Good science, bad science: Questioning research practices in psychological research

Bakker, M.

Appendix C

Dwelling on the past

We welcome the recommendations suggested in the target article by Asendorpf et al. (2013). The proposed changes in methodology 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 prominent enough to feature in course materials and popular science books. We discuss some examples to show that some of psychology's staple "facts" are actually no more than hypotheses with rather weak empirical support. We suggest various ways to remedy this situation.

This Appendix is published as: Bakker, M., Cramer, A. O. J., Matzke, D., Kievit, R. A., Van der Maas, H. L. J., Wagenmakers, E. J., Borsboom, D. (2013). Dwelling on the past (comment on Asendorpf et al.; 2013). *European Journal of Personality, 27,* 120-121. doi: 10.1002/per.1920

We support most of Asendorpf et al. (2013) proposed changes in the *modus operandi* of psychological research and, unsurprisingly perhaps, we are particularly enthusiastic about the idea to separate confirmatory from exploratory research (Wagenmakers et al., 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 detract from the importance of taking the measures that Asendorpf et al. propose, we also 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 (QRPs; 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. (2013) propose, 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 et al., 2012), even when it comes to textbook-level "facts".

For example, one area in psychology that has recently come under scrutiny is that of behavioral priming research (Yong, 2012). In one of the classic behavioral priming studies, (Bargh, Chen, & 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 gets 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 done, producing inconclusive results[21] (Cesario, Higgins, & Plaks, 2006; Hull, Slone, Meteyer, & Matthews, 2002), until two more recent, direct and well-powered replications failed to find the effect (Doyen, Klein, Pichon, & Cleeremans, 2012; Pashler, Harris, & Coburn, 2011).

[21] Hull et al. (2002) found the effect in two studies, but only for high 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.

As another example, the imitation of tongue gestures by young infants is mentioned in many recent books on developmental psychology (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 & Watson, 1981; Koepke, Hamm, Legerstee, & Russell, 1983) failed to replicate the original findings and a review by Anisfeld (1991) showed inconclusive results.

Even when some (approximately) direct replication studies are summarized in a 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 QRPs (John et al., 2012; Simmons et al., 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 et al., 2012; Shaffer & Kipp, 2009; Siegler et al., 2011), often referring to the meta-analysis of McCall and Carriger (1993). However, this meta-analysis suffers from publication bias (Bakker et al., 2012). At best, these results point to a weak relation between habituation and IQ, and possibly to no relation at all.

Using replications to distinguishing 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 (D. Campbell, 1997), and childrens 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 long-lasting influence far outside science even when the scientific controversy has largely died down.

Textbook-proof

The studies discussed above 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 a thorough examination, for example by an APA 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. (2013) and others. The work initiated by the Open Science Framework (OSF; http://opensciencefframework.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 improve its research methodology, and the procedures proposed by Asendorpf et al. (2013) 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.