© *Academy of Management Discoveries*
2019, Vol. 5, No. 2, 103–108.
Online only
https://doi.org/10.5465/amd.2019.0121

FROM THE EDITOR

ON THE REPLICABILITY OF ABDUCTIVE RESEARCH IN MANAGEMENT AND ORGANIZATIONS: INTERNAL REPLICATION AND ITS ALTERNATIVES

In one of this journal's first "From the Editors" (FTE) pieces, Chet Miller and I emphasized the importance of replication research and the priority placed on such articles by *AMD*. As we noted in that piece (Miller & Bamberger, 2016), "reproducibility is at the heart of the scientific enterprise and critical to the development of any scientific field." We also noted that, despite the centrality of replication to science, replication in social science is rare (Makel, Plucker, & Hegarty, 2012). Indeed, numerous reports have emerged in recent years about a replication or credibility "crisis" (Baker, 2012; Bergh, Sharp, Aguinis, & Li, 2017; Carpenter, 2012), with some noting that less than half of the relationships reported to be significant at $p < .05$ in original studies are found to be significant in replication studies (Open Science Collaboration, 2015). Unfortunately, 2 years since that call for replication was published, this journal has received less than a handful of replication study submissions, and the number of replication studies published in other leading management journals can similarly be counted on one hand.

The absence of replication research is a serious problem for any field. However, the absence of replication for studies grounded on abductive reasoning and empirical exploration is particularly concerning as such studies offer "first suggestions" of plausible links and explanations grounded on rigorous observations, rather than confirmations of theory-grounded hypotheses. In this regard, unlike deductive reasoning which can be evaluated on the basis of the logical connections between past results and extant theory on the one hand, and *a priori* predictions on the other (Hollenbeck & Wright, 2017), it is impossible to assess abductive reasoning on the basis of its consistency with past results and extant theory. Although knowledge claims based on abduction are weaker than those associated with induction or deduction (Behfar & Okhuysen, 2018), reflecting inferences drawn strictly from observation, such claims can, for a variety of reasons still be misinterpreted as confirmed truth (Maxwell, Lau, & Howard, 2015; Tversky & Kahneman, 1971). This is problematic as such observations, subject to the influence of a wide variety of sample-, measurement-, and model-related artifacts, cannot be easily contrasted with theory-grounded expectations. Moreover, these "first suggestions," as novel as they may be, may amount to little more than serendipity. Most problematic of all, because such findings often are novel, they tend to become the focus of media attention.

Given such concerns, *AMD* has, since its founding, pursued *reproducibility* ("the ability of other researchers to obtain the same results when they reanalyze the same data" [Kepes, Bennett, & McDaniel, 2014: 456]), largely by setting a high bar for methodological transparency and rigor. It has also striven for replicability (the ability to obtain the same pattern of findings in a separate sample drawn from the same [and/or different] population using the same [and/or different] procedures; Aguinis & Solarino, 2019) by encouraging authors to make every effort to demonstrate that their findings are not the result of some sample- or method-based artifact, and even occasionally requesting that authors of studies grounded on abduction consider self-replicating their work. Such internal replication or the "bundling of studies," with each sequentially building off the last, is common in other fields such as psychology (King, Goldfarb, & Simcoe, 2019). Accordingly, in this *FTE*, I wish to discuss the merits of self-replication in abductive research and—recognizing the concerns and challenges associated with such requests—offer some alternative avenues for addressing issues of replicability in abductive research.

THE CENTRALITY OF INTERNAL REPLICATION IN ABDUCTION (AND INDUCTION)

Most of the research on management and organizations appearing in scholarly journals remains consistent with the hypothetico-deductive model in that it is theory-grounded, with theory informing both deductive and inductive inquiries. In the case of deductive inquiries, the aim is to reason from the general to the specific and test with certainty whether a general principle, or some corollary drawn from it, explains a phenomenon in a sample drawn at random from the population to which that principle pertains. In the case of inductive inquiry, we seek to extend extant theory by gleaning insight into probable

underlying mechanisms and conditioning factors on the basis of reasoning from the specific to the general (Bamberger, 2018), using contrasts and comparisons to better understand the "why," "how," and "when" of the general principle underlying the inquiry.

In contrast to these two classic modes of inquiry, abductive reasoning is grounded on the principle of generating plausible, "first suggestions" about phenomena and their explanations on the basis of observations from one's data (Heckman & Singer, 2017; Peirce, 1883). Whereas, as with classic induction, abduction operates from the specific to the general, it is more consistent to what Leamer (1983) calls the Sherlock Holmes methodology, governed by the principle that "[i]t is a capital mistake to theorize before you have all the evidence" (Doyle, 1891). Accordingly, as noted by King et al. (2019: 12), "abduction provides nothing more than a means of 'guessing', and that the only truth claim that can be made is that a proposed supposition is a plausible one. To know anything further, this supposition must be "subject to further test" (Schurz, 2008)."

Given that abduction aims to make only the weakest of knowledge claims — surfacing a phenomenon that plausibly fails to meet the defining criteria for an extant construct or offering the "loveliest" (Lipton, 2004: 59), albeit merely plausible, explanation for some phenomena or a relationship poorly explained by the extant theory — one might argue that such further tests should come "down the road." And indeed, in the paper development workshops that my fellow editors and I have conducted around the world, we have often framed *AMD* as "pre-*AMR*" and "pre-pre-*AMJ*," arguing that *AMD* papers should offer criteria for down-the-road theorizing, with nascent theoretical insights flushed out in *AMR* and then tested in *AMJ*. However, papers submitted to these two outlets are generally very time-consuming to develop and subject to extremely high risk, with only the smallest proportion of submissions ultimately accepted for publication. Accordingly, any attempt to reduce that risk by moving the dial from "just plausible" toward "seemingly probable" within an *AMD* paper may be both prudent and laudable.

Internal replication offers one potential means of doing that, typically by engaging in some preliminary testing (yes, on the basis of the hypothetico-deductive model) of inferences drawn from an initial dataset. To the extent that evidence is found in a separate dataset supportive of these initial inferences, one could argue that the original inferences not only offer the loveliest plausible explanation, they also offer an explanation that we can accept with a somewhat greater degree of confidence. In this regard, the incorporation of internal replication into an abductive study can be seen as following Glaser and Strauss' (1967) notion of "constant comparison," the basis of grounded theory. Indeed, when applied to qualitative research, such comparison often implicitly encompasses empirical and conceptual replication as researchers use theoretical sampling to see where and when the same patterns emerge (and when and where they do not) using the same methods in a different population (empirical replication) or different methods with the same population (conceptual replication) (Aguinis & Solarino, 2019). Similarly, with regard to quantitative abduction, internal replication calls for testing whether inferences from one sample or using one set of procedures may be similarly drawn when applied to another independent but similar sample or when using a different set of procedures, and perhaps even going one or two steps further, pushing the boundaries and exploring the points at which replication is no longer possible.

Incorporating internal replication into abductive research offers a number of benefits. First, evidence that the same basic pattern observed in an abductive search is found in an independent dataset using some form of rigorous hypothesis testing boosts confidence that the initial findings are veridical and not simply serendipitous or artifactual. Such evidence is particularly meaningful in light of reports suggesting that many studies in strategic management fail to report information enabling replication (Aguinis & Solarino, 2019; Bergh et al., 2017) and that between a third and a half of significant findings in both macro (Bergh et al., 2017) and micro (Open Science Collaboration, 2015) research fail to replicate.

Second, internal replication may serve as a viable antidote to what Bliese and Wang (2019) term "origination bias," or in other words, "the practice of viewing findings from a single, original study as being almost sacred," even if these findings were exploratory in nature. Indeed, Francis (2012: 593) claims that, "there appears to be a tendency to believe that once an effect has been shown to be statistically significant, then its truth has been established." Origination bias makes replication research a highly risky venture, as successful replications tend to be rejected for lack of novelty, whereas failed replications tend to be rejected because they are demonstrating the null. In fact, Bettis (2012: 109) writes that, "professional norms generally preclude publication of replication studies and what are usually called 'non-results'." As external replication is therefore rare and generally not expected, Popper's notion of falsifiability (i.e., that failed confirmation indicates problematic or false premises) is largely ineffective as a mode of protection from the scientific equivalent of "fake news." Internal replicability provides a partial solution to

this problem, offering the consumers of research some assurance that the premises suggested by the observed patterns are indeed falsifiable and subject to fact-checking.

Third, because findings generated on the basis of abductive reasoning often emerge from studies in which the reported phenomena were not the intended focus of the initial inquiry, insufficient statistical power may heighten the risk of not detecting important conditioning effects or explanatory mechanisms. Internal replication, particularly if performed on the basis of a substantially larger sample size, may at least partially resolve this problem, not only reducing the risk of Type I error but in the process also lowering the rate of Type II errors (Hollenbeck & Wright, 2017; Maxwell et al., 2015).

Finally, as noted above, internal replication need not be exact (i.e., using identical measures to test inferences drawn from one sample on a separate, random sample drawn from the same population). Indeed, in the same way that many qualitative studies appraise the fit of an inference across a theoretical sample of participants, often using a variety of empirical approaches such as interviews, observations, and archival data, so may internal replication in quantitative research be conducted on the basis of this more playful, trial-and-error approach using empirical and conceptual replication, with the aim being to better understand where, when, and/or with respect to whom the inferences apply. That is, internal replication can be applied in an explorative, transparent manner to assess the internal and external validity of the nascent theoretical relationships emanating from the initial findings (Aguinis, Ramani, & Alabduljader, 2018).

CHALLENGES IN AND LIMITATIONS TO INTERNAL REPLICATION

Despite these benefits, internal replication is not without its challenges and limitations. First, it is possible, if not highly probably, that even an exact replication of the initial, exploratory study will fail to demonstrate the replicability of that study's central, significant finding, particularly if the number of observations on which that finding is based is limited (Maxwell et al., 2015). This is not surprising, given that studies with larger effects or sample sizes have more power (Francis, 2012). Indeed, Bliese and Wang (2019) demonstrate that even when $p < .05$, if the sample size is small, the probability of finding a similarly significant effect in a second, independent sample drawn at random from that same population may not be much greater than 50 percent. Furthermore, to the extent that a replication sample has an N no larger than the initial, exploratory study, there is little reason to believe that it will offer a more precise estimate of the true effect size (Maxwell et al., 2015). In fact, as noted by Francis (2012: 593), "when experiments have low or moderate power, there should frequently be experimental findings that fail to replicate a result, even if the effect is true." This is why, as noted by Hollenbeck and Wright (2017: 13), the magnitude and significance of most meta-analytically derived relationships is greater "relative to what was suggested in the original underpowered studies," and why Maxwell et al. (2015: 495) counsel that, "just as it may be unwise to consider a single original study as definitive, it may also be unwise to regard a single replication study as providing the final word."

A second limitation of exact, internal replications (i.e., replications applying the same measures and identical model specifications on an independent sample drawn at random from the same population as the one used in the initial exploratory study) is that they offer little insight into the internal or external validity of the novel finding beyond that offered by the initial, exploratory study. Because an exact replication applies the identical model specification and measures, it cannot rule out or make unlikely alternative explanations of the results, and, thus, is a fairly useless tool by which to assess internal validity. In addition, as the replication is on a sample drawn in the same way from the same population, it also provides no basis for assessing the finding's external validity.

Finally, although internal replication is an almost inherent element of experimental research and is increasingly popular with micro-oriented, descriptive field studies, it may be less feasible for researchers exploring expensive, one-off, archival datasets, or those whose findings are generated from longitudinal data collected over several months or years. For instance, how would one replicate findings regarding the emergence of a particular organizational form unique to 18th-century complex organizations generated from a one-of-a-kind, archival dataset of German railroads? Similarly, how realistic is it to expect scholars to replicate findings regarding the dynamic association between the retirement process and alcohol misuse generated on the basis of a study tracking older workers over 10–15 years as they transitioned through the various phases of retirement?

IMPLICATIONS FOR THOSE SEEKING TO PUBLISH IN *AMD*

The challenges to and limitations of replication noted previously are considerable (Maxwell et al., 2015). Accordingly, although *AMD* editors may in some cases suggest that authors consider internally replicating their study, failed replication will

never—on its own—serve as grounds for rejection. That does not mean that authors adopting this strategy should do so sloppily. Rather, as I note below, authors interested in pursuing internal replication may take a number of steps to improve the likelihood of replication and/or enrich the insights that might be drawn from such an effort. Moreover, as I also detail in the following paragraphs, authors may want to consider offering an important alternative to internal replication, namely, a simple but compelling discussion of the replicability of their findings.

Internal Replication

Authors of quantitative studies executing internal replications should try to ensure that the replication sample size is considerably larger than the sample studied in the initial exploratory analysis. This is because, as suggested earlier, the probability of replicability increases as a function of the estimate's associated *t* value. In addition, as noted by Bliese and Wang (2019), "standard errors decrease as sample sizes increase, so for a sample of 500 to generate a *t* value of (e.g.,) 2.149, the effect size observed from this sample has to be larger than the effect size from a sample of 5,000 that also generates a *t* value of 2.149." Simply put, conducting the replication on the basis of a larger sample size reduces the risk of Type II error or, in other words, the likelihood that the results of the first study will incorrectly be deemed serendipitous. This is particularly important in exploratory research where, as a result of not necessarily knowing what we are looking for, it is extremely difficult to take sample size and power considerations into account *a priori*.

Second, authors of quantitative studies should consider supplementing their initial study with a follow-up study that not only replicates the initial, exploratory finding but extends that finding, what Bettis, Helfat, and Shaver (2016: 2195) refer to as "quasi-replication." That is, beyond directly replicating their initial finding, authors may wish to incorporate into their follow-up design variables, measures, or manipulations aimed at ruling out alternative explanations and/or setting the basis for mid-range theorizing. For instance, by incorporating variables tapping alternative explanations and/or potential mediators and moderators, authors may both enhance the internal validity of their findings and offer a higher value-added contribution to theories of management and organization. In addition, (or instead), authors may want to demonstrate the robustness of their findings across different groups or contexts, thus enhancing the external validity of their initial, exploratory findings. Authors replicating with an eye toward external validity should draw from the notion of theoretical sampling, at the core of grounded theory (Glaser & Strauss, 1967). Using this approach, authors should offer a theory-grounded explanation for their choice in replication samples, recognizing that the transparent reporting of replication failure may be just as important as that of replication success as a basis for specifying theoretical boundary conditions.

In those cases in which the nature of the initial exploratory sample is such that conventional replication is not feasible (e.g., data are costly to obtain or come from a one-off, archival source), authors may want to consider the use of a "holdout" sample (Han, Kamber, & Pei, 2006), applying what Camstra and Boomsma (1992) refer to as cross-validation. Using this approach, researchers engaged in nonexperimental exploration split their sample using (ideally) a smaller "training" or calibration subsample as the basis for their exploratory analyses and then test any emergent inferences on the basis of their (ideally) larger holdout or validation sample. As Camstra and Boomsma (1992) note, there is no limitation on the number of validation samples one sets aside as long as the original dataset is large enough to support such multiple partitioning. However, it should be noted that unless one's initial sample is quite large, by parceling the sample, one effectively reduces its size, thereby increasing sampling error (Nunnally, 1978) and reducing the likelihood of replication.

Most importantly, regardless of the nature of the internal replication, authors are encouraged to be open and transparent about their findings. When replication fails to generate a significant effect, authors should make every effort to attempt to diagnose the "failure," recognizing that the lessons learned from such "failures" allow us to narrow the range of alternative explanations, and, hence, are central to abductive reasoning. In that regard, failed replications at *AMD* will be assessed on the basis of how compelling reviewers find (a) the inferences authors logically infer from their pattern of results and (b) the implications that they draw from them for down-the-road theorizing.

Discussions of Replicability as an Alternative

For quantitative studies grounded on abductive reasoning, a discussion of replicability can be helpful in that it can provide readers with a sense of where the observed effect falls on that imaginative dial noted earlier ranging from logically plausible to statistically probable. To the extent that a stronger case can be made for the potential replicability of a study's primary effects, others may be more motivated to engage in theorizing around that finding,

generating testable hypotheses grounded on such theorizing, and actually submitting those hypotheses to rigorous testing. To the extent that replicability is currently deemed low but insights are offered as to why that may be the case, others may be motivated to engage in further empirical exploration aimed at developing more sensitive instruments or more robust study designs, or determining the factors potentially conditioning the effect sizes.

Authors of quantitative studies may draw an inference of replicability on the basis of the *t* value associated with their estimate, recognizing that when the significance (i.e., *p*) associated with the *t* value is just a tad below 0.05, the probability of replicability is just a tad above 50 percent (Hoenig & Heisey, 2001). Bliese and Wang (2019) offer an alternative option grounded on resampling (i.e., bootstrapping). Using this approach, the researcher conducts multiple (indeed thousands of) direct replications, replacing random dropouts from the original sample with replacements also drawn at random from the original sample, thus "presumably holding everything constant except for minor variations in the subject pool" (Bliese & Wang, 2019). Unlike bootstrapping when applied to models of mediation (where we estimate the confidence interval associated with a particular model parameter), in this case, the bootstrap program is modified to count, "how often an effect is significant (e.g., has a *p* value less than .05)." The count of significant results (relative to the total number of resamples estimated) provides what Bliese and Wang (2019) refer to as a "*post hoc* estimate of the likelihood of replication," neatly doing so on the basis of the original data. Although bootstrapping is not without its critics (for a detailed critique, see Koopman, Howe, Hollenbeck and Sin [2015]), this approach offers a rather straightforward and intuitive means by which to ground a discussion of potential replicability.

Authors of qualitative studies may, where applicable, also wish to discuss the replicability of their findings. Indeed, Aguinis and Solarino (2019) offer some helpful insights on the replicability of qualitative research and how qualitative scholars authors may wish to demonstrate replicability in their own research. However, the relevance of replicability and just what it means for the kind of phenomenological and abductive qualitative research often published in *AMD* remains a more open question. Accordingly, although qualitative scholars are welcome and even encouraged to engage in a discussion of replicability as part of their broader discussion of the implications of their findings for down-the-road theorizing, we offer no specific "boilerplate" that they should feel obligated to follow.

CONCLUSION

Increased awareness of the challenges to our science posed by HARKing, p-hacking, and the absence of replication research (Bettis, 2012) has led an increasing number of scholars to accept the value of exploratory research grounded on abductive reasoning (Hollenbeck & Wright, 2017). However, as I have argued previously, the adoption of an abductive frame does not necessarily release researchers from the need to engage in a transparent discussion of the potentially serendipitous nature of their findings. Where *p* values are well below .01 and *t* values are well above 3, the likelihood of direct replicability is by definition quite high (Bliese & Wang, 2019). Yet, even in these cases, those grounding their analysis on abductive reasoning should go beyond simply reporting their significant effect and offer a compelling discussion of its replicability or perhaps even consider an internal replication. Indeed, as we have noted, such internal replication can serve as a useful means by which to demonstrate not merely the robustness of one's "discovery," but offer plausible insight into its potential boundaries and/or underlying mechanisms as well.

Peter A. Bamberger
Tel Aviv University

REFERENCES

Aguinis, H., Ramani, R. S., & Alabduljader, N. 2018. What you see is what you get? Enhancing methodological transparency in management research. ***Academy of Management Annals***, 12: 83–110.

Aguinis, H., & Solarino, A. M. 2019. Transparency and replicability in qualitative research: The case of interviews with elite informants. ***Strategic Management Journal***, 1–25. https://doi.org/10.1002/smj.3015.

Baker, M. 2012 Independent labs to verify high-profile papers. ***Nature***, doi:10.1038/nature.2012.11176. Available at: https://www.nature.com/news/independent-labs-to-verify-high-profile-papers-1.11176.

Bamberger, P. A. 2018. AMD—Clarifying what we are about and where we are going. ***Academy of Management Discoveries***, 4: 1–10.

Behfar, K., & Okhuysen, G. A. 2018. Perspective—Discovery within validation logic: Deliberately surfacing, complementing, and substituting abductive reasoning in hypothetico-deductive inquiry. ***Organization Science***, 29(2): 323–340.

Bergh, D. D., Sharp, B. M., Aguinis, H., & Li, M. 2017. Is there a credibility crisis in strategic management

research? Evidence on the reproducibility of study findings. ***Strategic Organization***, 15(3): 423–436.

Bettis, R. A. 2012. The search for asterisks: Compromised statistical tests and flawed theories. ***Strategic Management Journal***, 33(1): 108–113.

Bettis, R. A., Helfat, C. E., & Shaver, J. M. 2016. The necessity, logic, and forms of replication. ***Strategic Management Journal***, 37(11): 2193–2203.

Bliese, P. D., & Wang, M. 2019. ***What results from 10,000 alternative universes reveal about replicability, observed power and origination bias***. Working Paper.

Camstra, A., & Boomsma, A. 1992. Cross-validation in regression and covariance structure analysis: An overview. ***Sociological Methods & Research***, 21(1): 89–115.

Carpenter, S. 2012. Psychology's bold initiative. ***Science*** 335: 1558–1561.

Doyle, A. C. 1891. ***Sherlock Holmes: A Scandal in Bohemia***. London: G. Newnes.

Francis, G. 2012. The psychology of replication and replication in psychology. ***Perspectives on Psychological Science***, 7(6): 585–594.

Glaser, B. G., & Strauss, A. L. 1967. ***Discovery of grounded theory: Strategies for qualitative research***. Chicago: Aldine

Han, J., Kamber, M., & Pei, J. 2006. ***Data mining: Concepts techniques***. San Francisco, CA: Morgan Kaufmann Publishers.

Heckman, J. J., & Singer, B. 2017. Abducting economics. ***American Economic Review***, 107(5): 298–302.

Hoenig, J. M., & Heisey, D. M. 2001. The abuse of power: The pervasive fallacy of power calculations for data analysis. ***The American Statistician***, 55(1): 19–24.

Hollenbeck, R., & Wright, P. M. 2017. Harking, sharking, and tharking: Making the case for post hoc analysis of scientific data. ***Journal of Management***, 43(1): 5–18.

Kepes S., Bennett A., & McDaniel M. 2014. Evidence-based management and the trustworthiness of cumulative scientific knowledge: Implications for teaching, research and practice. ***Academy of Management Learning and Education***, 13: 446–466.

King, A., Goldfarb, B., & Simcoe, T. 2019. ***Learning from testimony on quantitative research in management***. Working Paper.

Koopman, J., Howe, M., Hollenbeck, J. R., & Sin, H. P. 2015. Small sample mediation testing: Misplaced confidence in bootstrapped confidence intervals. ***Journal of Applied Psychology***, 100(1): 194–202.

Leamer, E. E. 1983. Let's take the con out of econometrics. ***The American Economic Review***, 73(1): 31–43.

Lipton, P. 2004. ***Inference to the best explanation*** (2nd ed.). London: Routledge.

Makel, M. C., Plucker, J. A., & Hegarty, B. 2012. Replications in psychology research: How often do they really occur? ***Perspectives on Psychological Science***, 7(6): 537–542.

Maxwell, S. E., Lau, M. Y., & Howard, G. S. 2015. Is psychology suffering from a replication crisis? What does "failure to replicate" really mean? ***American Psychologist***, 70(6): 487–498.

Miller, C. C., & Bamberger, P. 2016. Exploring emergent and poorly understood phenomena in the strangest of places: The footprint of discovery in replications, meta-analyses, and null findings. ***Academy of Management Discoveries***, 2(4): 313–331.

Nunnally, J. C. 1978. ***Psychometric theory*** (2nd ed.). New York: McGraw-Hill.

Open Science Collaboration. 2015. Estimating the reproducibility of psychological science, ***Science***, 349: 1–8.

Peirce, C. S. 1883. A theory of probable inference. In ***The Johns Hopkins studies in logic***: 126–181. Boston, MA: Little, Brown and Company.

Schurz, G. 2008. Patterns of abduction. ***Synthese***, 164(2): 201–234.

Tversky, A., & Kahneman, D. 1971. Belief in the law of small numbers. ***Psychological Bulletin***, 76: 105–110.