10.1207/s15327957pspr0203\\_2](http://dx.doi.org/10.1207/s15327957pspr0203_2)
- Hitchcock, C., & Sober, E. (2004). Prediction versus accommodation and the risk of overfitting. *The British Journal for the Philosophy of Science*, 55, 1–34. <http://dx.doi.org/10.1093/bjps/55.1.1>
- 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, 5–18. <http://dx.doi.org/10.1177/0149206316679487>
- Howson, C. (1988). Accommodation, prediction and Bayesian confirmation theory. *Proceedings of the Biennial Meeting of the Philosophy of Science Association* (Vol. 2, pp. 381–392). Philosophy of Science Association.
- 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. <http://dx.doi.org/10.1177/0956797611430953>
- Kepes, S., & McDaniel, M. A. (2013). How trustworthy is the scientific literature in industrial and organizational psychology? *Industrial and Organizational Psychology: Perspectives on Science and Practice*, 6, 252–268. <http://dx.doi.org/10.1111/iops.12045>
- Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review*, 2, 196–217. [http://dx.doi.org/10.1207/s15327957pspr0203\\_4](http://dx.doi.org/10.1207/s15327957pspr0203_4)
- Kerr, N. L. (2011). HARK! A herald sings . . . but who's listening? In R. M. Arkin (Ed.), *Most underappreciated: 50 prominent social psychologists describe their most unloved work* (pp. 126–131). New York, NY: Oxford University Press. <http://dx.doi.org/10.1093/acprof:osobl/9780199778188.003.0024>
- Lakatos, I. (1970). Falsification and the methodology of scientific research programmes. In I. Lakatos & A. Musgrave (Eds.), *Criticism and the growth of knowledge* (pp. 96–191). London, England: Cambridge University Press.
- Lange, M. (2001). The apparent superiority of prediction to accommodation as a side effect: A reply to Maher. *British Journal for the Philosophy of Science*, 52, 575–588.
- Leung, K. (2011). Presenting post hoc hypotheses as a priori: Ethical and theoretical issues. *Management and Organization Review*, 7, 471–479. <http://dx.doi.org/10.1017/CBO9781139171434.009>
- Lindsay, D. S., Simons, D. J., & Lilienfeld, S. O. (2016, December). Research Preregistration 101. *APS Observer*. Retrieved from <https://www.psychologicalscience.org/observer/research-preregistration-101>
- Matsunaga, M. (2007). Familywise error in multiple comparisons: Disentangling a knot through a critique of O'Keefe's arguments against alpha adjustment. *Communication Methods and Measures*, 1, 243–265. <http://dx.doi.org/10.1080/19312450701641409>
- Maxwell, S. E., Lau, M. Y., & Howard, G. S. (2015). Is psychology suffering from a replication crisis? What does “failure to replicate” really mean? *American Psychologist*, 70, 487–498. <http://dx.doi.org/10.1037/a0039400>
- Mayo, D. G. (1991). Novel evidence and severe tests. *Philosophy of Science*, 58, 523–552. <http://dx.doi.org/10.1086/289639>
- Mayo, D. G. (1996). *Error and the growth of experimental knowledge*. Chicago: University of Chicago Press. <http://dx.doi.org/10.7208/chicago/9780226511993.001.0001>
- Mayo, D. G. (2008). How to discount double-counting when it counts: Some clarifications. *The British Journal for the Philosophy of Science*, 59, 857–879.
- Mayo, D. G. (2010). An ad hoc save of a theory of adhocness? In D. G. Mayo & A. Spanos (Eds.), *Error and inference: Recent exchanges on experimental reasoning, reliability, and the objectivity and rationality of science* (pp. 155–169). New York, NY: Cambridge University Press.
- Mayo, D. (2014). Some surprising facts about (the problem of) surprising facts (from the Dusseldorf Conference, February 2011). *Studies in History and Philosophy of Science*, 45, 79–86. <http://dx.doi.org/10.1016/j.shpsa.2013.10.005>
- Mayo, D. G., & Spanos, A. (2006). Severe testing as a basic concept in a Neyman–Pearson philosophy of induction. *The British Journal for the Philosophy of Science*, 57, 323–357. <http://dx.doi.org/10.1093/bjps/axl003>
- Mazzola, J. J., & Deuling, J. K. (2013). Forgetting what we learned as graduate students: HARKing and selective outcome reporting in I–O journal articles. *Industrial and Organizational Psychology: Perspectives on Science and Practice*, 6, 279–284. <http://dx.doi.org/10.1111/iops.12049>
- Motyl, M., Demos, A. P., Carsel, T. S., Hanson, B. E., Melton, Z. J., Mueller, A. B., . . . Skitka, L. J. (2017). The state of social and personality science: Rotten to the core, not so bad, getting better, or getting worse? *Journal of Personality and Social Psychology*, 113, 34–58. <http://dx.doi.org/10.1037/pspa0000084>
- Munafò, M. R., Nosek, B. A., Bishop, D. V., Button, K. S., Chambers, C. D., du Sert, N. P., . . . Ioannidis, J. P. (2017). A manifesto for reproducible science. *Nature Human Behaviour*, 1, 0021. <http://dx.doi.org/10.1038/s41562-016-0021>
- Murayama, K., Pekrun, R., & Fiedler, K. (2014). Research practices that can prevent an inflation of false-positive rates. *Personality and Social Psychology Review*, 18, 107–118. <http://dx.doi.org/10.1177/1088868313496330>
- Musgrave, A. (1974). Logical versus historical theories of confirmation. *The British Journal for the Philosophy of Science*, 25, 1–23. <http://dx.doi.org/10.1093/bjps/25.1.1>
- Nosek, B. A., Ebersole, C. R., DeHaven, A., & Mellor, D. (2017). *The preregistration revolution*. Retrieved from [osf.io/2dxu5](http://osf.io/2dxu5)
- O'Boyle, E. H., Jr., Banks, G. C., & Gonzalez-Mulé, E. (2017). The chrysalis effect: How ugly initial results metamorphosize into beautiful articles. *Journal of Management*, 43, 367–399.
- Open Science Collaboration. (2015, August 28). Estimating the reproducibility of psychological science. *Science*, 349, aac4716. <http://dx.doi.org/10.1126/science.aac4716>
- Parker, C. G. (2015). *Reframing the reproducibility crisis: Using an error-statistical account to inform the interpretation of replication results in psychological research*. Doctoral dissertation, Virginia Polytechnic Institute and State University. Retrieved from <https://vtechworks.lib.vt.edu/handle/10919/52963>
- Popper, K. R. (1979). *Objective knowledge: An evolutionary approach*. Oxford: New York, NY: Oxford University Press.
- Richards, T. (2016). *HARKing Back: Lessons in investing from science*. Retrieved from <https://seekingalpha.com/article/3895286-harking-back-lessons-investing-science>
- Rubin, M. (2017). Do *p* values lose their meaning in exploratory analyses? It depends how you define the familywise error rate. *Review of General Psychology*, 21, 269–275. <http://dx.doi.org/10.1037/gpr0000123>
- Schlesinger, G. N. (1987). Accommodation and prediction. *Australasian Journal of Philosophy*, 65, 33–42. <http://dx.doi.org/10.1080/00048408712342751>
- Schwab, A., & Starbuck, W. H. (2017). A call for openness in research reporting: How to turn covert practices into helpful tools. *Academy of Management Learning & Education*, 16, 125–141. <http://dx.doi.org/10.5465/amle.2016.0039>
- Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, 22, 1359–1366. <http://dx.doi.org/10.1177/0956797611417632>
- Simons, D. J., Shoda, Y., & Lindsay, D. S. (2017, April 17). Constraints on generality (COG): A proposed addition to all empirical papers. Retrieved from <http://osf.io/preprints/psyarxiv/w9e3r>
- Stroebe, W. (2016). Are most published social psychological findings false? *Journal of Experimental Social Psychology*, 66, 134–144. <http://dx.doi.org/10.1016/j.jesp.2015.09.017>

- Świątkowski, W., & Dompnier, B. (2017). Replicability crisis in social psychology: Looking at the past to find new pathways for the future. *Revue Internationale de Psychologie Sociale*, 30, 111–124. <http://dx.doi.org/10.5334/irsp.66>
- Szucs, D. (2016). A tutorial on hunting statistical significance by chasing *N*. *Frontiers in Psychology*, 7, 1444. <http://dx.doi.org/10.3389/fpsyg.2016.01444>
- Unkelbach, C. (2016). Increasing replicability. *Social Psychology*, 47, 1–3. <http://dx.doi.org/10.1027/1864-9335/a000270>
- van 't Veer, A. E., & Giner-Sorolla, R. (2016). Pre-registration in social psychology: A discussion and suggested template. *Journal of Experimental Social Psychology*, 67, 2–12. <http://dx.doi.org/10.1016/j.jesp.2016.03.004>
- Vazire, S. (2014). *Life after Bem* [Blog post]. Retrieved from <http://sometimesimwrong.typepad.com/wrong/2014/03/life-after-bem.html>
- Wagenmakers, E. J., Wetzels, R., Borsboom, D., Kievit, R., & van der Maas, H. L. (2015). A skeptical eye on psi. In E. C. May & S. B. Waraha (Eds.), *Extrasensory perception: Support, skepticism, and science*: (pp. 153–176). Santa Barbara, CA: Praeger.
- Wagenmakers, E. J., Wetzels, R., Borsboom, D., & van der Maas, H. L. (2011). Why psychologists must change the way they analyze their data: The case of psi: Comment on Bem (2011). *Journal of Personality and Social Psychology*, 100, 426–432. <http://dx.doi.org/10.1037/a0022790>
- 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, 632–638. <http://dx.doi.org/10.1177/1745691612463078>
- Worrall, J. (1985). Scientific discovery and theory-confirmation. In J. C. Pitt (Ed.), *Change and progress in modern science: Papers related to and arising from the Fourth International Conference on History and Philosophy of Science* (pp. 301–331). Dordrecht, the Netherlands: Reidel. [http://dx.doi.org/10.1007/978-94-009-6525-6\\_11](http://dx.doi.org/10.1007/978-94-009-6525-6_11)
- Worrall, J. (2010). Error, tests, and theory confirmation. In D. G. Mayo & A. Spanos (Eds.), *Error and inference: Recent exchanges on experimental reasoning, reliability, and the objectivity and rationality of science* (pp. 125–154). New York, NY: Cambridge University Press.
- Worrall, J. (2014). Prediction and accommodation revisited. *Studies in History and Philosophy of Science*, 45, 54–61. <http://dx.doi.org/10.1016/j.shpsa.2013.10.001>
- Zahar, E. (1973). Why did Einstein's programme supersede Lorentz's? (I). *The British Journal for the Philosophy of Science*, 24, 95–123. <http://dx.doi.org/10.1093/bjps/24.2.95>

Received May 1, 2017

Revision received August 4, 2017

Accepted August 30, 2017 ■

### E-Mail Notification of Your Latest Issue Online!

Would you like to know when the next issue of your favorite APA journal will be available online? This service is now available to you. Sign up at <https://my.apa.org/portal/alerts/> and you will be notified by e-mail when issues of interest to you become available!