European Journal of Personality, *Eur. J. Pers.* **27**: 120–144 (2013)
Published online in Wiley Online Library (wileyonlinelibrary.com) **DOI**: 10.1002/per.1920

OPEN PEER COMMENTARY

Dwelling on the Past

MARJAN BAKKER[1], ANGÉLIQUE O. J. CRAMER[1], DORA MATZKE[1], ROGIER A. KIEVIT[2], HAN L. J. VAN DER MAAS[1], ERIC-JAN WAGENMAKERS[1], DENNY BORSBOOM[1]

[1]*University of Amsterdam*
[2]*Medical Research Council, Cognition and Brain Sciences Unit, Cambridge, UK*
M.Bakker1@uva.nl

Abstract: We welcome the recommendations suggested by Asendorpf et al. Their proposed changes will undoubtedly improve psychology as an academic discipline. However, our current knowledge is based on past research. We therefore have an obligation to 'dwell on the past'; that is, to investigate the veracity of previously published findings—particularly those featured in course materials and popular science books. We discuss some examples of staple 'facts' in psychology that are actually no more than hypotheses with rather weak empirical support and suggest various ways to remedy this situation. Copyright © 2013 John Wiley & Sons, Ltd.

We support most of the proposed changes of Asendorpf et al. in the *modus operandi* of psychological research, and, unsurprisingly perhaps, we are particularly enthusiastic about the idea to separate confirmatory from exploratory research (Wagenmakers, Wetzels, Borsboom, Van der Maas, & Kievit, 2012). Nevertheless, perhaps we disagree with Asendorpf et al. on one point. Asendorpf et al. urge readers not to dwell '. . .on suboptimal practices in the past'. Instead, they advise us to look ahead: 'We do not seek here to add to the developing literature on identifying problems in current psychological research practice. [. . .] we address the more constructive question: How can we increase the replicability of research findings in psychology now?'

Although we do not want to diminish the importance of adopting the measures that Asendorpf et al. proposed, we think that, as a field, we have the responsibility to look back. Our knowledge is based on findings from work conducted in the past, findings that textbooks often tout as indisputable fact. Recent expositions on the methodology of psychological research reveal that these findings are based at least in part on questionable research practices (e.g. optional stopping, selective reporting, etc.). Hence, we cannot avoid the question of how to interpret past findings: Are they fact, or are they fiction?

Replications of the past

How can we evaluate past work? As Asendorpf et al. proposed, direct replication, possibly summarized in a meta-analysis, is one of the best ways to test whether an empirical finding is fact rather than fiction. Unfortunately, direct replication of findings is still uncommon in the psychological literature (Makel, Plucker, & Hegarty, 2012), even when it comes to textbook-level 'facts'.

For example, one area in psychology that has recently come under scrutiny is that of behavioural priming research (Yong, 2012). In one of the classic behavioural priming studies, Bargh, Chen, and Burrows (1996) showed that participants who were primed with words that supposedly activated elderly stereotypes walked more slowly than participants in the control condition. The Bargh et al. study is now cited over 2000 times and is described in various basic textbooks on (social) psychology, where it often has the status of fact (Augoustinos, Walker, & Donaghue, 2006; Bless, Fiedler, & Strack, 2004; Hewstone, Stroebe, & Jonas, 2012). However, only two relatively direct (but underpowered) replications had been performed, producing inconclusive results (Cesario, Plaks, & Higgins, 2006; Hull, Slone, Meteyer, & Matthews, 2002). Hull et al. (2002) found the effect in two studies, but only for highly self-conscious individuals. Cesario et al. (2006) established a partial replication in that some but not all of the experimental conditions showed the expected effects. Two more recent, direct, and well-powered replications failed to find the effect (Doyen, Klein, Pichon, & Cleeremans, 2012; Pashler, Harris, & Coburn, 2011).

As another example, imitation of tongue gestures by young infants is mentioned in many recent books on developmental psychology (e.g., Berk, 2013; Leman, Bremner, Parke, & Gauvain, 2012; Shaffer & Kipp, 2009; Siegler, DeLoache, & Eisenberg, 2011), and the original study by Meltzoff and Moore (1977) is cited over 2000 times. However, the only two direct replications (Hayes and Watson, 1981; Koepke, Hamm, Legerstee, & Rusell, 1983) failed to obtain the original findings, and a review by Anisfeld (1991) showed inconclusive results.

Even when some (approximately) direct replication studies are summarized in meta-analysis, we cannot be sure about the presence of the effect, as the meta-analysis may be contaminated by publication bias (Rosenthal, 1979) or the use of questionable research practices (John, Loewenstein, & Prelec, 2012; Simmons, Nelson, & Simonsohn, 2011). For example, many recent textbooks in developmental psychology state that infant habituation is a good predictor of later IQ (e.g., Berk, 2013; Leman, Bremner, Parke, & Gauvain, 2012; Shaffer & Kipp, 2009; Siegler, DeLoache, & Eisenberg, 2011), often referring to the meta-analysis of McCall and Carriger (1993). However, this meta-analysis suffers from publication bias (Bakker, van Dijk, & Wicherts, 2012). At best, these results point to a weak relation between habituation and IQ, and possibly to no relation at all.

Using replications to distinguish fact from fiction is important beyond the realms of scientific research and education. For instance, the (in)famous Mozart effect (Rauscher, Shaw, & Ky, 1993) suggested a possible 8–9 IQ point improvement in spatial intelligence after listening to classical music. Yet despite increasingly definite null replications dating back to 1995 (e.g., Newman et al., 1995; Pietschnig, Voracek, & Formann, 2010), the Mozart effect persists in the popular imagination. Moreover, the Mozart effect was the basis of a statewide funding scheme in Georgia (Cromie, 1999), trademark applications (Campbell, 1997), and children's products; for instance, Amazon.co.uk lists hundreds of products that use the name 'The Mozart Effect', many touting the 'beneficial effects on the babies brain'. Clearly, in addition to the scientific resources spent establishing whether the original claim was true, false-positive findings can have a long-lasting influence far outside science even when the scientific controversy has largely died down.

Textbook-proof

The studies discussed earlier highlight that at least some 'established findings' from the past are still awaiting confirmation and may very well be fictional. To resolve this situation, we need to dwell on the past, and several courses of action present themselves. First, psychology requires thorough examination, for example by an American Psychological Association taskforce, to propose a list of psychological findings that feature at the textbook level but in fact are still in need of direct replication. In a second step, those findings that are in need of replication can be reinvestigated in research that implements the proposals of Asendorpf et al. The work initiated by the Open Science Framework (http://openscienceframework.org/) has gone a long way in constructing a methodology to guide massive replication efforts and can be taken as a blueprint for this kind of work.

Psychology needs to improve its research methodology, and the procedures proposed by Asendorpf et al. will undoubtedly contribute to that goal. However, psychology also cannot avoid the obligation to look back and to find out which studies are textbook-proof and which are not. By implementing sensible procedures to further the veracity of our empirical work, psychologists have the opportunity to lead by example, an opportunity that we cannot afford to miss.

Minimal Replicability, Generalizability, and Scientific Advances in Psychological Science

JOHN T. CACIOPPO AND STEPHANIE CACIOPPO

University of Chicago
Cacioppo@uchicago.edu

Abstract: With the growing number of fraudulent and nonreplicable reports in psychological science, many question the replicability of experiments performed in laboratories worldwide. The focus of Asendorpf and colleagues is on research practices that investigators may be using to increase the likelihood of publication while unknowingly undermining replicability. We laud them for thoughtful intentions and extend their recommendations by focusing on two additional domains: the structure of psychological science and the need to distinguish between minimal replicability and generalizability. The former represents a methodological/statistical problem, whereas the latter represents a theoretical opportunity. Copyright © 2013 John Wiley & Sons, Ltd.

Although cases of outright fraud are rare and not unique to psychology, psychological science has been rocked in the past few years by a few cases of failed replications and fraudulent science. Among practices suggested by Asendorpf et al. as contributing to these outcomes are data selection and formulating decisions about sample size on the basis of statistical significance rather than statistical power. We laud Asendorpf et al. for their thoughtful and timely recommendations and hope their paper becomes required reading. We focus here on two domains they did not address: the structure of psychological science and the need to distinguish between minimal replicability and generalizability.

Publication of a new scientific finding should be viewed more as a promissory note than a final accounting. Science is not a solitary pursuit; it is a social process. If a scientific finding cannot be *independently* verified, then it cannot be regarded as an empirical fact. Minimal replicability, defined as an empirical finding that can be repeated by an independent investigator using the same operationalizations, situations, and time points in an independent sample of participants, is the currency of science.

Asendorpf et al. distinguish among reproducibility (duplication by an independent investigator analysing the same dataset), replicability (observation with other random samples), and generalizability (absence of dependence on an originally unmeasured variable). Issues of replicability and generalizability have been addressed before in psychology. Basic psychological research, with its emphasis on experimental control, was once criticized for yielding statistically reliable but trivial effects (e.g., Appley, 1990; Staats, 1989). Allport (1968) decades ago noted that scientific gains result from this hard-nosed approach, but he lamented the lack of generalizing power of many neat and elegant experiments: 'It is for this reason that some current investigations seem to end up in elegantly polished triviality—snippits of empiricism, but nothing more' (p. 68).

Many psychological phenomena, ranging from attention to racism, are multiply determined (Schachter,

DOI: 10.1002/per