HARKing: How Badly Can Cherry Picking and Question Trolling Produce Bias in Published Results?

Kevin R. Murphy
University of Limerick

Herman Aguinis
George Washington University

In Press
Journal of Business and Psychology

HARKing: How Badly Can Cherry Picking and Question Trolling Produce Bias in Published Results?

Abstract

The practice of hypothesizing after results are known (HARKing) has been identified as a potential threat to the credibility of research results. We conducted simulations using input values based on comprehensive meta-analyses and reviews in applied psychology and management (e.g., strategic management studies) to determine the extent to which two forms of HARKing behaviors might plausibly bias study outcomes and to examine the determinants of the size of this effect. When HARKing involves *cherry-picking*, which consists of searching through data involving alternative measures or samples to find the results that offer the strongest possible support for a particular hypothesis or research question, HARKing has only a small effect on estimates of the population effect size. When HARKing involves *question trolling*, which consists of searching through data involving several different constructs, measures of those constructs, interventions, or relationships to find seemingly notable results worth writing about, HARKing produces substantial upward bias particularly when it is prevalent and there are many effects from which to choose. Results identify the precise circumstances under which different forms of HARKing behaviors are more or less likely to have a substantial impact on a study's substantive conclusions and the field's cumulative knowledge. We offer suggestions for authors, consumers of research, and reviewers and editors on how to understand, minimize, detect, and deter detrimental forms of HARKing in future research.

HARKing: How Badly Can Cherry Picking and Question Trolling Produce Bias in Published Results?

The scientific method usually follows a hypothetico-deductive model, in which a researcher uses theory, existing research, past experience, or even conjecture to formulate hypotheses that are then tested. As Kerr (1998) noted, researchers sometimes follow a different path, looking at data and results first, generating a post-hoc hypothesis that fits these results, and using this same set of results to "test" those post-hoc hypotheses, a phenomenon he labelled hypothesizing after the results are known, or HARKing. In recent years, a number of studies have examined and commented on the incidence, probable causes, and implications for research and theory development of HARKing in management, applied psychology, and related fields (Aguinis, Cascio, & Ramani, 2017; Aguinis, Ramani, & Alabduljader, in press; Banks, O'Boyle et al., 2016; Banks, Rogelberg, Woznyj, Landis & Rupp, 2016; Bedeian, Taylor & Miller, 2010; Bettis, Ethiraj, Gambardella, Helfat, & Mitchell, 2016; Bosco, Aguinis, Field, Pierce, & Dalton, 2016; Fanelli, 2009; Grand et al., in press; Hitchcock & Sober, 2004; Hollenbeck & Wright, 2017; Loewenstein & Prelec, 2012; Lipton, 2005; Leung, 2011; O'Boyle, Banks & Gonzalez-Mulé, 2017; Shaw, 2017;White, 2003; Wright, 2016). The goal of our study is to understand the extent to which two forms of HARKing that appear to be particularly prevalent and troublesome produce bias in results and conclusions and the field's cumulative knowledge and to understand the determinants and boundary conditions of this bias.

HARKing is not merely a hypothetical concern. Over 90% of respondents in Bedeian et al.'s (2010) survey of research practices in management indicated they had knowledge of faculty members who had developed hypotheses after results were known. Results from a number of studies suggest that at least 30% of researchers admit to engaging in HARKing (Fanelli, 2009;

John et al., 2012). John et al. (2012) conducted a survey involving 2,155 academic psychologists regarding nine questionable research practices including "reporting an unexpected finding as having been predicted from the start." John et al. (2012) asked whether: (1) they had engaged in those practices (self-admission rate), (2) the percentage of other psychologists who had engaged in those practices (prevalence estimate), and (3) among those psychologists who had, the percentage that would admit to having done so (admission estimate). For this type of HARKing, the self-admission rate was about 30%, but the estimated prevalence rate was about 50%, and the admission estimate was about 90%. Finally, O'Boyle et al. (2016) presented particularly striking evidence of the frequency of HARKing, showing that dissertation hypotheses do not match their later published versions in as many as 70% of cases.

Because HARKing starts with results and then works backwards to develop a hypothesis, it is likely that the conclusions reached in that study will not be representative. That is, a researcher who scans the data to find a result that might become the genesis for an interesting hypothesis is unlikely to focus on trivial effects or null findings, and is more likely to seize upon results that appear potentially important. This process of selecting noteworthy (i.e., statistically significant, large) results as a basis for generating and purporting to test hypotheses should lead to a bias toward statistically significant as well as large rather than statistically non-significant and small effects. If HARKing is indeed at least somewhat common, the biases HARKing introduces to any particular study could lead to systematic biases in the cumulative knowledge of entire research literatures.

While concerns about HARKing are widespread, and there is growing empirical evidence regarding the frequency of HARKing, remarkably little is known about *how much difference* HARKing is likely to make, or the circumstances under which HARKing is likely to be a serious

systematic problem or simply an isolated nuisance. That is, we know that HARKing is an *ethical* concern (Honig, Lampel, Siegel & Drnevich, 2014; Leung, 2011; Wasserman, 2013), but it is not always clear whether, or under what conditions, different forms of HARKing are a *substantive* concern in terms of a study's results and conclusions and the field's cumulative knowledge.

Different Forms of HARKing

There are many forms of author misconduct that might lead to the publication of unrepresentative results, ranging from *p* fishing (i.e., trying multiple statistical tests to find one in which $p < .05$) to outright fabrication (Banks, Rogelberg, et al, 2016; Bettis, et al., 2016; Neuroskeptic, 2012). HARKing is often listed as an example of a questionable research practice. But, there are several forms of HARKing behavior, some of which might pose more significant concerns than others.

Table 1 summarizes a taxonomy including four types of HARKing. First, there is *hypothesis proliferation*, which occurs when authors add hypotheses to their study after the results have come in. This particular form of HARKing adds more luster to a result that was not part of the original conceptual design of a study, but it does not necessarily introduce bias into the research literature as long as the results in question are ones that would probably have been reported as part of the descriptive statistics customarily included in research reports (e.g., table including correlations between all study variables). Thus, this type of HARKing may not do much to distort the cumulative body of research in a field.

[Insert Table 1 about here]

Second, authors might engage in *THARKing* (transparently HARKing; Hollenbeck & Wright, 2017). For example, authors may describe in an article's discussion section that particular hypotheses were based on the data they collected for their study. Hollenbeck and

Wright (2017) argued that THARKing is not only ethical, but also likely to be beneficial, particularly if the hypotheses thus developed can subsequently be tested with an independent study.

Third, authors may engage in *cherry-picking*, which consists of searching through data involving alternative measures or samples to find the results that offer the strongest possible support for a particular hypothesis or research question a study was designed to investigate. The practice of data snooping in search of statistically significant or noteworthy results has been documented to be pervasive in many fields including finance (e.g., Lo & MacKinlay, 1990), applied psychology (Wing, 1982), international business (Aguinis et al., 2017), strategic management studies (e.g., Bergh, Sharp, Aguinis, & Li, 2017; Bettis et al., 2016), and other psychology subfields (Wilkinson, L., & Task Force on Statistical Inference, 1999). Cherry-picking the best possible outcome out of a set of multiple results all aimed at the same research question of hypothesis certainly presents ethical concerns, but it is not clear how much bias it might introduce.

Finally, HARKing can take the form we label *question trolling*, which consists of searching through data involving several different constructs, measures of those constructs, interventions, or relationships to find seemingly notable results worth writing about. The question trolling form of HARKing distorts the research process in two ways. First, rather than starting with a research question and proceeding to collect relevant data, researchers who engage in this form of HARKing allows the data to tell them which question is most likely to lead to a noteworthy result, thus bypassing the most important step in the research process—the conceptual development of a question that is worth attempting to answer. This type of HARKing

will also introduce bias into the cumulative scientific literature, potentially a good deal more bias compared to cherry picking.

Cherry picking can be thought of as a special case of question trolling. Cherry picking involves choosing the most favorable result out of a number of sample statistics that are all designed to estimate the same population parameter, whereas question trolling involves choosing the most favorable result out of a number of sample statistics that estimate several different population parameters (i.e., different relations among different sets of variables). In this sense, cherry picking involves selectively choosing from a set of values in which there is variability between studies all aimed at estimating the same underlying population value. On the other hand, question trolling involves selectively choosing among a set of values in which there is variability due to both differences in estimates of a population parameter in studies that all focus on the same question as well as differences in values of the population parameters that underlie the different questions, different variables, or different relationships.

Hypothesis proliferation and THARKing are not likely to introduce serious biases into the scientific literature, and hypothesis proliferation qualifies only as a questionable research practice due to lack of transparency. However, cherry picking and question trolling have the potential to be serious sources of bias due to a process of systematically and proactively selecting the largest possible effect size estimates from several available. In our study, we use simulations to estimate the amount of bias and distortion these two latter HARKing behaviors can reasonably be expected to create across a range of realistic situations.

Simulation is a particularly appropriate method for tackling this problem because HARKing can be difficult to identify and isolate in the field. Simulations can help researchers identify situations in which cherry picking and question trolling are or are not likely to seriously

distort research findings, and can help illustrate both the practical effects of these HARKing behaviors and the conditions under which these forms of HARKing are most likely to lead to systematic biases.

We designed a simulation study using input values derived from comprehensive meta-analyses to determine the extent to which cherry picking and question trolling could plausibly produce a meaningful bias in the conclusions reached in studies (i.e., estimated effect sizes) and our cumulative knowledge, and to investigate the determinants of this bias. Examining these issues will help researchers, consumers of research, and editors and reviewers understand the potential negative consequences of these forms of HARKing and determine the best practices and policies to prevent, deter, or at least minimize detrimental forms of HARKing in the future. More generally, an analysis of the factors that determine the amount of bias cherry picking and question trolling might reasonably introduce into the scientific literature may help in understanding the types of studies and research literatures that are particularly vulnerable to these forms of HARKing.

Determinants of Bias Produced by Cherry Picking and Question Trolling

Because both cherry picking and question trolling systematic capitalize on what are sometimes simple chance fluctuations in the effect size estimates produced by different samples or measures (Aguinis et al., 2017), some of the effects of these HARKing behaviors are likely to be systematically related to features of the studies themselves. In particular, we expect that the bias produces by cherry picking and question trolling will depend on four factors: (1) sample size, (2) the size of the pool of sample results the researcher has to choose from, (3) heterogeneity of the population effects that underlie that pool of sample statistics, and (4) the prevalence of the forms of HARKing behavior studied here.

Sample Size

The smaller the sample, the more variability one expects in sample statistics, and therefore, the larger the opportunity for a biased sampling procedure to lead to results that deviate sharply from the population effect. We expect the degree of overestimation of population effects to be strongly related to *N* and, in particular, to be a linear function of the reciprocal of the square root of *N* (i.e., $\frac{1}{\sqrt{N}}$). Our rationale is that the standard errors of most statistics, such as correlation and regression coefficients, that are used to test study hypotheses are a function of the square root of *N*. For example, our study examines the extent to which cherry picking and question trolling bias estimates of the population correlation between some pair of variables, X and Y. The standard error of the Fisher's z transformation of the correlation coefficient (in the range of values studied here, *r* is essentially identical to z in value) is $\frac{1}{\sqrt{N-3}}$. Similarly, if a study tested the hypothesis that the means in two samples (with sample variances of s_1^2 and s_2^2, respectively), the standard error of the difference between the two means would be $\sqrt{\frac{s_1^2}{N} + \frac{s_2^2}{N}}$. Many other statistics follow this same general form, in which the standard error is a function of the square root of *N*. As *N* decreases and the standard error increases, the likelihood that the most favorable study result (i.e., statistically significant and large) will deviate sharply from the population effect it estimates increases, meaning that there should be systematically more bias when *N* is small.

Size of the Pool of Sample Results from which to Choose

All other things being equal, a researcher who scans a set of 10 sample statistics before selecting one as the basis for his or her *post hoc* hypothesis will have more opportunities to seriously overestimate population effects than another researcher who scans a set of just three

sample statistics. We use the term "pool of results" to refer to the total number of statistical results available to a researcher from which to choose. We do not have a firm basis for predicting a specific functional form, but we do expect that cherry picking and question trolling effects will be a monotonic function of the size of the pool of results scanned by a researcher looking to generate a results-based hypothesis.

Heterogeneity of Population Effects

Cherry picking involves selectively reporting the most favorable results from different samples or measures that are all designed to address the same research question, which implies that the population effect all of these different sample results are designed to estimate is the same. That is, cherry picking involves *homogeneous effects* (i.e., one in which there is a single population parameter underlying all sample results, and in which the major source of variation in the results the researcher scans is sampling error). For example, suppose a researcher is interested in the correlation between individual self-esteem and commitment to the organization, and has three self-esteem measures and three commitment measures. This combination of measures would produce nine separate estimates of the same population correlation between self-esteem and commitment. If the researcher chose to discuss only the one pair of measures that led to the largest observed correlation, that choice would create an upward bias in estimating the population correlation between these two constructs.

Question trolling involves a more complex scenario in which a researcher scans findings from a dataset that covers a diverse set of constructs, measures, constructs, interventions, or relationships. Then, the researcher searches for the most favorable result and uses it to create a *post hoc* hypothesis. For example, a researcher working with survey data or an archival data set might have access to many variables, and if he or she scans the results of many different analyses

to choose the variables that seem most strongly related or the interventions that seem to have the largest effects to be the focus of his or her study, this choice will create systematic upward biases in estimating the relevant population parameters. In contrast to cherry picking, question trolling involves *heterogeneous effects* because even the choice of what to study is driven by sample results and in which the underlying population effects are heterogeneous (i.e., if the data set includes three variables, X, Y and Z, it is likely that ρ_{xy}, ρ_{xz}, and ρ_{yz} are not all identical values). The greater the heterogeneity of population effects underlying a set of sample results, the greater the opportunity to find some sample result that deviates far from (i.e., is much larger than) the average of the population effects (i.e., expected value).

Prevalence of Cherry Picking and Question Trolling

As noted earlier, several studies have presented evidence that some forms of HARKing are common (Bosco et al., 2016; Fanelli, 2009; John et al., 2012; O'Boyle et al., 2016) and estimates have typically ranged from 30%-70%. However, these studies have not been sufficiently specific in identifying the frequency of particular forms of HARKing, and it is unlikely that the precise prevalence of either cherry picking or question trolling in particular research literatures can be established with accuracy.

In simulating the possible effects of specific HARKing behaviors, it is useful to keep two facts in mind. First, unreported HARKing of any sort is usually regarded as counter-normative behavior (Hollenbeck & Wright, 2017), and therefore it is unlikely that any of the variations of HARKing are universal. Second, because most forms of HARKing are not approved of, the prevalence of specific type of HARKing in any particular situation may be difficult to determine with any precision. Simulations allow us to examine the potential biases that cherry picking and question trolling might plausibly create under a range of estimates of their prevalence.

Is a Simulation Really Needed Here?

It should be intuitively obvious to many readers that the four factors cited here, *N*, the size of the pool of results from which to choose, the heterogeneity of population effects, and the prevalence of specific forms of HARKing must all be related to the biases produced by these behaviors, which might lead to the question of why a simulation study is necessary at all. The answer is clear: No simulation is needed to show that these factors will influence the bias introduced by these HARKing behaviors. A simulation study is valuable, however, in addressing the impact of *how much bias* these HARKing behaviors are likely to introduce.

Suppose, for example, that the biases introduced by these behaviors were trivially small, regardless of sample size, prevalence, and other factors. This would tell us that while biases introduced by these HARKing behaviors could be a theoretical risk, they are not likely to have a meaningful effect on a field's cumulative knowledge. Simulations can also tell us whether these factors are likely to interact. If the effects of these four factors are largely independent, it will probably be easier to identify the situations in which these HARKing behaviors might to be a serious problem than if the effects of some parameters (e.g., sample size) depend on the values of other parameters (e.g., prevalence).

In sum, there is little doubt that each of the factors studied here will introduce some biases, but there is considerable uncertainty over the seriousness of these biases or over the range of circumstances under which different forms of HARKing behaviors will have minor or potentially substantial effects. The simulation study described below will help address these questions.

Method

We used the Monte Carlo method to evaluate the effects of the four parameters described

above on the biases created by cherry picking and question trolling, using **R** to create a set of simulations. The Monte Carlo method allowed us to obtain robust estimates of the bias that can be reasonably expected if various HARKing behaviors occur, as well as assessments of the conditions under which these biases are likely to be larger or smaller. Of course, the results of simulation studies depend substantially on the parameters and parameter values included in the simulation. Accordingly, we surveyed the relevant literature to choose values for all simulation parameters (e.g., *N*) that are consistent with those that might reasonably be encountered in the management and applied psychology literatures.

Simulation Scenarios

In all of our scenarios, a researcher starts out with a set of *k* sample correlations between some X and some Y variable, sampled from a population with known characteristics, and in which the researcher consistently chooses the strongest (i.e., most favorable) sample result as the basis for his or her "hypothesis," then uses support for this hypothesis as a basis for publishing or distributing this finding.

Recall that cherry picking involves scanning a set of *k* study results that estimate the same relation between two specific constructs, such as might be found if a researcher decided to study the relation between job satisfaction and job performance, but he or she has several different samples or measures, and using the strongest observed result as the sample-based estimate of the relationship between these two constructs in the population. And, recall that question trolling involves heterogeneous population effects, when there are multiple constructs or relationships among constructs in the set of results being examined and some population correlations are stronger than others. For example, suppose a researcher looks at the matrix of intercorrelations among four variables (e.g., general mental abilities, job satisfaction, job

performance, absenteeism) and he or she then chooses the largest of the six unique correlations (i.e., [k(k-1)]/2) as the basis for a post-hoc hypothesis.

Parameter Values

We used reviews of substantial bodies of research to choose realistic values for N, the size of the pool of results from which to choose, the variability of effect sizes that might reasonably be encountered in research across a wide range of domains, and the prevalence of HARKing. These choices enhance the generalizability of our results across fields including management (e.g., strategic management studies), organizational behavior, human resource management, and applied psychology.

***Sample size*.** Shen, Kiger, Davies, Rasch, Simon and Ones (2011) reviewed sample sizes in papers published in *Journal of Applied Psychology* from 1995-2008 (over 1,500 samples from 1,097 papers). They reported a median sample size of 172.5. Several studies had very large samples, meaning that there was a substantial spread between the mean (690) and median and substantial skew in the distribution. However, the spread between the 15th and 50th percentile in the sample size distribution (which would be approximately one standard deviation in a normal distribution) is approximately 100. In the domain of strategic management studies, Ketchen, Boyd, and Bergh (2008) reported that the median N, based on articles published in *Strategic Management Journal*, was 88 in the early 1980s, 142 in the early 1990s, and 234 in the early 2000s. In our simulations, we used sample sizes ranging from 100 to 280, in steps of 20 (i.e., 100, 120, 140…, 280).

Size of pool of effects from which to choose. With some exceptions (e.g., THARKing; Hollenbeck & Wright, 2017), researchers who develop hypotheses after results are known take pains to conceal their behaviors and, as a result, the precise processes that lead to HARKing are

not well understood. In our study, we examined a variety of potential cherry-picking and question-trolling behaviors that ranged from picking the best to two results to form the basis for a hypothesis to choosing the best of ten possible results. It is plausible that researchers who have access to large archival databases have many more than ten possible results to choose from (e.g., Ketchen, Ireland, & Baker, 2013), but many studies produce only a limited number of statistics that could form the basis for a post-hoc hypothesis.

Heterogeneity of effect sizes. Bosco, Aguinis, Singh, Field and Pierce (2015) reviewed 30 years of research reported in *Journal of Applied Psychology* and *Personnel Psychology*, and assembled a database of over 147,000 correlations. They used these results to establish baselines for the effect sizes most typical in applied psychology. They reported an average uncorrected correlation of .16 (absolute value), with the 33rd and 67th percentile values of .09 and .26, respectively. These values are similar to those reported by Aguinis, Dalton, Bosco, Pierce, and Dalton (2011) in their review of 5,581 effect sizes included in 196 meta-analyses published in *Academy of Management Journal*, *Journal of Applied Psychology*, *Journal of Management*, *Personnel Psychology*, and *Strategic Management Journal* from January 1982 through August 2009.

In our simulations, we set the population correlation $\rho = .20$. We chose a population value of .20 rather than the exact meta-analytic mean (e.g. $r = .16$ in Bosco et al., 2015) to make it easier for readers to quickly and simply evaluate the amount of bias introduced by HARKing. As we note in a later section, the biases introduced by HARKing are essentially independent of the mean of the effect size distribution, meaning that results based on $\rho = .20$ can be readily generalized to $\rho = .16$. O'Boyle et al. (2015) analyzed over 140,000 correlations from studies published in leading journals in applied psychology and management. The standard deviation of

these correlations was approximately .15. In our simulations, we used this figure as an upper bound, as it would represent the heterogeneity expected if researchers sampled haphazardly from the entire literature in their field, and estimated the effects of question trolling in populations of effect sized with a mean of .20 and standard deviations of .05, .10 and .15.

Prevalence of HARKing. As noted earlier, estimates of the prevalence of HARKing behaviors in the published literature have typically ranged from 30%-70%, depending on the particular situations and the particular definitions of HARKing used (Bosco et al.., 2016; Fanelli, 2009; John et al., 2012; O'Boyle et al., 2016). In our study, we estimated the biases that would be expected if 20%, 40%, 60% or 80% of the authors of studies in a cumulative literature engaged in either cherry picking or question trolling.

Simulation Design

We conducted 1,000 replications of each of the simulation conditions, and report results averaging across those 1,000 replications. We started each replication by generating a 10 x 9 matrix of effect size estimates (Fisher z' transformation of correlation coefficients). Each value in this matrix represented transformed correlation coefficients sampled from a normally distributed population with: (1) a mean that corresponds to the constant z' transformed ρ value (homogeneous case) or with transformed ρ values that vary (heterogeneous case) with the same mean as the constant transformed ρ value in the homogeneous case, and (2) a standard deviation that depends on the sample size ($\frac{1}{\sqrt{N-3}}$).

To create each row of this matrix, we generated a set of ten correlations, each drawn from the population described above given a particular *N*. We then populated each column of this matrix by: (1) randomly sampling 2, 3, …10 values from this group of 10 correlations to create the set of values a researcher would examine to find a favorable result that could start the

HARKing process, and (2) selecting the largest value from this set as the basis for his or her *post hoc* hypothesis. Thus, the first row in this 10 X 9 matrix consisted of Fisher z' transformations of correlations drawn from a population with a mean of ρ = .20 and a standard deviation of .1015 (i.e. $\frac{1}{\sqrt{N-3}}$), where N = 100 and each column represented the largest value from a set of k values (where k = 2, 3…, 10) randomly sampled from this initial group of 10 correlations. Before analyzing trends in effect size estimates, Fisher z' values were transformed back into correlations, and the results shows in the tables and figures are in a correlation coefficient metric.

To estimate the effects of different levels of prevalence of cherry picking and question trolling, we created a mixture model in which the expected value of the effect size in each cell represented a weighted linear combination of the expected value that would be obtained from combinations of two populations, one of which involved no HARKing (in this population, ρ = .20) and one of all studies involved HARKing. By varying the weights applied to these two population means, we were able to estimate the biases expected in a cumulative literature if there are different levels of prevalence of cherry picking and question trolling.

Dependent Variable

The dependent variable is the expected value of the correlation coefficient over 1,000 replications for each combination of sample size, number of effect sizes from which to choose, (in the heterogeneous case) SD of the distribution of ρ values (i.e., SD_ρ), and prevalence levels. If there is no HARKing, each of these correlations would be approximately .20 (the population value in the homogeneous case and the average of the population values in the heterogeneous case).

R code for the cherry picking and question trolling simulations is included in the Appendix. This code makes explicit a notion discussed earlier, that cherry picking represents a

special case of question trolling in which there is no variability in the population parameter being estimated by the different sample estimates included in the sets of values the researcher chooses from (i.e., $SD_\rho = 0$).

Results

Before discussing specific findings, it is useful to note that except at the extremes of the distribution (i.e., absolute values of ρ near 0 or 1.0), the biases produced by either cherry picking or question trolling are independent of the size of ρ. The independence of bias from the mean effect size can be confirmed by running the code included in the Appendix with population effect size estimates that vary. That is, in the range of effect sizes typically seen in management and applied psychology (See Bosco et al., 2015 for a detailed discussion of the distribution of effect size estimates in these literatures), the difference between the HARKed estimate of ρ and the actual value of ρ will remain essentially constant in any given scenario, regardless of whether the population correlation one is trying to estimate is relatively large or small. Comparing the values obtained from our simulations to the value expected in the absence of HARKing allows readers to estimate both the absolute size of the bias introduced by HARKing (i.e., value from simulation minus .20) and the relative size of this bias (i.e., value from simulation divided by .20)

Cherry Picking

Table 2 presents results emerging out of studies that engaged in cherry picking. On the whole, most of the values presented in Table 2 are relatively close to .20, the value that would be expected if there was no HARKing. The only cases in which the difference between the HARKed mean and the population value were greater than .10 were those in which the incidence of cherry picking was very high (80%), the sample size was small, and the pool sizes also large (i.e., the authors chose the highest of 8 to 10 correlations to report). These results suggest that

cherry-picking the most favorable outcome from a set of sample statistics that are all designed to estimate the same population parameter is likely to substantially bias the research literature only when this type of HARKing is very prevalent and there are many effects from which to choose. Clearly, cherry-picking results might be a problem in an individual study and it might present ethical problems regardless of its practical effects, but our results suggest that this type of cherry-picking of results is unlikely to lead to a substantial bias in the cumulative conclusions reached in a body of research.

[Insert Table 2 about here]

We expected each of the parameters varied in this simulation to have monotonic effects on bias, and results supported this expectation. Bias is larger when *N* is small, when the pool of effect sizes is large, and when the prevalence of cherry picking is high. Furthermore, the effects of these three variables are largely independent. We estimated all possible interaction effects, but found only one small two-way interaction in our analysis of bias as a function of *N*, pool size and prevalence; the effect of pool size gets slightly larger as the prevalence of cherry picking goes up (η^2 =.05).

Question Trolling

Table 2 presents results regarding question trolling, which occurs when researchers scan results from a diverse set of constructs and relations between constructs and then choose the strongest sample result as the basis for a *post hoc* hypothesis. From a mathematical perspective (and as is implicit in the Appendix), cherry picking can be thought of as a special case of question trolling, in which there is a single population parameter underlying all sample results, which means that the standard deviation of the distribution of rho is zero.

Results in Table 2 show that question trolling can lead to substantial overestimates of the

population effect, but in almost three quarters of the question trolling scenarios shown in this table, this questionable research practice leads to bias that is less than .10 correlation units. However, if question trolling is highly prevalent (i.e., 60 to 80%), the pool of effects from which to choose is large (i.e., 6 to 10 sample results), and the set of population parameters that underlie the data examined to generate a *post hoc* hypothesis is quite heterogeneous, HARKed estimates can exceed .30 and even .40, which is a considerable difference in relation to the value of .20 expected in the absence of HARKing. On the other hand, if the prevalence of question trolling is lower than 60%, it will almost always have a fairly small effect.

As in the case of cherry picking, we found only two nontrivial two-way interactions among the full set of ten possible interactions involving the four variables manipulated in this simulation. First, the effect of pool size is slightly larger as prevalence goes up (η^2 =.04). Second, the effect of heterogeneity is slightly larger as prevalence goes up (η^2 =.03).

Discussion

Bias is always a concern when evaluating research, and even if its effects are small, many types of HARKing can have adverse effects on the scientific enterprise by undermining our confidence in the credibility and trustworthiness of scientific reports (e.g., Bosco et al., 2016). That being said, our results suggest that cherry picking and question trolling are likely to have a substantial biasing effect on the conclusions of the cumulative research literature in certain specific situations only.

First, as Table 2 suggests, if a researcher chooses the most favorable out of several alternative ways of measuring or studying the same relationship (cherry picking), HARKing is likely to have a small effect, especially in *N* is reasonably large. Even if cherry picking is quite widespread (e.g., incidence of 80%), the difference between the value expected with no

HARKing and a HARKed estimate does not exceed .10 correlation units for the pool of effects sizes studied here if *N* is much larger than 200. Cherry picking cannot be dismissed as a problem, but our results do suggest that to the extent cherry picking mirrors the behaviors simulated here (i.e., reviewing a set of *k* different results of tests designed to address the same research question and selecting the most favorable one for discussion), cherry picking is unlikely to be a major concern in our research literature.

Second, a more worrisome case is the form of HARKing involving question trolling because the very definition of the research question itself is retroactively engineered based on a set of sample statistics. At least in the homogeneous case that defines cherry picking, the researcher knows what the research question is before starting to develop some sort of *post hoc* hypothesis. In the heterogeneous case, the research question is up for grabs, and bias will no longer be a simple function of sampling error. As Table 2 shows, bias due to question trolling can be substantial when the underlying ρ values vary substantially. Even where the sample size is large based on typical *N* values reported in the applied psychology and management literatures, bias can be substantial if the prevalence of question trolling is high or if the heterogeneity of the population of effect sizes researchers sample from is large. However, Table 2 also gives some room for optimism. For example, if the prevalence of question trolling is 40% or lower, question trolling rarely leads to inflation of effect size estimates as large as .10. In virtually all cases where the population of effect sizes sampled from is less heterogeneous than the entire field of applied psychology and management (i.e., SD_ρ across these entire fields is approximately .15; O'Boyle et al., 2015), bias was .10 correlation units or less.

Third, we emphasize that there are some cases where the bias introduced by question trolling is quite large and important. For example, a researcher who scans a large dataset in

search of a question to answer might deal with sets of relationships considerably larger than the ones examined here, and the potential for substantial inflation is considerable. In general, the results presented in Table 3 suggest that question trolling has the potential to be a significant concern, except in cases where samples are quite large and trolling in minimized. The effects of question trolling are a function of two variables whose values are difficult to estimate: The actual prevalence of question trolling and its degree of brazenness. Question trollers who sample from a large and heterogeneous set of results, forming an *a priori* hypothesis around whatever result seems especially striking, have the potential to do substantial harm to the research enterprise.

Finally, we noted earlier that the degree of bias produced by cherry picking or question trolling is essentially independent of the size of the population parameter being estimated. This implies that HARKing might be a larger problem in areas where population effects are likely to be small than in areas where population effects are likely to be large. For example, if the $\rho = .20$, our results suggest a range of scenarios in which cherry picking or question trolling could lead to estimates in the .30s (i.e., artificial inflation of 50%), and this bias is proportionately more serious that a scenario in which $\rho = .50$ and HARKed estimates are in the .60s (i.e., artificial inflation of 20%).

Detrimental Effects of HARKing on Cumulative Knowledge

So far, we have described our results in terms of differences between true and HARKed effect size estimates. In other words, the way we described the size of the HARKing effect has been in terms of correlation coefficient unit differences. As has been documented based on a database of about 150,000 correlations reported in *Journal of Applied Psychology* and *Personnel Psychology* over the past 30 years, the average uncorrected correlation is .16 (Bosco et al., 2015). This large-scale effort aimed at documenting the typical size of effects has shown that Cohen's

(1988) often-invoked traditional benchmarks of correlations of .10 being "small" and .30 "medium" were clearly overestimates and are not applicable to the applied psychology and management literatures.

The reality that a large number of correlations are in the .10s means that a HARKing effect of about .10 correlation units could essentially double the observed effect. In other words, there is a potential for 100% inflation in the observed effect when the population effect is itself small. So, although our results show that in many situations the HARKing effect is "small" in terms of correlation unit differences (i.e., in the .10s), this size of HARKing bias can be sufficiently large to artificially inflate effects and affect cumulative knowledge in many fields and areas of study. For example, Bosco et al. (2015) reported that the median correlation between job performance and personal characteristics is .09, the median correlation between movement behavior (e.g., turnover, job search behaviors) and psychological characteristics is .11, and the median correlation between job performance and organization attitudes is .16. In these particular research domains, an effect of question trolling of .10 correlation units can have a very sizable and arguably detrimental effect on our understanding of relations between constructs.

As an illustration in the domain of strategic management studies, consider the domain of top management teams (TMT). Meta-analytically derived correlations between firm performance and the variables most frequently studied in TMT research such CEO tenure, TMT tenure, TMT size, TMT diversity, board size, board independence, and board leadership structure firm performance are also in the .10s or even smaller (Bergh et al., 2016). Accordingly, our results show that HARKing can also have a very important and noticeable biasing effect in this research domain.

In short, considered in absolute correlation coefficient units, the effect of HARKing on the estimated size of effects seems small. However, the implications of our results regarding the damaging effects of HARKing for theory as well as practice necessitates that we place results within specific fields and research domains because in many cases "small" effects make a big difference (Aguinis, Werner, Abbott, Angert, Park, & Kohlhausen, 2010; Cortina & Landis, 2009).

Implications for Researchers, Research Consumers, and Editors and Reviewers: Putting HARKing in Context

HARKing is often discussed as a form of author misconduct and, sometimes, this label is entirely appropriate. However, HARKing is not solely a function of author misconduct. One reason authors HARK is in reaction to reviewers' negative reactions to non-supported hypotheses (Edwards & Berry, 2010; Hubbard & Armstrong, 1997; Orlitzky, 2012; Pfeffer, 2007). In fact, manuscript reviewers are the ones who often suggest that hypotheses be added *post hoc* during the peer review process (Bedeian et al., 2010). For example, Bosco et al. (2016) conducted a survey of authors who had published in *Journal of Applied Psychology* and *Personnel Psychology* and found that 21% reported that at least one hypothesis change had occurred as a result of the review process. Although reviewer suggestions about the post hoc inclusion of hypotheses may be motivated by authors' implicit reference to them, this phenomenon is also likely attributable to the "theory fetish" in organizational research (Hambrick, 2007, p. 1346).

We should also recognize that using data to develop hypothesis is not necessarily a bad thing. Both inductive and abductive research methods can lead to important discoveries and be helpful in developing models and theories (Aguinis & Vandenberg, 2014; Bamberger & Ang,

2016; Fisher & Aguinis, 2017; Locke, 2007). While inductive research is generally known by applied psychology and management researchers, abductive reasoning is likely less familiar. In a nutshell, it is a form of reasoning that attempts to make sense of puzzling facts, as in the case of medical diagnosis, fault diagnosis, and archaeological reconstruction, by searching for the simplest and most likely explanation (Fisher & Aguinis, 2017). Locke, Golden-Biddle, and Feldman (2008, p. 907) contrasted abduction with other forms of reasoning by pointing out that "deduction proves that something must be; induction shows that something actually is operative; abduction merely suggests that something may be." HARKing differs from inductive and abductive research because it uses the same data to both generate and test hypotheses. For example, in a genuinely inductive approach to research, data can drive the formation of hypotheses, but these same data are typically not used to "test" that hypotheses. In properly conducted inductive research, once hypotheses are generated, follow-up work is conducted to replicate, verify, or confirm findings and learn whether they were produced by chance alone. Absent this level of transparency regarding the process that led to a study's conclusions, it would be reasonable for the consumers of this research to believe that results are more robust that the may be, and that there is not any special need for replication or verifiability, as compared to other studies where results were entirely deductive and theory-driven (Aguinis et al., in press; Bosco et al., 2016).

Our taxonomy of HARKing behaviors and the results presented in this study have a number of conclusions and practical implications. Table 3 summarizes several conclusions about HARKing. First, it is important to recognize that there are different forms of HARKing and they differ regarding their effects. Specifically, they range from actually desirable (THARKing) to highly damaging to a field's cumulative knowledge and the credibility of science (e.g., some

instances of question trolling). Second, our simulation results showed that cherry picking's effect was overall small, but the impact of question trolling can be quite substantial. For example, question trolling is most damaging when there is greater variability in the population parameters underlying the results. Third, there are several actions that can be taken to minimize the detrimental effects of HARKing. For example, large sample sizes reduce the bias produced by cherry picking.

[Insert Table 3 about here]

Clearly, HARKing is the result of complex processes involving the current reward and incentive system which motivates researchers to publish in top journals (Aguinis et al., in press; Aguinis, Shapiro, Antonacopoulou, & Cummings, 2014). Describing statistically significant effects that are as large as possible, which is done more easily by HARKing, seems to help in this regard. Also, another factor that leads to HARKing is researchers' and reviewers' lack of awareness regarding HARKing's detrimental effects.

Inadvertent Cherry Picking and Question Trolling: How Multivariate Procedures Produce Comparable Biases

An additional implication of our results is that there are several widely-used statistical procedures that involve the same forms of HARKing we examined, but in these cases the biases introduced by these statistical optimization procedures are not the result of a conscious choice by the researcher to use study results to present the most optimistic picture possible. The use of these procedures is not in and of itself unethical, but the incautious use of statistical methods that maximize the fit of models to sample data or that maximize the predictive power of statistical models can cause precisely the same sorts of problems cherry picking and question trolling cause. The use of statistical maximization is not thought of as unethical, and therefore little detail

often is provided in published articles (Aguinis et al., in press), but its effects could in the end be considerably more insidious than the effects of HARKing studied here. Consider, for example, the use of statistical criteria to drive key decisions about the contents of regression models and structural models or relationships among constructs.

Multiple regression. If there are *p* variables that might be used to predict a criterion variable via multiple regression, it is not unusual to use some type of stepwise selection procedure to identify the subset of variables that yields the largest R^2, given the number of variables in the model. There are many variable selection methods available to researchers, including forward selection algorithms (where variables are added to a regression model until they fail to lead to incremental increases in R^2), backward selection algorithms (where one starts with a model with all *p* variables, then drops variables that make the smallest contribution to the predictive power of the model until one reaches the point where dropping the next variable will lead to a meaningful decrease in R^2), or true stepwise selection procedures in which the decision to include specific variables in a model is re-evaluated each time a new variable is added.

Although these variable-selection methods are widely used (e.g., forward selection is the default method in the SPSS Regression procedure), serious concerns have been raised about the effects of building prediction models on the basis of statistical criteria. For example, these methods: (1) overestimate the values of both R^2 and the regression coefficients, (2) make unpredictable and potentially arbitrary choices of which variables to include and which to exclude from prediction models, (3) capitalize on chance fluctuations in sample data, and (4) produce results that do not reliably replicate (Derksen & Keselman, 1992; Harrell, 2011; Judd & McClelland, 1989).[2]

A prevalent yet not acknowledged instance of this mechanism is the choice of a particular set of control variables from several that are examined in a given study (Bernerth & Aguinis, 2016). In most published studies, there is little information on the particular rationale for the inclusion of a set of control variables and, in many cases, it is likely that several combinations of controls are examined until a particular set leads to the largest possible R^2. But, the trial-and-error process that led to choosing the final set is usually not described in the published article (Bernerth & Aguinis, 2016). So, the selection of control variables is often essentially the same process involved in question trolling.

These procedures combine some of the potentially worrisome features of both cherry picking and question trolling because the choice of variables to study because the choice of variables to include in a model and the weights assigned to those variables are explicitly chosen to maximize the predictive power of the model. It is not uncommon for researchers to use formulas to estimate shrinkage when this sample equation is applied in a broader population, but even when these formulas are used, the distorting effects of selecting variables to study and assigning weights to those variables solely on the basis of maximizing R^2 are potentially worrisome.

Structural equation modeling. When evaluating structural equation models, it is common to use modification indices to make decisions about which links between variables should or should not be included in a model. Many authors have expressed serious reservations about using modification indices to guide important decisions about structural models (e.g., Bollen, 1989; Hayduk, 1987; Kline, 2005; Sörbom, 1989). In particular, if these indices are used at all, it is usually recommended that they be relied on only when the changes they suggest are: (1) theoretically sensible, (2) minor, and (3) few in number.

As with stepwise selection procedures in multiple regression, the use of modification indices in altering the structure or meaning of structural models is a worrisome practice. First, it blindly elevates sample statistics to a position that might not be warranted on conceptual grounds. Researchers using SEM may examine dozens of possible relations involving dozens of indicators and their relations with other indicators as well as latent constructs—as well as relations among latent constructs and residual terms (Landis, Edwards, & Cortina, 2009), and reliance on statistical criteria to determine the structure and contents of models has the effect of. driving out logic, theory, past research and even sound scientific judgment in the blind pursuit of improving results in one's current sample.

Data mining. The "big data" movement has given new life to an approach that can be found in the management and applied psychology literatures in one way or another for decades – i.e., data mining, in the form of using exploratory factor analysis of data with the hope of finding new insights into human behavior. Starting in the 1930s, this method became the basis for developing structural models in many domains, notably the structure of cognitive abilities (Jensen, 1980; Thurstone, 1934). As confirmatory factor analysis became more accessible (e.g., via LISREL, EQS, and other off-the-shelf software), factor-analytic studies increasingly shifted from an exploratory to confirmatory mode, in which *a prio*ri theory rather than statistical results in a sample formed the basis for posing questions about underlying structures (Cortina, Aguinis, & DeShon, 2017). The decline of exploratory factor analysis can be thought of as the end of the first wave of fascination with data mining.

In recent years, data mining has come back with a vengeance. One of the fastest growing occupations is "Data Scientist" and much of what data scientists do is search for patterns and regularities in data. It has been claimed that the combination of big data and emerging data

mining techniques represents a revolution in the scientific method (Pigliucci, 2009) that will replace the older method of forming hypotheses prior to analyzing data.

Contemporary data mining techniques, where paired with the very large data sets to which they are designed to be applied, arguably avoids one of the problems highlighted by our results (i.e., the instability of small-sample results). The shortcoming that big data and rigorous data-mining techniques cannot overcome is that the results are necessarily driven by what is and what is not measured. For example, some objective performance measures are relatively easy to collect, but are generally regarded as deficient measures of job performance (Murphy & Cleveland, 1995). A big-data approach to studying performance evaluations would almost certainly depend on these objective measures, despite their questionable adequacy. In general, a reliance on big data and data mining can lead to a constriction of research questions, where only those questions that can be answered by data that are easily collected and assembled into massive datasets are pursued.

Data mining explicitly capitalizes on one of the key principles of both cherry picking and question trolling – i.e., that if a researcher looks at enough sample results, he or she is bound to eventually find *something* that looks interesting. By using very large samples, this method solves one of the problems noted here (i.e., that HARKing effects can be particularly large when *N* is small). But, in the end, its effects on the scientific method may be substantially more pernicious than the damage done by the researcher who occasionally scans several results before arriving at a hypothesis. By driving deduction, scientific judgment, and a consideration of existing research from any consideration of what to study or what it means, we run the risk of magnifying every shortcoming of "dustbowl empiricism" to an unprecedented degree.

Detecting and Deterring HARKing

Our results, combined with a review of research on HARKing, led to a number of suggestions for detecting and deterring HARKing. First, Occam's Razor is an essential tool for detecting HARKing. As Hollenbeck and Wright (2017) noted, HARKed hypotheses often involve convoluted reasoning or counterfactual assumptions. If the conceptual case for a hypothesis seems unlikely to hold up to scrutiny, or does not seem to emerge organically from the literature and theory the author cites, this is one potential indicator of HARKing.

Second, it is useful to have a healthily skeptical attitude; stories that are too good to be true may not be true. In our combined experience, we have supervised hundreds of theses, dissertation and research studies and have carried out a large number of studies ourselves. It is unusual for every prediction and every hypothesis to be supported in a study, and when reviewing a manuscript in which every prediction is supported, skepticism is warranted.

Third, an additional critical point is the need to reduce the temptation to engage in HARKing. This can be done in two ways. First, HARKing is not simply a problem of author misbehavior. It is common for reviewers and editors to encourage authors to drop hypotheses and analyses that do not pan out, and this creates problems that have a good deal in common with HARKing. It is critical for reviewers and editors to realize that their efforts to tidy up otherwise messy research (especially by encouraging authors to drop or modify hypotheses that are not supported) can have the same effect as HARKing. Editors and reviewers who encourage authors to change or drop hypotheses run the risk of distorting the scientific enterprise in ways that present an overly optimistic and neat picture of what is inherently a complex process of discovery. We believe that journal reviewers and particularly journal editors could help to reduce the incidence of HARKing by resisting the urge to tidy up articles by trimming and modifying

hypotheses so the resulting manuscript conveys a neater, more straight forward, more coherent and even "interesting" story (Aguinis et al., in press).

Editors and reviewers can reduce the incentive to HARK by encouraging and rewarding replications. The temptation to artificially inflate one's results, including cherry picking and question trolling, would be smaller is researchers believed that subsequent attempts at replication would quickly expose the unrepresentative nature of the results they published. Replication is a critical component of the standard scientific method, and if authors saw a chance to publish their work in top journals by replicating public studies, they would also know that if they published inflated results, the chance that they this would be detected would be high. So, as noted by Aguinis et al. (in press), solutions to many "research performance problems" such as HARKing need to focus not only on what knowledge and skills need to have to conduct replicable research, but also their motivation to do so.

Summary and Conclusions

Our results suggest that two forms HARKing, cherry picking and question trolling, can lead to biases in estimates of the effects of the variables we study in management, applied psychology, and related fields. The artificial inflation in effect sizes due to cherry picking can be mitigated by using large sample sizes, but the bias due to question trolling cannot be eliminated simply by insisting on large samples when the pool of effects from which to choose is large, which is a common situation particularly today given the availability of archival databases, web scraping, and the big data movement (Landers, Brusso, Cavanaugh, & Collmus, 2016; Tonidandel, King, & Cortina, 2016).

An argument can be made that the effects of cherry picking or question trolling are least worrisome when the population effects being studied are large. That is, a study that reports a

correlation of .65 when .55 might have been a more representative value might not do all that much harm because the core finding is pretty much the same with or without HARKing – a very strong effect. It is clear, however, that this is not a situation we normally deal with in management (e.g., strategic management studies) and applied psychology research. For example, Bosco et al. (2015) and Aguinis et al. (2011) have shown that the great majority of relations are in a range that would best be described as small to modest. For example, the 80th percentile of all of the uncorrected correlations (more than 147,000) examined by Bosco et al. (2015) is .36, and half are in the range of .07 to .32. This is, unfortunately, a range of values where cherry picking and question trolling have the greatest potential to materially distort results.

In closing, we distinguished various forms of HARKing and our results offer information about when and why various forms of HARKing are more or less likely to produce bias in published results. This information is useful for authors as well as editors and reviewers in terms of which forms of HARKing are more or less detrimental and therefore which forms of HARKing to avoid in which specific situations. We also offer suggestions for consumers of research (e.g., skepticism in the presence of all supported hypotheses—especially when sample size is small). Overall, we hope our work will help the field move forward regarding the important conversation of the credibility and trustworthiness of our research results.

References

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.

Aguinis, H., Dalton, D. R., Bosco, F. A., Pierce, C. A., & Dalton, C. M. (2011). Meta-analytic choices and judgment calls: Implications for theory building and testing, obtained effect sizes, and scholarly impact. *Journal of Management*, *37*, 5-38.

Aguinis, H., Ramani, R. S., & Alabduljader, N. (in press). What you see is what you get? Enhancing methodological transparency in management research. *Academy of Management Annals*. doi: 10.5465/annals.2016.0011

Aguinis, H., Shapiro, D. L., Antonacopoulou, E., & Cummings, T. G. (2014). Scholarly impact: A pluralist conceptualization. *Academy of Management Learning and Education*, *13*, 623-639.

Aguinis, H., & Vandenberg, R. J. (2014). An ounce of prevention is worth a pound of cure: Improving research quality before data collection. *Annual Review of Organizational Psychology and Organizational Behavior, 1,* 569-595.

Aguinis, H., Werner, S., Abbott, J. L., Angert, C., Park, J. H., & Kohlhausen, D. (2010). Customer-centric science: Reporting significant research results with rigor, relevance, and practical impact in mind. *Organizational Research Methods*, *13*, 515-539.

Bamberger, P., & Ang, S. (2016). The quantitative discovery: What is it and how to get it published. *Academy of Management Discoveries*, *2*, 1-6.

Banks, G. C., O'Boyle, E. H., 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.

Banks, G.C., Rogelberg, S.G., Woznyj, H.M., Landis, R.S. & Rupp, D.E. (2016). Editorial: Evidence on questionable research practices: The good, the bad and the ugly. *Journal of Business and Psychology, 31,* 323-338.

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.

Bettis, R. A., Ethiraj, S., Gambardella, A., Helfat, C., & Mitchell, W. (2016). Creating repeatable cumulative knowledge in strategic management: A call for a broad and deep conversation among authors, referees, and editors. *Strategic Management Journal*, *37*, 257-261.

Bergh, D. D., Aguinis, H., Heavey, C. Ketchen, D. J., Boyd, B. K., Su, P., Lau, C., & Joo, H. (2016). Using meta-analytic structural equation modeling to advance strategic management research: Guidelines and an empirical illustration via the strategic leadership-performance relationship. *Strategic Management Journal, 37*, 477-497.

Bergh, D. D., Sharp, B. M., Aguinis, H., & Li, M. (2017). Is there a credibility crisis in strategic management research? Evidence on the reproducibility of study findings. *Strategic Organization, 15*, 423-436.

Bernerth, J. & Aguinis, H. (2016). A critical review and best-practice recommendations for control variable usage. *Personnel Psychology, 69,* 229-283.

Bollen, K. A. (1989). *Structural equations with latent variables.* New York: John Wiley.

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

Bosco, F. A., Aguinis, H. Singh, K., Field, J. G., & Pierce, C. A. (2015). Correlational effect size benchmarks. J*ournal of Applied Psychology, 100,* 431-449.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Cortina, J. M., Aguinis, H., & DeShon, R. P. (2017). Twilight of dawn or of evening? A century of research methods in the Journal of Applied Psychology. *Journal of Applied Psychology, 102,* 274-290.

Cortina, J. M., & Landis, R. S. (2009). When small effect sizes tell a big story, and when large effect sizes don't. In C. E. Lance & R. J. Vandenberg (Eds.), *Statistical and methodological myths and urban legends: Doctrine, verity, and fable in the organizational and social sciences* (pp. 287–308). New York, NY: Routledge.

Derksen, S. & Keselman, H.J. (1992). Backward, forward and stepwise automated subset selection algorithms: Frequency of obtaining authentic and noise variables. *British Journal of Mathematical and Statistical Psychology*, 45, 265-282.

Fanelli, D. (2009). How many scientists fabricate and falsify research? A systematic review and meta-analysis of survey data. *PLoS ONE,* 4, e5738.

Fisher, G., & Aguinis, H. (2017). Using theory elaboration to make theoretical advancements. *Organizational Research Methods, 20,* 438-464.

Grand, J. A., Rogelberg, S. G., Allen, T. D., Landis, R. S., Reynolds, D. H., Scott, J. C., Tonidandel, S., & Truxillo, D. M. (in press). A systems-based approach to fostering robust science in industrial-organizational psychology. *Industrial and Organizational Psychology: Perspectives on Science and Practice*.

Harrell, H. (2011). *Regression modeling strategies with applications to linear models, logistic regression and survival analysis*. New York: Springer-Verlag.

Hayduk, L.A. (1987). *Structural equation modeling with LISREL*: *Essentials and advances*. Baltimore: Johns Hopkins University Press.

Hitchcock, C., & Sober, E. (2004). Prediction versus accommodation and the risk of overfitting. *British Journal for the Philosophy of Science, 55*, 1-34.

Hollenbeck, J. H. & Wright, P. M. (2017). Harking, sharking, and tharking: Making the case for post hoc analysis of scientific data. *Journal of Management, 43*, 5-18.

Honig, B., Lampel, J., Siegel, D. & Drnevich, P. (2014). Ethics in the production and dissemination of management research: Institutional failure or individual fallibility. *Journal of Management Studies, 51*, 118-142.

Jensen, A. (1980). *Bias in mental testing*. New York: Free Press.

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.

Judd, C.M., & McClelland, G. H. (1989). *Data analysis: A model comparison approach.* New York: Harcourt

Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality & Social Psychology Review*, *2*, 196.

Ketchen, D. J., Boyd, B. K., & Bergh, D. D. (2008). Research methodology in strategic management past accomplishments and future challenges. *Organizational Research Methods, 11*, 643-658.

Ketchen, D. J., Ireland, R. D., & Baker, L. T. (2013). The use of archival proxies in strategic management studies: Castles made of sand? *Organizational Research Methods, 16*, 32-42.

Kline, R. B. (2005). *Principles and practice of structural equation modeling* (2nd ed.). New York: Guilford Press.

Landers, R. N., Brusso, R. C., Cavanaugh, K. J., & Collmus, A. B. (2016). A primer on theory-driven web scraping: Automatic extraction of big data from the Internet for use in psychological research. *Psychological Methods, 21*, 475-492.

Landis, R. S., Edwards, B. D., & Cortina, J. M. (2009). On the practice of allowing correlated residuals among indicators in structural equation models. In C. E. Lance & R. J. Vandenberg (Eds.), *Statistical and methodological myths and urban legends: Doctrine, verity and fable in the organizational and social sciences* (pp. 193-214). New York: Routledge/Taylor & Francis Group.

Leung, K. (2011). Presenting post hoc hypotheses as a priori: Ethical and theoretical issues. *Management and Organization Review, 7*, 471-479.

Lipton, P. (2005). Testing hypotheses: Prediction and prejudice. *Science, 307*, 219-221.

Lo, A. W., & MacKinlay, A. C. (1990). Data-snooping biases in tests of financial asset pricing models. *Review of Financial Studies, 3*, 431–467.

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

Locke, K., Golden-Biddle, K., & Feldman, M. S. (2008). Perspective-making doubt generative: Rethinking the role of doubt in the research process. *Organization Science, 19*, 907-918.

Murphy, K.R. & Cleveland, J.N. (1995). *Understanding performance appraisal: Social, organizational and goal-oriented perspectives*. Newbury Park, CA: Sage.

Neuroskeptic. (2012). The nine circles of scientific hell. *Perspectives on Psychological Science, 7,* 643–644.

O'Boyle, E. H., Banks, G. C., & Gonzalez-Mulé, E. (2017). The chrysalis effect: How ugly initial results metamorphosize into beautiful articles. *Journal of Management, 43,* NPi.*doi*: 10.1177/0149206314527133.

Pigliucci, M. (2009). The end of theory in science? *EMBO Reports, 10*, 534.

Shaw, J. B. (2017). Advantages of starting with theory. *Academy of Management Journal, 60,* 819-822.

Shaw, J. B., & Riskind, J. H. (1983). Predicting job stress using data from the Position Analysis Questionnaire. *Journal of Applied Psychology, 68*, 253-261.

Shen, W., Kiger, T. B., Davies, S. E., Rasch, R. L., Simon, K. M., & Ones, D. S. (2011). Samples in applied psychology: Over a decade of research in review. *Journal of Applied Psychology, 96,* 1055-1064.

Sörbom, D. (1989). Model modification. *Psychometrika, 54,*371-384.

Srinivasan, V., & Weinstein, A. G. (1973). Effects of curtailment on an admissions model for a graduate management program. *Journal of Applied Psychology*, *58,* 339-346.

Thurstone, L.L. (1934). The vectors of the mind. *American Psychologist, 41*, 1-32.

Tonidandel, S., King, E. B., & Cortina, J. M. (Eds.) (2016). *Big data at work: The data science revolution and organizational psychology*. New York: Routledge.

Wasserman, R. (2013). Ethical issues and guidelines for conducting data analysis in psychological research. *Ethics and Behavior*, *23*, 3-15.

Wilkinson, L., & Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54,* 594-604.

Wing, H. (1982). Statistical hazards in the determination of adverse impact with small samples. *Personnel Psychology, 35,* 153-162.

Wright, P. M. (2016). Ensuring research integrity: An editor's perspective. J*ournal of Management*, *42,* 1037-1043.

Footnotes

[1] When ρ is very large, ceiling effects can limit the biases produced by HARKing. When ρ is equal to or very near to zero, bias is limited because the largest effect is equally likely to be negative as it is to be positive. In addition, when ρ = 0, HARKing will produce a distribution of sample effects whose mean is not changed, but whose standard deviation is inflated.

[2] Although this method is rarely encountered in the research literature, several software packages (e.g., NCSS, JMP) include an even more aggressive option – i.e., one that evaluates all possible regression models, starting with models that include 2 variables and examining every possible combination of predictors until the full p-variable model is tested.

[3] Although this method is rarely encountered in the research literature, several software packages (e.g., NCSS, JMP) include an even more aggressive option – i.e., one that evaluates all possible regression models, starting with models that include 2 variables and examining every possible combination of predictors until the full *p*-variable model is tested.

Table 1

A Taxonomy of HARKing Behaviors

Less Problematic – Little Potential to Bias Cumulative Knowledge

1. *Hypothesis Proliferation:* An author adds hypotheses to a study after data are collected and analyzed to place added emphasis on a result that was not part of the original conceptual design but was nevertheless going to be reported in the manuscript (e.g., correlation table).
2. *THARKing*: an author is likely to transparently HARK in the discussion section of a paper by forming new hypotheses on the basis of results obtained (Hollenbeck & Wright (2017).

More Problematic – Great Potential to Bias Cumulative Knowledge

3. *Cherry-Picking:* An author searches through data involving alternative measures or samples to find the results that offer the strongest possible support for a particular hypothesis or research question.
4. *Question Trolling*: An author searches through data involving several different constructs, measures of those constructs, interventions, or relationships to find seemingly notable results worth writing about.

Table 2

HARKed Estimates of the Population Correlation

Cherry Picking

Prevalence	Pool size	2	4	6	8	10
	N					
20%						
	100	.212	.219	.223	.226	.229
	140	.209	.217	.220	.223	.225
	180	.208	.215	.218	.220	.222
	220	.207	.213	.216	.218	.219
	260	.207	.212	.215	.217	.218
	280	.206	.212	.214	.216	.217
40%						
	100	.223	.239	.246	.252	.258
	140	.218	.233	.241	.246	.249
	180	.216	.229	.235	.240	.244
	220	.215	.227	.232	.236	.239
	260	.213	.224	.229	.233	.237
	280	.212	.224	.229	.233	.234
60%						
	100	.235	.258	.269	.278	.287
	140	.226	.250	.261	.269	.274
	180	.224	.244	.253	.260	.265
	220	.222	.240	.248	.254	.258
	260	.220	.237	.244	.250	.255
	280	.218	.235	.243	.249	.252
80%						
	100	.247	.277	.293	.304	.316
	140	.235	.267	.282	.292	.299
	180	.232	.258	.271	.280	.287
	220	.229	.253	.264	.272	.278
	260	.226	.249	.258	.267	.273
	280	.224	.247	.258	.265	.269

Table 2 (Cont.)

Question Trolling

Heterogeneity

SD =.05

20%					
100	.213	.222	.226	.230	.232
140	.210	.220	.224	.227	.228
180	.210	.218	.222	.225	.226
220	.210	.216	.220	.223	.225
260	.209	.216	.220	.222	.220
280	.209	.216	.219	.222	.220
40%					
100	.225	.243	.253	.261	.264
140	.220	.240	.247	.253	.257
180	.219	.235	.243	.249	.252
220	.219	.233	.241	.247	.250
260	.218	.232	.239	.244	.246
280	.218	.231	.239	.243	.247
60%					
100	.238	.265	.279	.291	.297
140	.230	.259	.271	.280	.285
180	.229	.253	.265	.274	.279
220	.229	.249	.261	.270	.276
260	.226	.247	.259	.266	.269
280	.227	.247	.258	.265	.270
80%					
100	.251	.287	.305	.321	.329
140	.240	.279	.294	.307	.314
180	.238	.271	.286	.299	.305
220	.238	.266	.281	.293	.301
260	.235	.263	.279	.287	.293
280	.236	.263	.277	.286	.294

Table 2 (Cont)

SD=.10

20%					
100	.217	.230	.237	.241	.245
140	.216	.228	.235	.239	.242
180	.216	.228	.234	.238	.241
220	.215	.227	.233	.237	.239
260	.215	.226	.232	.237	.239
280	.215	.227	.233	.236	.240
40%					
100	.234	.260	.273	.283	.290
140	.231	.256	.270	.277	.285
180	.232	.255	.267	.277	.281
220	.229	.253	.266	.274	.279
260	.230	.253	.264	.274	.278
280	.231	.254	.265	.273	.279
60%					
100	.250	.290	.310	.324	.334
140	.247	.285	.305	.316	.327
180	.247	.283	.301	.315	.322
220	.244	.280	.299	.311	.318
260	.245	.279	.296	.311	.318
280	.246	.281	.298	.309	.319
80%					
100	.267	.320	.347	.365	.379
140	.262	.313	.340	.355	.370
180	.263	.310	.334	.353	.362
220	.259	.307	.332	.348	.358
260	.260	.306	.328	.347	.357
280	.262	.308	.331	.345	.359

Table 3 (cont)

SD=.15					
20%					
100	.220	.238	.246	.252	.254
140	.219	.237	.244	.250	.254
180	.220	.237	.245	.249	.254
220	.219	.237	.243	.248	.253
260	.220	.236	.242	.249	.253
280	.219	.235	.243	.247	.252
40%					
100	.239	.275	.292	.303	.309
140	.238	.274	.287	.299	.309
180	.239	.274	.289	.298	.307
220	.238	.274	.287	.297	.305
260	.239	.272	.285	.297	.306
280	.237	.270	.285	.294	.305
60%					
100	.259	.313	.337	.355	.363
140	.258	.310	.331	.349	.363
180	.259	.311	.334	.347	.361
220	.257	.311	.330	.345	.358
260	.259	.308	.327	.346	.359
280	.256	.305	.328	.342	.357
80%					
100	.279	.351	.383	407	.417
140	.277	.347	.375	.399	.417
180	.279	.345	.378	.398	.413
220	.276	.348	.374	.394	.411
260	.278	.343	.370	.395	.410
280	.275	.340	.371	.389	.409

Note. Absent HARKing, ρ = .20. Cherry picking can be thought of as a special case of question trolling in which there is only one population parameter describing the relationship being studied, so that SD=0. N = sample size, pool size = number of results available in a particular study and dataset from which a researcher can choose, ρ = underlying population effect, prevalence = proportion of studies in which cherry picking or question trolling occurs. Because the values in adjacent cells of our simulation design were highly similar, this table presents only a subset of all of the values we calculated; full tables are available from the authors upon request.

Table 3

Conclusions Regarding HARKing

Different Forms of HARKing

1. *Not all forms of HARKing produce biased results.* As summarized in Table 1, some forms of HARKing such as THARKing can actually be desirable because they can lead to important discoveries.

2. *Cherry picking and question trolling will always produce biased results*. Because these both represent biased searches for the most favorable results, these forms of HARKing will, by definition, introduce biases. Under some circumstances (e.g., small samples, large sets of variables to be examined, heterogeneous population parameters), these biases can be substantial, both in absolute and relative terms.

3. *HARKing is not always a case of author misconduct*. Reviewers and editors often suggest substantial revisions to the paper they review, and they need to be on the lookout for pushing authors to create hypotheses that did not exist prior to the submission of their manuscript.

Which Forms of HARKing Produce the Largest Bias

4. *Cherry picking's impact is generally small*. Except when HARking is very prevalent and sample size is small, cherry-picking results have a small biasing impact on effect size estimates.

5. *Question trolling's impact can be very large*. Question trolling can have a large effect when this behavior is pervasive and if the set of results the author chooses from is highly heterogeneous (i.e., the variability in population parameters underlying the results that are scanned approaches the variability across the entire field of study of applied psychology).

6. *How we measure HARKing's bias matters*. Biases produced by cherry picking and question trolling are generally small when measured in correlation units (i.e., usually in the .10s). But, this same amount of bias seems to be much larger and impactful if measured in percent increase compared to the true effect sizes (i.e., 50% and even 100%).

Minimizing the Detrimental Forms of HARKing

7. *Increase sample size.* Large samples (e.g., samples larger than 200) help to minimize the biases associated with cherry picking and question trolling. Some of the biases that are introduced by these HARKing behaviors are the result of taking systematic advantage of random fluctuations in data, and large samples help mitigate this concern.

8. *Decrease the prevalence of HARKing*. Decreasing the prevalence of HARKing may be sufficient to decrease its cumulative effects. Across all of the simulations we performed, the biases produced by cherry picking or question trolling were generally small if the prevalence of HARKing was less than 40%. It may not be necessary to eliminate all HARKing to keep its detrimental effects in check.

How Can We Detect and Deter Detrimental Forms of HARKing?

9. *Use Occam's Razor*. HARKed hypotheses often involve convoluted reasoning or counterfactual assumptions. If the conceptual case for a hypothesis seems unlikely to hold up to scrutiny, or does not seem to emerge organically from the literature and theory the author cites, this is one potential indicator of HARKing.

10. *Have a healthily skeptical attitude*. Stories that are too good to be true may not be true. In our combined experience, we have supervised hundreds of theses, dissertation and research studies and have carried out a large number of studies ourselves. It is unusual for every prediction and every hypothesis to be supported in a study, and when reviewing a manuscript in which every prediction is supported, skepticism is warranted.

11. *Reduce the temptation to HARK*. HARKing is not simply a problem of author misbehavior—reviewers and editors' requests that authors tidy up otherwise "messy" research by encouraging authors to drop or modify hypotheses that are not supported have the same detrimental effect as HARKing.

12. *Encourage and reward replications*. The temptation to artificially inflate one's results, including cherry picking and question trolling, would be smaller is researchers believed that subsequent attempts at replication would quickly expose the unrepresentative nature of the results they published.

Appendix: R Code Used in Simulation Studies

The code below calculates the expected results if 100% of studies engage in either cherry picking or question trolling. The final estimates of the values expected if some proportion of all studies involve either cherry picking or question trolling. Is obtained by calculating a weighted average (weighted by estimated prevalence) of the values produced by the code below and the expected value of r = .20 if there is no either cherry picking or question trolling.

R Code for Cherry Picking

```
ES<-.20
outvector<-1:1000
harkvector<-1:10
nvector<-c(100,120,140,160,180,200,220,240, 260, 280)
harkarray<-array(1:100, dim=c(10,10))
for (l in 1:10){
se<-1/(sqrt(nvector[l]-3))
harkarray[l,1]<-nvector[l]
for (j in 2:10) {
for (i in 1:1000) {
R1<- rnorm(1,ES,se)
R2<- rnorm(1,ES,se)
R3<- rnorm(1,ES,se)
R4<- rnorm(1,ES,se)
R5<- rnorm(1,ES,se)
R6<- rnorm(1,ES,se)
R7<- rnorm(1,ES,se)
R8<- rnorm(1,ES,se)
R9<- rnorm(1,ES,se)
R10<- rnorm(1,ES,se)
rvector<-c(R1,R2,R3,R4,R5,R6,R7,R8,R9,R10)
rselvector<-sample(rvector,j)
                    }
                    #find z prime value needed for significance for each N
                    #
                    for (j in 1:10){
                    se<-1/(sqrt(nvector[j]-3))
                    minval[j]<-1.96*se
                    }

rval<-max(rselvector)
outvector[i]<- fisherz2r(rval)
}
harkarray[l,j]<-mean(outvector)
}
}
```

write.csv(harkarray,file="File Location")

R Code for Question Trolling

```
ES<-.203
N<-100
outvector<-1:1000
escvector<-c(.10, .14, .18, .19, .20, .20, .21, .23, .25, .29)
esvector<- fisherz (escvector)
indexvector<-c(1,2,3,4,5,6,7,8,9,10)
nvector<-c(100,120,140,160,180,200,220,240, 260, 280)
harkarray<-array(1:100, dim=c(10,10))
for (l in 1:10){
se<-1/(sqrt(nvector[l]-3))
harkarray[l,1]<-nvector[l]
for (j in 2:10) {
for (i in 1:1000) {
subsvector<-sample(indexvector)
i1<-subsvector[1]
i2<-subsvector[2]
i3<-subsvector[3]
i4<-subsvector[4]
i5<-subsvector[5]
i6<-subsvector[6]
i7<-subsvector[7]
i8<-subsvector[8]
i9<-subsvector[9]
i10<-subsvector[10]
R1<- rnorm(1,esvector[i1],se)
R2<- rnorm(1,esvector[i2],se)
R3<- rnorm(1,esvector[i3],se)
R4<- rnorm(1,esvector[i4],se)
R5<- rnorm(1,esvector[i5],se)
R6<- rnorm(1,esvector[i6],se)
R7<- rnorm(1,esvector[i7],se)
R8<- rnorm(1,esvector[i8],se)
R9<- rnorm(1,esvector[i9],se)
R10<- rnorm(1,esvector[i10],se)
rvector<-c(R1,R2,R3,R4,R5,R6,R7,R8,R9,R10)
rselvector<-sample(rvector,j)
rval<-max(rselvector)
rval1= fisherz2r(rval)
outvector[i]<-rval1
```

```
}
harkarray[l,j]<-mean(outvector)
}
}
```

write.csv(harkarray,file="/Users/kevinmurphy/Desktop/hark.csv")

Note: the code above is for a distribution with a mean of .20 and SD_ρ of 05.
For SD_ρ = .10, use

```
escvector<-c(.03,.09, .14, .18,.20,.21, .23, .27, .32, .39)
```

For SD=.15, use

```
escvector<-c(-.06,.02, .07, .15, .19, .21, .24, .33, .35, .46)
```