PERSONNEL PSYCHOLOGY
THE STUDY OF PEOPLE AT WORK

PERSONNEL PSYCHOLOGY
2016, 69, 709–750

HARKING'S THREAT TO ORGANIZATIONAL RESEARCH: EVIDENCE FROM PRIMARY AND META-ANALYTIC SOURCES

FRANK A. BOSCO
Virginia Commonwealth University

HERMAN AGUINIS
Indiana University

JAMES G. FIELD
Virginia Commonwealth University

CHARLES A. PIERCE
University of Memphis

DAN R. DALTON
Indiana University

We assessed presumed consequences of hypothesizing after results are known (HARKing) by contrasting hypothesized versus nonhypothesized effect sizes among 10 common relations in organizational behavior, human resource management, and industrial and organizational psychology research. In Study 1, we analyzed 247 correlations representing 9 relations with individual performance in 136 articles published in *Journal of Applied Psychology* and *Personnel Psychology* and provide evidence that correlations are significantly larger when hypothesized compared to nonhypothesized. In Study 2, we analyzed 281 effect sizes from a meta-analysis on the job satisfaction–job performance relation and provide evidence that correlations are significantly larger when hypothesized compared to nonhypothesized. In addition, in Study 2, we documented that hypothesized variable pairs are more likely to be mentioned in article titles or abstracts. We also ruled out 13 alternative explanations to the presumed HARKing effect pertaining to methodological (e.g., unreliability, publication year, research setting, research design, measure contextualization, publication source) and substantive (e.g., predictor–performance pair, performance measure, satisfaction measure,

The first and second authors contributed equally to this research.

We thank Chad Van Iddekinge, two *Personnel Psychology* anonymous reviewers, Keith Hutchison, Keith Leavitt, Barry Nathan, and Rick Block for highly constructive feedback that allowed us to improve our manuscript substantially. In addition, we thank Christian J. Calderon and Kathy Ma for assistance with coding Study 1 articles regarding research setting and type of performance measure.

Correspondence and requests for reprints should be addressed to Frank A. Bosco, Department of Management, School of Business, Virginia Commonwealth University, P.O. Box 844000, Richmond, VA 23284-4000; fabosco@vcu.edu.

doi: 10.1111/peps.12111

occupation, job/task complexity) issues. Our results suggest that HARKing seems to pose a threat to research results, substantive conclusions, and practical applications. We offer recommended solutions to the HARKing threat.

Hypothesizing after results are known (HARKing; Kerr, 1998) refers to the questionable research practice of retroactive hypothesis inclusion of an unexpected finding or exclusion of a "failed" prediction. The practice of HARKing, also referred to as accommodational hypothesizing (Hitchcock & Sober, 2004) and presenting post hoc hypotheses as a priori (Leung, 2011), has been admitted by about 30% of researchers (Fanelli, 2009; John, Loewenstein, & Prelec, 2012).

Although HARKing is considered a "questionable" research practice, the following fundamental questions remain: What are the effects of HARKing, if any? Does HARKing affect research results and substantive conclusions or is it simply a nuisance? Our article reports two studies whose purpose is to provide evidence regarding the extent to which HARKing is associated with changes in effect size estimates. To this end, we implement an indirect methodological approach for assessing HARKing's impact because authors do not describe the process of hypothesis generation in their articles. Moreover, HARKing is a sensitive topic—for authors, journal editors, and reviewers. Thus, we are not able to study the phenomenon in real time, and therefore we examine it post hoc. Our logic is that, if hypothesized relations are stronger than nonhypothesized relations, the difference is likely due to HARKing. In our studies, we document the magnitude of the HARKing effect by comparing hypothesized versus nonhypothesized published effect sizes. In addition, we ask whether hypothesized relations are more visible (e.g., mentioned in article abstracts) than nonhypothesized relations. Importantly, we also rule out 13 alternative explanations for the relation between hypothesized status and effect size estimates. Because of the nonexperimental nature of our research design, the HARKing effect we document should be interpreted as the "presumed" HARKing effect.

Epistemological Background of HARKing

HARKing has long been a topic of debate among philosophers of science, who distinguish between hypotheses built as predictions (i.e., a priori) versus accommodations (i.e., a posteriori; e.g., Harker, 2008; Lipton, 2001; White, 2003). In fact, for some epistemologists (e.g., Lipton, 2005), whether hypotheses are constructed before versus after examining the data is a pivotal distinction. However, hypothesis origin information is rarely available to, and therefore rarely considered by, consumers of science (Gardner, 1982).

The following scenario (adapted from Hitchcock & Sober, 2004) illustrates the distinction between prediction (a priori hypothesizing) and accommodation (i.e., HARKing) with two hypothetical researchers: Penny Predictor and Annie Accommodator. Imagine that Penny Predictor hypothesizes a priori (i.e., predicts) that openness to experience and employee turnover will be related. Penny tests and rejects the null hypothesis and reports an effect size between the variable pair, r_{Penny}. The other researcher, Annie Accommodator, hypothesizes a relation between extraversion and employee turnover. She also successfully rejects her null hypothesis. However, after analyzing the data, Annie discovers that a different variable, openness to experience, also predicts turnover, and thus she builds an accommodating hypothesis, a theoretical rationale for it, and reports an effect size between the accommodated pair, r_{Annie}. Still other researchers might have removed the openness to experience–turnover hypothesis from their manuscript had they failed to observe a significant relation yet still possibly reported the effect size (e.g., in a correlation matrix involving all study variables; Dalton, Aguinis, Dalton, Bosco, & Pierce, 2012). Is Penny's hypothesis or result about the openness to experience-turnover relation more credible than Annie's? Has Annie created a needlessly complex hypothesis or model, thus complicating the theoretical landscape unnecessarily (Leavitt, Mitchell, & Peterson, 2010)? Will Annie's hypothesis have less predictive success in the future? If so, what are the ramifications for scientific progress?

For philosophers of science, debate on Annie's and Penny's situation has ensued for more than a century and a half (e.g., Mill, 1843). The view that Penny's hypothesis has an advantage over Annie's, by dint of having predicted the outcome, is labeled *predictivism* (also known as the *advantage thesis*). Proponents of this view (e.g., Hitchcock & Sober, 2004) argue that hypothesis accommodation (i.e., HARKing) leads to overfitting of data and impedes a theory's potential for predictive precision. In contrast, proponents of the alternative view, *accommodationism*, are agnostic to the difference between Penny's and Annie's hypotheses. They argue that no privileged status should be afforded to Penny's hypothesis. Indeed, "Mill (1843) claimed that no serious scientific mind could grant more than a psychological distinction between prediction and accommodation" (Hitchcock & Sober, 2004, p. 2).

HARKing Mechanisms

Prevalence of and Motivation for HARKing

HARKing's prevalence was demonstrated by a recent content analysis of hypothesis statements in dissertation—later published article pairs (O'Boyle, Banks, & Gonzalez-Mulé, in press). In this study, the supported-to-nonsupported hypothesis ratio was significantly larger for published articles compared to that of the dissertations on which they relied, roughly 2 to 1 and 1 to 1, respectively. According to O'Boyle, Banks, and Gonzalez-Mulé (in press), this finding is driven by authors' removal of nonsupported hypotheses (most common); addition of new, supported hypotheses (less common); and reversing directional hypotheses (least common). In addition, Fanelli (2009) reported that 34% of scientists admitted to HARKing—findings were "'mined' to find a statistically significant relation . . . then presented as the original target of the study" (p. 1). Similarly, John et al. (2012) reported a HARKing frequency of 27%. Other evidence indicates that researchers admit to knowledge of their colleagues' HARKing and, less frequently, "massaging" data (e.g., De Vries, Anderson, & Martinson, 2006; Steneck, 2006). Thus, the extant literature indicates that HARKing is quite common.

One reason why authors HARK involves reviewers' negative reactions to nonsupported 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 a posteriori during the peer review process (Bedeian, Taylor, & Miller, 2010). 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). In addition, there are other explanations for the prevalence of HARKing that are specific to organizational research such as the infrequent implementation of experimental designs (Aguinis, Pierce, Bosco, & Muslin, 2009; Scandura & Williams, 2000). Indeed, compared to passive observational (i.e., correlational) research, relatively fewer HARKing opportunities are present in experimental research environments where hypotheses are often linked a priori to independent variable manipulations. Typically, an experiment involves one or two manipulations, and dropping them from a manuscript would mean that there is little information remaining to report. Finally, much organizational research is conducted by those who seek to confirm their own theories using null tests (Leavitt et al., 2010). In contrast, strong inference, which pits theories against each other (Edwards & Berry, 2010; Platt, 1964), is based on an experimental design paradigm, infrequent in organizational research and therefore offers relatively fewer opportunities for HARKing.

Overfitting, Complexity, and Predictive Precision

Hitchcock and Sober (2004) argued that the severity of HARKing's consequences depends on the presence of safeguards for overfitting data. Overfitting refers to an increase in model complexity beyond some criterion of incremental variance explanation. Any set of data may be perfectly fit (e.g., $R^2 = 1.00$) with a model of n-1 parameters, where n represents the number of observations. However, a line must be drawn between variance explained and parsimony. This is because overly complex models lack predictive precision (Hitchcock & Sober, 2004). As an illustration, imagine that a researcher is conducting a structural equation modeling analysis and sifts through a library of data containing several predictors of some outcome variable. Ritualistic tinkering might occur by adding some variables and removing others. At the end of the exercise, a model is presented with n degrees of freedom along with several fit statistics. However, as Babyak (2004, p. 416) noted, “Although it may look like we have not used many degrees of freedom in the final model, we have actually used up a whole passel of them along the way during the selection process. These phantom degrees of freedom just happen to be hidden from us at the end stage.” The end result is a model whose fit estimates are artificially inflated.

HARKing, Results Visibility, and Effect Size Estimates

Even if safeguards for overfitting were present, HARKing has another potential consequence. Specifically, HARKing results in the emphasis of supported findings through retroactive hypothesis inclusion and de-emphasis of unsupported findings through retroactive hypothesis exclusion. If hypothesized relations are more likely to be mentioned in article titles and abstracts, such findings become easier to locate and become more prominent and visible than unsupported findings. Similarly, smaller effect size estimates associated with nonsupported and removed hypotheses become more difficult to locate and also become less prominent and visible. Indeed, as Bem (2002) instructed, “data may be strong enough to justify recentering your article around . . . new findings and subordinating or even ignoring your original hypotheses” (p. 3). This presents a concern particularly for subsequent narrative literature reviews and also meta-analyses. Given that literature reviews often rely on electronic searches of titles, abstracts, and keywords, results run the risk of upward bias brought by HARKing’s promotion of larger and significant findings and demotion of smaller and nonsignificant ones. This is likely to be the case in spite of recent technological advancements and the recommendation that electronic searches involved in a meta-analysis rely on articles’ full text (Dalton et al., 2012), which often return more false positives than hits. Furthermore, because results from narrative and meta-analytic literature reviews are reproduced in textbooks and reach a broad audience that includes practitioners, HARKing has the potential to widen the science–practice gap and hamper evidence-based management (Cascio & Aguinis, 2008).

Research Questions

We are not able to determine unequivocally whether a given relation was the product of prediction or HARKing. We do, however, posit that the comparison of effect sizes across levels of hypothesized status (e.g., hypothesized vs. nonhypothesized) is a useful indicator of HARKing's presumed effects and potential for downstream impact, particularly when several other possible reasons and competing explanations for this effect are ruled out.

Consider that nonsupported hypotheses are often removed by authors and that supported, a posteriori hypotheses are frequently born from "incidentally" observed findings (e.g., Type I errors; O'Boyle et al., in press). Holding sample size constant, the degree of support of a hypothesis depends on the size of the relation in the population. All else being equal, then, removed hypotheses should be associated with smaller effect sizes than those belonging to original or added hypotheses. Given that many researchers, by their own admission (e.g., Fanelli, 2009), engage in these behaviors, what downstream effects might we expect? First, we might expect that many small and nonsignificant findings are hidden within articles (i.e., removed from hypotheses and deemphasized). Second, unexpectedly significant (i.e., larger) effect sizes are given additional attention through the addition of hypotheses and promotion in salient article texts (e.g., abstract or title). Provided that hypotheses are a major component of an article's purpose and message, there exists the potential for a large-scale disconnect between existing research findings and their salient summaries. As one route to ascertain the possible downstream effects of HARKing—whether HARKing actually matters—we examine the magnitude of the presumed HARKing effect. Specifically, our first research question is as follows:

Research Question 1: To what extent are hypothesized status and effect size related?

In addition, we investigate the extent to which bivariate relations' hypothesized status is related to article centrality. This is an important consideration for literature reviews because, as stated in the sixth edition of the *American Psychological Association*'s publication manual, there is a need to "Include in the abstract only the four or five most important concepts, findings, or implications. Use the specific words in your abstract that you think your audience will use in their electronic searches" (p. 26). To the extent that hypothesis-relevant variables are relatively central to an article's message, authors are able to manipulate variable centrality through HARKing. Because hypothesis-relevant variables play relatively more central roles in research articles, it is reasonable to expect that they will benefit from greater prominence and visibility in articles. Thus, our second research question is as follows:

Research Question 2: Do hypothesized variable pairs appear more frequently in article titles or abstracts compared to nonhypothesized variables pairs?

Study 1

We examined 247 effect sizes for relations between job performance and nine other constructs (i.e., agreeableness, autonomy, conscientiousness, emotional stability, extraversion, self-efficacy, leader–member exchange [LMX], distributive justice, procedural justice) reported in *Journal of Applied Psychology* (*JAP*) and *Personnel Psychology* (*PPsych*) from 1980 to 2010. In addition, we estimated HARKing self-admittance frequency by contacting a sample of authors of articles included in the study and requesting that they share hypothesis modification information. In addition, we tested alternative and competing explanations for the presumed HARKing effect such as type of relation (i.e., performance with each of the nine constructs), measure unreliability, publication year, research setting (i.e., lab or field), performance measure type (i.e., objective or subjective rating of performance and job or training performance), type of occupation (i.e., managerial, skilled/semiskilled, student, sales, professional, police, or other), measure contextualization (i.e., contextualized or noncontextualized), task complexity (i.e., low, medium, or high), and type of self-efficacy measure (i.e., specific, generalized, or specific/generalized composite).

Method

Data set. We used correlation coefficients reported in correlation matrices in articles published in *JAP* and *PPsych* from 1980 to 2010 as made available by an early version of the database created by Bosco, Aguinis, Singh, Field, and Pierce (2015). In total, the database contains 174,576 rows of data, with 148,739 rows representing distinct bivariate correlations, and the remainder (25,837) representing information on the variables themselves (e.g., names, mean, *SD*, reliability, sample size). This is a large database that is currently being expanded to other journals and is publicly available at http://www.frankbosco.com/data/CorrelationalEffectSizeBenchmarks.html. The database can be used for many different purposes, such as locating studies and correlations or conducting meta-analyses (Bosco et al., 2015). In our particular study, we used it to locate relations of interest (e.g., autonomy–employee performance), although some of those variables may have played a minimal role in the original study (e.g., as a control variable for another relation of interest).

Using extant taxonomies of topical research areas in organizational behavior/human resource management (OBHRM) and industrial-organizational (I-O) psychology as our guide (Cascio & Aguinis, 2008; Crampton & Wagner, 1994), we searched for the most commonly reported bivariate relations in our database using an automated contingent matching search algorithm. In this way, we were able to enter the two search criteria (each variable) into the software and view all results where that pair had been reported in the database. We limited the search to those containing one variable traditionally used as a predictor and one traditionally used as a criterion (e.g., conscientiousness-performance). Our search returned nine common bivariate relations with 10 or more samples each with employee performance: agreeableness, conscientiousness, emotional stability, extraversion, LMX, distributive justice, procedural justice, autonomy, and self-efficacy. For these relations, if more than one performance criterion was included in the article (e.g., sales volume and supervisor ratings), we combined the results before submitting the effect size to the analysis by calculating the mean of the two effect sizes (in the case of equal sample sizes) or sample size weighted the effect sizes (in the case of unequal sample sizes) using bare-bones meta-analytic procedures (Hunter & Schmidt, 2004). We focused on in-role performance rather than a variety of performance constructs (e.g., helping behaviors, organizational citizenship behavior, counterproductive behavior, deviant behavior, creative performance, adaptive performance) because our goal was to foster as much control as possible, and this involved holding the criterion constant (this same rationale guided our selection of effect size estimates from a limited year range and also from a limited set of journals). In addition, we chose in-role performance as the focal criterion because it is the most frequently assessed type of performance.

We extracted 192 correlations from 106 unique articles in *JAP* and 77 correlations from 38 unique articles in *PPsych*, for a total of 269 correlations from 144 unique articles. Similar to other reviews and syntheses of correlations (e.g., Aguinis, Dalton, Bosco, Pierce, & Dalton, 2011; Bosco et al., 2015), we conducted our analyses at the effect size level because we were interested in unique bivariate relations and substantive relations. For example, if an article reported relations between conscientiousness–performance and agreeableness–performance, we did not aggregate these correlations because they address different types of relations. The exception was the very few cases in which a sample was associated with a relation between the same two variables over time such as autonomy–performance (Time 1) and autonomy–performance

(Time 2), in which case we combined according to the approach described earlier. Hence, our results were not affected by possible differences between singular versus composite correlations. In addition, we analyzed raw *r*s, rather than absolute value *r*s, because all the summary estimates pertaining to the relations that we examined demonstrated positive relations with performance.

Hypothesis status coding procedure. The first and third author coded each of the 269 effect sizes independently. To maintain coder blindness, hypothesized status coding was performed in a spreadsheet that did not contain effect size estimates. We extracted effect size, sample size, reliability, and hypothesized status information from the original sources. Except for relations that were not hypothesized, both of the variables in the investigated pair must have been stated as related in a single statement or model for it to have been coded as hypothesized. The variable pair could be coded as one of the following: (a) nonhypothesized (e.g., exploratory study), (b) hypothesized to be related (i.e., main, moderating, or mediating effect), or (c) hypothesized to be weaker or stronger than another relation (e.g., strength contrast hypothesis; autonomy will predict performance to a greater degree than engagement). Because moderating effects are symmetrical (Aguinis, 2004), effect sizes classified as belonging to a "moderation hypothesis" refer to either the bivariate relation X_1-Y moderated by X_2 or the bivariate relation X_2-Y moderated by $X_{1.}$ We used a similar approach to classify cases as belonging to a mediating hypothesis. Thus, for the relation $X \rightarrow Z \rightarrow Y$, the X-Y and Z-Y bivariate relations were candidates for coding as belonging to a mediation hypothesis.

We excluded relation strength contrast hypotheses ($k = 22$ or 8% of the 269 effect sizes) due to a limited sample for these contrasts; this was especially the case within each of the nine relations. Our analyzable sample thus contained 178 correlations from 101 unique articles in *JAP* and 69 correlations from 35 unique articles in *PPsych*, for a total of 247 correlations from 136 unique articles. Although we originally coded for six levels of hypothesis type because these were the most frequent ones (i.e., nonhypothesized, main effect, moderating effect, mediating effect, "stronger" relation contrast, and "weaker" relation contrast), our analyses used a dichotomous code: nonhypothesized or hypothesized (a combination of main, moderating, and mediating hypotheses). The complete data set, with all original codes, is available from the authors upon request.

For articles that did not state formal hypotheses, we searched for several keywords reflective of informal hypotheses, but stated predictions nonetheless. Specifically, we used a document search process to locate the letter strings "hypo," "expect," "predict," and "anticipate." Instances of article text such as, "we predict A to be related to B," without being labeled explicitly as a hypothesis, were coded as hypotheses. For articles wherein hypotheses were not found after the letter string search process, we scanned the paragraphs preceding the beginning of the Method section for such statements. Finally, for articles that tested a model and did not present formally stated hypotheses, we coded the relation as portrayed by the model. As an example, if a model portrayed an interactive relation between X_1 and X_2 with Y, but did not present it in text as a formal hypothesis, the case was coded as belonging to a moderation hypothesis.

Article centrality coding procedure. We coded each of the 247 correlations for their variables' presence or absence in its salient article search text. To do so, we searched the title and abstract text of each article for the variables involved in the bivariate relation. Correlations were coded as *central* if both variable terms appeared in the title or abstract or *peripheral* if neither variable terms were contained in the title or abstract. For cases in which one variable was mentioned in the title and the other mentioned in the abstract, the relation was coded as central. Because the keyword coding relied on a simple letter string matching, the coding was conducted by only the first author.

Study setting and performance measure objectivity and type. Two management doctoral students who were naive to our study hypotheses independently coded articles with respect to the study's setting (i.e., lab vs. field) and measure of performance (i.e., objective vs. subjective rating and job vs. training performance). These variables were added to our study during the review process, and, therefore, we used coders who were uninvolved with our research (as requested by the review team).

HARKing prevalence. To assess the extent to which our sample reflects admitted HARKing rates reported elsewhere, we emailed corresponding authors of all articles in our data set published from 2005 through 2010 (62, or 46% of the 136 articles in our data set). We chose 2005 as a cutoff year because the duration between initial hypothesis formulation for 2005 articles approached 10 years as we were writing this manuscript; authors of papers published in earlier years may not be able to recall whether or how the hypotheses may have changed. We asked authors whether any changes in hypotheses had occurred between the completion of data collection and subsequent publication, and to describe any such changes. We received responses from 53 of the 62 authors, a response rate of 85%. Responses were content analyzed by the first and third author according to four variables, all coded as "yes" or "no" in terms of (a) whether any hypothesis changes were recalled, (b) whether any hypothesis changes were recalled as initiated by the authors, (c) whether any hypothesis changes were recalled as suggested by manuscript reviewers and/or the editor, and (d) whether the respondent indicated that he or she could not recall whether or how the hypotheses changed.

Results and Discussion

Agreement assessment. The first and third author independently coded each of the 247 effect sizes and reached acceptable levels of agreement for sample size (96%), effect size (97%), reliability (96%), and hypothesis status (94%). Regarding the HARKing admittance data supplied by corresponding authors, of the 212 codes (53 responses by four questions), the coding process resulted in 13 disagreements (94% agreement), each of which was resolved by discussion. The coding regarding study setting and performance measure was conducted by two management doctoral students who were blind to our study hypotheses. First, they each coded five of the articles independently. The raters only disagreed in their coding of one article for the objective versus subjective distinction (i.e., 93% agreement). This one disagreement was easily resolved. Each coder then independently coded 10 additional articles. The raters only disagreed in their coding of two articles for the objective versus subjective variable and one article for the job versus training categorization (i.e., 90% agreement). These three disagreements were also easily resolved. In sum, the two coders independently coded the 15 articles with 91.1% agreement. Subsequently, after additional coding training and given the high level of agreement, the two coders each independently coded half of the remaining articles.

HARKing prevalence. Twenty of the 53 respondents (38%) reported that at least one hypothesis had changed between the completion of data collection and publication, 12 (23%) reported that at least one hypothesis change was initiated by the author(s), 11 (21%) reported that at least one hypothesis change occurred as a result of the review process, and 15 (28%) used phrases indicating they were unable to recall whether or how the hypotheses changed. Within the set of respondents using phrases indicative of lack of recall, 5 of the 15 respondents (33%) reported that at least one hypothesis had changed between the completion of data collection and publication, 2 (13%) reported that at least one hypothesis change was initiated by the author(s), and 4 (27%) reported that at least one hypothesis change occurred as a result of the review process. Finally, within the set of respondents who did not use phrases indicative of lack of recall, 15 of the 38 respondents (39%) reported that at least one hypothesis had changed between the completion of data collection and publication, 10 (26%) reported that at least one hypothesis change was initiated by the author(s), and 7 (18%) reported that at least one hypothesis change occurred as a result of the review process. John et al.'s (2012) questionnaire findings indicate a self-admission rate of "reporting an unexpected finding as having been predicted from the start" (i.e., HARKing) of 27%. Thus, the level of self-admitted HARKing in our sample is similar to or greater than that reported in previous research (e.g., Fanelli, 2009; John et al., 2012).

Research Question 1: To what extent are hypothesized status and effect size related?

Table 1 shows meta-analytic results for the complete set of 247 effect sizes and each of the nine relations, corrected and uncorrected for unreliability. An omnibus meta-analytic test for moderation, with all nine relations combined, revealed that hypothesized relations (mean r = .20; 95% CI [.17, .22]; k = 141; N = 30,175) are larger than nonhypothesized relations (mean r = .09; 95% CI [.07, .11]; k = 106; N = 25,171; Q_b = 166.08, $p < .01$), a difference of .11. Note that Hunter and Schmidt (2004) do not favor the Q statistic because it "has all the flaws of any significance test" (p. 416). However, Sagie and Koslowsky (1993) conducted a Monte Carlo simulation study and concluded that the Q test had power rates above .80 and Type I error rates below 10%. Hence, our tables include Q statistic results. However, the tables also include the correlation for each subgroup.

It is possible that hypothesized relations are larger not due to HARKing but because they may be assessed with higher-quality measures compared to nonhypothesized relations. Accordingly, to assess the extent to which differential measurement error may account for the presumed HARKing effect, we corrected each effect size for predictor and criterion unreliability. We obtained predictor reliability estimates for 209 (85%) of the 247 effect sizes and criterion reliability information for 157 (64%) of the 247 effect sizes. We did not code for type of criterion reliability estimate given that the vast majority were internal consistency coefficients (i.e., alpha). We did not see the need to code for which type of reliability was used because several reviews have documented the prevalence of alpha. For example, Köhler, Cortina, Kurtessis, and Gölz (in press) counted reliability coefficients reported in articles published in *Academy of Management Journal* and *JAP* between 2004 and 2011, and found that approximately 90% of the criterion reliability coefficients were alpha reliability estimates. We imputed missing predictor reliability values, based on the sample-weighted mean of the available reliability values, within each of the nine relations. Criterion reliability values were imputed based on the complete set of 157 reliability values. With effect sizes corrected individually for measurement error in each variable, hypothesized relations (mean r = .24; 95% CI [.21, .27]; k = 141; N = 30,175) were larger than nonhypothesized relations (mean r = .11; 95% CI [.08, .13]; k = 106; N = 25,171; Q_b = 175.79, $p < .01$), a difference of .13, which is similar to the .11 increase observed for uncorrected correlations.

In addition, we addressed our first research question within each of the nine relations with the caveat that we conducted some of these tests

TABLE 1
Study 1: Presumed Effects of HARKing on Predictor–Job Performance Correlations

Predictor	k	N	*Mean* r (ρ)	SDr ($SD\rho$)	95% CI (L) for r (ρ)	95% CI (U) for r (ρ)	Q_b for r (ρ)
Complete set	247	55,346	.15 (.18)	.15 (.13)	.13 (.16)	.17 (.20)	166.08 ($p < .01$)
Hypothesized	141	30,175	.20 (.24)	.15 (.13)	.17 (.21)	.22 (.27)	(175.79 ($p < .01$))
Nonhypothesized	106	25,171	.09 (.11)	.12 (.10)	.07 (.08)	.11 (.13)	
Agreeableness	26	6,514	.05 (.06)	.08 (.06)	.02 (.02)	.08 (.10)	.07 ($p = .79$)
Hypothesized	7	1,037	.04 (.06)	.12 (.09)	−.05 (−.06)	.13 (.17)	(.04 ($p = .84$))
Nonhypothesized	19	5,477	.05 (.07)	.07 (.05)	.02 (.02)	.08 (.10)	
Autonomy	11	2,080	.14 (.18)	.14 (.12)	.06 (.07)	.23 (.28)	2.65 ($p = .10$)
Hypothesized	9	1,810	.16 (.19)	.14 (.12)	.06 (.08)	.25 (.30)	(2.36 ($p = .12$))
Nonhypothesized	2	270	.05 (.07)	.11 (.07)	−.11 (−.14)	.20 (.27)	
Conscientiousness	44	8,264	.16 (.20)	.11 (.08)	.13 (.16)	.19 (.24)	1.14 ($p = .28$)
Hypothesized	30	6,214	.15 (.19)	.12 (.09)	.11 (.14)	.19 (.24)	(1.48 ($p = .22$))
Nonhypothesized	14	2,050	.18 (.23)	.08 (.03)	.14 (.17)	.22 (.29)	
Distributive justice	12	4,858	.10 (.12)	.14 (.13)	.02 (.03)	.18 (.22)	21.07 ($p < .01$)
Hypothesized	4	875	.24 (.28)	.23 (.22)	.02 (.02)	.47 (.53)	(21.25 ($p < .01$))
Nonhypothesized	8	3,983	.07 (.09)	.09 (.07)	.01 (.01)	.13 (.16)	
Emotional stability	33	7,237	.12 (.09)	.09 (.06)	.09 (.08)	.15 (.16)	25.67 ($p < .01$)
Hypothesized	14	4,742	.16 (.06)	.06 (.02)	.13 (.13)	.19 (.21)	(24.59 ($p < .01$))
Nonhypothesized	19	2,495	.04 (.09)	.09 (.00)	.00 (.00)	.08 (.10)	
Extraversion	30	5,887	.13 (.16)	.11 (.08)	.10 (.11)	.17 (.21)	6.21 ($p = .01$)
Hypothesized	15	3,665	.16 (.18)	.09 (.07)	.11 (.13)	.21 (.24)	(6.58 ($p = .01$))
Nonhypothesized	15	2,223	.09 (.11)	.11 (.08)	.03 (.03)	.15 (.19)	

(continued)

TABLE 1 (continued)

Predictor	k	N	*Mean* r (ρ)	SDr ($SD\rho$)	95% CI (L) for r (ρ)	95% CI (U) for r (ρ)	Q_b for r (ρ)
LMX	18	5,212	.23 (.26)	.08 (.06)	.19 (.21)	.27 (.30)	2.87 ($p = .09$)
Hypothesized	16	3,315	.25 (.28)	.09 (.06)	.21 (.23)	.29 (.33)	(4.73 ($p = .03$))
Nonhypothesized	2	1,898	.20 (.22)	.07 (.06)	.11 (.12)	.29 (.31)	
Procedural justice	18	5,972	.05 (.06)	.16 (.15)	−.03 (−.03)	.13 (.15)	8.50 ($p < .01$
Hypothesized	6	1,347	.12 (.13)	.20 (.19)	−.04 (−.05)	.28 (.32)	(7.46 ($p < .01$))
Nonhypothesized	12	4,625	.03 (.04)	.14 (.13)	−.05 (−.06)	.11 (.13)	
Self-efficacy	55	9,323	.28 (.33)	.15 (.14)	.24 (.27)	.32 (.38)	13.56 ($p < .01$)
Hypothesized	40	7,171	.30 (.35)	.16 (.14)	.25 (.28)	.35 (.42)	(20.72 ($p < .01$))
Nonhypothesized	15	2,152	.21 (.24)	.11 (.08)	.15 (.18)	.27 (.30)	

Note. k = number of samples, N = number of observations, r = sample size-weighted correlation, ρ = effect size corrected for predictor and criterion reliability, SD = standard deviation, 95% CI = 95% confidence interval, U = upper, L = lower, Q_b = χ^2-based test for significance of moderation, LMX = leader–member exchange. Q_b for ρ calculated according to Aguinis, Sturman, and Pierce (2008).

using a small sample of studies. As shown in Table 1, uncorrected effect sizes pertaining to five of the nine relations (i.e., performance with distributive justice, emotional stability, extraversion, procedural justice, and self-efficacy) presented with significantly larger effect sizes when hypothesized compared to nonhypothesized. Four of the nine comparisons were not statistically significant (i.e., performance with agreeableness, autonomy, conscientiousness, and LMX). Analyses with effect sizes corrected for predictor and criterion unreliability revealed a similar pattern, with one additional relation (i.e., LMX–performance) that reached significance, resulting in six of the nine comparisons being statistically significant.

Furthermore, we conducted meta-regression analyses on the 247 effect sizes to address our first research question while assessing publication year, relation type, performance measure objectivity (i.e., subjective = 0; objective = 1), research setting (i.e., lab = 0; field = 1), and performance type (i.e., training performance = 0; job performance = 1) as alternative explanations. We used the metafor 1.9–3 package for R (Viechtbauer, 2010), which implements the meta-regression procedures proposed by Knapp and Hartung (2003). Our choice was guided by Monte Carlo simulation results indicating that this approach is able to control Type I error rate at the prespecified level, which is not the case with the standard meta-regression method applied in most meta-analyses to date (Viechtbauer, López-López, Sánchez-Meca, & Marín-Martínez, in press). We used the likelihood ratio test (LRT) to compare fit of contrasting models, with a significant LRT indicating that the full model accounts for additional residual heterogeneity compared to the reduced model.

As shown in Table 2, we assessed the possible effect of publication year in Step 1, which did not significantly explain variance in effect sizes. Next, we assessed the effect of relation type by entering eight dummy vectors representing the nine bivariate relations included in our data set in Step 2, which significantly improved model fit (LRT = 82.98, $p < .01$). Next, to assess the possible effect of research setting, performance measure objectivity, and performance type, we entered three dichotomous dummy vectors in Step 3, which did not significantly explain variance in effect sizes. Finally, in Step 4, we added one dichotomous vector representing hypothesized status (i.e., 0 = nonhypothesized; 1 = hypothesized), which significantly improved model fit beyond Step 3, LRT = 7.41, $p < .01$, $\beta = .05$ ($SE = .02$). As shown in Table 2, analyses conducted with effect sizes corrected for predictor and criterion unreliability revealed a similar pattern.

To rule out additional competing explanations for the effects of HARKing, we conducted two more meta-regression analyses pertaining to the following specific relation subsets given their larger number of studies compared to other relations: emotional stability–performance,

TABLE 2

Study 1: Results of Hierarchical Meta-Regression Analysis Assessing Competing Explanations for the Presumed Effects of HARKing: Omnibus Analysis for All Nine Relations With Performance

	Model 1		Model 2		Model 3		Model 4	
Variable	*β*	*SE*	*β*	*SE*	*β*	*SE*	*β*	*SE*
Publication year	−.00 (−.00)	.00 (.00)	−.00 (−.00)	.00 (.00)	−.00 (−.00)	.00 (.00)	−.00 (−.00)	.00 (.00)
Relation D1 (agreeableness)			−.03 (−.04)	.04 (.05)	−.02 (−.03)	.04 (.05)	−.03 (−.04)	.04 (.05)
Relation D2 (autonomy)			.12 (.13)	.04 (.05)	.12 (.13)	.04 (.05)	.09 (.09)	.04 (.06)
Relation D3 (conscientiousness)			.08 (.10)	.04 (.05)	.08 (.09)	.04 (.04)	.06 (.06)	.04 (.05)
Relation D4 (emotional stability)			.02 (.02)	.04 (.05)	.02 (.02)	.04 (.05)	.01 (.00)	.04 (.05)
Relation D5 (extraversion)			.04 (.05)	.04 (.05)	.04 (.05)	.04 (.05)	.03 (.03)	.04 (.05)
Relation D6 (self-efficacy)			.23 (.28)	.04 (.04)	.21 (.26)	.04 (.05)	.19 (.22)	.04 (.05)
Relation D7 (LMX)			.19 (.20)	.04 (.05)	.19 (.20)	.04 (.05)	.15 (.15)	.04 (.05)
Relation D8 (distributive justice)			.04 (.04)	.05 (.06)	.04 (.04)	.05 (.06)	.04 (.04)	.04 (.05)
Performance measure objectivity (objective)					−.03 (−.04)	.03 (.04)	−.03 (−.04)	.03 (.04)
Research setting (field)					−.05 (−.07)	.04 (.05)	−.04 (−.06)	.04 (.05)
Performance type (job performance)					−.02 (−.01)	.03 (.04)	−.02 (−.01)	.03 (.04)
Hypothesized status (hypothesized)							.05 (.07)	.02 (.02)
Log likelihood	84.89 (27.82)		126.38 (64.98)		128.19 (66.47)		131.89 (70.52)	
Likelihood ratio test			82.98** (74.32**)		3.61 (2.96)		7.41** (8.11**)	

Note. $k = 247$, LMX = leader–member exchange. Values in parentheses are based on unreliability-corrected *r*s. Values outside of parentheses are based on uncorrected *r*s. D = dummy vector. For each variable, the category in parentheses is the level that received a code of 1.

$^{*}p < .05$. $^{**}p < .01$.

extraversion–performance, and self-efficacy–performance. Results are shown in Table 3 (emotional stability and extraversion) and Table 4 (self-efficacy). Regarding the emotional stability–performance and extraversion–performance relations, the first and third author independently coded effect sizes for occupation type (i.e., managerial, skilled/semiskilled, student, sales, professional, police, or other) and measure contextualization (i.e., contextualized or noncontextualized). We chose to include these particular factors and levels for these factors as competing explanations given their research attention, as indicated by their coverage in existing meta-analyses (Barrick & Mount, 1991; Barrick, Mount, & Judge, 2001; Shaffer & Postlethwaite, 2012). Coders agreed in 96% of cases and resolved discrepancies as needed. Regarding the self-efficacy–performance relations, the first and third author independently coded effect sizes for task complexity (i.e., low, medium, or high) and self-efficacy measure type (specific, generalized, or specific/generalized composite). Again, these variables and their levels were chosen based on existing meta-analytic coverage (Stajkovic & Luthans, 1998). Coders agreed in 88% of cases and resolved discrepancies as needed. As shown in Table 3, the addition of the hypothesized status dummy code significantly improved model fit for uncorrected emotional stability–performance and extraversion–performance effect sizes above and beyond publication year, occupation type, subjective versus objective performance, lab versus field setting, training versus job performance, and measure contextualization (LRT = 4.07, $p < .05$). Similarly, as shown in Table 4, the addition of the hypothesized status dummy code significantly improved model fit for uncorrected self-efficacy–performance effect sizes above and beyond publication year, task complexity, type of self-efficacy measure, subjective versus objective performance, lab versus field setting, and training versus job (LRT = 6.79, $p < .01$). In each case, a similar pattern was observed with effect sizes corrected for predictor and criterion unreliability.

Research Question 2: Do hypothesized variable pairs appear more frequently in article titles or abstracts compared to nonhypothesized variables pairs?

Of the 141 hypothesized pairs, 110 (78%) were central and 31 (22%) were peripheral. Of the 106 nonhypothesized bivariate pairs, 77 (73%) were presented as central and 29 (27%) were peripheral. Thus, compared to nonhypothesized pairs, hypothesized pairs were descriptively more likely to be presented as central (odds ratio = 1.34; 95% CI [.75, 2.40]); however, this contrast did not reach statistical significance (χ^2 [1, N = 247] = .95, p = .33). At the finer level of analysis offered by our data set, we observed odds ratios greater than 1.0 in four of the eight relation types (M = 2.03) and odds ratios equal to or less than 1.0 in four relation

TABLE 3

Study 1: Results of Hierarchical Meta-Regression Analysis Assessing Competing Explanations for the Presumed Effects of HARKing: Emotional Stability–Performance and Extraversion–Performance Relations

	Model 1		Model 2		Model 3		Model 4	
Variable	*β*	*SE*	*β*	*SE*	*β*	*SE*	*β*	*SE*
Publication year	.00 (.00)	.00 (.00)	.00 (.00)	.00 (.00)	.00 (.00)	.00 (.00)	.00 (.00)	.00 (.00)
Relation type (extraversion)			.02 (.03)	.03 (.03)	.03 (.03)	.02 (.03)	.03 (.03)	.02 (.03)
Occupation type D1 (managerial)					.03 (.04)	.07 (.08)	.01 (.01)	.07 (.08)
Occupation type D2 (skilled-semiskilled)					.10 (.12)	.07 (.08)	.07 (.09)	.07 (.08)
Occupation type D3 (student)					−.07 (−.10)	.09 (.10)	−.07 (−.10)	.08 (.10)
Occupation type D4 (sales)					.04 (.02)	.08 (.10)	.02 (−.00)	.08 (.09)
Occupation type D5 (professionals)					−.08 (−.10)	.08 (.10)	−.09 (−.10)	.08 (.10)
Occupation type D6 (police)					.06 (.06)	.09 (.12)	.01 (−.01)	.10 (.12)
Performance measure objectivity (objective)					.01 (.04)	.05 (.06)	.02 (.04)	.05 (.06)
Research setting (field)					−.17 (−.22)	.06 (.07)	−.15 (−.18)	.06 (.07)
Performance type (job performance)					−.04 (−.02)	.03 (.04)	−.02 (−.01)	.03 (.04)
Measure contextualization (contextualized)					.05 (.04)	.04 (.06)	.03 (.02)	.04 (.06)
Hypothesized status (hypothesized)							.05 (.06)	.03 (.03)
Log likelihood	45.59 (30.48)		45.99 (30.88)		59.70 (41.80)		61.73 (43.80)	
Likelihood ratio test			.81 (.80)		27.42** (21.83*)		4.07* (4.00*)	

Note. $k = 63$. Values in parentheses are based on unreliability-corrected *r*s. Values outside of parentheses are based on uncorrected *r*s. D = dummy vector. For each variable, the category in parentheses is the level that received a code of 1.

$^{*}p < .05$. $^{**}p < .01$.

TABLE 4

Study 1: Results of Hierarchical Meta-Regression Analysis Assessing Competing Explanations for the Presumed Effects of HARKing: Self-Efficacy–Performance Relations

	Model 1		Model 2		Model 3	
Variable	*β*	*SE*	*β*	*SE*	*β*	*SE*
Publication year	−.00 (.00)	.00 (.00)	−.00 (.00)	.00 (.00)	−.00 (.00)	.00 (.00)
Task complexity D1 (low)			−.04 (−.04)	.11 (.14)	−.04 (−.03)	.10 (.13)
Task complexity D2 (medium)			−.04 (−.04)	.11 (.15)	−.03 (−.03)	.11 (.14)
Self-efficacy measure type D1 (specific)			.09 (.13)	.09 (.12)	.13 (.18)	.09 (.12)
Self-efficacy measure type D2 (general)			−.04 (−.05)	.13 (.18)	.05 (.09)	.13 (.17)
Performance measure objectivity (objective)			−.05 (−.04)	.09 (.11)	−.04 (−.02)	.08 (.11)
Research setting (field)			.06 (.04)	.10 (.13)	.02 (−.01)	.10 (.13)
Performance type (job performance)			−.17 (−.18)	.13 (.17)	−.13 (−.11)	.13 (.16)
Hypothesized status D1 (hypothesized)					.14 (.19)	.05 (.06)
Log likelihood	13.91 (−4.42)		16.29 (−2.77)		19.69 (.86)	
Likelihood ratio test			4.77 (3.31)		6.79** (7.26**)	

Note. $k = 55$. Values in parentheses are based on unreliability-corrected *r*s. Values outside of parentheses are based on uncorrected *r*s. D = dummy vector. For each variable, the category in parentheses is the level that received a code of 1.

$^{*}p < .05$. $^{**}p < .01$.

types ($M = .77$); in each of the eight cases, nonsignificant χ^2 values were observed (all ps $> .14$). Note that lack of statistical significance may be due to small sample size (i.e., mean $k = 27$).

In sum, results of Study 1 provide evidence that, for the majority of comparisons, effect sizes are larger when they are hypothesized compared to nonhypothesized. Moreover, results also show that, after implementing best practice recommendations that involved controlling for the effects of several methodological and substantive competing explanations (Bernerth & Aguinis, in press), the presumed effects of HARKing still remain. However, we did not detect evidence of a relation between hypothesized status and article centrality. One potential reason is because the term performance was included in a title or abstract for most of the present samples.

A limitation of Study 1 is its reliance on a database containing nine distinct bivariate relations with individual performance. In particular, some degree of external validity and generalizability is afforded by the array of relations, but some of the analyses were necessarily conducted using a small sample of studies. We addressed this limitation in Study 2.

Study 2

The purpose of Study 2 was to address our research questions with a large database of correlations pertaining to a single bivariate relation while simultaneously controlling for methodological and substantive factors that may serve as competing and alternative explanations for the presumed effects of HARKing. We specifically selected the job satisfaction–job performance relation because it has long been central to OBHRM and I-O psychology research (Judge, Thoresen, Bono, & Patton, 2001; Thorndike, 1917). We present analyses across all measures of job satisfaction and also for the two most frequently used measures: the 72-item Job Descriptive Index (JDI; Smith, Kendall, & Hulin, 1969) and the 20-item Minnesota Satisfaction Questionnaire Short Form (MSQ-SF; Weiss, Dawis, & England 1967). Finally, we also conducted analyses across the nine levels of occupation type used by Judge et al. (2001) in their meta-analysis.

Method

Data set. We extracted Judge et al.'s (2001, pp. 403–407) list of primary sources on the job satisfaction-performance relation. This meta-analysis is among the most comprehensive conducted on a single bivariate relation in organizational research to date and is considered an exemplar of best practices in terms of how to conduct a meta-analysis (Kepes, McDaniel, Brannick, & Banks, 2013). In addition, job satisfaction–performance effect size estimates are relatively homogenous.

We located 294 of the 312 (94%) samples included in Judge et al.'s (2001) meta-analysis. Twelve of the nonlocated samples are contained in noncirculating theses or dissertations, and six are unpublished manuscripts that we were unable to locate. Bare-bones meta-analytic estimates for our located set (mean $r = .179$; 95% CI [.163, .195]; $k = 294$; $N = 51{,}023$) were nearly identical to those of the original, complete set reported by Judge et al. (mean $r = .180$; 95% CI [.165, .196]; $k = 312$; $N = 54{,}391$), confirming the integrity of our data.

As in Study 1, we excluded effect sizes associated with relation strength contrasts (13, or 4% of the 294 samples), resulting in 281 analyzable effect sizes. Bare-bones meta-analytic estimates for the set of 281 effect sizes (mean $r = .184$; 95% CI [.167, .200]; $k = 281$; $N = 48{,}470$) were nearly identical to the complete set. Unlike Study 1, we did not correct for criterion unreliability because 85% of Judge et al.'s (2001) criterion reliability estimates were imputed based on an external estimate (Viswesvaran, Ones, & Schmidt, 1996).

Procedure. All procedural and meta-analytic approach details were identical to Study 1. However, in contrast to Study 1, we extracted information for each effect size (i.e., *N*, *r*, reliability) and for the variables that may serve as alternative and competing explanations for the presumed effects of HARKing from Judge et al.'s (2001) appendix. Specifically, Judge et al. (2001) included (a) four levels for publication source based on journal quality ratings for published sources (top-tier publication, other ranked publication, unranked publication, nonpublished/dissertation); (b) three types of job performance measures (supervisory ratings, objective records, peer/subordinate ratings, or other—no self-ratings of performance were included by Judge et al., and pairwise comparisons among the three sources of performance ratings were statistically nonsignificant); (c) three types of job satisfaction measures (global, facet composite, unknown/not specified); (d) two levels of study design (cross-sectional, longitudinal), (e) three levels of job/task complexity based on Roos and Treiman's (1980) job title ratings (high: 1 *SD* or more above mean, low: 1 *SD* or more below mean, and medium for all others); and (f) nine levels of occupation type (scientists-engineers, sales, teachers, managers/supervisors, accountants, clerical workers/secretaries, unskilled and semiskilled laborers, nurses, and miscellaneous/mixed). For this study, the first and third authors independently coded hypothesized status information and particular job satisfaction scale and scale length from each original source. As in Study 1, the first author coded article centrality information based on letter string match.

Results and Discussion

Agreement assessment. The first and third author coded the articles' hypothesized status information independently. The coders met to resolve discrepancies, and in cases where we could not reach agreement (5, or 2% of the effect sizes), they met with the fourth author to resolve the discrepancy. Agreement assessments after removing relation strength contrast hypotheses ($k = 13$ or 4%) were nearly identical. Recoding the six levels of hypothesis type into a dichotomous code, hypothesized (main; moderating; mediating) or nonhypothesized, resulted in 95% agreement. For the subset of effect sizes reporting original reliability information, coders agreed in 93% of cases.

Research Question 1: To what extent are hypothesized status and effect size related?

As show in Table 5, meta-analytic results indicate that hypothesized job satisfaction–job performance relations (mean uncorrected $r = .22$; 95% CI [.19, .24]; $k = 136$; $N = 20{,}079$) are larger than nonhypothesized job satisfaction–job performance relations (mean uncorrected $r = .16$; 95% CI [.14, .18]; $k = 145$; $N = 28{,}391$; $Q_b = 45.70$, $p < .01$), a difference of .06. We observed the same pattern among effect sizes corrected for unreliability in job satisfaction with imputation. Specifically, hypothesized job satisfaction–job performance relations (mean corrected $r = .26$; 95% CI [.23, .29]; $k = 136$; $N = 20{,}079$) were larger than nonhypothesized job satisfaction–job performance relations (mean corrected $r = .19$; 95% CI [.16, .22]; $k = 145$; $N = 28{,}391$; $Q_b = 47.60$, $p < .01$), a difference of .07. We also addressed our first research question among samples pertaining to specific measures of job satisfaction. As shown in Table 5, hypothesized job satisfaction–job performance relations were significantly larger for the 72-item JDI (Smith et al., 1969) for uncorrected effect sizes (i.e., $r = .21$ vs. $r = .06$, $Q_b = 19.42$, $p < .01$) and corrected effect sizes (i.e., $\rho = .24$ vs. $\rho = .07$, $Q_b = 21.97$, $p < .01$), a difference of .15 and .17, respectively. For the 20-item MSQ-SF (Weiss et al., 1967), hypothesized effect sizes were larger than nonhypothesized effect sizes for uncorrected (i.e., $r = .26$ vs. $r = .15$, $Q_b = 11.88$, $p < .01$) and corrected (i.e., $\rho = .30$ vs. $\rho = .20$, $Q_b = 11.24$, $p < .01$) relations, a difference of .11 and .10, respectively.

Table 6 includes results pertaining to meta-regression analyses addressing our first research question while also assessing the effect of methodological and substantive competing explanations for the presumed effects of HARKing. Publication year was entered at Step 1 and did not significantly explain variance in effect sizes. Consistent with our previous results that job satisfaction–job performance effect size estimates

TABLE 5

Study 2: Presumed Effects of HARKing on Job Satisfaction–Job Performance Correlations

Predictor	k	N	*Mean* r (ρ)	SDr ($SD\rho$)	95% CI (L) for r (ρ)	95% CI (U) for r (ρ)	Q_b for r (ρ)
Complete set	281	48,470	.18 (.22)	.14 (.12)	.17 (.20)	.20 (.24)	45.70 ($p < .01$)
Hypothesized	136	20,079	.22 (.26)	.15 (.13)	.19 (.23)	.24 (.29)	(47.60 ($p < .01$))
Nonhypothesized	145	28,391	.16 (.19)	.13 (.11)	.14 (.16)	.18 (.22)	
MSQ-SF 20-item only	19	5,601	.18 (.22)	.11 (.09)	.13 (.14)	.22 (.30)	11.88 ($p < .01$)
Hypothesized	6	1,224	.26 (.30)	.08 (.05)	.20 (.21)	.33 (.40)	(11.24 ($p < .01$))
Nonhypothesized	13	4,377	.15 (.20)	.10 (.08)	.10 (.09)	.20 (.21)	
JDI 72-item only	27	4,973	.18 (.20)	.11 (.09)	.13 (.15)	.22 (.25)	19.42 ($p < .01$)
Hypothesized	20	3,692	.21 (.24)	.09 (.05)	.17 (.19)	.25 (.28)	(21.97 ($p < .01$))
Nonhypothesized	7	1,101	.06 (.07)	.11 (.08)	−.02 (−.03)	.14 (.16)	

Note. k = number of samples, N = number of observations, r = sample size-weighted correlation, ρ = effect size corrected for predictor and criterion reliability, SD = standard deviation, 95% CI = 95% confidence interval, L = lower, U = upper, $Q_b = \chi^2$-based test for significance of moderation, JDI = job descriptive index, MSQ-SF = Minnesota Satisfaction Questionnaire Short Form. All JDS samples were nonhypothesized. Q_b for ρ calculated according to Aguinis et al. (2008).

TABLE 6

Study 2: Results of Hierarchical Meta-Regression Analysis Assessing Competing Explanations for the Presumed Effects of HARKing: Omnibus Analysis for All Job Satisfaction–Performance Relations

	Model 1		Model 2		Model 3	
Variable	*β*	*SE*	*β*	*SE*	*β*	*SE*
Publication year	−.00 (−.00)	.00 (.00)	−.00 (−.00)	.00 (.00)	.00 (.00)	.00 (.00)
Publication source D1 (top-tier)			.01 (.01)	.03 (.05)	.02 (.02)	.03 (.05)
Publication source D2 (other ranked)			.01 (.03)	.03 (.06)	.01 (.02)	.03 (.06)
Publication source D3 (unranked)			−.03 (−.06)	.04 (.07)	−.01 (−.05)	.04 (.06)
Measure of job performance D1 (supervisory ratings)			−.04 (−.12)	.04 (.07)	−.03 (−.11)	.04 (.07)
Measure of job performance D2 (peer/subordinate/other)			−.12 (−.20)	.07 (.13)	−.12 (−.20)	.07 (.13)
Measure of job performance D3 (objective records)			−.06 (−.13)	.05 (.09)	−.05 (−.13)	.05 (.09)
Measure of job satisfaction D1 (facet composite)			−.07 (−.08)	.03 (.06)	−.08* (−.10)	.03 (.06)
Measure of job satisfaction D2 (unknown)			−.08 (−.13*)	.04 (.06)	−.07 (−.12)	.04 (.06)
Research design D1 (cross-sectional)			.01 (.00)	.04 (.08)	.03 (.02)	.04 (.08)
Job/task complexity D1 (low)			.03 (.04)	.04 (.08)	.04 (.06)	.04 (.08)
Job/task complexity D2 (medium)			−.03 (−.07)	.03 (.06)	−.03 (−.07)	.03 (.06)
Occupation D1 (salespersons)			−.01 (−.02)	.08 (.14)	−.03 (−.04)	.08 (.14)
Occupation D2 (miscellaneous/mixed)			−.06 (−.12)	.08 (.13)	−.04 (−.09)	.08 (.13)
Occupation D3 (laborers)			−.07 (−.05)	.09 (.16)	−.08 (−.07)	.09 (.16)

(continued)

TABLE 6 (continued)

	Model 1		Model 2		Model 3	
Variable	β	*SE*	β	*SE*	β	*SE*
Occupation D4 (scientist-engineers)			−.03 (.11)	.09 (.16)	−.02 (.12)	.09 (.15)
Occupation D5 (clerical-secretaries)			−.03 (−.02)	.08 (.14)	−.01 (.01)	.08 (.14)
Occupation D6 (manager-supervisors)			.01 (.03)	.08 (.13)	.02 (.05)	.08 (.13)
Occupation D7 (nurses)			−.11 (−.18)	.09 (.15)	−.10 (−.17)	.08 (.15)
Occupation D8 (accountants)			−.04 (−.09)	.10 (.17)	−.03 (−.06)	.09 (.17)
Hypothesized status D1 (hypothesized)					.08** (.12**)	.02 (.04)
Log likelihood	40.66 (−82.48)		48.40 (−82.48)		52.39 (−68.75)	
Likelihood ratio test			15.47 (20.90)		7.98** (6.45*)	

Note. $k = 281$. Values in parentheses are based on unreliability-corrected *r*s. Values outside of parentheses are based on uncorrected *r*s. D = dummy vector. For each variable, the category in parentheses is the level that received a code of 1.
$^{*}p < .05$. $^{**}p < .01$.

are relatively homogeneous, the seven control variables entered simultaneously did not significantly improve model fit (LRT = 15.47, *ns*). Finally, a model with a dummy vector representing hypothesized status (i.e., 0 = nonhypothesized; 1 = hypothesized) at Step 3 significantly improved model fit, LRT = 7.98, $p < .01$, $\beta = .08$, $SE = .02$. As shown in Table 6, a similar pattern was observed with the 281 effect sizes individually corrected for unreliability.

We conducted similar meta-regression analysis with the sample limited to the 72-item JDI ($k = 27$) and 20-item MSQ-SF ($k = 19$). These meta-regression analyses addressed the potential competing explanation that type of measure may account for the effects of HARKing such that older and more established measures may be associated with larger effect sizes. In the case of the JDI, publication year did not significantly predict variance in effect sizes (Step 1); the publication source, measure of job performance, and job/task complexity moderators did not significantly improve model fit (LRT = 9.42, *ns*); and the addition of the hypothesized status vector significantly improved model fit (LRT = 8.65, $p <$ .01, $\beta = .13$, $SE = .04$). Regarding the MSQ-SF, publication year did not significantly predict variance in effect sizes, the substantive competing explanations did improve model fit beyond publication year (LRT = 26.67, $p < .01$), and the hypothesized status vector significantly improved model fit in the final step (LRT = 4.16, $p < .05$, $\beta = .30$, $SE =$ .14) when analyses were based on uncorrected correlations but not when the analyses were based on unreliability-corrected correlations ($k = 19$, LRT = .07, *ns*).

Table 7 shows a test of our first research question across the nine levels of occupation type used by Judge et al. (2001). In four of the nine occupation groups (clerical/secretaries, managers/supervisors, skilled or semiskilled laborers, miscellaneous and mixed), we observed larger hypothesized (compared to nonhypothesized) effect sizes for the uncorrected and corrected effect sizes. In two cases (salespersons, scientists/engineers), either the corrected or uncorrected effect sizes—but not both—presented with the effect. Finally, in three cases (accountants, nurses, and teachers), we observed a statistically nonsignificant relation between hypothesized status and effect size.

Research Question 2: Do hypothesized variable pairs appear more frequently in article titles or abstracts compared to nonhypothesized variables pairs?

Regarding our second research question, we observed that hypothesized variable pairs were more likely to appear in article titles or abstracts than nonhypothesized variables pairs. Specifically, of the 136 hypothesized pairs, 120 (88%) were presented as central and 16 (12%)

TABLE 7

Study 2: Effect of Hypothesized Status on Job Satisfaction–Job Performance Correlations by Occupation (k = 281)

Predictor	*k*	N	*Mean r* (ρ)	*SDr*	95% CI (L) for *r* (ρ)	95% CI (U) for *r* (ρ)	Q_b for *r* (ρ)
Complete set	281	48,470	.18 (.22)	.14	.17 (.20)	.20 (.24)	45.70 ($p < .01$)
Hypothesized	136	20,079	.22 (.26)	.15	.19 (.23)	.24 (.29)	(47.60 ($p < .01$))
Nonhypothesized	145	28,391	.16 (.19)	.13	.14 (.16)	.18 (.22)	
Accountants	7	1,240	.17 (.26)	.10	.10 (.14)	.24 (.38)	.91 ($p = .34$)
Hypothesized	2	354	.21 (.33)	.08	.10 (.16)	.32 (.50)	(1.30 ($p = .25$))
Nonhypothesized	5	886	.15 (.24)	.10	.06 (.09)	.24 (.38)	
Clerical-secretaries	18	3,019	.19 (.34)	.14	.13 (.23)	.25 (.46)	7.49 ($p < .01$)
Hypothesized	8	1,335	.25 (.49)	.15	.14 (.31)	.35 (.68)	(28.06 ($p < .01$))
Nonhypothesized	10	1,684	.15 (.22)	.10	.08 (.13)	.21 (.32)	
Manager-supervisors	31	4,276	.21 (.34)	.14	.16 (.26)	.26 (.42)	12.29 ($p < .01$)
Hypothesized	14	1,568	.28 (.47)	.19	.18 (.31)	.38 (.62)	(9.29 ($p < .01$))
Nonhypothesized	17	2,708	.17 (.27)	.09	.13 (.20)	.21 (.34)	
Nurses	13	2,129	.12 (.19)	.10	.07 (.11)	.18 (.28)	.28 ($p = .60$)
Hypothesized	7	993	.14 (.21)	.13	.04 (.06)	.23 (.37)	(.32 ($p = .57$))
Nonhypothesized	6	1,136	.11 (.18)	.06	.06 (.10)	.16 (.26)	
Salespersons	22	4,384	.19 (.28)	.10	.15 (.22)	.23 (.34)	1.84 ($p = .17$)
Hypothesized	17	3,458	.20 (.30)	.10	.15 (.23)	.25 (.37)	(3.99 ($p < .05$))
Nonhypothesized	5	926	.15 (.21)	.07	.09 (.11)	.21 (.30)	

(continued)

TABLE 7 (continued)

Predictor	k	N	*Mean* r (ρ)	*SDr*	95% CI (L) for r (ρ)	95% CI (U) for r (ρ)	Q_b for r (ρ)
Scientist-engineers	17	2,192	.19 (.46)	.12	.13 (.27)	.24 (.65)	8.11 ($p < .01$)
Hypothesized	10	592	.29 (.46)	.09	.23 (.38)	.34 (.55)	(2.86 ($p = .09$))
Nonhypothesized	7	1,600	.15 (.46)	.11	.07 (.11)	.23 (.80)	
Teachers	6	665	.23 (.40)	.13	.13 (.21)	.34 (.58)	1.06 ($p = .30$)
Hypothesized	4	506	.26 (.44)	.14	.12 (.20)	.39 (.68)	(2.36 ($p = .12$))
Nonhypothesized	2	159	.16 (.26)	.04	.10 (.16)	.22 (.36)	
Laborers	24	3,120	.16 (.25)	.20	.08 (.12)	.24 (.38)	21.45 ($p < .01$)
Hypothesized	17	2,136	.21 (.34)	.18	.13 (.20)	.30 (.48)	(15.02 ($p < .01$))
Nonhypothesized	7	984	.04 (.06)	.18	−.10 (−.16)	.17 (.27)	
Misc and mixed	143	27,445	.18 (.30)	.14	.16 (.27)	.21 (.34)	15.74 ($p < .01$)
Hypothesized	57	9,137	.22 (.35)	.15	.18 (.29)	.26 (.41)	(11.10 ($p < .01$))
Nonhypothesized	86	18,308	.17 (.28)	.14	.14 (.23)	.20 (.33)	

Note. k = number of samples, N = number of observations, r = sample size-weighted correlation, ρ = effect size corrected for predictor and criterion reliability, *SD* = standard deviation, 95% CI = 95% confidence interval, U = upper, L = lower; Q_b = χ^2-based test for significance of moderation; LMX = leader–member exchange. Q_b for ρ calculated according to Aguinis et al. (2008).

were peripheral. Of the 145 nonhypothesized bivariate pairs, 105 (72%) were presented as central and 40 (28%) were peripheral. Thus, compared to nonhypothesized pairs, hypothesized pairs were more likely to be presented as central (odds ratio = 2.86; χ^2 [1, $N = 281$] = 11.01, $p < .01$). We observed a similar pattern within the 72-item JDI; all of the 20 hypothesized pairs (100%) were presented as central, and 4 of the 7 nonhypothesized bivariate pairs (57%) were presented as central (χ^2 [1, $N = 27$] = 9.64, $p < .01$). Finally, for the MSQ-SF sample, 5 of the 6 hypothesized pairs (83%) were presented as central and 5 of the 13 (38%) nonhypothesized pairs were presented as central (odds ratio = 8.00; χ^2 (1, $N = 19$) = 3.32, $p = .07$).

Supplemental analyses. We conducted additional analyses to examine whether differences in the reliability of scores may serve as an alternative explanation for the presumed effects of HARKing. To do so, we implemented reliability generalization, which is a method used to meta-analyze reliability estimates rather than the more typical meta-analysis focused on correlation coefficients (Rodriguez & Maeda, 2006). As noted by Vacha-Haase (1998), reliability generalization is a procedure used to understand "the typical reliability of scores for a given test across studies" (p. 6). Of the 281 ESs, 162 (58%) reported reliability estimates and scale length. The mean alpha for nonhypothesized effect sizes ($\alpha = .83$; 95% CI [.83, .83]; $k = 97$; $N = 20{,}524$) was significantly larger than the mean alpha for hypothesized effect sizes ($\alpha = .81$; 95% CI [.80, .81]; $k = 65$; $N = 11{,}286$), although the difference is only .02. In short, measure reliability did not account for the presumed effects of HARKing.

General Discussion

HARKing's prevalence has been acknowledged by authors and editors (e.g., Bedeian et al., 2010; Fanelli, 2009), authors' evolving hypothesis statements within sources over time (O'Boyle et al., in press), and suspiciously low hypothesis falsification rates (e.g., Fanelli, 2010; Francis, Tanzman, & Matthews, 2014; Leung, 2011). We developed a protocol for identifying HARKing's consequences. Although the evidence in each case is indirect, as for a smoking gun, we submit that the present comparisons across levels of hypothesized status, coupled with results ruling out multiple alternative methodological and substantive explanations, provide an informative proxy. At present, there is no known way to conduct an experiment on HARKing by, for example, randomly assigning researchers to HARKing and non-HARKing conditions. Hence, given that an experimental design is not possible to answer our questions, we followed recommendations by Shadish, Cook, and Campbell (2002) and conducted multiple tests to rule out competing explanations.

Table 8 includes a description of each of the 13 alternative explanations, results assessing each, and interpretation of results. In Study 1, we ruled out explanations pertaining to type of relation, measure unreliability, publication year, research setting, performance measure type, type of occupation, measure contextualization, task complexity, and type of self-efficacy measure. In Study 2, we ruled out explanations related to publication source, type of measure of job performance, type of measure of job satisfaction, research design, job/task complexity, and type of occupation. Finally, using a recent sample of studies from Study 1, we ascertained that HARKing admittance rates were similar to those reported in previous investigations. Taken together, our two studies involving common bivariate relations in OBHRM and I-O psychology research provide evidence regarding the presumed effects of HARKing.

Interpretation of the Magnitude of HARKing's Impact

As one lens through which to describe the impact of HARKing from our meta-analytic results, Study 1 findings indicate HARKing effects in six of the nine relation groups, with hypothesized effect sizes up to about .20 correlation units larger than nonhypothesized effect sizes. In addition, hypothesized job satisfaction–job performance relations were also larger than nonhypothesized relations (Study 2). Contextualizing the size of these effects in light of a recent review of correlational effect sizes reported in *PPsych* and *JAP* (Bosco et al., 2015) leads to the conclusion that these effects are medium to large in size.

As a second lens through which to interpret the magnitude of the presumed HARKing effect, consider that the mean unreliability-corrected effect size for the omnibus analysis in Study 1 is $r = .18$. Consider further that, after ruling out competing explanations for our results, the hypothesized status–effect size relation is $\beta = .06$ ($SE = .02$; see Table 3). Thus, this coefficient translates to a contrast of $r = .15$ (nonhypothesized) and $r = .21$ (hypothesized), a .06 increase in effect size (holding other variables constant). In Study 2, the mean unreliability-corrected effect size for the omnibus analysis is $r = .22$. As in Study 1, after ruling out a variety of alternative explanations, we observed a significant relation between hypothesized status and effect size ($\beta = .08$, $SE = .02$; see Table 6), translating to a contrast of $r = .18$ (nonhypothesized) and $r = .26$ (hypothesized). Comparing these findings to those from Study 1, the Study 2 contrast yielded practically identical results: a .08 increase in effect size.

Taken together, in Study 1 and Study 2, the hypothesized status contrast accounted for a statistically significant and substantial proportion of variance in effect sizes. Coupled with the result that hypothesized relations are presented as more central to articles compared to nonhypothesized relations (Study 2), our studies are the first to provide empirical evidence

TABLE 8

Results of Tests of 13 Alternative Explanations for Relation Between Hypothesized Status and Effect Size

Alternative explanation	Test	Result	Interpretation
1. Hypothesized status covaries with measure reliability (e.g., nonhypothesized variables may be associated with lower reliability and, thus, lower effect sizes)	Study 1 and Study 2 include meta-analytic and meta-regression analyses conducted with corrected and uncorrected effect size estimates	Hypothesized relations are significantly larger than nonhypothesized relations for corrected and uncorrected effect sizes	Measure reliability does not account for HARKing's effects
	Study 2 includes reliability generalization analyses conducted across levels of hypothesized status	Reliability is higher in the nonhypothesized subsample compared to the hypothesized subsample	
2. Hypothesized status covaries with publication year (e.g., relations may be hypothesized less frequently over time) and effect sizes exhibit temporal trends downward (i.e., decline effect)	Study 1 and Study 2 include source publication year as a control variable in meta-regression analyses	In Study 1 and Study 2, hypothesized status improved the fit of the model predicting effect sizes after controlling for publication year	Publication year does not account for HARKing's effects
3. Hypothesized status covaries with measure dimensionality (i.e., there may be differences in effect sizes based on the use of global vs. facet-composite measures of job satisfaction)	Study 2 includes job satisfaction dimensionality as a control in meta-regression analyses	Hypothesized status improved the fit of the model predicting effect sizes after controlling for dimensionality	Measure dimensionality does not account for HARKing's effects
	Study 2 includes reliability generalization analyses conducted across levels of dimensionality	Reliability does not vary across levels of dimensionality	

(continued)

TABLE 8 (continued)

Alternative explanation	Test	Result	Interpretation
4. Hypothesized status covaries with publication tier (e.g., there may be smaller effect sizes and lower likelihood of hypothesis inclusion for lower-ranked outlets)	Study 1 includes only top-tier publications Study 2 includes publication tier (four levels) as a control variable in meta-regression analyses	Hypothesized relations are significantly larger than nonhypothesized relations Hypothesized status improved the fit of the model predicting effect sizes after controlling for publication tier	Publication tier does not account for HARKing's effects
5. Hypothesized status covaries with study design (e.g., longitudinal studies may be associated with smaller effect sizes)	Study 2 includes study design as a control variable in meta-regression analyses	Hypothesized status improved the fit of the model predicting effect sizes after controlling for study design	Study design does not account for HARKing's effects
6. Hypothesized status covaries with occupation type (i.e., effect sizes may be smaller for certain occupations, e.g., extraversion performance is less commonly hypothesized for technical occupations)	Study 2 includes occupation type as a control variable in meta-regression analyses	Hypothesized status improved the fit of the model predicting effect sizes after controlling for occupation type	Occupation type does not account for HARKing's effects
7. Hypothesized status covaries with level of task complexity (i.e., effect sizes may be larger for high-complexity jobs)	Study 2 includes task complexity as a control variable in meta-regression analyses	Hypothesized status improved the fit of the model predicting effect sizes after controlling for complexity	Task complexity does not account for HARKing's effects

(continued)

TABLE 8 (continued)

Alternative explanation	Test	Result	Interpretation
8. HARKing did not actually occur in the present sample and findings are driven by something other than HARKing	Study 1 includes an estimate of the admitted HARKing frequency in a sample of the analyzed articles	The 2005–2010 article sample was associated with an overall rate of 38% (23% self-initiated) HARKing self-admittance rate, similar to John et al.'s (2012) self-reported estimate of 27%	The level of HARKing's self-admittance rate in our sample is similar to or greater than those reported elsewhere.
9. HARKing effect is spurious or relation-specific (e.g., only for job satisfaction–job performance)	Study 1 includes a comparison of effect size estimates across levels of hypothesized status for nine distinct bivariate relations	Hypothesized relations are significantly larger than nonhypothesized relations for six of the nine relation types	HARKing effect is generalizable; larger effect size estimates for hypothesized compared to nonhypothesized relations were observed for 7 of the 10 relations included in these studies
	Study 1 controls for nine bivariate relation types with meta-regression analyses	Hypothesized status improved the fit of the model predicting effect sizes after controlling for relation type	
	Study 2 includes a large sample of effect sizes pertaining to a single bivariate relation	Hypothesized relations are significantly larger than nonhypothesized relations for job satisfaction–job performance relations	
10. Hypothesized status covaries with lab/field setting (i.e., larger effects in lab settings)	Study 1 and Study 2 control for study setting with meta-regression analyses	Hypothesized status improved the fit of the model predicting effect sizes after controlling for study setting	Study setting does not account for HARKing's effects

(continued)

TABLE 8 (continued)

Alternative explanation	Test	Result	Interpretation
11. Hypothesized status covaries with measure specificity (i.e., general vs. specific self-efficacy)	Study 1 controls for measure specificity with meta-regression analyses (self-efficacy subsample)	Hypothesized status improved the fit of the model predicting effect sizes after controlling for measure specificity	Measure specificity does not account for HARKing's effects
12. Hypothesized status covaries with measure contextualization (i.e., contextualized vs. noncontextualized personality measures)	Study 1 controls for measure contextualization with regression analyses (emotional stability and extraversion subsamples)	Hypothesized status improved the fit of the model predicting effect sizes after controlling for measure contextualization	Measure contextualization does not account for HARKing's effects
13. Hypothesized status covaries with type of performance measure (i.e., subjective vs. objective and training performance vs. job performance)	Study 1 controls for type of performance measure with meta-regression analyses (emotional stability, extraversion, and self-efficacy subsamples)	Hypothesized status improved the fit of the model predicting effect sizes after controlling for performance measure type	Type of performance measure does not account for HARKing's effects

regarding HARKing's downstream impact. Specifically, hypothesis-relevant effect sizes were larger and more likely to be presented as central in journal articles. Consequently, literature reviews run the risk of overlooking peripheral, smaller relations that are not prominent within articles (e.g., through HARKing). That is, although the effect size of interest might be presented in the correlation matrix, researchers conducting narrative or quantitative reviews would encounter difficulty in locating smaller relations between variable pairs that were HARKed by subtraction and encounter relative ease in locating larger relations that were HARKed by addition.

Implications for Researchers and Practitioners and Strategies for Reducing HARKing

It seems likely that HARKing makes summaries of findings appear larger than they are in actuality. In turn, scientific progress is slowed by overfitting, lack of falsification, increased theoretical complexity (Hitchcock & Sober, 2004), and positively biased literature review conclusions. Through modifications to literature search processes (e.g., relying less on the content of article abstracts), meta-analysts are likely to locate a larger sample of effect sizes and also locate effect sizes that might have played ancillary study roles (e.g., control variables).

In addition, HARKing can lead to less-than-ideal management practices because effect size estimates are the central input to estimates of practical significance (Aguinis et al., 2010; Bosco et al., 2015). For example, they play a central role as a key input value in utility calculations in personnel selection. In sum, as effect sizes become increasingly inflated from HARKing, scientific understanding and practical significance estimates become overly optimistic. Unfortunate consequences for practitioners include failure to replicate findings in organizational settings, practitioners' unmet effectiveness expectations, and a widening of the science–practice gap (Cascio & Aguinis, 2008).

Recommendations for reducing HARKing at the individual (i.e., author) level include promoting the application of strong inference testing (Leung, 2011). As noted by Edwards and Berry (2010), Leavitt et al. (2010), and Aguinis and Edwards (2014), increased application of strong inference is likely to foster scientific progress. Although individual solutions (e.g., research ethics education) may be intuitively appealing, such approaches are only marginally trustworthy in research environments wherein reward structures make HARKing a "rational choice" (Kerr, 1998, p. 213). In addition, such interventions are likely futile without corresponding structural changes in university performance management systems (Aguinis, Shapiro, Antonacopoulou, & Cummings, 2014; Tsui, 2013).

Suggestions for structural modifications are also numerous and exist at higher levels of the research community. For example, effects of HARKing might be addressed in professional codes of conduct (Colquitt, Kozlowski, Morgeson, Rogelberg, & Rupp, 2012; Kerr, 1998), such as those set forth by the Academy of Management and the American Psychological Association. Other promising solutions include a field's collective promotion of replication studies, decreasing the overemphasis on hypothesis and theory testing and legitimizing inductive research (Aguinis & Vandenberg, 2014), making HARKing a basis for manuscript rejection, legitimizing exploratory or descriptive research, delegitimizing admitted post hoc hypotheses (Kerr, 1998), and insisting on the use of registries in which study details are posted before being conducted (Brandt et al., 2014). Similarly, Leung (2011) argued that reviewers should resist negative reactions to nonsupported hypotheses. However, these approaches rely on policing, policy setting, and attitude change. Furthermore, if successful, these changes would ultimately require a great deal of time to be realized. We hope that the availability of our results will motivate professional organizations and journal editors to change policies addressing these issues.

We think that perhaps the most promising route to reducing HARKing lies in modifications to journals' manuscript peer review processes, perhaps the ultimate impetus for the researcher's choice to HARK. Indeed, as described earlier, manuscript reviewers react negatively to nonsupported hypotheses (Bedeian et al., 2010; Kerr, 1998). Kepes and McDaniel (2013) proposed that the peer review process proceed in two stages. In particular, preliminary editorial decisions (i.e., accept or reject) could be formed prior to reviewers' and editors' knowledge of results and discussion sections. The argument rests on the assumption that the purpose of the peer review process is to screen out poorly conducted or marginally relevant studies, not to judge whether the findings or conclusions are palatable to the prevailing zeitgeist. In addition, Schminke (2010) argued that data could be collected by authors after a conditional acceptance by a journal, resulting in less time wasted with flawed methodologies or less critical research questions. As another option, if a time lag between editorial decision and data collection was undesirable, we propose that results and discussion sections could be submitted simultaneously, in a separate password-protected document. In turn, following a favorable editorial decision, manuscript authors could submit the password.

Limitations and Directions for Future Research

Although we ascertained that HARKing occurred at typical rates in a recent sample of articles from Study 1, we remain uncertain as to the extent HARKing actually occurred in each of our analyzed sources across these two studies. In addition, we remain uncertain as to the proportion of findings that are not included by meta-analysts, which, if small, could suggest only a small HARKing effect. As one possible future research direction, researchers could consider investigating HARKing in environments where its detection is more certain, as in O'Boyle et al. (in press). Similar comparisons could be made by contrasting publications to their earlier conference papers or grant proposals. However, these approaches would provide more certain estimates of HARKing's prevalence, but they would not necessarily be informative regarding HARKing's relation with research findings and its downstream effects.

As a second limitation, Study 1 included nine distinct bivariate relations and, thus, our ability to control for alternative explanations within each of the nine relations was naturally limited by small sample sizes. Although this limitation was addressed in Study 2, our Study 1 findings, although they provide a glimpse of approximately how widespread HARKing's downstream effects might be, remain open to alternative explanations.

An anonymous reviewer noted that, for some articles, there may have been no reason to offer a hypothesis given the particular goals of the study. In other words, the argument is that the failure to offer a hypothesis may not be due to HARKing but dictated by the goals of the study. As noted by this anonymous reviewer, addressing this point requires answering the following question: "Given the substantive focus of the study, would one have expected authors to offer a particular hypothesis or not?" Clearly, the process of gathering data regarding this issue involves many ambiguities. For example, coders would have to read the articles and make a judgment call to determine the extent to which a hypothesis should or should not have been included based on the study's goals. Another possibility would be to conduct an ethnographic study in real time while a team of researchers is in the process of generating hypotheses for their study to understand the extent to which the researchers think a certain hypothesis is needed or not based on the study's goals. Both of these possibilities highlight the complexities in studying a sensitive topic such as HARKing and the need for future research involving novel designs and protocols.

In terms of additional research directions, future work could address an assessment of whether the difference between hypothesized versus nonhypothesized relations may be smaller for constructs whose predictive validity tends to be more generalizable. In other words, is it possible that there may be a greater opportunity for HARKing in domains with greater effect size variability? We conducted an initial assessment of this possibility by calculating $SD\rho$ values (i.e., an index of dispersion of the population estimates) for each of the 10 relations in our two studies. The *SDρ* values ranged from .06 to .19. We calculated the difference between unreliability-corrected hypothesized versus nonhypothesized correlations and then correlated them with *SDρ v*alues, resulting in $r = .26$. Although this correlation is not statistically significant given the small $k = 10$ and *t*-statistic with only 8 *df*, this result points to the possibility that there is more opportunity to HARK relations that are more variable across studies, and this issue could be examined in future research (a table with detailed results regarding this analysis is available from the authors).

Conclusion

To revisit the Annie Accommodator versus Penny Predictor debate, our research provides empirical evidence that, in contrast to Mill's (1843) perspective, the distinction between prediction (i.e., a priori hypothesizing) and accommodation (i.e., HARKing) is more than psychological. Indeed, HARKing appears to be more than a nuisance and, instead, poses a potential threat to research results, substantive conclusions, and practical applications. Specifically, effect sizes are larger when the focal variables are hypothesized to be related compared to when these same variables are not hypothesized to be related. We demonstrated this effect among 10 central relations in OBHRM and I-O psychology research: 247 effect sizes representing nine common bivariate relations with individual performance and 281 effect sizes representing the job satisfaction–employee performance relation while simultaneously ruling out 13 alternative explanations for the presumed effects of HARKing. Importantly, the magnitude of the difference in effect sizes is large in relation to typical effects reported in OBHRM and I-O psychology research (Bosco et al., 2015). Finally, in Study 2, variables included in hypothesized relations are more likely to appear in article titles or abstracts, demonstrating that HARKing has the potential to lead to potentially biased literature searches, thus threatening the validity of narrative and meta-analytic review findings and practitioner perceptions regarding the efficacy of evidence-based practice. We hope that our results will lead to increased awareness about the deleterious impact of HARKing, further research on this phenomenon, and the implementation of our proposed solutions to reduce or eliminate HARKing.

REFERENCES

Aguinis H. (2004). *Regression analysis for categorical moderators*. New York, NY: Guilford.

Aguinis H, Edwards JR. (2014). Methodological wishes for the next decade and how to make wishes come true. *Journal of Management Studies*, *51*, 143–174. doi: 10.1111/joms.12058

Aguinis H, Pierce, CA, Bosco, FA, Muslin, IS. (2009). First decade of *Organizational Research Methods*: Trends in design, measurement, and data-analysis topics. *Organizational Research Methods*, *12*, 69–112. doi: 10.1177/1094428108322641

Aguinis H, Vandenberg RJ. (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. doi:10.1146/annurev-orgpsych-031413-091231

Aguinis H, Sturman MC, Pierce CA. (2008). Comparison of three meta-analytic procedures for estimating moderating effects of categorical variables. *Organizational Research Methods*, *11*, 9–34. doi: 10.1177/1094428106292896

Aguinis H, Werner S, Lanza AJ, Angert C, Joon HP, Kohlhausen D. (2010). Customer-centric science: Reporting significant research results with rigor, relevance, and practical impact in mind. *Organizational Research Methods*, *13*, 515–539. doi: 10.1177/1094428109333339

Aguinis H, Dalton DR, Bosco FA, Pierce CA, Dalton CM. (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. doi: 10.1177/0149206310377113

Aguinis H, Shapiro DL, Antonacopoulou EP, Cummings TG. (2014). Scholarly impact: A pluralist conceptualization. *Academy of Management Learning & Education*, *13*, 623–639. doi: 10.5465/amle.2014.0121

Babyak MA. (2004). What you see may not be what you get: A brief, nontechnical introduction to overfitting in regression-type models. *Psychosomatic Medicine*, *66*, 411–421. doi: 10.1097/00006842-200405000-00021

Barrick MR, Mount MK. (1991). The Big Five personality dimensions and job performance: A meta-analysis. PERSONNEL PSYCHOLOGY, *44*, 1–26. doi: 10.1111/j.1744-6570.1991.tb00688.x

Barrick MR, Mount MK, Judge TA. (2001). Personality and performance at the beginning of the new millennium: What do we know and where do we go next? *International Journal of Selection and Assessment*, *9*, 9–30. doi: 10.1111/1468-2389.00160

Bedeian AG, Taylor SG, Miller AN. (2010). Management science on the credibility bubble: Cardinal sins and various misdemeanors. *Academy of Management Learning & Education*, *9*, 715–725. doi: 10.5465/amle.2010.56659889

Bem DJ. (2002). Writing the empirical journal article. In Darley JM, Zanna MP, & Roediger III HL (Eds.), *The compleat academic: A career guide* (pp. 3–26). Washington, DC: American Psychological Association.

Bernerth JB, Aguinis H. (in press). A critical review and best-practice recommendations for control variable usage. PERSONNEL PSYCHOLOGY. doi: 10.1111/peps.12103

Bosco FA, Aguinis H, Singh K, Field JG, Pierce CA. (2015). Correlational effect size benchmarks. *Journal of Applied Psychology*, *100*, 431–449. doi: 10.1037/a0038047

Brandt MJ, Ijzerman H, Dijksterhuis A, Farach FJ, Geller J, Giner-Sorolla R, . . . van ‘t Veer A. (2014). The replication recipe: What makes for a convincing replication? *Journal of Experimental Social Psychology*, *50*, 217–224. doi: http://dx.doi.org/10.1016/j.jesp.2013.10.005

Cascio WF, Aguinis H. (2008). Research in industrial and organizational psychology from 1963 to 2007: Changes, choices, and trends. *Journal of Applied Psychology*, *93*, 1062–1081. doi: 10.1037/0021-9010.93.5.1062

Colquitt JA, Kozlowski SWJ, Morgeson FP, Rogelberg SG, Rupp DE. (2012). *Journal editor ethics*. Retrieved June 10, 2012, from: http://editorethics.uncc.edu/

Crampton SM, Wagner JA. (1994). Percept-percept inflation in microorganizational research: An investigation of prevalence and effect. *Journal of Applied Psychology*, *79*, 67–76. doi: 10.1037/0021-9010.79.1.67

Dalton DR, Aguinis H, Dalton CM, Bosco FA, Pierce CA. (2012). Revisiting the file drawer problem in meta-analysis: An empirical assessment of published and nonpublished correlation matrices. PERSONNEL PSYCHOLOGY, *65*, 221–249. doi: 10.1111/j.1744-6570.2012.01243.x

De Vries, R, Anderson MS, Martinson BC. (2006). Normal misbehaviour: Scientists talk about the ethics of research. *Journal of Empirical Research on Human Research Ethics: An International Journal*, *1*, 43–50. doi: 10.1525/jer.2006.1.1.43

Edwards JR, Berry JW. (2010). The presence of something or the absence of nothing: Increasing theoretical precision in management research. *Organizational Research Methods*, *13*, 668–689. doi: 10.1177/1094428110380467

Fanelli D. (2009). How many scientists fabricate and falsify research? A systematic review and meta-analysis of survey data. *PLoS ONE*, *4*, e5738. doi: 10.1371/journal.pone.0005738

Fanelli D. (2010). "Positive" results increase down the hierarchy of the sciences. *PLoS ONE*, *5*, e10068. doi: 10.1371/journal.pone.0010068

Francis G, Tanzman J, Matthews WJ. (2014). Excess success for psychology articles in the journal Science. *PLoS ONE*, *9*, e114255. doi: 10.1371/journal.pone.0114255

Gardner MR. (1982). Predicting novel facts. *The British Journal for the Philosophy of Science*, *33*, 1–15. doi: 10.2307/687237

Hambrick DC. (2007). The field of management's devotion to theory: Too much of a good thing? *Academy of Management Journal*, *50*, 1346–1352. doi: 10.2307/20159476

Harker D. (2008). On the predilections for predictions. *British Journal for the Philosophy of Science*, *59*, 429–453. doi: 10.1093/bjps/axn017

Hitchcock C, Sober E. (2004). Prediction versus accommodation and the risk of overfitting. *British Journal for the Philosophy of Science*, *55*, 1–34. doi: 10.2307/3541832

Hubbard R, Armstrong JS. (1997). Publication bias against null results. *Psychological Reports*, *80*, 337–338. doi: 10.2466/pr0.1997.80.1.337

Hunter JE, Schmidt FL. (2004). *Methods of meta-analysis: Correcting error and bias in research findings* (2nd ed.). New York, NY: Academic Press.

John LK, Loewenstein G, Prelec D. (2012). Measuring the prevalence of questionable research practices with incentives for truth-telling. *Psychological Science*, *23*, 524–532. doi: 10.1177/0956797611430953

Judge TA, Thoresen CJ, Bono JE, Patton GK. (2001). The job satisfaction–job performance relationship: A qualitative and quantitative review. *Psychological Bulletin*, *127*, 376–407. doi: 10.1037/0033-2909.127.3.376

Kepes S, McDaniel MA. (2013). How trustworthy is the scientific literature in industrial and organizational psychology? *Industrial & Organizational Psychology*, *6*, 252–268. doi: 10.1111/iops.12045

Kepes S, McDaniel M, Brannick M, Banks G. (2013). Meta-analytic reviews in the organizational sciences: Two meta-analytic schools on the way to MARS (the meta-analytic reporting standards). *Journal of Business & Psychology*, *28*, 123–143. doi: 10.1007/s10869-013-9300-2

Kerr NL. (1998). HARKing: Hypothesizing after the results are known. *Personality & Social Psychology Review*, *2*, 196–217. doi: 10.1207/s15327957pspr0203_4

Knapp G, Hartung J. (2003). Improved tests for a random effects meta-regression with a single covariate. *Statistics in Medicine*, *22*, 2693–2710. doi: 10.1002/sim.1482

Köhler T, Cortina JM, Kurtessis JN, Gölz M. (in press). Are we correcting correctly? Interdependence of reliabilities in meta-analysis. *Organizational Research Methods*. doi: 10.1177/1094428114563617

Leavitt, K, Mitchell TR, Peterson J. (2010). Theory pruning: Strategies to reduce our dense theoretical landscape. *Organizational Research Methods*, *13*, 644–667. doi: 10.1177/1094428109345156

Leung K. (2011). Presenting post hoc hypotheses as a priori: Ethical and theoretical issues. *Management and Organization Review*, *7*, 471–479. doi: 10.1111/j.1740-8784.2011.00222.x

Lipton P. (2001). Inference to the best explanation. In Newton-Smith WH (Ed.): *A companion to the philosophy of science* (pp. 184–193). Malden, MA: Blackwell.

Lipton P. (2005). Testing hypotheses: Prediction and prejudice. *Science*, *307*, 219–221. doi: 10.2307/3840099

Mill JS. (1843). *A system of logic*. London, UK: Routledge.

O'Boyle EH, Banks GC, Gonzalez-Mulé E. (in press). The chrysalis effect: How ugly initial results metamorphosize into beautiful articles. *Journal of Management*. doi: 10.1177/0149206314527133

Orlitzky M. (2012). How can significance tests be deinstitutionalized? *Organizational Research Methods*, *15*, 199–228. doi: 10.1177/1094428111428356

Pfeffer J. (2007). A modest proposal: How we might change the process and product of managerial research. *Academy of Management Journal*, *50*, 1334–1345. doi: 10.2307/20159475

Platt JR. (1964). Strong inference. *Science*, *146*, 347–353. doi: 10.2307/1714268

Rodriguez MC, Maeda Y. (2006). Meta-analysis of coefficient alpha. *Psychological Methods*, *11*, 306–322. doi: 10.1037/1082-989x.11.3.306

Roos PA, Treiman DJ. (1980). Worker functions and worker traits for the 1970 U.S. census classification. In Miller AR, Treiman DJ, Cain PS, Roos PA (Eds.). *Work, jobs, and occupations: A critical review of the Dictionary of Occupational Titles* (pp. 336–389). Washington, DC: National Academy Press.

Sagie A, Koslowsky M. (1993). Detecting moderators with meta-analysis: An evaluation and comparison of techniques. PERSONNEL PSYCHOLOGY, *46*, 629–640. doi: 0.1111/j.1744-6570.1993.tb00888.x

Scandura TA, Williams EA. (2000). Research methodology in management: Current practices, trends, and implications for future research. *Academy of Management Journal*, *43*, 1248–1264. doi: 10.2307/1556348

Schminke M. (2010, October). *Enhancing research integrity: A modest proposal* [PowerPoint slides]. Presented at the annual conference of the Society for Organizational Behavior, Binghamton, NY.

Shadish WR, Cook TD, Campbell DT. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. New York, NY: Cengage.

Shaffer JA, Postlethwaite BE. (2012). A matter of context: A meta-analytic investigation of the relative validity of contextualized and noncontextualized personality measures. PERSONNEL PSYCHOLOGY, *65*, 445–494. doi: 10.1111/j.1744-6570.2012.01250.x

Smith PC, Kendall LM, Hulin CL. (1969). *The measurement of satisfaction in work and retirement*. Chicago, IL: Rand McNally.

Stajkovic AD, Luthans F. (1998). Self-efficacy and work-related performance: A meta-analysis. *Psychological Bulletin*, *124*, 240–261. doi: 10.1037/0033-2909.124.2.240

Steneck NH. (2006). Fostering integrity in research: Definitions, current knowledge, and future directions. *Science & Engineering Ethics*, *12*, 53–74. doi: 10.1007/PL00022268

Thorndike EL. (1917). The curve of work and the curve of satisfyingness. *Journal of Applied Psychology*, *1*, 265–267. doi: 10.1037/h0074929

Tsui AS. (2013). Editorial: The spirit of science and socially responsible scholarship. *Management and Organization Review*, *9*, 375–394. doi: 10.1111/more.12035

Vacha-Haase T. (1998). Reliability generalization: Exploring variance in measurement error affecting score reliability across studies. *Educational and Psychological Measurement*, *58*, 6–20. doi: 10.1177/0013164498058001002

Viechtbauer W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, *36*, 1–48.

Viechtbauer W, López-López JA, Sánchez-Meca J, Marín-Martínez F. (in press). A comparison of procedures to test for moderators in mixed-effects meta-regression models. *Psychological Methods*. doi: 10.1037/met0000023

Viswesvaran C, Ones DS, Schmidt FL. (1996). Comparative analysis of the reliability of job performance ratings. *Journal of Applied Psychology*, *81*, 557–574. doi: 10.1037/0021-9010.81.5.557

Weiss DJ, Dawis RV, England GW, Lofquist, LH. (1967). Manual for the Minnesota satisfaction questionnaire. *Minnesota Studies in Vocational Rehabilitation*, *22*. Retrieved from http://vpr.psych.umn.edu/assets/pdf/Monograph%20XXII%20-%20Manual%20for%20the%20MN%20Satisfaction%20Questionnaire.pdf

White R. (2003). The epistemic advantage of prediction over accommodation. *Mind*, *112*, 653–683. doi: 10.1093/mind/112.448.653