The Costs of HARKing

Mark Rubin
The University of Newcastle, Australia

Citation: Rubin, M. (2022). The costs of HARKing. *British Journal for the Philosophy of Science*, *73*(2), 535-560. https://doi.org/10.1093/bjps/axz050

Abstract

Kerr ([1998]) coined the term 'HARKing' to refer to the practice of 'hypothesizing after the results are known'. This questionable research practice has received increased attention in recent years because it is thought to have contributed to low replication rates in science. The present article discusses the concept of HARKing from a philosophical standpoint and then undertakes a critical review of Kerr's ([1998]) twelve potential costs of HARKing. It is argued that these potential costs are either misconceived, misattributed to HARKing, lacking evidence, or that they do not take into account pre- and post-publication peer review and public availability to research materials and data. It is concluded that it is premature to assume that HARKing has led to low replication rates.

Keywords: accommodation, falsification, HARKing, prediction, questionable research practices, replication crisis

Correspondence concerning this article should be addressed to Mark Rubin at the School of Psychology, Behavioural Sciences Building, The University of Newcastle, Callaghan, NSW 2308, Australia.
E-mail: Mark.Rubin@newcastle.edu.au Web: http://bit.ly/rubinpsyc

1 Introduction

Kerr ([1998]) coined the term 'HARKing' to refer to the practice of 'hypothesizing after the results are known'. As he explained, 'HARKing may be defined as presenting a *post hoc* hypothesis in the introduction of a research report as if it were an *a priori* hypothesis' (p. 197). HARKing may also involve failing to disclose an *a priori* hypothesis in a research report.

Kerr's ([1998]) article has experienced a large increase in citations over the last few years (see also Rubin [2017b], Figure 1). Based on data from Google Scholar, the average annual citation

rate for Kerr's article during the period of 2000–2010 was 5.2 citations per year. However, during the following period of 2011–2018, this citation rate increased to 75.1 citations per year, and in the period of 2017–2018, it was 155.5 citations per year. As Rubin ([2017b]) noted, this dramatic increase in citations to Kerr's article may be because HARKing is regarded as a 'questionable research practice' (John, Loewenstein, and Prelec [2012]), and questionable research practices are thought to be one of the causes of low replication rates in science (for example, Mazzola and Deuling [2013]; Aguinis, Cascio, and Ramani [2017]; Hollenbeck and Wright [2017]).

Given the increased attention to the concept of HARKing, and its potential implication for replicability, it is important to understand the nuances and complexities of this questionable research practice. In this context, some commentators have considered the conditions under which it might be appropriate and inappropriate to engage in HARKing (Leung [2011]; Rubin [2017b]; Vancouver [2018]). Rubin ([2017b]) distinguished between three different types of HARKing: (1) constructing hypotheses after the results are known (CHARKing), (2) retrieving hypotheses word-for-word from published literature after the results are known (RHARKing), and (3) suppressing *a priori* hypotheses after the results are known (SHARKing). I concluded that it is never acceptable to CHARK, always acceptable to RHARK, and only acceptable to SHARK if the suppressed hypotheses are both poorly tested *and* unrelated to the research conclusions. Vancouver ([2018]) and Leung ([2011]) agreed that SHARKing is acceptable if the suppressed hypotheses are unrelated to the research conclusions. Vancouver ([2018], p. 78) also proposed that CHARKing is acceptable 'if revised logic and evidence for a set of hypotheses are sound and the method allows for the elimination of most of the alternative explanations'. However, he qualified that HARKing may not be appropriate under these conditions if the study is investigating a predictor's validity or, potentially, an intervention.

Rubin ([2017b]), Vancouver ([2018]), and Leung ([2011]) have opened up an important discussion about when HARKing may be appropriate and inappropriate. However, they did not provide a detailed consideration of the potential costs of HARKing. Kerr ([1998], p. 211) detailed twelve potential costs:

1. Translating Type I errors into hard-to-eradicate theory
2. Propounding theories that cannot (pending replication) pass Popper's disconfirmability test
3. Disguising *post hoc* explanations as *a priori* explanations (when the former tend also be [sic] more *ad hoc*, and consequently, less useful)
4. Not communicating valuable information about what did not work
5. Taking unjustified statistical licence
6. Presenting an inaccurate model of science to students
7. Encouraging 'fudging' in other grey areas
8. Making us less receptive to serendipitous findings
9. Encouraging adoption of narrow, context-bound new theory
10. Encouraging retention of too-broad, disconfirmable old theory
11. Inhibiting identification of plausible alternative hypotheses
12. Implicitly violating basic ethical principles

No prior work has provided a critical evaluation of these twelve potential costs of HARKing. It is important to undertake this critical evaluation in order to understand the risks involved in HARKing and the potential role of HARKing in relation to replicability and scientific progress more broadly.

The current article is divided into two main sections. The first section provides a discussion about how scientists evaluate hypotheses, and it then considers how HARKing might affect hypothesis evaluation. The second section provides a critical review of Kerr's ([1998]) twelve costs of HARKing.

2 Hypothesis Evaluation and HARKing

2.1 Hypothesis evaluation

Scientists do not make absolute, dichotomous judgements about theories and hypotheses being 'true' or 'false'. Instead, they make relative, continuous judgements about theories and hypotheses being more or less true that other theories and hypotheses in accounting for certain phenomena. This process of 'inference to the best explanation' (for example, Mackonis [2013]) involves researchers and their readers judging which theories and hypotheses are 'truer' than other theories and hypotheses. Judgements about the relative truth-likeness of theories and hypotheses can be described as 'estimated relative verisimilitude' (Popper [1985], p. 58; Meehl [1990]; Cevolani and Festa [2018]).

The concept of verisimilitude faces the epistemic problem of estimating an unknown quantity: the truth! Nonetheless, as Meehl ([2004], p. 626) observed, practicing scientists often speak of improving their theories and of favouring one imperfect theory over another. Hence, notwithstanding this epistemic problem, scientists do make estimates of what they perceive to be the relative verisimilitude of theories and hypotheses. Furthermore, recent conceptualizations of verisimilitude have attempted to overcome its epistemic problem by redefining it as an inextricable mix of truth-likeness and informativeness (Cevolani and Festa [2018]).

During the research process, the estimated relative verisimilitude of a hypothesis is based on two sources of information. The first source is the deduction of the hypothesis from *a priori* theory and evidence. This 'theoretical rationale' can be used to make an initial estimate of the verisimilitude of the hypothesis relative to other hypotheses. The theory and theoretical rationale may be judged in terms of several 'theoretical virtues' (for example, Ladyman [2002], p. 181; Mackonis [2013], p. 978). For example, Mackonis ([2013]) described four theoretical virtues. 'Coherence' refers to the consistency of the theoretical rationale with the primary theory, background theories and knowledge, meta-theories, and prior evidence, and it provides an indication of the theoretical plausibility of the hypothesis. Note that a theoretical rationale that is inconsistent with the theory from which it is purported to be derived will lack coherence. 'Breadth' indicates the extent to which a theoretical rationale might explain a wide variety of effects, and it relates to the predictive informativeness and explanatory power of a hypothesis. 'Depth' indicates the extent to which a theoretical rationale has the potential to provide a fundamental account of explanatory mechanisms. Finally, 'parsimony' indicates a theoretical rationale's efficiency in explaining an effect with the smallest number of explanatory constructs (causes, assumptions, principles, and so on). All other things being equal, theories and theoretical rationales that are coherent, broad, deep, and parsimonious will produce hypotheses that have higher estimated verisimilitude than theories and rationales that are incoherent, narrow, shallow, and complex.

Note that an initial estimate of relative verisimilitude does not need to be a formal quantitative estimate (for example, a Bayesian prior probability distribution). It can be as simple and informal as a belief about the theory and hypothesis relative to other competing theories and

hypotheses (for example, 'All things considered, I think that this theory and hypothesis has greater verisimilitude than these other theories and hypotheses').

The second source of information that is used to make judgements about the relative verisimilitude of a hypothesis is the relative degree of consistency between the hypothesis and the current result. Here, we are interested in the extent to which the hypothesis is more or less consistent with the current result compared to other relevant hypotheses. These other hypotheses include auxiliary hypotheses about the reliability and validity of the research design and measurements. Hence, judgements of relative verisimilitude take into account what Mayo ([1996], [2018]) terms the 'severity' of the test. All other things being equal, if low quality research design and measures cannot explain the current result, and all other relevant alternative hypotheses are less consistent with the current result than the primary hypothesis, then the primary hypothesis will be judged as being relatively verisimilar.

Researchers and readers of research reports use information about the current result to make *changes* to their initial estimate of the relative verisimilitude of a hypothesis. An initial estimate of the relative verisimilitude is based on the coherence, breadth, depth, and parsimony of the theoretical rationale for the hypothesis. This estimate is then updated with information about the relative consistency between the hypothesis and the current result relative to other relevant hypotheses. The outcome is an increase or decrease in the estimated relative verisimilitude of the hypothesis and its associated theory.

Importantly, according to the principle of 'use novelty' (Worrall [1985], [2014]; see also Musgrave [1974]; Mayo [1996], [2008]; Rubin [2017b]), a current result can only be used to *change* the initial estimate of relative verisimilitude if that result has not already been *used as the basis for* that initial estimate. If a current result has been used to generate an initial estimate of the relative verisimilitude of a hypothesis, then, logically, it cannot be used again to increase or decrease that estimate. This approach would represent 'double counting' (Mayo [2008]).

The use novelty principle implies an important distinction between 'prediction', in which hypotheses are deduced from *a priori* theory and evidence, and 'accommodation', in which hypotheses are induced from current results. In the absence of any other information, only prediction meets the use novelty principle because only prediction allows researchers and readers to derive an estimate of relative verisimilitude from a source that is separate from the current result (namely, *a priori* theory and evidence). In contrast, accommodation violates the use novelty principle because it uses the current result as the basis for an initial estimate of relative verisimilitude.

2.2 HARKing

A key criticism of HARKing is that it represents accommodation and, consequently, it violates the use novelty principle (for example, Kerr [1998], p. 206). However, HARKing is more likely to take the form of '*post hoc* prediction' than *ad hoc* accommodation and, consequently, it is consistent with the use novelty principle.

To explain, HARKers are obliged to present a theoretical rationale for their HARKed hypothesis in the Introduction section of their research report (Kerr [1998], p. 197). In this theoretical rationale, HARKers usually explain how they deduced their hypothesis from *a priori* theory and evidence, and they include citations and references to this theory and evidence. Despite its *post hoc* generation, researchers and readers can use this theoretical rationale to make an initial estimate of the relative verisimilitude of the HARKed hypothesis independent from the current research result. They can then make a valid update of this estimate based on the current result

without violating the use novelty principle. Hence, a HARKer's *post hoc* theoretical rationale provides the basis for an initial estimate of the relative verisimilitude of the hypothesis, and the current result provides a second, independent basis for increasing or decreasing this initial estimate (for a similar view, see Oberauer and Lewandowsky [2019]).

This approach to estimating the relative verisimilitude of a hypothesis with and without information provided by the current result is best characterized as the test of a *post hoc* prediction that has been deduced from *a priori* theory and evidence rather than as the generation of an *ad hoc* accommodation that has been induced from the current result. Importantly, this test remains valid even if readers have been misled about when the theoretical rationale was developed (that is, after rather than before the researcher knew the current result).

Critics might question whether *post hoc* prediction really adheres to the use novelty principle. Surely, if a current result motivates and/or guides the construction of a hypothesis, then it forfeits its use novelty with respect to that hypothesis? However, this question assumes that information that motivates or guides the construction of a hypothesis is also necessary for the initial (pre-result) evaluation of that hypothesis. Contrary to this assumption, if a hypothesis can be deduced from *a priori* theory and evidence, then an initial estimate of the hypothesis' relative verisimilitude can be based on this deduction, and information from the current result is not necessary to make this initial estimate. Consequently, although a current result may influence the construction of a hypothesis, the current result can be considered to be 'epistemically independent' from that hypothesis if it is not required in order to make an initial estimate of the hypothesis' relative verisimilitude. For example, if the story is to be believed, a falling apple motivated and guided Newton's construction of the theory of gravity. Nonetheless, from a use novelty perspective, that same falling apple is not necessary in order to make an initial estimate of the relative verisimilitude of Newton's theory. Consequently, the result represented by that particular falling apple may be used to increase support for Newton's theory if an initial estimate of the relative verisimilitude of that theory is based on *a priori* theory and evidence that excludes that result. In summary, a current result may inspire a search for a hypothesis that is consistent with that result, and it may even guide the construction of that hypothesis, but neither of these points prevent the information about the result from being excluded from an initial estimate of the relative verisimilitude of that hypothesis and then being used to change that estimate.

To illustrate, imagine that a researcher performs a study in which university students are randomly assigned to eat an apple every day for a week or, in a control condition, not to eat any apples during that week. The researcher then measures a variety of different outcome variables and finds that participants who eat an apple every day show a significant improvement in their mood compared to participants in the control condition. Further imagine that this result motivates the researcher to construct a hypothesis that is consistent with this result. After searching the literature, the researcher finds that (a) *a priori* theory predicts that 'Vitamin C improves mood', and (b) *a priori* evidence shows that 'apples are rich in Vitamin C'. Combining this *a priori* theory and evidence, the researcher deduces the hypothesis that 'eating apples improve mood'. They then explain this reasoning in the Introduction section of their research report, together with relevant citations and references to the *a priori* theory and evidence that they have used. However, they do not disclose the timing of their hypothesis construction (that is, after they knew the current result). Given that the researcher explains their hypothesis in the Introduction of their research report, readers may incorrectly assume that the hypothesis was constructed *before* the researcher knew the research result. Despite this incorrect assumption, and the undisclosed fact that the current result motivated a search for the hypothesis, the researcher and their readers can make an initial

estimate about the verisimilitude of the hypothesis that 'eating apples improve mood' based on the theoretical rationale that is presented in the research report, which is itself based on *a priori* theory and evidence regarding Vitamin C, mood, and the Vitamin C content of apples. Critically, readers can estimate the relative verisimilitude of the hypothesis (a) without taking the current result into account and (b) after taking the current result into account, even if they have been misled about when the researcher constructed the hypothesis. Hence, they are able to undertake a valid evaluation of the hypothesis even though HARKing has occurred.

The above conception of HARKing is inspired by Bayesian hypothesis testing. Notably, Kerr ([1998], p. 206) considered a similar Bayesian approach in his article:

> One could, in principle, counterfactually estimate the prior probability of that [*post hoc*] 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).

Similar to Kerr ([1998]), I assume that researchers and readers can use *a priori* theory and evidence to estimate the relative verisimilitude (prior probability) of a HARKed hypothesis separate from the new results. They can then use the new results to support the HARKed hypothesis (that is, increase its estimated relative verisimilitude or posterior probability). The Bayesian perspective on this process of 'counterfactual updating' is described in detail by Howson ([1984], [1985]; see also Howson and Urbach [1993], pp. 403–8).

Kerr ([1998], p. 206) discussed this Bayesian approach to HARKing when comparing the concepts of accommodation and prediction. However, Kerr's Bayesian approach does not actually contrast accommodation with prediction. In the Bayesian case, accommodation would occur if the new results were used to estimate the prior probability of the hypothesis. This use of the new results to form 'data-dependent' or 'empirical priors' makes the results logically ineligible as a basis for updating the prior probability to compute a posterior probability. As Berger ([2006], p. 399) noted, the use of such empirical priors 'results in an undesirable double use of the data'. In other words, it violates the use novelty principle. Critically, however, Kerr's Bayesian approach does not represent *ad hoc* inductive accommodation because the prior probability is not estimated on the basis of the new results. Instead, the prior probability is estimated on the basis 'of all evidence available *except for those new results*' (Kerr [1998], p. 206, my emphasis). Consequently, the use novelty principle is not violated, and it is legitimate to use the new results to update the prior probability to compute the posterior probability.

Kerr's ([1998]) Bayesian approach serves to contrast two forms of deductive prediction: (a) '*a priori* prediction' based on *a priori* theory and evidence and (b) '*post hoc* prediction' based on *a priori* theory and evidence. The only difference between *a priori* and *post hoc* prediction is the timing at which the researcher deduces the prediction. *A priori* predictions are deduced *before* the researcher knows the current result, and *post hoc* predictions are deduced *after* the researcher knows the current result. In contrast, *ad hoc* accommodation is always induced on the basis on the current result.

Kerr ([1998], pp. 206–7) dismissed the Bayesian approach to HARKing because he felt that it was susceptible to the hindsight bias and other similar biases. In particular, he concluded that 'reconstructing unbiased prior probabilities after one knows the result may be difficult or even impossible' ([1998], p. 207). Certainly, the hindsight and other biases may alter a researcher's subjective feelings about the prior probability of the associated hypothesis, and, if the researcher's estimated prior probability is based on these subjective feelings, then the hindsight and other biases may affect this estimate and the updating process (namely, double counting). However, scientists

are usually required to provide a transparent and objectively verifiable explanation to support their beliefs about a hypothesis (Gelman and Hennig [2017]). Even Bayesian 'subjective' priors need to have some reasonable, objectively verifiable basis that is independent from the current result (for example, meta-analytic evidence, a survey of expert opinion, and so on; Van de Schoot *et al.* [2014], pp. 845, 857; Gelman and Hennig [2017], pp. 974–6). Hence, although the hindsight bias may secretly affect researchers' private prior probabilities, it cannot secretly affect their public scientific prior probabilities, because the reasoning that is used to construct public prior probabilities is made objectively verifiable. Importantly, it is these public scientific prior probabilities that are used during hypothesis evaluation.

Of course, the hindsight bias may secretly motivate and guide researchers' selection, interpretation, and combination of *a priori* theory and evidence that they use to deduce a hypothesis. Nonetheless, readers will remain able to make an initial estimate of the relative verisimilitude of the researcher's hypothesis independent from the current result and to update that estimate with information about the relative consistency between the hypothesis and the result. Again, Newton may have constructed his theory of gravity in hindsight, after the apple fell on his head. But this hindsight bias does not prevent readers from making an initial estimate of the relative verisimilitude of Newton's theory of gravity based on *a priori* theory and evidence that omits that particular falling apple and then updating their estimate with knowledge of that falling apple.

3 A Critical Review of Kerr's Twelve Costs of Harking

3.1 Cost 1: Translating Type I errors into theory

Kerr's ([1998]) first potential cost of HARKing is that it may translate Type I errors into hard-to-eradicate theory. In particular, Kerr was concerned about the potential implications of HARKing after a Type I error has been made. As he explained, 'when such a Type I error is followed by HARKing, then "theory" is constructed to account for what is, in fact, an illusory effect' (p. 205). Of course, both *a priori* and *post hoc* hypotheses may account for illusory effects. Hence, Kerr ([1998], p. 205) asked 'are the costs of making such an error any greater when an author has HARKed than under other circumstances?'. He suspected so, arguing that

> […] an explicitly *post hoc* hypothesis implicitly acknowledges its dependence upon the result in hand as a cornerstone (or perhaps, the entirety) of its foundation, and thereby sensitizes the reader to the vulnerability of the hypothesis to the risks of an immediate Type I error. (p. 205)

Kerr is concerned about *overfitting* here (for example, Hitchcock and Sober [2004]; see also Kerr [1998], pp. 206–7). Overfitting occurs when a hypothesis is constructed in order to accommodate a collection of current results, some of which are representative of the population and some of which are spurious Type I errors that are specific to the particular sample in hand. In this case, the relative verisimilitude of the hypothesis will be overestimated because false positive results will be mistaken for true positive results (that is, Type I errors). This overestimation of relative verisimilitude will be greater when the sample is less representative of the population (for example, when the sample is smaller and/or less randomly selected from the population).

Although overfitting is a problem for *ad hoc* accommodation, it is not a problem in the case of HARKing that uses *post hoc* prediction. In this latter case, researchers and readers base their initial judgement of relative verisimilitude on a theoretical rationale that is deduced from *a priori* theory and evidence. Consequently, their initial estimate of relative verisimilitude is not

based on the current results (including spurious, sample-specific Type I errors), and so no overfitting can occur.

Critics might argue that HARKing allows researchers to deduce a potentially infinite number of hypotheses from *a priori* theory and evidence, and that these hypotheses can then be used to fit any collection of current results (for example, Kerr [1998], p. 210). This is true. However, this criticism refers to 'flexible theorizing' rather than overfitting, and the concern about flexible theorizing is mitigated by the fact that science progresses via inference to the 'best' explanation rather than inference to 'any' explanation. Researchers and readers do not ask 'whether' a current result can be explained but rather 'how well' it can be explained. Consequently, if flexible theorizing leads to a low quality theoretical rationale (that is, incoherent, narrow, shallow, and complex), then the associated hypothesis will be negatively evaluated as having relatively low verisimilitude even if it has a relatively high degree of consistency with the current results (for a related discussion, see Szollosi and Donkin [2019]). Importantly, readers are able to identify deficiencies in the quality of a theoretical rationale, as presented in a research report, even if they are misled about when that rationale was developed. Hence, although flexible theorizing may occur during HARKing, readers can identify any associated low quality theorizing and take this into account in their estimates of relative verisimilitude.

Kerr ([1998]) also associated HARKing with data snooping in large data sets without appropriate experimentwise Type I error rate protection. There are two issues to consider here. First, researchers are unlikely to be interested in the experimentwise error rate unless they are interested in the experimentwise null hypothesis, and they are unlikely to be interested in the experimentwise null hypothesis in experiments that include multiple theoretically unrelated tests (for similar views, see Rothman [1990]; Savitz and Olshan [1995]; Cook and Farewell [1996], pp. 101–2; Perneger [1998]; Bender and Lange [2001]; Hewes [2003]; O'Keefe [2003]; Schulz and Grimes [2005]; Matsunaga [2007]; Morgan [2007]; Rothman, Greenland, and Lash [2008]; Parascandola [2010]; Hurlbert and Lombardi [2012]; Armstrong [2014]; Rubin [2017a], p. 271). For example, imagine a study that investigates differences in mathematics performance as a function of participants' gender, age, and social class. In this case, the two-sided experimentwise joint null hypothesis would be that 'men, older adults, and people from a lower social class background do not have either better or worse mathematics performance than women, younger adults, and people from a higher social class background, respectively'. If researchers undertake union-intersection testing (Kim *et al.* [2004]), then they will reject this experimentwise null hypothesis if any one of the three constituent two-sided tests yields a significant result. It is this 'multiple testing' issue that provides the rationale for lowering the significance threshold. For example, a Bonferroni correction would lower the conventional two-sided significance threshold from 0.050 to 0.016 in the case of the mathematics study (that is, 0.050/3 tests). Critically, however, even if gender, age, and social class are all expected to affect mathematics performance, the theoretical reasons for these associations will be different in each case (for example, gender stereotyping in the case of gender, cognitive ability in the case of age, and education levels in the case of social class). Consequently, it would be inappropriate to treat these three demographic variables as theoretically exchangeable predictors of mathematics performance, which is how they would be treated if a significant result for 'any one' of them was sufficient to reject the experimentwise null hypothesis. As a result, researchers should not be interested in testing theoretically impotent experimentwise null hypotheses or in computing associated experimentwise error rates. Instead, they should be more interested in testing theoretically meaningful (one-sided) individual null hypotheses (for example, 'men do not perform better than women in mathematics')

and theoretically meaningful (one-sided) joint null hypotheses (for example, 'men do not perform better than women in either (a) algebra or (b) calculus'). Note that the alpha level for individual tests of individual hypotheses does not require an adjustment, even if many such tests are performed within the same experiment (Tukey [1953]; Savitz and Olshan [1995], p. 906; Cook and Farewell [1996]; Matsunaga [2007]; Hurlbert and Lombardi [2012]; Rubin [2017a]).

Second, failing to disclose the timing of the construction of a HARKed but theoretically meaningful joint null hypothesis is a separate issue from failing to disclose the multiple testing of that hypothesis. Hence, a researcher can HARK a theoretically meaningful joint null hypothesis (for example, 'gender is not associated with performance in either (a) algebra or (b) calculus') and then transparently disclose and adjust for each of the multiple tests of this hypothesis (for example, using a significance threshold of 0.025 instead of 0.050). Mayo ([1996], p. 312) might describe this researcher as an 'honest hunter' because, although the timing of the hypothesis construction has been concealed (HARKing), the multiple testing has been disclosed and the significance threshold has been lowered accordingly. Note that the practice of making research materials and data publically available assists with the detection of any undisclosed multiple testing of joint null hypotheses (Rubin [2017a], p. 273).

3.2 Cost 2: Failing Popper's criterion of disconfirmability

Cost 2 is that HARKing propounds theories that cannot (pending replication) pass Popper's disconfirmability test. Specifically, Kerr ([1998]) argued that 'any hypothesis that could never fail (to be confirmed) can never succeed (as a scientifically testable explanation). A HARKed hypothesis fails this criterion, at least in a narrow, temporal sense' ([1998], pp. 205–6; see also Leung [2011], p. 475). However, as Kerr ([1998], p. 198) later discussed in relation to Cost 11, researchers can construct 'one or more hypotheses known *post hoc* to be contradicted by the data'. Hence, contrary to Kerr's Cost 2 assumption, it is possible to HARK a disconfirmed hypothesis. Rubin ([2017b], p. 318) considered HARKed disconfirmations and proposed that they were acceptable for RHARKed hypotheses but not for CHARKed hypotheses, because 'CHARKed hypotheses lack independence from the observed evidence'. However, elsewhere in this article, I argued that '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' ([2017b], p. 312). Hence, Rubin ([2017b]) had an inconsistent position on this issue. The concern that 'CHARKed hypotheses lack independence from the observed evidence' (p. 318) did not consider the concept of epistemic independence, and it confounded hypothesis construction with hypothesis evaluation.

Contrary to Kerr ([1998]) and Rubin ([2017b]), HARKed hypotheses pass Popper's disconfirmability test. To explain, it is necessary to understand Popper's views on falsifiability. Popper was concerned with a 'logical asymmetry' between verification and falsification (Earp and Trafimow [2015], p. 6). His critical argument is that 'a set of singular observation statements[…]may at times falsify or refute a universal law; but it cannot possibly verify a law, in the sense of establishing it' (Popper [1985], p. 181). So, for example, repeatedly observing black swans (a set of singular observations) may falsify the substantive hypothesis (universal law) that 'all swans are white' via the logical syllogism of *modus tollens*: If all swans are white, then no black swans should be observed. Black swans are observed. Therefore, not all swans are white. In contrast, the repeated observation of black swans cannot verify or confirm the substantive hypothesis that 'all swans are black'. If all swans are black, then black swans should be observed. Black swans are observed. However, these observations do not confirm that all swans are black. It

could be that only some swans are black, and we have sampled those black swans in our observations. We would be making the logical fallacy of affirming the consequent if we were to argue that we had verified that 'all swans are black' based on our observations. The problem of induction means that it is impossible to exhaustively test and confirm (verify) the universal law that 'all swans are black'.

A key implication of Popper's logical asymmetry between falsification and verification is that researchers should attempt to falsify substantive null hypotheses (universal laws) in order to provide piecemeal support for corresponding substantive alternative hypotheses. However, Popper conceded that these attempted falsifications could never be certain due to the problem of underdetermination (for a discussion, see Earp and Trafimow [2015]). In particular, it is not possible to falsify a substantive null hypothesis conclusively because it is always possible for this hypothesis to be true and for the observed research results to be explained by the falsity of some other (auxiliary) substantive hypothesis (for example, the hypothesis that the measuring instruments are valid; Earp and Trafimow [2015], p. 7). As Popper ([1985], p. 186) put it, 'we may have made a mistake'. Hence, observing several black swans only falsifies the substantive null hypothesis that 'all swans are white' *if* we have not made a mistake in defining and measuring the blackness and swanness of the birds that we observed. Give this problem of underdetermination, Popper distinguished between two senses of falsification: the logical sense and the practical sense ('naïve falsification'; Popper [1985], pp. xxxiv–v). As he explained,

> I have always stressed that even a theory which is obviously falsifiable in the first sense is never falsifiable in this second sense….Although the first sense refers to the logical possibility of a falsification in principle, the second sense refers to a conclusive practical experimental proof of falsity. But anything like conclusive proof to settle an empirical question does not exist (Popper [1985], p. xxii).

Despite Popper's clarifications on this point, Kerr's ([1998]) Popperian critique of HARKing is based on conclusive practical falsification rather than logical falsification. Hence, Kerr argued that, 'when the investigator knows the results of the study in advance and HARKs a hypothesis consistent with those results, *no immediate possibility of disconfirmation exists*' (p. 206, my emphasis). However, even Popper conceded that 'conclusive practical experimental proof of falsity' is impossible, because 'we may have made a mistake' (Popper [1985], p. xxii, p. 186; see also Earp and Trafimow [2015]). Hence, although Kerr's ([1998]) point is correct, it is correct for both HARKed and non-HARKed hypotheses. Consequently, it does not serve as a useful criticism of HARKing. HARKed hypotheses can never be conclusively disconfirmed, but neither can non-HARKed hypotheses.

To be fair, Kerr ([1998]) was aware of his oversimplification of Popper's position. As he explained, 'because either theories or research findings can be equivocal, certain hypotheses can be neither indisputably consistent nor inconsistent with the results in hand' (p. 197). However, he continued 'but for the present purposes, let us assume a simpler, more definitive world' (p. 197). The problem is that researchers and their readers do not live in this simpler, more definite world of naïve falsification (Popper [1985], pp. xxxiv–v). They live in a world in which the problem of underdetermination makes it impossible for a research finding to falsify a hypothesis conclusively. Ignoring this point does not provide a valid basis for making realistic claims about the costs of HARKing.

It is important to note that HARKing does not preclude Popper's second, more important meaning of falsification, which refers to the logical asymmetry between falsifiability and verifiability. This asymmetry is based on the content of hypotheses, rather than the timing of their

construction. For example, the hypothesis that 'all swans are white' is logically falsifiable, whereas the hypothesis that 'some swans are white' is not logically falsifiable. This asymmetry persists even if these two hypotheses are constructed (a) after knowing the research results and (b) on the basis of the research results. Hence, contrary to Kerr ([1998]), HARKing does not propound theories or hypotheses that fail to pass Popper's disconfirmability test.

3.3 Cost 3: Disguising accommodation as prediction

Cost 3 refers back to readers' underappreciation of the risk of overfitting when researchers disguise accommodation as prediction. However, if the act of disguising accommodation as prediction entails the deduction of hypotheses from *a priori* theory and evidence, then this act produces a genuine prediction that precludes the cost of overfitting. To explain, imagine that a researcher conducts a study and finds that participants who eat an apple a day report better mood than those in a control condition. The researcher may then accommodate this current result by inducing the *ad hoc* hypothesis that 'eating apples improves mood'. In the absence of any other information, the relative verisimilitude of this *ad hoc* hypothesis can only be estimated on the basis of the current result. However, now imagine that the researcher 'disguises' this accommodation as prediction by explaining in the Introduction section of their research report that (a) *a priori* theory predicts that 'Vitamin C improves mood', (b) *a priori* evidence shows that 'apples are rich in Vitamin C', and so (c) 'eating apples should improve mood'. In this case, the act of disguising *ad hoc* accommodation based on a current result as an *a priori* prediction based on *a priori* theory and evidence provides the researcher and their readers with a new basis for making an initial estimate of the relative verisimilitude of the hypothesis that is independent from the current result. In turn, this act allows them to use the current result to increase or decrease that initial estimate.

In summary, Cost 3 assumes that the act of disguising accommodation as prediction hides the costs associated with accommodation (namely, double-counting, overfitting). However, if this act entails the public deduction of the hypothesis from *a priori* theory and evidence, then it actually eliminates the costs associated with accommodation because it provides a valid independent basis for making an initial estimate of the relative verisimilitude of the hypothesis. In effect, the act of 'disguising' accommodation as prediction turns *ad hoc* accommodation into *post hoc* prediction.

3.4 Cost 4: Not communicating information about what did not work

Cost 4 is not communicating valuable information about disconfirmed hypotheses. Indeed, as discussed below (Cost 6), there is some evidence that researchers suppress hypotheses that are disconfirmed (Mazzola and Deuling [2013]). However, as other commentators have pointed out, a key distinction here is between disconfirmed hypotheses that are relevant and irrelevant to the research conclusions (Leung [2011]; Rubin [2017b]; Vancouver [2018]).

Obviously, it is biased for researchers to suppress *a priori* hypotheses that are relevant to their research conclusions. However, this bias is likely to be identified and addressed during two stages of peer review. First, if the suppressed hypotheses are obviously relevant to the research conclusions, then they are likely to be suggested as alternative explanations during the pre-publication peer review process (see Cost 11). As Kerr ([1998], p. 208) asked, 'if one's original hypotheses had a sufficient rationale (theoretical, empirical, or even intuitive) to recommend itself to one researcher, why would it not also occur to others?'. Second, if suppressed relevant *a priori* hypotheses are overlooked during pre-publication peer review, then they may be generated by interested readers during post-publication peer review. Furthermore, if research materials and data are made publically available, then these hypotheses may be tested by interested readers.

In contrast to the suppression of *a priori* hypotheses that are relevant to the research conclusions, the suppression of *a priori* hypotheses that are irrelevant to the research conclusions will not bias research conclusions (Leung [2011]; Rubin [2017b]; Vancouver [2018]). Nonetheless, Rubin ([2017b], p. 316) argued that even the suppression of irrelevant *a priori* hypotheses can 'bias information about the size and replicability of effects in meta-analyses'. In addition, the suppression of irrelevant *a priori* hypotheses that yield inconclusive (null) findings prevents other researchers from identifying and avoiding nondiagnostic methods in their future research. Again, however, making research materials and data publically available helps to alleviate these problems. Specifically, readers, meta-analysts, and other researchers may access publically available research materials and data in order to formulate and test their own hypotheses. For an illustration, please see my use of Janke, Daumiller, and Rudert's ([2019]) publically available research materials and data to test a hypothesis that was irrelevant to Janke *et al.*'s conclusions but relevant to Kerr's ([1998]) Cost 7 below.

In summary, the suppression of disconfirmed hypotheses may lead to biased research conclusions, missing null results, and/or unreported failed methods. However, these problems are reduced through pre- and post-publication peer review and the public availability of research materials and data.

3.5 Cost 5: Taking unjustified statistical license

Cost 5 refers to researchers taking unjustified statistical license, such as using one-tailed tests and contrast analyses that are supposed to be based on *a priori* theory. Certainly, in the absence of a theoretical rationale for a hypothesis, a two-sided (nondirectional) test is more appropriate than a one-sided test, because the researcher has no reason to predict that the effect will be positive or a negative (Cortina and Dunlap [1997], p. 168). However, a one-sided test is appropriate if it is accompanied by a theoretical rationale that specifies a directional prediction based on *a priori* theory and evidence (Cortina and Dunlap [1997], p. 168). Importantly, the timing of the construction of this theoretical rationale is not relevant. A theoretical rationale may warrant a directional test regardless of whether that rationale is constructed before or after a researcher conducts the test. For example, imagine that a researcher uses a one-sided test and obtains a significant positive correlation ($\alpha = 0.050$). In the absence of a theoretical rationale, this directional test is unwarranted. However, further imagine that the test result inspires the researcher to deduce a *post hoc* hypothesis from *a priori* theory and evidence that predicts a positive correlation. Now, the hypothesis provides a rationale for using the one-sided test. Critics may argue that an alpha adjustment is required in this case (for example, $\alpha = 0.025$) because the researcher could have conducted two one-sided tests: one to test for a positive correlation and the other to test for a negative correlation. However, the researcher's hypothesis only makes a single, directional, one-sided prediction about the presence of a positive correlation (and, consequently, the absence of a negative correlation). It does not make a joint, bidirectional, two-sided prediction about the presence of either a positive or a negative correlation. Consequently, it does not warrant a union-intersection test of the joint null hypothesis of 'no positive or negative correlation' (Cortina and Dunlap [1997], p. 168; Kim *et al.* [2004]), and no alpha adjustment for the familywise error rate of this joint hypothesis is required, because this hypothesis is not being tested.

Similarly, a HARKed hypothesis may predict a specific pattern of results that can be tested using contrast analyses. Again, the statistical license for these analyses is the theoretical rationale for the hypothesis, as explained in the research report. This license does not become unjustified

merely because the theoretical rationale was conceived after the researcher knew the test result. Instead, it becomes unjustified if there is no theoretical rationale for the directional prediction.

In summary, researchers who present a theoretical rationale for a hypothesis based on *a priori* theory and evidence are justified in using statistical tests that assess the directional predictions made by that hypothesis regardless of whether that rationale was developed before or after the test result was known to them.

3.6 Cost 6: Presenting an inaccurate model of science

Cost 6 is presenting an inaccurate and distorted model of science to students. In particular, Kerr ([1998], p. 208) argued that HARKing presents a 'too-rosy picture' of how science normally proceeds because it gives the impression that researchers are usually correct in their predictions. However, Cost 6 is caused by 'publication bias' rather than HARKing. In particular, Cost 6 arises because, all other things being equal, professional research journals tend to favour the publication of significant results and reject the publication of nonsignificant null findings. HARKing arises as a consequence of this publication bias: Researchers suppress *a priori* hypotheses that are associated with nonsignificant results and HARK hypotheses that fit significant results. Consistent with this view, in a survey of 156 behavioural scientists, Kerr and Harris ([1998], as cited in Kerr [1998], p. 202) found that respondents believed that a HARKed article had a greater chance of publication than a non-HARKed article. A more recent study by Mazzola and Deuling ([2013]) reinforced the idea that publication bias causes HARKing. These researchers analysed 215 published journal articles and 127 unpublished PhD dissertations in the area of industrial and organizational psychology during the period 2010–2012. They found a significantly higher percentage of supported hypotheses and a significantly lower percentage of unsupported hypotheses in the published journal articles compared to the unpublished PhD dissertations. Hence, it appears that researchers' awareness of publication bias caused them to HARK when writing journal articles.

Eliminating HARKing would not eliminate the publication bias that lies behind it. If researchers did not HARK, then their null findings would continue to be rejected by journals for publication, and these journals would continue to present a distorted 'too-rosy picture' of science. The way to address this distorted view is to address publication bias, not HARKing.

It is also important to educate students and researchers about HARKing and the conditions under which it may and may not be appropriate (see also Kerr [1998], p. 214; Rubin [2017b], p. 311). In particular, it is important to understand the potential meanings of phrases such as 'we predicted that' and 'as hypothesized'. A common interpretation of these phrases is that they imply that researchers constructed their hypothesis before they observed their research results. However, as explained earlier, phrases such as these may also denote *post hoc* predictions that researchers have deduced from *a priori* theory and evidence (Worrall [1985], [2014]; Rubin [2017b], p. 311). Under this latter interpretation, the phrase 'we hypothesized that…' retains its scientific meaning in 'the epistemically important sense' (Worrall [2014], p. 55) even if researchers constructed their hypotheses after observing their current results.

Finally, it is important to appreciate the epistemic source of predictions. Again, the common view is that researchers, rather than hypotheses, do the predicting. However, this view is misleading because we should be more interested in the predictive accuracy of hypotheses than of researchers. In other words, we should be concerned about 'hypothesis testing' than 'researcher testing'. As Vancouver ([2018], p. 77) explained,

the hypothesis is *not* about the author. It is about the theory or arguments brought to bear to the research question constrained by the methods used to test the theory. These theories or arguments might be brought to bear after the results are known; but, regardless of order, the hypotheses follow from these arguments, and that is how to state it (e.g., 'this leads to the following hypothesis').

3.7 Cost 7: Encouraging 'fudging' in other grey areas

Cost 7 is the encouragement of 'fudging' in other grey areas. Kerr ([1998], pp. 209–10) explained that 'HARKing may fall into that grey area of "questionable scientific practices" where the line between clearly appropriate and clearly inappropriate behavior is indistinct'. Consistent with this view, John, Loewenstein, and Prelec ([2012]) categorized HARKing as one of ten questionable research practices. However, as John, Loewenstein, and Prelec ([2012], p. 531) noted, there is a great degree of variability in the 'questionability' of these research practices. Some research practices may be considered to be acceptable under some conditions (for example, 'failing to report all of a study's dependent measures'), whereas others are clearly inappropriate under all conditions (for example, 'falsifying data'; for a further discussion, see Pickett and Roche [2018], pp. 152–3).

According to Kerr ([1998], p. 210), Cost 7 is that the informal acceptance of HARKing by the scientific community encourages fudging in other grey areas. In particular, the acceptance of HARKing by editors and peer reviewers 'encourages ignoring or blurring other such lines when convenient' and is 'greasing the slippery slope' towards the use of inappropriate research practices (p. 210). If Kerr's slippery slope hypothesis is correct, then there should be a positive correlation between HARKing and other questionable research practices. To test this prediction, I analysed data that was made publically available by Janke, Daumiller, and Rudert ([2019]; https://journals.sagepub.com/doi/suppl/10.1177/1948550618790227). These researchers sampled 217 German psychology doctoral candidates and postdocs and measured questionable research practices using John, Loewenstein, and Prelec's ([2012]) ten items. The HARKing question was: 'How often have you reported an unexpected finding as having been predicted from the start?' Participants responded to this and the other nine items using a five-point scale ranging from 'never' to 'very frequently'. The average correlation between HARKing and the other research practices was small ($r = 0.06$). The three highest (and only significant) correlations were between HARKing and (a) failing to report all of a study's dependent measures ($r = 0.35$), (b) selectively reporting studies that 'worked' ($r = 0.28$), and (c) deciding whether to exclude data after looking at the impact of doing so on the results ($r = 0.26$). Note that these three research practices are associated with suppressing hypotheses after the results are known, which is a form of HARKing (Kerr [1998]; Rubin [2017b]). Hence, there are theoretical grounds for expecting these positive correlations. Also, note that each of these research practices are justifiable under certain conditions. For example, failing to report all of a study's dependent measures may not bias the research conclusions if the measures yielded results that are irrelevant to the research conclusions (see Cost 4) and are presented to participants at the end of a study and so had no chance of affecting results from previously presented dependent measures. Similarly, failing to report an entire study may not bias the research conclusions if the study produced irrelevant results and/or suffered from a fatal flaw (for example, an invalid experimental manipulation or very low sample size) that rendered its results inconclusive. Finally, deciding whether to exclude data after looking at the impact of doing so on the results may be acceptable if the researcher transparently describes and

justifies this approach in their research report and reports any substantive discrepancies in their pattern of results when the data is included.

HARKing had weak and nonsignificant correlations with three other questionable research practices, including optional stopping of data collection in order to achieve significant results (*r*s = 0.06 and 0.10) and failing to report all of a study's conditions ($r = 0.01$). Importantly, HARKing also had weak nonsignificant correlations with research practices that are clearly inappropriate, including reporting a p value of 0.054 as $p < 0.05$ ($r = 0.05$), claiming results are unaffected by demographic variables when one is unsure or knows that they are affected ($r = 0.09$), and falsifying data ($r = -0.06$).

In summary, Janke, Daumiller, and Rudert's ([2019]) data showed limited support for Kerr's ([1998]) slippery slope hypothesis. HARKing was only associated with three out of nine questionable, incorrect, or fraudulent research practices, and these three research practices can be interpreted as being appropriate under certain conditions. Hence, there is no clear support for the idea that HARKing is associated with 'fudging' in other areas.

3.8 Cost 8: Making researchers less receptive to serendipitous findings

Cost 8 is that the current scientific system, with its emphasis on *a priori* confirmation and prediction, may adversely affect scientific discovery by making researchers less receptive to serendipitous findings (see also Hollenbeck and Wright [2017], p. 12). As Kerr ([1998]) noted, 'this is less a consequence of HARKing per se than of one of its primary causes: insistence on confirmation of *a priori* hypotheses as the only route to new knowledge' (p. 210). Hence, lack of receptivity to serendipitous findings is a cost of the scientific system's insistence of confirmation of *a priori* hypotheses. It is not a cost of HARKing. Indeed, HARKing is a solution to this problem, because it is a means by which researchers can publically report serendipitous findings in a scientific system that values *a priori* prediction over *post hoc* prediction.

3.9 Cost 9: Encouraging the adoption of narrow, context-bound theory

Costs 9, 10, and 11 refer to the type of theory development that might occur if HARKing was commonplace. Cost 9 is the encouragement of the adoption of narrow, context-bound new theory. This cost refers back to Cost 1 (translating Type I errors into theory) and the concern about overfitting. As Kerr ([1998], p. 210) explained,

> HARKing may foster preoccupation with explaining an effect in hand with an attendant lack of attention to the broader set of potentially relevant prior findings. If HARKing were commonplace, much of our new theory would have such a genesis – being developed to explain some specific effect observed in some specific experimental context and paradigm.

Cost 9 assumes that HARKing produces low quality ('narrow, context-bound') theoretical rationales and hypotheses. To my knowledge, there is no evidence that HARKed hypotheses are perceived to be of poorer quality than non-HARKed hypotheses. Indeed, given that HARKing is often encouraged by expert peer reviewers (Kerr and Harris [1998], as cited in Kerr [1998]; Leung [2011]; Hollenbeck and Wright [2017]; Rubin [2017b]), it may actually improve the quality of hypotheses relative to unadulterated *a priori* hypotheses.

However, even if HARKing does lower the quality of hypotheses, readers can identify this low quality in the theoretical rationale that is presented in the Introduction section of the research report and take it into account when making their initial estimates of relative verisimilitude. For example, the initial estimated relative verisimilitude of a hypothesis that is based on a narrow, context-bound theoretical rationale will be low, and it will remain relatively low after being

updated by information from the current result, unless that current result provides particularly compelling evidence.

In summary, HARKing may not necessarily lead to narrow, context-bound theoretical rationales. However, if it does, then readers are able to take such low quality theorizing into account in their initial estimates of relative verisimilitude.

3.10 Cost 10: Encouraging the retention of too-broad theory

Cost 10 is the retention of too-broad disconfirmable old theory. As Kerr ([1998], p. 210) explained, this includes

> theory with which one can predict nearly any pattern of results in nearly any context. Such theory is likely to be vague in its concepts, to incorporate multiple (sometimes opposing) processes, and to have many auxiliary assumptions that are not firmly fixed and can be used in an *ad hoc* fashion to accommodate nearly any observation.

Like Cost 9, Cost 10 assumes that theories are judged in terms of their empirical support and not in terms of their theoretical quality. Indeed, at another point in his article, Kerr ([1998]) lamented that 'we trivialize the term *theory* and overcredit apparent confirmations when we draw no distinction between hypotheses derived from poor theory and those from better theory' (p. 215). Certainly, Costs 9 and 10 assume that researchers and their readers trivialize theory and overcredit confirmation. But is this characterization correct? Do researchers and readers really fail to distinguish between poor theory and better theory? I would argue that they do not. Instead, researchers and readers regularly criticize research in terms of the quality of the theoretical ideas, referring to theoretical virtues such as coherence, breadth, depth, and parsimony (Ladyman [2002]; Meehl [2004]; Mackonis [2013]). Hence, to return to Kerr's example, a person who perceives a theory to be vague in its concepts, to incorporate multiple opposing processes, and to have many flexible auxiliary assumptions is likely to regard that theory as being lower in coherence, depth, and parsimony than a theory that is well-specified, internally consistent, and less flexible.

3.11 Cost 11: Inhibiting the identification of alternative hypotheses

Cost 11 is the lack of identification of plausible alternative hypotheses. Specifically, Kerr ([1998]) argued that researchers tend to consider a confirmed hypothesis to be a correct hypothesis, and they do not tend to consider the possibility that other hypotheses might have predicted the same result. This 'satisficing bias' may apply to both *a priori* and *post hoc* hypotheses. However, Kerr suggested that it is particularly problematic when researchers HARK, because HARKed hypotheses always meet researchers' expectations, and researchers are less likely to consider alternative explanations when their expectations are met. As he explained, 'HARKing guarantees the provision of one or more hypotheses sufficient to account for a result in hand. If readers or the investigator himself or herself acted like such satisficers, they would not need to consider alternative explanations' (p. 210).

Cost 11 does not take into account the realities of the scientific publication system. Editors and peer reviewers of research reports often encourage researchers to compare the primary explanation for their results with alternative explanations (for example, in the 'Limitations' and 'Directions for Future Research' sections of a research report). Furthermore, editors may reject reports from publication or request a revision if alternative explanations are not considered. Hence, researchers are required to engage in a documented process of inference to the best explanation for their results, and this process includes a discussion of alternative explanations. HARKers are no less likely to face this requirement than non-HARKers. However, in a system that discourages

the publication of *post hoc* hypothesis testing, HARKers have a greater potential to publish tests of alternative explanations, because they are able to disguise *post hoc* alternative explanations as *a priori* alternative explanations. Hence, contrary to Kerr ([1998]), and in a reversal of Cost 4 (suppressing relevant *a priori* hypotheses), HARKing may actually facilitate the publication of tests of alternative explanations.

Kerr ([1998], p. 198) briefly considered this possibility when he mentioned that researchers can construct 'one or more hypotheses known *post hoc* to be contradicted by the data'. He noted that 'this approach could create an illusion of competitive hypothesis testing' (p. 198). Here, the 'illusion' refers to the reader's misguided belief that the researchers were unaware of the outcome of the competitive hypothesis test prior to conducting the test. Despite this erroneous belief, the competitive hypothesis test facilitates valid estimates of relative verisimilitude in a process of inference to the best explanation. For example, if a positive correlation is observed, and *a priori* theory and evidence is then used to independently deduce (a) a *post hoc* hypothesis that predicts this positive correlation and (b) a *post hoc* hypothesis that predicts a negative correlation, then the estimated verisimilitude of the first hypothesis can be increased relative to second.

3.12 Cost 12: Violating the ethical principles of honesty and openness

Cost 12 refers to the violation of ethical principles. Specifically, Kerr ([1998], p. 209) argued that 'HARKing violates a fundamental ethical principle of science: the obligation to communicate one's work honestly and completely'. However, he provided some nuance to this point: 'HARKing can entail concealment. The question then becomes whether what is concealed in HARKing can be a useful part of the "truth" […] or is instead basically uninformative (and may, therefore, be safely ignored at an author's discretion)' ([1998], p. 209).

Hence, as a general principle, researchers should be honest and open when reporting their research. However, researchers do not need to be honest and open about facts that are 'basically uninformative' and do not form 'a useful part of the "truth"' (Kerr [1998], p. 209). For example, it is usually unnecessary for researchers to disclose the colour of the socks that they were wearing when they conducted their study! Similarly, as I have argued throughout this article, although the epistemic independence between a hypothesis and a current result is 'a useful part of the "truth"' (Kerr [1998], p. 209), the timing of the construction of the hypothesis is 'basically uninformative'. Hence, concealing the timing of hypothesis construction may not necessarily be regarded as being unethical, because it does not prevent an accurate assessment of epistemic independence (for a similar view, see Vancouver [2018], pp. 77–8).

4 Summary and Conclusions

What are the costs of harking? Kerr ([1998]) proposed that they are plentiful, and recent discussions have argued that they may have led to low replication rates in science (John, Loewenstein, and Prelec [2012]; Mazzola and Deuling [2013]; Aguinis, Cascio, and Ramani [2017]; Hollenbeck and Wright [2017]). The present article challenges this perspective. Table 1 provides a summary of Kerr's ([1998]) twelve costs of HARKing and some responses.

By definition, HARKing conceals the timing of hypothesizing. However, it does not conceal the quality of (a) the hypothesizing, (b) the research methodology, or (c) the statistical analysis. Readers can make judgements about the quality of each these aspects of the research without knowing the timing of the hypothesizing. In particular, even if readers are unaware that a

Table 1
Kerr's ([1998]) Twelve Costs of HARKing and Some Responses

Cost	Kerr ([1998])	Response
1	Translating Type I errors into hard-to-eradicate theory.	The overfitting of *post hoc* hypotheses to Type I errors is not possible when those hypotheses are deduced from *a priori* theory and evidence. Flexible theorizing is possible, but it can be identified and taken into account in estimates of relative verisimilitude.
2	Propounding theories that cannot (pending replication) pass Popper's disconfirmability test.	Popper's disconfirmability test refers to the content of hypotheses, not the timing of the construction of those hypotheses. HARKed hypotheses can pass this test.
3	Disguising *post hoc* explanations as *a priori* explanations (when the former tend also be [sic] more *ad hoc*, and consequently, less useful).	Deducing *post hoc* explanations from *a priori* theory and evidence provides an epistemically independent basis for estimating relative verisimilitude that prevents overfitting and validates *post hoc* predictions.
4	Not communicating valuable information about what did not work.	The potential costs associated with biased research conclusions, missing null results, and/or unreported failed methods are alleviated via pre- and post-publication peer review and the public availability of research materials and data.
5	Taking unjustified statistical licence.	*Post hoc* theoretical rationales that are deduced from *a priori* theory and evidence can provide valid justifications for directional statistical tests.
6	Presenting an inaccurate model of science to students.	Publication bias presents an inaccurate model of science to students. HARKing is only a response to publication bias.
7	Encouraging 'fudging' in other grey areas.	There is limited evidence that HARKing encourages the use of inappropriate research practices.
8	Making us less receptive to serendipitous findings.	Lack of receptiveness to serendipitous findings is caused by the scientific community valuing *a priori* prediction over *post hoc* prediction. HARKing solves this problem rather than causes it.
9	Encouraging adoption of narrow, context-bound new theory.	Even if HARKing is associated with low quality theorizing, this problem is taken into account in readers' estimates of relative verisimilitude.
10	Encouraging retention of too-broad, disconfirmable old theory.	Too-broad, disconfirmable theory will receive low ratings of relative verisimilitude.
11	Inhibiting identification of plausible alternative hypotheses.	The scientific publication system requires researchers to consider alternative explanations in their research reports. HARKing facilitates this process in a system that discourages the publication of *post hoc* hypothesis testing.
12	Implicitly violating basic ethical principles.	The ethical principle of honesty and openness applies to information that forms a useful part of the truth. The timing of the construction of a hypothesis does not form a useful part of the truth.

hypothesis has been HARKed, they are still able to criticize (a) the theoretical quality of HARKed hypotheses (too-broad theorizing, too-narrow theorizing, no alternative explanations; Costs 9, 10, and 11), (b) the appropriateness of the methodology for testing those hypotheses (test severity; Rubin [2017b]; Mayo [2018]) and (c) the appropriateness of the statistical analyses (lack of correction for multiple testing, lack of justification for directional tests; Costs 1 and 5). Hence, Costs 1, 5, 9, 10, and 11 are misattributed to HARKing when they are actually criticisms of the quality of the hypothesizing and data analysis. Similarly, Costs 6 and 8 are misattributed to HARKing when they are actually due to publication bias (Cost 6) and the scientific community's preference for *a priori* prediction over *post hoc* prediction (Cost 8).

Cost 2 is misconceived: HARKed hypotheses can be disconfirmed, and they do not necessarily fail Popper's disconfirmability test. In addition, Costs 1 and 3 do not recognize that the act of disguising *ad hoc* accommodation as prediction by deducing hypotheses from *a priori* theory and evidence precludes the cost of overfitting.

The suppression of *a priori* hypotheses (Cost 4) may lead to biased research conclusions, missing null results, and/or unreported failed methods. However, biased conclusions can be addressed through pre- and post-publication peer review, and missing information can be addressed by making research materials and data publically available.

Cost 7 lacks empirical evidence: There is no clear support for Kerr's ([1998]) slippery slope argument that HARKing encourages fudging in other areas. Finally, HARKing cannot be considered to be unethical if it conceals information that is uninformative, and the timing of the hypothesizing may be considered to be scientifically uninformative (Cost 12).

Given the potentially limited costs of HARKing to the scientific process, it is premature to conclude that HARKing is an important contributor to low replication rates. However, it is also premature to conclude that HARKing has no costs to scientific progress. Future research in this area might profit from testing the potential costs of HARKing. For example, based on the current review, it may be useful to investigate whether HARKing leads to a failure to disclose multiple testing of HARKed hypotheses (Cost 1), the suppression of relevant *a priori* hypotheses (Cost 4), the encouragement of inappropriate research practices (Cost 7), the production of low quality hypotheses (Cost 9), and the inhibition of alternative hypotheses (Cost 11). These potential costs of HARKing need to be balanced against the potential costs of proposed solutions to HARKing, such as preregistration, in order to arrive at the most efficacious approach to scientific progress.

References

Aguinis, H. Cascio, W. F. and Ramani, R. S. [2017]: 'Science's Reproducibility and Replicability Crisis: International Business is Not Immune', *Journal of International Business Studies,* **48,** pp. 653–63. http://dx.doi.org/10.1057/s41267-017-0081-0

Armstrong, R. A. [2014]: 'When to Use the Bonferroni Correction', *Ophthalmic and Physiological Optics,* **34,** pp. 502–8. http://dx.doi.org/10.1111/opo.12131

Bender, R. and Lange, S. [2001]: 'Adjusting for Multiple Testing—When and How?' *Journal of Clinical Epidemiology,* **54**, pp. 343–9. http://dx.doi.org/10.1016/S0895-4356(00)00314-0

Berger, J. [2006]: 'The Case for Objective Bayesian Analysis', *Bayesian Analysis,* **1,** pp. 385–402. http://dx.doi.org/10.1214/06-BA115

Cevolani, G. and Festa, R. [2018]: 'A Partial Consequence Account of Truthlikeness', *Synthese,* available at http://dx.doi.org/10.1007/s11229-018-01947-3

Cook, R. J. and Farewell, V. T. [1996]: 'Multiplicity Considerations in the Design and Analysis of Clinical Trials', *Journal of the Royal Statistical Society: Series A,* **159,** pp. 93–110. http://dx.doi.org/10.2307/2983471

Cortina, J. M. and Dunlap, W. P. [1997]: 'On the Logic and Purpose of Significance Testing', *Psychological Methods,* **2,** pp. 161–72. http://dx.doi.org/10.1037/1082-989X.2.2.161

Earp, B. D. and Trafimow, D. [2015]: 'Replication, Falsification, and the Crisis of Confidence in Social Psychology', *Frontiers in Psychology,* **6**, pp. 1–11. http://dx.doi.org/10.3389/fpsyg.2015.00621

Gelman, A. and Hennig, C. [2017]: 'Beyond Subjective and Objective in Statistics', *Journal of the Royal Statistical Society: Series A,* **180,** pp. 967–1033. http://dx.doi.org/10.1111/rssa.12276

Hewes, D. E. [2003]: 'Methods as Tools: A Response to O'Keefe', *Human Communication Research*, **29,** pp. 448–54. http://dx.doi.org/10.1111/j.1468-2958.2003.tb00847.x

Hitchcock, C. and Sober, E. [2004]: 'Prediction versus Accommodation and the Risk of Overfitting', *British Journal for the Philosophy of Science,* **55,** pp. 1–34. http://dx.doi.org/10.1093/bjps/55.1.1

Hollenbeck, J. R. and Wright, P. M. [2017]: 'Harking, Sharking, and Tharking: Making the Case for Post Hoc Analysis of Scientific Data', *Journal of Management,* **43,** pp. 5–18. http://dx.doi.org/10.1177/0149206316679487

Howson, C. [1984]: 'Bayesianism and Support by Novel Facts', *British Journal for the Philosophy of Science,* **35,** pp. 245–51. 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,** pp. 305–9. https://www.jstor.org/stable/687574

Howson, C. and Urbach, P. [1993]: '*Scientific Reasoning: The Bayesian Approach*', Chicago, Il: Open Court.

Hurlbert, S. H. and Lombardi, C. M. [2012]: 'Lopsided Reasoning on Lopsided Tests and Multiple Comparisons', *Australian and New Zealand Journal of Statistics,* **54,** pp. 23–42. http://dx.doi.org/10.1111/j.1467-842X.2012.00652.x

Janke, S. Daumiller, M. and Rudert, S. C. [2019]: 'Dark Pathways to Achievement in Science: Researchers' Achievement Goals Predict Engagement in Questionable Research Practices', *Social Psychological and Personality Science,* **10,** pp. 783–91. http://dx.doi.org/10.1177/1948550618790227

John, L. K. Loewenstein, G. and Prelec, D. [2012]: 'Measuring the Prevalence of Questionable Research Practices with Incentives for Truth Telling', *Psychological Science,* **23,** pp. 524–32. http://dx.doi.org/10.1177/0956797611430953

Kerr, N. L. [1998]: 'HARKing: Hypothesizing After the Results Are Known', *Personality and Social Psychology Review,* **2,** pp. 196–217. http://dx.doi.org/10.1207/s15327957pspr0203_4

Kim, K. Zakharkin, S. O. Loraine, A. and Allison, D. B. [2004]: 'Picking the Most Likely Candidates for Further Development: Novel Intersection-Union Tests for Addressing Multi-Component Hypotheses in Comparative Genomics', in *Proceedings of the American Statistical Association, Alexandria, VA: American Statistical Association* pp. 1396–402. http://www.uab.edu/cngi/pdf/2004/JSM%202004%20-IUTs%20Kim%20et%20al.pdf

Ladyman, J. [2002]: *Understanding Philosophy of Science,* London: Routledge. http://dx.doi.org/10.4324/9780203463680

Leung, K. [2011]: 'Presenting Post Hoc Hypotheses as A Priori: Ethical and Theoretical Issues', *Management and Organization Review,* **7,** pp. 471–9. http://dx.doi.org/10.1017/CBO9781139171434.009

Mackonis, A. [2013]: 'Inference to the Best Explanation, Coherence and Other Explanatory Virtues', *Synthese,* **190,** pp. 975–95. http://dx.doi.org/10.1007/s11229-011-0054-y

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,** pp. 243–65. http://dx.doi.org/10.1080/19312450701641409

Mayo, D. G. [1996]: *Error and the Growth of Experimental Knowledge,* Chicago, Il: Chicago University Press.

Mayo, D. G. [2008]: 'How to Discount Double-Counting When it Counts: Some Clarifications', *British Journal for the Philosophy of Science,* **59,** pp. 857–79. https://doi.org/10.1093/bjps/axn034

Mayo, D. G. [2018]: *Statistical Inference as Severe Testing,* Cambridge: Cambridge University Press.

Mazzola, J. J. and 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,** pp. 279–84. http://dx.doi.org/10.1111/iops.12049

Meehl, P. E. [1990]: 'Appraising and Amending Theories: The Strategy of Lakatosian Defense and Two Principles That Warrant it', *Psychological Inquiry,* **1,** pp. 108–41. http://dx.doi.org/10.1207/s15327965pli0102_1

Meehl, P. E. [2004]: 'Cliometric Metatheory III: Peircean Consensus, Verisimilitude and Asymptotic Method', *British Journal for the Philosophy of Science,* **55,** pp. 615–43. http://dx.doi.org/10.1093/bjps/55.4.615

Morgan, J. F. [2007]. '*P* value fetishism and use of the Bonferroni adjustment', *Evidence-Based Mental Health*, **10**, pp. 34-35. http://dx.doi.org/10.1136/ebmh.10.2.34

Musgrave, A. [1974]: 'Logical versus Historical Theories of Confirmation', *British Journal for the Philosophy of Science,* **25,** pp. 1–23. http://dx.doi.org/10.1093/bjps/25.1.1

Oberauer, K. and Lewandowsky, S. [2019]: 'Addressing the Theory Crisis in Psychology', *Psychonomic Bulletin and Review*. http://dx.doi.org/10.3758/s13423-019-01645-2

O'Keefe, D. J. [2003]: 'Colloquy: Should Familywise Alpha Be Adjusted?' *Human Communication Research,* **29,** pp. 431–47. http://dx.doi.org/10.1111/j.1468-2958.2003.tb00846.x

Parascandola, M. [2010]: 'Epistemic Risk: Empirical Science and the Fear of Being Wrong', *Law, Probability and Risk,* **9,** pp. 201–14. http://dx.doi.org/10. 1093/lpr/mgq005

Perneger, T. V. [1998]: 'What's Wrong with Bonferroni Adjustments', *British Medical Journal,* **316,** pp. 1236–8. https://doi.org/10.1136/bmj.316.7139.1236

Pickett, J. T. and Roche, S. P. [2018]: 'Questionable, Objectionable or Criminal? Public Opinion on Data Fraud and Selective Reporting in Science', *Science and Engineering Ethics,* **24,** pp. 151–71. http://dx.doi.org/10.1007/s11948-017-9886-2

Popper, K. [1985]: *Realism and the Aim of Science: From the Postscript to the Logic of Scientific Discovery,* London: Routledge.

Rothman, K. J. [1990]: 'No Adjustments Are Needed for Multiple Comparisons', *Epidemiology,* **1,** pp. 43–6. https://www.jstor.org/stable/20065622

Rothman, K. J. Greenland, S. and Lash, T. L. [2008]: *Modern Epidemiology*, New York: Lippincott Williams and Wilkins.

Rubin, M. [2017a]: 'Do *p* Values Lose Their Meaning in Exploratory Analyses? It Depends How You Define the Familywise Error Rate', *Review of General Psychology*, **21,** pp. 269–75. http://dx.doi.org/10.1037/gpr0000123

Rubin, M. [2017b]: 'When Does HARKing Hurt? Identifying When Different Types of Undisclosed Post Hoc Hypothesizing Harm Scientific Progress', *Review of General Psychology*, **21,** pp. 308–20. http://dx.doi.org/10.1037/gpr0000128

Savitz, D. A. and Olshan, A. F. [1995]: 'Multiple Comparisons and Related Issues in the Interpretation of Epidemiologic Data', *American Journal of Epidemiology*, **142,** pp. 904–8. http://dx.doi.org/10.1093/oxfordjournals.aje.a117737

Schulz, K. F. and Grimes, D. A. [2005]: 'Multiplicity in Randomised Trials I: Endpoints and Treatments', *The Lancet*, **365,** pp. 1591–5. http://dx.doi.org/10.1016/S0140-6736(05)66461-6

Szollosi, A. and Donkin, C. [2019]: '*Arrested Theory Development: The Misguided Distinction between Exploratory and Confirmatory Research*', available at http://dx.doi.org/10.31234/osf.io/suzej

Tukey, J. W. [1953]: *The Problem of Multiple Comparisons*, Princeton, NJ: Princeton University Press.

Vancouver, J. B. [2018]: 'In Defense of HARKing', *Industrial and Organizational Psychology*, **11,** pp. 73–80. http://dx.doi.org/10.1017/iop.2017.89

Van de Schoot, R. Kaplan, D. Denissen, J. Asendorpf, J. B. Neyer, F. J. and Van Aken, M. A. [2014]: 'A Gentle Introduction to Bayesian Analysis: Applications to Developmental Research', *Child Development*, **85,** pp. 842–60. http://dx.doi.org/10.1111/cdev.12169

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*, Dordrecht: Reidel, pp. 301–31. http://dx.doi.org/10.1007/978-94-009-6525-6_11

Worrall, J. [2014]: 'Prediction and Accommodation Revisited', *Studies in History and Philosophy of Science*, **45,** pp. 54–61. http://dx.doi.org/10.1016/j.shpsa.2013.10.001

Funding

The author declares no funding sources.

Conflict of Interest

The author declares no conflict of interest.