European Journal of Personality, *Eur. J. Pers.* **27**: 108–119 (2013)
Published online in Wiley Online Library (wileyonlinelibrary.com) **DOI**: 10.1002/per.1919

Recommendations for Increasing Replicability in Psychology†

JENS B. ASENDORPF[1]*, MARK CONNER[2], FILIP DE FRUYT[3], JAN DE HOUWER[4], JAAP J. A. DENISSEN[5], KLAUS FIEDLER[6], SUSANN FIEDLER[7], DAVID C. FUNDER[8], REINHOLD KLIEGL[9], BRIAN A. NOSEK[10], MARCO PERUGINI[11], BRENT W. ROBERTS[12], MANFRED SCHMITT[13], MARCEL A. G. VANAKEN[14], HANNELORE WEBER[15] and JELTE M. WICHERTS[5]

[1]*Department of Psychology, Humboldt University Berlin, Berlin, Germany*
[2]*Institute of Psychological Sciences, University of Leeds, Leeds, UK*
[3]*Department of Developmental, Personality and Social Psychology, Ghent University, Ghent, Belgium*
[4]*Department of Experimental Clinical and Health Psychology, Ghent University, Ghent, Belgium*
[5]*School of Social and Behavioral Sciences, Tilburg University, Tilburg, The Netherlands*
[6]*Department of Psychology, University of Heidelberg, Heidelberg, Germany*
[7]*Max Planck Institute for Research on Collective Goods, Bonn, Germany*
[8]*Department of Psychology, University of California at Riverside, Riverside, CA USA*
[9]*Department of Psychology, University of Potsdam, Potsdam, Germany*
[10]*Department of Psychology, University of Virginia, Charlottesville, VA USA*
[11]*Department of Psychology, University of Milano-Bicocca, Milan, Italy*
[12]*Department of Psychology, University of Illinois, Chicago, IL USA*
[13]*Department of Psychology, University of Koblenz–Landau, Landau, Germany*
[14]*Department of Psychology, Utrecht University, Utrecht, The Netherlands*
[15]*Department of Psychology, University of Greifswald, Greifswald, Germany*

Abstract: Replicability of findings is at the heart of any empirical science. The aim of this article is to move the current replicability debate in psychology towards concrete recommendations for improvement. We focus on research practices but also offer guidelines for reviewers, editors, journal management, teachers, granting institutions, and university promotion committees, highlighting some of the emerging and existing practical solutions that can facilitate implementation of these recommendations. The challenges for improving replicability in psychological science are systemic. Improvement can occur only if changes are made at many levels of practice, evaluation, and reward.

Key words: replicability; confirmation bias; publication bias; generalizability; research transparency

PREAMBLE

The purpose of this article is to recommend sensible improvements that can be implemented in future research without dwelling on suboptimal practices in the past. We believe the suggested changes in documentation, publication, evaluation, and funding of research are timely, sensible, and easy to implement. Because we are aware that science is pluralistic in nature and scientists pursue diverse research goals with myriad methods, we do not intend the recommendations as dogma to be applied rigidly and uniformly to every single study, but as ideals to be recognized and used as criteria for evaluating the quality of empirical science.

*Correspondence to: Jens B. Asendorpf, Department of Psychology, Humboldt University, Unter den Linden 6, 10099 Berlin, Germany.
E-mail: jens.asendorpf@online.de
†This target paper is the result of an Expert Meeting on 'Reducing non-replicable findings in personality research' in Trieste, Italy, July 14–16, 2012, financed by the European Association of Personality Psychology (EAPP) in the recognition of the current debate on insufficient replicability in psychology and medicine. The participants of this Expert Meeting served as authors of the current article (the organizer of the meeting as the first author) or as its editor.

MOVING BEYOND THE CURRENT REPLICABILITY DEBATE

In recent years, the replicability of research findings in psychology (but also psychiatry and medicine at large) has been increasingly questioned (Ioannidis, 2005; Lehrer, 2010; Yong, 2012). Whereas current debates in psychology about unreplicable findings often focus on individual misconduct or even outright frauds that occasionally occur in all sciences, the more important questions are which specific factors and which incentives in the system of academic psychology might contribute to the problem (Nosek, Spies, & Motyl, 2012). Discussed are, among others, an underdeveloped culture of making data transparent to others, an overdeveloped culture of encouraging brief, eye-catching research publications that appeal to the media, the absence of incentives to publish high-quality null results, failures to replicate earlier research even when based on stronger data or methodology, and contradictory findings within studies.

Whatever the importance of each such factor might be, current psychological publications are characterized by strong orientation towards confirming hypotheses. In a comparison of publications in 18 empirical research areas, Fanelli (2010) found rates of confirmed hypotheses ranging from 70% (space

science) to 92% (psychology and psychiatry), and in a study of historic trends across sciences, Fanelli (2012) reported a particularly sharp increase of the rate for psychology and psychiatry between 1990 and 2007. The current confirmation rate of 92% seems to be far above rates that should be expected, given typical effect sizes and statistical power of psychological studies (see section on Increase Sample Sizes). The rate seems to be inflated by selective nonreporting of nonconfirmations as well as *post hoc* invention of hypotheses and study designs that do not subject hypotheses to the possibility of refutation. In contrast to the rosy picture presented by publications, in a recent worldwide poll of more than 1000 psychologists, the mean subjectively estimated replication rate of an established research finding was 53% (Fuchs, Jenny, & Fiedler, 2012).

Among many other factors, two widespread habits seem to contribute substantially to the current publication bias: excessive flexibility in data collection and in data analysis. In a poll of more than 2000 psychologists, prevalences of 'Deciding whether to collect more data after looking to see whether the results were significant' and 'Stopping data collection earlier than planned because one found the result that one had been looking for' were subjectively estimated at 61% and 39%, respectively (John, Loewenstein, & Prelec, 2012). And it is all too easy to apply multiple methods and then selectively pick those generating hypothesis confirmation or interesting findings (e.g. selection of variables and inclusion of covariates, transformation of variables, and details of structural equation models; Simmons, Nelson, & Simonsohn, 2011).

The question of whether there might be something fundamentally wrong with the mainstream statistical null-hypothesis testing approach is more difficult. This has perhaps been best highlighted by publication of the highly implausible precognition results in volume 100 of *JPSP* (Bem, 2011) that, according to the editor, could not be rejected because this study was conducted according to current methodological standards. In response to this publication, some critics called for Bayesian statistics relying on *a priori* probabilities (Wagenmakers, Wetzels, Borsboom, & van der Maas, 2011). This is not the only solution, however; treating stimuli as random factors (sampled from a class of possible stimuli, just as participants are sampled from a population) also leaves Bem's findings nonsignificant (refer to Judd, Westfall, & Kenny, 2012, and the later section on a Brunswikian approach to generalizability).

We do not seek here to add to the developing literature on identifying problems in current psychological research practice. Because replicability of findings is at the heart of any empirical science and because nonreplicability is the common thread that runs through most of the current debate, we address the following more constructive question: How can we increase the replicability of research findings in psychology now?

First, we define replicability and distinguish it from data reproducibility and generalizability. Second, we address the replicability concept from a more detailed methodological and statistical point of view. Third, we offer recommendations for increasing replicability at various levels of academic psychology: How can authors, reviewers, editors, journal policies, departments, and granting agenciescontribute to improving replicability, what incentives would encourage achieving this goal, what are the implications for teaching psychological science, and how can our recommendations be implemented in everyday practice?

DATA REPRODUCIBILITY, REPLICABILITY, AND GENERALIZABILITY

Given that replicability is not precisely defined in psychology, we propose a definition based on Brunswik's notion of a representative design (Brunswik, 1955) and distinguish the *replicability of a research finding* from its *reproducibility from the same data set* as well as from its *generalizability*.

Reproducibility of a research finding from the same data set is a necessary requirement for replicability. *Data reproducibility* means that Researcher B (e.g. the reviewer of a paper) obtains exactly the same results (e.g. statistics and parameter estimates) that were originally reported by Researcher A (e.g. the author of that paper) from A's data when following the same methodology.[1] To check reproducibility, Researcher B must have the following: (a) the raw data; (b) the code book (variable names and labels, value labels, and codes for missing data); and (c) knowledge of the analyses that were performed by Researcher A (e.g. the syntax of a statistics program). Whereas (c) can be described to some extent in the method section of a paper, (a), (b), and more details on (c) should either be available on request or, preferably, deposited in an open repository (an open-access online data bank; see www.opendoar.org for an overview of quality-controlled repositories).

Replicability means that the finding can be obtained with other random samples drawn from a multidimensional space that captures the most important facets of the research design. In psychology, the facets typically include the following: (a) individuals (or dyads or groups); (b) situations (natural or experimental); (c) operationalizations (experimental manipulations, methods, and measures); and (d) time points. Which dimensions are relevant depends on the relevant theory: What constructs are involved, how are they operationalized within the theory underlying the research, and what design is best suited to test for the hypothesized effects? Replication is obtained if differences between the finding in the original Study A and analogous findings in replication Studies B are insubstantial and due to unsystematic error, particularly sampling error, but not to systematic error, particularly differences in the facets of the design.

The key point here is that studies do not sample only participants; they also often sample situations, operationalizations, and time points that can also be affected by sampling error that should be taken into account. By analogy with analysis of variance, *all* design facets might be considered for treatment as random factors. Although there are sometimes good reasons to assume that a facet is a fixed factor, the alternative of treating it as a random factor is often not even considered (see Judd et al., 2012, for a recent discussion concerning experimental stimuli). Brunswikian replicability

[1]Our use of the term reproducibility is aligned with the use in computational sciences but not in some other sciences such as biological science applications where reproducibility is more akin to the concept of replicability used in psychology. Nevertheless, we use the term reproducibility to distinguish it from replicability.

DOI: 10.1002/per

requires that researchers define not only the population of participants but also the universe of situations, operationalizations, and time points relevant to their designs. Although such specification is difficult for situations and operationalizations, specification of any facet of the design is helpful for achieving replicability; the less clear researchers are about the facets of their designs, the more doors are left open for nonreplication.

Generalizability of a research finding means that it does not depend on an originally unmeasured variable that has a systematic effect. In psychology, generalizability is often demonstrated by showing that a potential moderator variable has no effect on a group difference or correlation. For example, student samples often contain a high proportion of women, leaving it unclear to what extent results can be generalized to a population sample of men and women. Generalizability requires replicability but extends the conditions to which the effect applies.

To summarize, data reproducibility is necessary but not sufficient for replicability, and replicability is necessary but not sufficient for generalizability. Thus, if I am claiming a particular finding, it is necessary for reproducibility that this finding can be recovered from my own data by a critical reviewer, but this reviewer may not replicate the finding in another sample. Even if this reviewer can replicate the finding in another sample from the same population, attaining replication, this does not imply that the finding can be easily generalized to other operationalizations of the involved constructs, other situations, or other populations.

Sometimes, replicability is dismissed as an unattainable goal because strict replication is not possible (e.g. any study is performed in a specific historic context that is always changing). This argument is often used to defend business as usual and avoid the problem of nonreplication in current research. But replication, as we define it, is generalization in its most narrow sense (e.g. the findings can be generalized to another sample from the same population). If not even replicability can be shown, generalizability is impossible, and the finding is so specific to one particular circumstance as to be of no practical use. Nevertheless, it is useful to distinguish between 'exact' replicability and 'broader' generalizability because the latter 'grand perspective' requires many studies and ultimately meta-analyses, whereas replicability can be studied much more easily as a first step towards generalizability. In the following, we focus on the concept of 'exact' replicability.

RECOMMENDATIONS FOR STUDY DESIGN AND DATA ANALYSIS

Increasing replicability by decreasing sources of error

Scientists ideally would like to make no errors of inference, that is, they would like to infer from a study a result that is true in the population. If the result is true in the population, a well-powered replication attempt (as discussed later) will likely confirm it. The issue of replicability can thus be approached by focusing on the status of the inference in the initial study, the logic being that correct inferences are likely to be replicated in subsequent studies.

Within a null-hypothesis significance testing approach that is only concerned with whether an effect can be attributed to chance or not, there are two types of errors: rejecting the null hypothesis when it is true (false positive, α) and failing to reject it when it is false (false negative, β). These two types of errors can be best understood from the perspective of power (Cohen, 1988). The power of a statistical test is the probability of rejecting the null hypothesis when it is false, or the complement of the false-negative error ($1 - \beta$). Its value depends on sample size, effect size, and α level. Within this framework, there is a negative relation between the two types of error: Given effect and sample sizes, reducing one type of error comes at the cost of increasing the other type of error. This may give the misleading impression that one has to choose between the two types of errors when *planning* a study. Instead, it is possible to minimize *both* types of errors simultaneously by increasing statistical power (Maxwell, Kelley, & Rausch, 2008). Replicable results are more likely when power is high, so the key question becomes identifying the factors that increase statistical power. The answer is simple: For any chosen α level, statistical power goes up as effect sizes and sample sizes increase.

Instead of the null-hypothesis significance testing, one can adopt a statistical approach emphasizing parameter estimation. Within this alternative approach, there is a third type of error: inaccuracy of parameter estimation (Kelley & Maxwell, 2003; Maxwell et al., 2008). The larger the confidence interval (CI) around a parameter estimate, the less certain one can be that the estimate approximates the corresponding true population parameter. Replicable effects are more likely with smaller CIs around the parameter estimates in the initial study, so the key question becomes identifying the factors that decrease CIs. Again the answer is simple: The width of a CI increases with the standard deviation of the parameter estimate and decreases with sample size (Cumming & Finch, 2005).

Increase sample size

These considerations have one clear implication for attempts to increase replicability. All else equal, statistical power goes up and CI width goes down with larger sample size. Therefore, results obtained with larger samples are more likely to be replicable than those obtained with smaller ones. This has been said many times before (e.g. Cohen, 1962; Tversky & Kahneman, 1971), but reviews have shown little improvement in the typical sample sizes used in psychological studies. Median sample sizes in representative journals are around 40, and average effect sizes found in meta-analyses in psychology are around $d = 0.50$, which means that the typical power in the field is around .35 (Bakker, Van Dijk, & Wicherts, 2012). These estimates vary, of course, with the subdiscipline. For example, Fraley and Marks (2007) did a meta-analysis of correlational personality studies and found the median effect size to be $r = .21$ ($d = 0.43$) for a median of 120 participants, resulting in a power of .65, a little better, but still far from ideal.

Consequently, if all effects reported in published studies were true, only 35% would be replicable in similarly underpowered studies. However, the rate of confirmed hypotheses in current psychological publications is above 90% (Fanelli, 2010). Among other factors, publishing many low-powered studies contributes to this excessive false-positive bias. It cannot be stressed enough that researchers should collect bigger sample sizes, and editors, reviewers, and readers should insist on them.

DOI: 10.1002/per

Planning a study by focusing on its power is not equivalent to focusing on its accuracy and can lead to different results and decisions (Kelley & Rausch, 2006). For example, for regression coefficients, precision of a parameter estimate depends on sample size, but it is mostly unaffected by effect size, whereas power is affected by both (Kelley and Maxwell, 2003; Figure 2). Therefore, a focus on power suggests larger sample sizes for small effects and smaller ones for large effects compared with a focus on accuracy. The two approaches emphasize different questions (Can the parameter estimate be confidently tested against the null hypothesis? Is the parameter estimate sufficiently accurate?). Both have merits, and systematic use would be an important step in increasing replicability of results. An optimal approach could be to consider them together to achieve both good statistical power and CIs that are sufficiently narrow.

Last but not least, this emphasis on sample size should not hinder exploratory research. Exploratory studies can be based on relatively small samples. This is the whole point, for example, of pilot studies, although studies labelled as such are not generally publishable. However, once an effect is found, it should be replicated in a larger sample to provide empirical evidence that it is unlikely to be a false positive and to estimate the involved parameters more accurately.

Increase reliability of the measures

Larger sample size is not the only factor that decreases error. The two most common estimators of effect size (Cohen's *d* and Pearson's *r*) both have standard deviations in their denominators; hence, all else equal, effect sizes go up *and* CIs and standard errors down with decreasing standard deviations. Because standard deviation is the square root of variance, the question becomes how can measure variance be reduced without restricting true variation? The answer is that measure variance that can be attributed to error should be reduced. This can be accomplished by increasing measure reliability, which is defined as the proportion of measure variation attributable to true variation. All else equal, more reliable measures have less measurement error and thus increase replicability.

Increase study design sensitivity

Another way of decreasing error variance without restricting true variation is better control over methodological sources of errors (study design sensitivity, Lipsey & Hurley, 2009). This means distinguishing between systematic and random errors. Random errors have no explanation, so it is difficult to act upon them. Systematic errors have an identifiable source, so their effects can potentially be eliminated and/or quantified. It is possible to reduce systematic errors using clear and standardized instructions, paying attention to questionnaire administration conditions and using stronger manipulations in experimental designs. These techniques do, however, potentially limit generalizability.

Increase adequacy of statistical analyses

Error can also be decreased by using statistical analyses better suited to study design. This includes testing appropriateness of method-required assumptions, treating stimuli as random rather than fixed factors (Judd et al., 2012), respecting dependence within the data (e.g. in analyses of dyads, Kenny, Kashy, & Cook, 2006, or hierarchically nested data, Hox, 2010), and removing the influences of covariates, given appropriate theoretical rationale (Lee, 2012).

Avoid multiple underpowered studies

It is commonly believed that one way to increase replicability is to present multiple studies. If an effect can be shown in different studies, even though each one may be underpowered, many readers, reviewers, and editors conclude that it is robust and replicable. Schimmack (2012), however, has noted that the opposite can be true. A study with low power is, by definition, unlikely to obtain a significant result with a given effect size. Unlikely events sometimes happen, and underpowered studies may occasionally obtain significant results. But a series of such results begins to strain credulity. In fact, a series of underpowered studies with the same result are so unlikely that the whole pattern of results becomes literally 'incredible'. It suggests the existence of unreported studies showing no effect. Even more, however, it suggests sampling and design biases. Such problems are very common in many recently published studies.

Consider error introduced by multiple testing

When a study involves many variables and their interrelations, following the aforementioned recommendations becomes more complicated. As shown by Maxwell (2004), the likelihood that some among multiple variables will show significant relations with another variable is higher with underpowered studies, although the likelihood that any specific variable will show a significant relation with another specific variable is smaller. Consequently, the literature is scattered with inconsistent results because underpowered studies produce different sets of significant (or nonsignificant) relations between variables. Even worse, it is polluted by single studies reporting overestimated effect sizes, a problem aggravated by the confirmation bias in publication and a tendency to reframe studies *post hoc* to feature whatever results came out significant (Bem, 2000). The result is a waste of effort and resources in trying and failing to replicate a certain result (Maxwell, 2004, p. 160), not to mention the problems created by reliance on misinformation.

Contrary to commonly held beliefs, corrections for multiple testing such as (stepwise) Bonferroni procedures do not solve the problem and may actually make things worse because they diminish statistical power (Nakagawa, 2004). Better procedures exist and have gained substantial popularity in several scientific fields, although still very rarely used in psychology. At an overall level, random permutation tests (Sherman & Funder, 2009) provide a means to determine whether a set of correlations is unlikely to be due to chance. At the level of specific variables, false discovery rate procedures (Benjamini & Hochberg, 1995) strike better compromises between false positives and false negatives than Bonferroni procedures. We recommend that these modern variants also be adopted in psychology. But even these procedures do not completely solve the problem of multiple testing. Nonstatistical solutions are required such as the explicit separation of *a priori* hypotheses preregistered in a repository from exploratory *post hoc* hypotheses (section on Implementation).

DOI: 10.1002/per

Is a result replicated?

Establishing whether a finding is quantitatively replicated is more complex than it might appear (Valentine et al., 2011). A simple way to examine replicability is to tabulate whether the key parameters are statistically significant in original and replication studies (*vote counting*). This narrow definition has the advantage of simplicity but can lead to misleading conclusions. It is based on a coarse dichotomy that does not acknowledge situations such as $p = .049$ (initial study) and $p = .051$ (second study). It can also be misleading if replication studies are underpowered, making nonreplication of an initial finding more likely. A series of underpowered or otherwise faulty studies that do not replicate an initial finding do not allow the conclusion that the initial finding was not replicable. Moreover, statistical significance is not the only property involved. The size of the effect matters too. When two studies both show significant effects, but effect sizes are very different, has the effect been replicated?

More useful from a replicability perspective is a quantitative comparison of the CIs of the key parameters. If the key parameter (e.g. a correlation) of the replication study falls within the CI of the initial study (or if the two CIs overlap substantially, Cumming & Finch, 2005), one can argue more strongly that the result is replicated. But again, the usefulness of this method depends on study power, including that of the initial study. For instance, suppose that an initial study with 70 participants has found a correlation between two measures of $r = .25$ [0.02, 0.76], which is significant at $p = .037$. A high-powered replication study of 1000 participants finds a correlation of $r = .05$ [−0.01, 0.11], which besides being trivial is not significant ($p = .114$). A formal comparison of the two results would show that the correlation in the second study falls within the CI of the first study ($Z = 1.63$, $p = .104$). One might therefore conclude that the initial result has been replicated. However, this has only occurred because the CI of the initial study was so large. In this specific case, a vote counting approach would be better.

The logic of quantitative comparison can be pushed further if effect sizes from more than two studies are compared (Valentine et al., 2011, p. 109). This basically means running a small meta-analysis in which the weighted average effect size is calculated and study heterogeneity is examined; if heterogeneity is minimal, one can conclude that the subsequent studies have replicated the initial study. However, the statistical power of heterogeneity tests is quite low for small samples, so the heterogeneity test result should be interpreted cautiously. Nonetheless, we recommend the meta-analytic approach for evaluation of replicability even when not many replication studies exist because it helps to focus attention on the size of an effect and the (un)certainty associated with its estimate.

In the long run, psychology will benefit if the emphasis is gradually shifted from whether an effect exists (an initial stage of research) to the size of the effect (a hallmark of a cumulative science). Given that no single approach to establish replicability is without limits, however, the use of multiple inferential strategies along the lines suggested by Valentine et al. (2011, especially Table 1) is a better approach. In practice, this means summarizing results by answering four questions: (a) Do the studies agree about direction of effect? (b) What is the pattern of statistical significance? (c) Is the effect size from the subsequent studies within the CI of the first study? (d) Which facets of the design should be considered fixed factors, and which random factors?

RECOMMENDATIONS FOR THE PUBLICATION PROCESS

Authors

Authors of scientific publications often receive considerable credit for their work but also take responsibility for the veracity of what is reported. Authors should also, in our view, take responsibility for assessing the replicability of the research they publish. We propose that an increase in replicability of research can be achieved if, in their role as prospective authors of a scientific article, psychologists address the following two main questions: (1) How does our treatment of this research contribute to increasing the transparency of psychological research? (2) How does this research contribute to an acceleration of scientific progress in psychology? We propose that answering these questions for oneself become an integral part of one's research and of authoring a scientific article. We briefly elaborate on each question and propose steps that could be taken in answering them. Implementing some of these steps will require some cooperation with journals and other publication outlets.

Increasing research transparency

(a) *Provide a comprehensive (literature) review*. We encourage researchers to report details of the replication status of key prior studies underlying their research. Details of 'exact' replication studies should be reported whether they did or did not support the original study. Ideally, this should include information on pilot studies where available.

(b) *Report sample size decisions*. So that the research procedure can be made transparent, it is important that researchers provide *a priori* justification for sample sizes used. Examples of relevant criteria are the use of power analysis or minimum sample size based on accepted good practice (see for further discussion Tressoldi, 2012). The practice of gradually accumulating additional participants until a statistically significant effect is obtained is unacceptable given its known tendency to generate false-positive results.

(c) *Preregister research predictions*. Where researchers have strong predictions, these and the analysis plan for testing them should be registered prior to commencing the research (section on Implementation). Such preregistered predictions should be labelled as such in the research reports and might be considered additional markers of quality. Preregistration is, for example, a precondition for publication of randomized controlled trials in major medical journals.

(d) *Publish materials, data, and analysis scripts*. Most of all, we recommend that researchers think of publication as requiring more than a PDF of the final text of an article. Rather, a publication includes all written materials, data,

DOI: 10.1002/per

and analysis scripts used to generate tables, figures, and statistical inferences. A simple first step in improving trust in research findings would be for all authors to indicate that they had seen the data. If practically possible, the materials, data, and analysis scripts should be made available in addition to the final article so that other researchers can reproduce the reported findings or test alternative explanations (Buckheit & Donoho, 1995). The information can be made available through open-access sources on the internet. There is a broad range of options: repositories housed at the author's institution or personal website, a website serving a group of scientists with a shared interest, or a journal website (section on Implementation). Options are likely to vary in degree of technical sophistication.

Accelerate scientific progress

(a) *Publish working papers*. We recommend that authors make working papers describing their research publically available along with their research materials. To increase scientific debate and transparency of the empirical body of results, prepublications can be posted in online repositories (section on Implementation). The most prominent preprint archive related to psychology is the Social Science Research Network (http://ssrn.com/).
(b) *Conduct replications*. Where feasible, researchers should attempt to replicate their own findings prior to first publication. 'Exact' replication in distinct samples is of great value in helping others to build upon solid findings and avoiding dead ends. Replicated findings are the stuff of cumulative scientific progress. Conducting generalizability studies is also strongly encouraged to establish theoretical understanding. Replication by independent groups of researchers is particularly encouraged and can be aided by increasing transparency (see the earlier recommendations).
(c) *Engage in scientific debate in online discussion forums*. To increase exchange among individual researchers and research units, we advocate open discussion of study results both prior to and after publication. Learning about each other's results without the publication time lag and receiving early feedback on studies create an environment that makes replications easy to conduct and especially valuable for the original researchers. After study publication, such forums could be places to make additional details of study design publicly available. This proposal could be realized in the same context as recommendation 1(d).

Reviewers, editors, and journals

Researchers do not operate in isolation but in research environments that can either help or hinder application of good practices. Whether they will adopt the recommendations in the previous section will depend on whether the research environments in which they operate reinforce or punish these practices. Important aspects of the research landscape are the peer reviewers and editors that evaluate research reports and the journals that disseminate them. So that replicability can be increased, reviewers, editors, and journals should allow for and encourage the implementation of good research practices.

Do not discourage maintenance of good practices

Reviewers and editors should accept not only papers with positive results that perfectly confirm the hypotheses stated in the introduction. Holding the perfectly confirmatory paper as the gold standard impedes transparency regarding nonreplications and encourages use of data analytic and other techniques that contort the actual data, as well as study designs that cannot actually refute hypotheses. Reviewers and editors should publish robustly executed studies that include null findings or results that run counter to the hypotheses stated in their introductions.

Importantly, such tolerance for imperfection can augment rather than detract from the scientific quality of a journal. Seemingly perfectly consistent studies are often less informative than papers with occasional unexpected results if they are underpowered. When a paper contains only one perfect but underpowered demonstration of an effect, high-powered replication studies are needed before much credibility can be given to the observed effect. The fact that a paper contains many underpowered studies that all perfectly confirm the hypotheses can be an indication that something is wrong (Schimmack, 2012).

For example, if an article reports 10 successful confirmations of a (actually true) finding in studies, each with a power of .60, the probability that all of the studies could have achieved statistical significance is less than 1%. This probability is itself a 'significant' result that, in a more conventional context, would be used to reject the hypothesis that the result is plausible (Schimmack, 2012).

We do not mean to imply that reviewers and editors should consistently prefer papers with result inconsistencies. When effects are strong and uniform, results tend to be consistent. But most psychological effects are *not* strong *or* uniform. Studies with result inconsistencies help to identify the conditions under which effects vary. Low publication tolerance for them impedes scientific progress, discourages researchers from adopting good research practices, and ultimately reduces a journal's scientific merits.

There are several other subtle ways in which actions of reviewers, editors, and journals can discourage researchers from maintaining good practices. For instance, because of copyright considerations, some journals might prevent authors from making working papers freely available. Such policies hinder transparency.

Proactively encourage maintenance of good practices

Journals could allow reviewers to discuss a paper openly with its authors (including access to raw data). Reviewers who do so could be given credit (e.g. by mentioning the reviewer's name in the publication). Journals could also give explicit credit (e.g. via badges or special journal sections) to authors who engaged in good practices (e.g. preregistration of hypotheses). Also, they could allow authors to share their reviews with editors from other journals (and vice versa). This encourages openness and debate. It is likely to improve the review process by giving editors immediate access to

DOI: 10.1002/per

prior reviews, helping them to decide on the merits of the work or guiding collection of additional reviews.

As part of the submission process, journals could require authors to confirm that the raw data are available for inspection (or to stipulate why data are not available). Likewise, co-authors could be asked to confirm that they have seen the raw data and reviewed the submitted version of the paper. Such policies are likely to encourage transparency and prevent cases of data fabrication by one of the authors. Related to this, reviewers and editors can make sure that enough information is provided to allow tests of reproducibility and replicability. To facilitate communication of information and minimize journal space requirements, authors can be allowed to refer to supplementary online materials.

Journals could also explicitly reserve space for reports of failures to replicate existing findings. At minimum, editors should revoke any explicit policies that discourage or prohibit publication of replication studies. Editors should also recognize a responsibility to publish important replication studies, especially when they involve studies that were originally published in their journals. Editors and journals can go even further by launching calls to replicate important but controversial findings. To encourage researchers to respond to such calls, editors can offer guarantees of publication (i.e. regardless of results) provided that there is agreement on method before the study is conducted (e.g. sufficient statistical power).

Recommendations for teachers of research methods and statistics

A solid methodological education provides the basis for a reliable and valid science. At the moment, (under)graduate teaching of research methods and statistics in psychology is overly focused on the analysis and interpretation of single studies, and relatively little attention is given to the issue of replicability. Specifically, the main goals in many statistical and methodological textbooks are to teach assessing the validity of and analysing the data from individual studies using null-hypothesis significance testing. Undergraduate and even graduate statistical education are based almost exclusively on rote methods for carrying out this framework. Almost no conceptual background is offered, and rarely is it mentioned that null-hypothesis testing is controversial and has a chequered history and that other approaches are available (Gigerenzer et al., 1989).

We propose that an increase in research replicability can be achieved if, in their role as teachers, psychologists pursue the following goals (in order of increasing generality): (1) introduce and consolidate statistical constructs necessary to understand the concept of replicable science; (2) encourage critical thinking and exposing hypotheses to refutation rather than seeking evidence to confirm them; and (3) establish a scientific culture of 'getting it right' instead of 'getting it published'. This will create a basis for transparent and replicable research in the future. In the following, we describe each of these goals in more detail and propose exemplary steps that could be taken.

Establish a scientific culture of 'getting it right' in the classroom

The most important thing that a supervisor/teacher can do is establish a standard of good practice that values soundness of research over publishability. This creates a research environment in which reproducible and replicable findings can be created (Nosek et al., 2012).

Teach concepts necessary to understand replicable science

(a) *Teach and practice rigorous methodology by focusing on multiple experiments*. This entails stressing the importance of *a priori* power estimates and sizes of effects in relation to standard errors (i.e. CIs) rather than outcomes of significance testing. Students should also learn to appreciate the value of nonsignificant findings in sufficiently powerful and rigorously conducted studies. Finally, students need to realize that multiple studies of the same effect, under the same or highly similar designs and with highly similar samples, may have divergent outcomes simply as a result of chance but also because of substantively or methodologically important differences.

(b) *Encourage transparency*. To stimulate accurate documentation and reproducibility, students should be introduced to online systems to archive data and analysis scripts (section on Implementation) and taught best practices in research (Recommendations for Authors section). So that the the value of replication of statistical analyses can be taught, students should reanalyse raw data from published studies.

(c) *Conduct replication studies in experimental methods classes*. One practical way to increase awareness of the importance of transparent science and the value of replications is to make replication studies essential parts of classes. By conducting their own replication studies, students have the chance to see which information is necessary to conduct a replication and experience the importance of accuracy in setting up, analysing, and reporting experiments (see Frank & Saxe, 2012, for further discussion of the advantages that accompany implementation of replication studies in class). Any failures to replicate the experience will reinforce its importance.

Critical thinking

(a) *Critical reading*. Learning to see the advantages and also flaws of a design, analysis, or interpretation of data is an essential step in the education of young researchers. Teachers should lead their students to ask critical questions when reading scientific papers (i.e. Do I find all the necessary information to replicate that finding? Is the research well embedded in relevant theories and previous results? Are methods used that allow a direct investigation of the hypothesis? Did the researchers interpret the results appropriately?). To develop skills to assess research outcomes of multiple studies critically, students should be taught to review well-known results from the literature that were later replicated successfully and unsuccessfully.

(b) *Critical evaluation of evidence (single-study level)*. Students should become more aware of the degree to which sampling error affects study outcomes by learning

DOI: 10.1002/per

how to interpret effect sizes and CIs correctly by means of examples. A didactical approach focused on multiple studies is well suited to explaining relevant issues of generalizability, statistical power, sampling theory, and replicability even at the undergraduate level. It is important to make clear that a single study generally represents only preliminary evidence in favour of or against a hypothesized effect.

Students should also become aware that statistical tools are not robust to the following: (1) optional stopping (adding more cases depending on the outcome of preliminary analyses); (2) data fishing; (3) deletion of cases or outliers for arbitrary reasons; and (4) other common 'tricks' to reach significance (Simmons et al., 2011).

(i) *Critical evaluation of evidence (multistudy level)*. At the graduate level, students should be taught the importance of meta-analysis as a source for effect size estimates and a tool to shed light on moderation of effects across studies and study homogeneity. Problems associated with these estimates (e.g. publication biases that inflate outcomes reported) must also be discussed to promote critical evaluation of reported results.

RECOMMENDATIONS FOR INSTITUTIONAL INCENTIVES

The recommended changes described earlier would go a long way to changing the culture of psychological science if implemented voluntarily by psychological scientists as researchers, editors, and teachers. If researchers adopt good research practices such as being more transparent in approach, submitting and tolerating more null findings, focusing more on calibrating estimation of effects rather than null-hypothesis significance testing, and communicating the need for doing so to students, the culture will naturally accommodate the new values. That said, we are sceptical that these changes will be adopted under the current incentive structures. Therefore, we also call upon the key institutions involved in the creation, funding, and dissemination of psychological research to reform structural incentives that presently support problematic research approaches.

Focus on quality instead of quantity of publications

Currently, the incentive structure primarily rewards publication of a large number of papers in prestigious journals. The sheer number of publications and journal impact factors often seem more important to institutional decisions than their content or relevance. Hiring decisions are often made on this basis. Grant awards are, in part, based on the same criteria. Promotion decisions are often predicated on publications and the awarding of grants. Some might argue that research innovation, creativity, and novelty are figured into these incentives, but if judgment of innovativeness, creativity, and novelty is based on publications in journals that accept questionable research practices, then publication quantity is the underlying indirect incentive. Given its current bias against producing null findings and emphasis on flashy and nonreplicable research, this does not serve our science well.

Therefore, we believe that the desirable changes on the parts of researchers, reviewers/editors/journals, and teachers that we described earlier need to be supplemented by changes in the incentive structures of supporting institutions. We consider incentives at three institutional levels: granting agencies, tenure committees, and the guild of psychologists itself.

Use funding decisions to support good research practices

Granting agencies could carry out the first, most effective change. They could insist upon direct replication of research funded by taxpayer money. Given the missions of granting agencies, which are often to support genuine (and thus reliable) scientific discoveries and creation of knowledge, we believe that granting agencies should not only desire but also promote replication of funded research.

One possibility is to follow an example set in medical research, where a private organization has been created with the sole purpose of directly replicating clinically relevant findings (Zimmer, 2012). Researchers in medicine who discover a possible treatment pay a small percentage of their original costs for another group to replicate the original study. Given the limited resources dedicated to social science research, a private endeavour may not be feasible. However, granting agencies could do two things to facilitate direct replication. First, they could mandate replication, either by requiring that a certain percentage of the budget of any given grant be set aside to pay a third party to replicate key studies in the programme of research or by funding their own consortium of researchers contracted to carry out direct replications. Second, granting agency decisions should be based on quality-based rather than quantity-based assessment of the scientific achievements of applicants. Junior researchers would particularly benefit from a policy that focuses on the quality of an applicant's research and the soundness of a research proposal. The national German funding agency recently changed its rules to allow not more than five papers to be cited as reference for evaluation of an applicant's ability to do research.

Additionally, attention should be paid to the publication traditions in various subdisciplines. Some subdisciplines are characterized by a greater number of smaller papers, which may inflate the apparent track records of researchers in those areas relative to those in subdisciplines with traditions of larger and more theoretically elaborated publications.

Revise tenure standards

We recommend that tenure and promotion policies at universities and colleges be changed to reward researchers who emphasize both reproducibility and replication (King, 1995). Some may argue that tenure committees do weigh quality of research in addition to overall productivity. Unfortunately, quality is often equated with journal reputation. Given that many of the most highly esteemed journals in our field openly disdain direct replication, discourage publication of

DOI: 10.1002/per

null findings, tolerate underpowered research, and/or rely on short reports, one can question whether journal reputation is a sound quality criterion. Because number of publications weighted by journal reputation is also used in evaluating grants, it also promotes another widely accepted criteria for promotion—acquisition of external funding.

King (1995) argued that researchers should also get credit for creating and disseminating data sets in ways that the results can be replicated and extended by other researchers (also King, 2006). To the extent that research becomes more replicable and replication is rewarded, tenure committees could also consider the extent to which researchers' work is replicated by others (Hartshorne & Schachner, 2012).

Conversely, tenure and promotion committees should not punish assistant professors for failing to replicate irreproducible research. If a young assistant professor is inspired by a recent publication to pursue a new line of research only to find that the original result cannot be replicated because the study was unsound, most evaluation committees will see this as a waste of time and effort. The assistant professor will look less productive than others, who, ironically, may be pursuing questionable research strategies to produce the number of publications necessary for tenure. The tragedy of the current system is that years of human capital and knowledge are spent on studies that produce null findings simply because they are based on studies that should not have been published in the first place. The problem here lies not with the replication efforts. On the contrary, creatively disconfirming existing theoretical ideas based on nonreplicable findings is at least as important as producing new ideas, and universities and colleges could acknowledge this by rewarding publication of null findings as much as those of significance.

One consequence of these proposed incentives for promotion and tenure would be to change the way tenure committees go about their work. Rather than relying on cursory reviews by overworked letter writers or arbitrary criteria, such as numbers of publications in the 'top' journals, tenure committees may have to spend more time reading a candidate's actual publications to determine their quality. For example, Wachtel (1980) recommended that researchers be evaluated on a few of their best papers, rather than CV length. This type of evaluation would, of course, demand that the members of tenure committees be sufficiently knowledgeable about the topic to discuss the nature and approach of the research described.

Change informal incentives

Finally, informal incentives within our guilds need to change for our scientific practices to change. When we discuss problematic research, we are not referring to abstract cases, but rather to the research of colleagues and friends. Few researchers want to produce research that contradicts the work of their peers. For that matter, few of us want to see failures to replicate our own research. The situation is even worse for assistant professors or graduate students. Should they even attempt to publish a study that fails to replicate an eminent scientist's finding? The scientist who one day will most likely weigh in on their tenure prospects? In the current research environment, that could indeed hamper their careers. Unless our entire guild becomes more comfortable with nonreplicated findings as an integral part of improving future replicability, the disincentives to change will outweigh the incentives. We hope that one effect of this document is to increase the value of identifying replicable research.

IMPLEMENTATION

Recommendations aim for implementation. However, even when awareness of importance is high and practical improvements identified, changing behaviour is hard. This is particularly true if implementing improvements adds time, effort, and resources to existing workflow. Researchers are already busy, and incentive structures for how to spend one's time are well defined. They are unlikely to invest in additional work unless that work is essential for desired rewards. However, strong incentives for good research practices can be implemented. For example, funders have strong leverage. If they require publishing data in repositories as a condition of funding, then researchers will follow through because earning grants is a strong incentive for researchers. Likewise, journals and editors can impose improvements. They may not be able to do so singlehandedly though. If the resource costs imposed exceed the perceived value of publishing in a journal, authors may abandon that journal and publish elsewhere.

Practical improvements cannot rely solely on appealing to scientists' values or pressures imposed by institutions. A researcher might agree that sharing data and study materials is a good thing, but if sharing is difficult to achieve, then it is not in the researcher's self-interest to do it. Practicalities affect success in implementing individual behavioural change. Ensuring success thus requires attention to the infrastructure and procedures required to implement the improvements.

The Internet is a mechanism for sharing of materials and data that address some of the practical barriers. But its existence is not sufficient. A system is needed that does the following: (a) makes it extremely simple to archive and document research projects and data; (b) provides a shared environment so that people know where to go to deposit and retrieve the materials; (c) integrates with the researchers' own documentation, archiving, and collaboration practices; and (d) offers flexibility to cover variation in research applications and sensitivity to ethical requirements. This might include options of no sharing, sharing only with collaborators, sharing by permission only, and sharing publicly without restriction.

Ways to accomplish this are emerging rapidly. They differ in scope, degree of organization, technical sophistication, long-term perspective, and whether they are commercial or nonprofit ventures. We present a few of them at different levels of scope, without any claim of comprehensive or representative coverage. They illustrate the various levels of engagement already possible.

In Europe, there are two large projects with the mission to enable and support digital research across all of the humanities and social sciences: Common Language

DOI: 10.1002/per

Resources and Technology Infrastructure (http://www.clarin.eu/), financed by the European Seventh Framework programme, and Digital Research Infrastructure for the Arts and the Humanities (http://www.dariah.eu/). These aim to provide resources to enhance and support digitally enabled research, in fields including psychology. The goal of these programmes is to secure long-term archiving and access to research materials and results.

Unconstrained topically and geographically, the commercial venture Figshare (http://figshare.com/) offers an easy user interface for posting, sharing, and finding research materials of any kind. Likewise, public ventures such as Dataverse (http://thedata.org/) address parts of the infrastructure challenges by making it easy to upload and share data. And the for-profit Social Science Research Network (http://www.ssrn.com/) is devoted to the rapid dissemination of social science research manuscripts.

There are study registries, such as http://clinicaltrials.gov/, but they are mostly available for clinical trial research in medicine thus far. The fMRI Data Center (http://www.fmridc.org/f/fmridc) in neurosciences and CHILDES (http://childes.psy.cmu.edu/) for child-language development provide data sharing and aggregation solutions for particular subdisciplines. There are also groups organized around specific topics (e.g. on cognitive modelling, http://www.cmr.osu.edu/). Finally, many researchers pursue open access for papers and research materials by posting them on their own institutional websites.

We highlight a project that aspires to offer most of the aforementioned options within a single framework: the Open Science Framework (http://opensciencefframework.org/). The Open Science Framework is an open solution developed by psychological scientists for documenting, archiving, sharing, and registering research materials and data. Researchers create projects and drag-and-drop materials from their workstations into the projects. Wikis and file management offer easy means of documenting the research; version control software logs changes to files and content. Researchers add contributors to their projects, and then the projects show up in the contributors' own accounts for viewing and editing. Projects remain private for their collaborative teams until they decide that some or all of their content should be made public. Researchers can 'register' a project or part of a project at any time to create a read-only, archived version. For example, researchers can register a description of a hypothesis, the research design, and analysis plan prior to conducting data collection or analysis. The registered copy is time stamped and has a unique, permanent universal resource locator that can be used in reporting results to verify prior registration.[2]

Many emerging infrastructure options offer opportunities for implementing the improvements we have discussed. The ones that will survive consider the daily workflow of the scientist and are finding ways to make it more efficient while simultaneously offering opportunities, or nudges, towards improving scientific rigour.

[2]Neither this nor any other system prevents a researcher from registering a hypothesis after having performed the study and conducted the analysis. However, doing this is active fraud.

CONCLUSION

A well-known adage of psychometrics is that measures must be reliable to be valid. This is true for the overall scientific enterprise as well; only, the reliability of results is termed replicability. If results are not replicable, subsequent studies addressing the same research question with similar methods will produce diverging results supporting different conclusions. Replicability is a prerequisite for valid conclusions. This is what we meant by our opening statement that 'replicability of findings is at the heart of any empirical science'. We have presented various proposals to improve the replicability of psychology studies. One cluster of these proposals could be called technical: improve the replicability of our findings through larger samples and more reliable measures, so that CIs become smaller and estimates more precise. A second cluster of proposals pertains more to the culture within academia: Researchers should avoid temptation to misuse the inevitable 'noise' in data to cherry-pick results that seem easily publishable, for example because they appear 'sexy' or unexpected. Instead, research should be about interpretation of broad and robust patterns of data and about deriving explanations that have meaning within networks of existing theories.

Some might say that the scientific process (and any other creative process) has Darwinian features because it consists of two steps (Campbell, 1960; Simonton, 2003): blind variation and selective retention. Like genetic mutations, this means that many research results are simply not very useful, even if they are uncovered using perfect measures. No single study 'speaks for itself': Findings have to be related to underlying ideas, and their merits discussed by other scientists. Only the best (intellectually fittest) ideas survive this process. Why then bother with scrutiny of the replicability of single findings, one may ask?

The answer is pragmatic: Publishing misleading findings wastes time and money because scientists as well as the larger public take seriously ideas that should not have merited additional consideration, based on the way they were derived. Not realizing that results basically reflect statistical noise, other researchers may jump on a bandwagon and incorporate them in planning follow-up studies and setting up new research projects. Instead of this, we urge greater continuity within broad research programmes designed to address falsifiable theoretical propositions. Such propositions are plausibly strengthened when supportive evidence is replicated and should be reconsidered when replications fail. Strong conceptual foundations therefore increase the information value of failures to replicate, provided the original results were obtained with reliable methods. This is the direction that psychology as a field needs to take.

We argue that aspects of the culture within psychological science have gradually become dysfunctional and have offered a hierarchy of systematic measures to repair them. This is part of a self-correcting movement in science: After long emphasizing large numbers of 'sexy' and 'surprising' papers, the emphasis now seems to be shifting towards 'getting it right'. This shift has been caused by systemic shocks, such as the recent fraud scandals and the publication of papers

DOI: 10.1002/per

deemed lacking in seriousness. We hope that this movement will be sustained and lead to an improvement in the way our science is conducted.

Ultimately, every scientist is responsible for the choices that he or she makes. In addition to the external measures that we propose in this article, we appeal to scientists' intrinsic motivation. Desire for precise measurements and curiosity to make deeper sense of incoherent findings (instead of cherry-picking those that seem easy to sell) are the reasons many of us have chosen a scholarly career. We hope that future developments will create external circumstances that are better aligned with these intrinsic inclinations and help the scientific process to become more accurate, transparent, and efficient.

REFERENCES

Bakker, M., Van Dijk, A., & Wicherts, J. M. (2012). The rules of the game called psychological science. *Perspectives on Psychological Science*, *7*, 543–554.

Bem, D. J. (2000). Writing an empirical article. In R. J. Sternberg (Ed.), *Guide to publishing in psychology journals* (pp. 3–16). Cambridge: Cambridge University Press.

Bem, D. J. (2011). Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology*, *100*, 407–426.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B (Methodological) 57*, 289–300.

Brunswik, E. (1955). Representative design and probabilistic theory in a functional psychology. *Psychological Review*, *62*, 193–217.

Buckheit, J., & Donoho, D. L. (1995). Wavelab and reproducible research. In A. Antoniadis (ed.). *Wavelets and statistics* (pp. 55–81). New York, NY: Springer-Verlag.

Campbell, D. T. (1960). Blind variation and selective retention in creative thought as in other knowledge processes. *Psychological Review*, *67*, 380–400.

Cohen, J. (1962). The statistical power of abnormal–social psychological research: A review. *Journal of Abnormal and Social Psychology*, *65*, 145–153.

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

Cumming, G., & Finch, S. (2005). Inference by eye: Confidence intervals, and how to read pictures of data. *The American Psychologist*, *60*, 170–180.

Fanelli, D. (2010). "Positive" results increase down the hierarchy of the sciences. *PLoS One*, *5*, e10068.

Fanelli, D. (2012). Negative results are disappearing from most disciplines and countries. *Scientometrics*, *90*, 891–904.

Fraley, R. C., & Marks, M. J. (2007). The null hypothesis significance-testing debate and its implications for personality research. In R. W. Robins, R. C. Fraley, & R. F. Krueger (Eds.), *Handbook of research methods in personality psychology* (pp. 149–169). New York: Guilford.

Frank, M. C., & Saxe, R. (2012). Teaching replication. *Perspectives on Psychological Science*, *7*, 600-604.

Fuchs, H., Jenny, M., & Fiedler, S. (2012). Psychologists are open to change, yet wary of rules. *Perspectives on Psychological Science*, *7*, 639-642.

Gigerenzer, G., Swijink, Z., Porter, T., Daston, L., Beatty, J., & Krüger, L. (1989). *The empire of chance: How probability changed science and everyday life*. Cambridge: Cambridge University Press.

Hartshorne, J. K., & Schachner, A. (2012). Tracking replicability as a method of post-publication open evaluation. *Frontiers in Computational Science*, *6*, 1–14.

Hox, J. J. (2010). *Multilevel analysis* (2nd ed.). New York, NY: Routledge.

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, *2*, e124.

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth-telling. *Psychological Science*, *23*, 524–532.

Judd, C. M., Westfall, J., & Kenny, D. A. (2012). Treating stimuli as a random factor in social psychology: A new and comprehensive solution to a pervasive but largely ignored problem. *Journal of Personality and Social Psychology*, *103*, 54–69.

Kelley, K., & Maxwell, S. E. (2003). Sample size for multiple regression: Obtaining regression coefficients that are accurate, not simply significant. *Psychological Methods*, *8*, 305–321.

Kelley, K., & Rausch, J. R. (2006). Sample size planning for the standardized mean difference: Accuracy in parameter estimation via narrow confidence intervals. *Psychological Methods*, *11*, 363–385.

Kenny, D. A., Kashy, D. A., & Cook, W. L. (2006). *Dyadic data analysis*. New York, NY: Guilford Press.

King, G. (1995). Replication, replication. *PS: Political Science and Politics*, *28*, 443–499.

King, G. (2006). Publication, publication. *PS: Political Science and Politics*, *34*, 119–125.

Lee, J. J. (2012). Correlation and causation in the study of personality. *European Journal of Personality*, *26*, 372–390.

Lehrer, J. (2010). *The truth wears off: Is there something wrong with the scientific method? The New Yorker*, December 13.

Lipsey, M. W., & Hurley, S. M. (2009). Design sensitivity: Statistical power for applied experimental research. In L. Bickman, & D. J. Rog (Eds.), *The SAGE handbook of applied social research methods* (pp. 44–76). Los Angeles, CA: SAGE Publications.

Maxwell, S. E., Kelley, K., & Rausch, J. R. (2008). Sample size planning for statistical power and accuracy in parameter estimation. *Annual Review of Psychology*, *59*, 537–563.

Nakagawa, S. (2004). A farewell to Bonferroni: The problems of low statistical power and publication bias. *Behavioral Ecology*, *15*, 1044–1045.

Nosek, B. A., Spies, J. R., & Motyl, M. (2012). Scientific utopia: II. Restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science 7*, 615–631.

Schimmack, U. (2012). The ironic effect of significant results on the credibility of multiple-study articles. *Psychological Methods 17*, 551–566.

Sherman, R. A., & Funder, D. C. (2009). Evaluating correlations in studies of personality and behavior: Beyond the number of significant findings to be expected by chance. *Journal of Research in Personality*, *43*, 1053–1061.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*, 1359–1366.

Simonton, D. K. (2003). Scientific creativity as constrained stochastic behavior: The integration of product, person, and process perspectives. *Psychological Bulletin*, *129*, 475–494.

Tressoldi, P. E. (2012). Replication unreliability in psychology: Elusive phenomena or "elusive" statistical power? *Frontiers in Psychology*, *3*. doi: 10.3389/fpsyg.2012.00218

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

Valentine, J. C., Biglan, A., Boruch, R. F., Castro, F. G., Collins, L. M., Flay, B. R., . . . Schinke, S. P. (2011). Replication in prevention science. *Prevention Science*, *12*, 103–117.

DOI: 10.1002/per

Wachtel, P. L. (1980). Investigation and its discontents: Some constraints on progress in psychological research. *The American Psychologist*, *5*, 399–408.

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., & van der Maas, H. L. J. (2011). Why psychologists must change the way they analyze their data: The case of psi: Comment on Bem (2011). *Journal of Personality and Social Psychology*, *100*, 426–432.

Yong, E. (2012). Bad copy: In the wake of high-profile controversies, psychologists are facing up to problems with replication. *Nature*, *485*, 298–300.

Zimmer, C. (2012). *Good scientist! You get a badge*. *Slate*, August 14 (on-line). http://www.slate.com/articles/health_and_science/science/2012/08/reproducing_scientific_studies_a_good_housekeeping_seal_of_approval_.html

DOI: 10.1002/per

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

Christenfeld, Ravina, & Bilous, 1991). This multiply determined nature of many psychological phenomena calls for the parsing of big research questions into smaller, tractable series of research questions that ultimately constitute systematic and meticulous programmes of research. Where to parse a phenomenon may not be obvious without empirical evidence, however. Therefore, the generalizability problem, as Allport referred to it, may represent a theoretical rather than methodological problem when investigating phenomena that are multiply determined. For instance, four decades ago, concerns that experimental research on attitude change was not replicable or generalizable existed because the same experimental factors (e.g., source credibility) were found to produce different outcomes in different studies. Rather than treat this as a statistical or methodological problem, we identified two distinct mechanisms (routes) through which attitude change could occur, and we specified the theoretical conditions in which a given factor would trigger each route. The resulting elaboration likelihood model (Petty & Cacioppo, 1981, 1986) made sense of what had appeared to be conflicting results, generated predictions of new patterns of data that have since been verified, and remains a staple in the field.

Multiple determinism includes parallel determinism (more than one antecedent condition can alone be *sufficient* to produce the outcome of interest) and convergent determinism (two or more antecedent conditions are *necessary* to produce an outcome). A lack of generalizing power in studies of the role of single factors is a predictable property of multiply determined phenomena. Because it is rare for a single factor or determinant to be a necessary *and* sufficient cause for a psychological phenomenon, the failure to find generalizability raises the theoretical question of whether multiple parallel or convergent determinism exist and, if so, under what conditions each antecedent may be operating and what other factors may also be operating.

To be specific, let ψ represent a psychological phenomenon of interest, let τ represent a factor or treatment whose effect on ψ is of interest, and let $\widetilde{\tau}$ ('not τ') represent all other antecedents of ψ, known or unknown. Carefully conceived, statistically powered, and controlled experimentation on the role of τ in producing ψ can be denoted as $P(\psi/\tau)$. When multiple factors are *sufficient* to produce the psychological outcome (i.e. parallel determinism), then $P(\psi/\tau) > 0$, but because other factors can also affect the outcome, $P(\psi/\widetilde{\tau}) > 0$. Only when other sufficient causes of ψ have been controlled in a particular experimental paradigm or assessment context will $P(\psi/\tau)$ approach 1. $P(\psi/\widetilde{\tau})$ will approach 0 in a given experimental context by virtue of experimental control—because all other determinants of to ψ have been eliminated or controlled in the experimental setting. Because ψ is multiply determined, however, $P(\psi/\widetilde{\tau}) > 0$ and may be much greater than 0 when aspects of the design, sample, operationalizations, or context are changed. This 'generalizing problem' need not reflect a methodological quagmire but rather can represent a theoretical challenge; it can lead to new insights into and research on the boundary conditions for theories, the operation of additional antecedents, and the specification of new theoretical organizations (Cacioppo & Berntson,1992).

In sum, attention to study details, from conceptualization, statistical power, and execution to analysis and interpretation, increases the likelihood that the empirical results constitute replicable scientific facts upon which one can solidly build. Asendorpf et al. argue that the facets of a research design relevant for replicability include individuals, situations, operationalizations, and time points. If psychological phenomena in principle had singular antecedents, this would be sufficient. This is not the only possible definition of replicability, however. In a complex science such as psychology, in which phenomena of interest can be multiply determined, *minimal replicability* refers to the same observation by an independent investigator using the same operationalizations, situations, and time points in an independent sample from the same population. Such minimal replications suggest that an empirical fact has been established, and failures to replicate the finding using different operationalizations, situations, time points, or populations suggest the operation of potentially important moderator variables (and, thus, generate theoretical questions) rather than methodological problems. To the extent that psychological phenomena are multiply determined, therefore, a failure to replicate a phenomenon across these facets of a research design may more productively be viewed as a failure to generalize and may trigger a search for the operation of, for instance, a previously unrecognized determinant.

From Replication Studies to Studies of Replicability

MICHAEL EID

Freie University Berlin
eid@zedat.fu-berlin.de

Abstract: Some consequences of the Asendorpf et al. concept of replicability as a combination of content validity and reliability are discussed. It is argued that theories about relevant facets have to be developed, that their approach requires multiple sampling within the same study, and that the proposed aggregation strategies should not be applied blindly. Studies on the conditions of replicability should complement pure replication studies. Copyright © 2013 John Wiley & Sons, Ltd.

DOI: 10.1002/per

Asendorpf et al. assume that research designs are characterized by multifaceted structures. Facets are individuals, situations, operationalizations, and time points. These facets are considered random factors. To ensure replicability, random samples must be drawn from these populations. Unlike *ad hoc* samples of individuals, test items, situations, and time points, random sampling will usually result in greater variance of the traits considered. Although this variance is necessary for representativeness, the authors consider it error variance that should be decreased in a next step through aggregation. Aggregations across individuals, items, and so on reduce standard errors andincrease reliability. Thus, they propose a two-step procedure to ensure replicability characterized by random sampling (to ensure content validity) in combination with aggregation (to increase precision and reliability). In contrast to generalizability, the concept of replicability does not focus on the variances of the facets *per se*, as the variances have to be considered to obtain unbiased aggregated scores. Consequently, replicability depends on content validity and reliability.

This conceptualization of replicability has some important consequences:

1. The facet populations have to be known. This might be the case for the population of individuals. But it might not be the case for the other facets such as items, methods, situations, and so on. In many areas of psychological research, theories are missing about the universes of stimuli, items, and methods. Taking methods as an example, with respect to Campbell and Fiske's (1959) seminal paper on convergent validity, Sechrest, Davis, Stickle, and McKnight (2000) noted that 'method variance was never defined in any exact way' (p. 63), but they added, 'and we will not do so here' (p. 63). It seems to be difficult to define the population of methods. This is also true for the other facets. Psychological theories often are not clear about these methodological aspects. The definition of Asendorpf et al. shows that we should focus much more on the development of theories about the different facets (e.g., methods and situations) that might play roles in our research designs. We must understand the variances of the facets not only to get rid of them by aggregation but also to understand the phenomenon we are studying and to guide the sampling process. Decreasing the standard error by sampling more individuals might not be appropriate for increasing replicability if they are sampled from the 'wrong' population. Increasing reliability by adding items that are reformulations of other items in the scale might not ensure replicability. All the statistical recommendations of the authors have their basis in the appropriate theoretical underpinning of the facets that are considered. These theoretical ideas have to be communicated to plan replication studies. In many research areas, however, they have to be developed first because theories often do not integrate theories about situations, methods, and so on.
2. Their two-step approach of replicability requires that there is a sampling process not only across different studies but also *within* a study. If in one study only a single example of a facet has been considered and in another study only a different example, replicability is not ensured. Although Asendorpf et al. focus on the increase of sample size for individuals and items, this concerns the other facets as well. This is in contrast to research practice in many areas of psychology where mono-method, mono-situation, and cross-sectional studies are predominant. However, it is unrealistic that in each and every study, random samples of individuals, items, stimuli, methods, and so on be drawn. Moreover, it might not be necessary if the variance due to facets is low. Planning replication studies requires available knowledge about which facets are relevant and which facets are random and not fixed. In many research areas, there is no knowledge about the importance of different facets. Systematic replicability studies that focus on different facets and the conditions of replicability are necessary. Examples are 'generalizability' studies using the definition from generalizability theory (Cronbach, Gleser, Nanda & Rajaratnam, 1972).
3. According to Asendorpf et al., aggregation is an important further step. Aggregation is an appropriate method for reducing variability due to random facets. If items, stimuli, individuals, and so on are considered interchangeable, aggregation is an efficient way to get rid of the resulting 'error variance'. However, if the elements of a facet are not interchangeable, aggregation might reduce relevant information and might not be appropriate. For example, if raters are interchangeable (e.g., students rating the quality of a course), aggregation can be used to obtain more precise estimates (e.g., mean quality of the course). However, if raters are not interchangeable (e.g., children and their parents rating the suicide risk of the child), aggregation might not be appropriate (the mean suicide risk score across raters might not be a valid indicator of the child's true suicide risk; Eid, Geiser, & Nussbeck, 2009). The recommendations of Asendorpf et al. for increasing replicability by decreasing sources of error are closely linked to their concept of facets as *random* factors. These factors may not be random in all applications. However, their approach clarifies that researchers have to think more closely about what are sources of error variance that should be eliminated and what are sources of substantive variance that should not be eliminated. This again requires theories about the nature of the facets, whether they are random or fixed. Aggregating across structurally different subpopulations (of methods, individuals, items, situations, etc.) might not be appropriate to enhance replicability even if this might increase reliability and power (Eid et al., 2008). Aggregation might be too often used blindly. The recommendation of Asendorpf et al. is linked to random facets that are linked to replicability. Fixed factors are related to generalizability.

It is the merit of the Asendorpf et al. concept of replicability that it makes clear that it is the combination of content validity (representativeness) and reliability (reduction of error variances) that should guide research. Moreover, their distinction between replicability (random

DOI: 10.1002/per

facets) and generalizability (fixed facets) is important for choosing appropriate research strategies and methods of data analyses (Eid et al., 2008). Their ideas require research programmes built on theories of facets. Consequently, research should move from pure replication studies to studies on replicability.

interest in negative results. Surprisingly, however, negative results in psychology and other disciplines are cited just as much as positives, suggesting that the source of bias might have less to do with file drawers and more with confirmation biases of psychologists themselves (Fanelli, 2012a). Another example is the recommendation to

Only Reporting Guidelines Can Save (Soft) Science

DANIELE FANELLI

University of Edinburgh
dfanelli@exseed.ed.ac.uk

Abstract: Some of the unreplicable and biased findings in psychological research and other disciplines may not be caused by poor methods or editorial biases, but by low consensus on theories and methods. This is an epistemological obstacle, which only time and scientific debate can overcome. To foster consensus growth, guidelines for best research practices should be combined with accurate, method-specific reporting guidelines. Recommending greater transparency will not suffice. A scientific system that enforced reporting requirements would be easily implemented and would save scientists from their unconscious and conscious biases. Copyright © 2013 John Wiley & Sons, Ltd.

Asendorpf et al. propose an excellent list of recommendations that may increase the likelihood that findings are true, by improving their replicability. However, these recommendations might still be too generic and fail to account for peculiarities and limitations that psychology and other social and biological disciplines face in their quest for truth. Some of the initiatives they suggested could be impractical or even counterproductive in psychology, which would make greater and easier progress if it shifted attention from what researchers do to what they report.

Psychology, like many other social and biological sciences, appears to be 'soft' (Fanelli, 2010; Simonton, 2004). It deals with extremely complex phenomena, struggling against an enormous amount of diversity, variables and noise, in conditions where practical and ethical considerations impede optimal study design. These characteristics—no doubt, with great variability among individual fields—probably render data in psychology relatively unable to 'speak for themselves', hampering scholars' ability to reach consensus on the validity of any theory, method, or finding and therefore to build upon them. In such conditions, scientists inevitably have many 'degrees of freedom' to decide how to conduct and interpret studies, which increases their likelihood to 'find' what they expect. Bias and false positives, in other words, are to some extent physiological to the subject matter, and no amount of statistical power, quantitative training, and reduced pressures to publish will remove them completely.

Although publication bias in the psychological literature is superficially similar to that observed in biomedical research, its causes might be partly different and thus require different solutions. Existing guidelines for best research practices tend to overlook this, and so do Asendorpf et al. For example, they recommend that editors and reviewers learn to accept negative results for publication, which buys into the standard (biomedical) explanation that publication bias is caused by a file drawer problem, created by lack of

preregister experimental hypotheses, in analogy to what is attempted with clinical trials. This suggestion fails to take into account the low predictive ability of psychological theories and low truth value of published research findings. Psychology is not astrophysics. Most of its predictions may rest on shaky grounds, and the same study could both support and falsify them depending on subtle changes in design and interpretation (LeBel & Peters, 2011; Weisburd & Piquero, 2008; Yong, 2012). Forcing psychologists to predeclare what they intend to test will push them, I fear, to either formulate more generic and less falsifiable hypotheses or 'massage' their findings even more.

In sum, although I support most of the recommendations of Asendorpf et al., I believe that they do not fully accommodate the fact that psychology has lower theoretical and methodological consensus than much biomedical research, let alone most physical sciences. Scientific consensus will hopefully grow over time, but only if we allow it to harden through an extended, free, and fair war of ideas, approaches, and schools of thought. Good research practices are the essential weapons that scientists need, but fairness and freedom in battle are only guaranteed by complete transparency and clarity of reporting.

What makes some of human knowledge scientific is not the superior honesty or skills of those who produced it, but their willingness to share all relevant information, which allowed truth to emerge by a collective process of competition, criticism, and selection. There is nothing wrong in conducting exploratory analyses, trying out several statistical approaches, rethinking one's hypothesis after the data have been collected, dropping outliers, and increasing one's sample size half-way through a study *as long as this is made known* when results are presented. These behaviours might increase false-positive ratios but will also increase the likelihood of discovering true patterns and new methods, and psychology seems to be still in a phase where it can benefit from all discovery attempts.

DOI: 10.1002/per

Good research practices notwithstanding, therefore, the keys to good science are good *reporting* practices, which, interestingly, are much easier to ensure. Indeed, reporting guidelines are rapidly being adopted in biomedical and clinical research, where initiatives such as the EQUATOR Network (http://www.equator-network.org) and Minimum Information about a Biomedical or Biological Investigation (http://mibbi.sourceforge.net) publish updated lists of details that authors need to provide, depending on what methodology they used. Major journals have adopted these guidelines spontaneously because doing so improves their reputation. If authors do not comply, their papers are rejected.

This approach could easily be exported to all disciplines and, if it became central to the way we do science, it would bring many collateral advantages. Peer reviewers, for example, instead of spending more of their precious time checking results as Asendorpf et al. suggest, could specialize in assessing papers' compliance with objective reporting guidelines. Indeed, peer reviewing could finally become a career option in itself, separate from teaching and doing research. Journals could decide on initial acceptance on the basis of the accuracy of methods—that is, blindly to outcome—and only later ask active researchers to assess the results and discussion. Strictness of reporting requirements could become a measure of a journal's quality, quite independent of impact factor. Moreover, reporting guidelines would provide clear and objective standards for teachers, students, and officers faced with allegations of misconduct (Fanelli, 2012b).

In conclusion, Asendorpf et al. make important recommendations. I highlight those of funding replication studies, emphasizing effect sizes, and rewarding replicated results. But the key to saving psychologists (and all other scientists) from themselves is ensuring the transparency of their work. Daryl Bem's evidence of precognition is problematic mainly because we lack information on all tests and analyses that were carried out before and after his experiments (LeBel & Peters, 2011). Diederik Stapel's fraudulent career might have never taken off if he had been forced to specify where and how he had collected his data (Levelt Committee, Noort Committee, & Drenth Committee, 2012).

We Don't Need Replication, but We Do Need More Data

GREGORY FRANCIS

Purdue University
gfrancis@purdue.edu

Abstract: Although the many valuable recommendations Asendorpf et al. are presented as a way of increasing and improving replication, this is not their main contribution. Replication is irrelevant to most empirical investigations in psychological science, because what is really needed is an accumulation of data to reduce uncertainty. Whatever criterion is used to define success or failure of a replication is either meaningless or encourages a form of bias that undermines the integrity of the accumulation process. Even though it is rarely practised, the fixation on replication actively hurts the field. Copyright © 2013 John Wiley & Sons, Ltd.

Asendorpf et al. present many recommendations that will likely improve scientific practice in psychology. Despite the good advice, many of the recommendations are based on fundamental misunderstandings about the role of replication in science. As Asendorpf et al. emphasize, replication is commonly viewed as a foundation of every empirical science. Experimental results that successfully replicate are interpreted to be valid, whereas results that fail to replicate are considered invalid. Although replication has worked wonderfully for fields such as physics and chemistry, the concept of replication is inappropriate for a field like experimental psychology.

The problem for psychology is that almost all experimental conclusions are based on statistical analyses. When statistical noise is large relative to the magnitude of the effect being investigated, then the conclusion is uncertain. This uncertainty is often a characteristic of what is being measured. The call to 'increase replicability' is a strange request because it asks for certainty where it cannot exist: No one would complain that coin flips are unreliable because they do not always land on the same side. In a similar way, uncertainty is often part of what is being investigated in psychological experiments.

Even when we try to measure a fixed effect, replication is irrelevant. Suppose scientist A rejects the null hypothesis for an experimental finding. Scientist B decides to repeat the experiment with the same methods. There are two possible outcomes for scientist B's experiment.

1. Successful replication: the replication experiment rejects the null hypothesis.
2. Failure to replicate: the replication experiment does not reject the null hypothesis.

How should the scientists interpret the pair of findings? For Case 1, it seems clear that a good scientific strategy is to use meta-analytic methods to pool the findings across the experiments and thereby produce a more precise estimate of the effect.

For Case 2, it may be tempting to argue that the failure to replicate invalidates the original finding; but such a claim requires a statistical argument that is best made with meta-analysis. These methods appropriately weight the experimental findings by the sample sizes and variability. Scientist B's finding will dominate the meta-analysis if it is based on a much larger sample size than scientist A's finding.

DOI: 10.1002/per

Importantly, the recommended scientific strategy for both successful and unsuccessful replication outcomes is to use meta-analysis to pool the experimental findings. Indeed, if the experimental methods are equivalent, then pooling the data with meta-analysis is always the recommended action. Scientists should not focus on an outcome that makes no difference. Rather than being a foundation of the scientific method, the concept of replication is irrelevant.

This claim is not just semantics. A fixation on replication success and failure, combined with misunderstandings about statistical uncertainty, likely promotes some of the problems described by Asendorpf et al., such as *post hoc* theorizing. A researcher who expects almost every experiment to successfully demonstrate a true effect can easily justify generating a theoretical difference between two closely related experiments that give different outcomes. The researcher can always point to some methodological or sampling difference as an explanation for the differing outcomes (e.g., time of day or subject motivation). Statisticians call this 'fitting the noise', and it undermines efforts to build coherent and generalizable theories. It is no wonder that journal editors, reviewers, and authors do not encourage replications: Replications rarely resolve anything.

This all sounds very bleak. If replication is not a useful concept for psychological science, then how should the field progress? First, researchers must have an appropriate recognition of the uncertainty that is inherent in their experimental studies. There is nothing fundamentally wrong with drawing a conclusion from a hypothesis test that just barely satisfies $p < .05$, but the conclusion should be tentative rather than definitive. Confidence in the conclusion increases by gathering additional data that promote meta-analysis. We need more data, not more replication.

Second, although exploratory work is fine, scientific progress often requires testing and refining quantitative theories. A *quantitative* theory is necessary because it provides a way to interpret measurements and to predict experimental outcomes. In contrast, a verbal theory claiming that one condition should have a bigger mean than another condition is only useful for exploratory work. Contrary to the claim made in many experimental papers, such verbal theories cannot predict the outcome of a hypothesis test because they do not provide a predicted effect size, which is necessary to estimate power. Discussions and debates about quantitative theories will identify where resources should be applied to precisely measure important experimental effects.

None of this is easy. When we determine whether experimental results should be pooled together or kept separate, equivalent methods trump measurement differences (even statistically significant ones); but such methodological equivalence often depends on a theoretical interpretation. Likewise, modifying a theory so that it better reflects experimental findings requires consideration of the uncertainty in the measurements, so data and theory go back and forth in a struggle for coherence and meaning. Researchers will have to chase down details of experimental methods to determine whether reported differences are meaningful or due to random sampling. Proper application of the scientific method to psychology takes a tremendous amount of work, and it cannot be reduced to the outcome of a statistical analysis.

Replication is often touted as the heart of the scientific method, but experimental psychologists trying to put it to practise quickly discover its inadequacies. Perhaps many researchers have intuitively recognized replication's irrelevance, and this is why the field praises replication but does not practise it. When combined with unrealistic interpretations about the certainty of conclusions and a lack of quantitative models, confusion about replication likely contributes to the current crisis in psychological sciences. It is a positive sign that, despite these difficulties, Asendorpf et al. were able to generate many valuable recommendations on how to improve the field. Most of their recommendations will be even better by shifting the emphasis from the concept of replication and towards gathering additional data to reduce uncertainty and promote development of quantitative theories.

Calls for Replicability Must Go Beyond Motherhood and Apple Pie

EARL HUNT

University of Washington
ehunt@u.washington.edu

Abstract: I expand on the excellent ideas of Asendorpf et al. for improving the transparency of scientific publications. Although their suggestions for changing the behaviour of authors, editors, reviewers, and promotion committees seem useful, in isolation, some are simply rallying cries to 'be good'. Other suggestions might attack the replicability problem but would create serious side effects in both the publication process and the academic endeavour. Moreover, we should be concerned primarily for replication of key, highly cited results, not for replication of the entire body of literature.

Asendorpf et al. wish to increase both transparency and replicability of scientific studies. Who would object? However, calls for a desirable goal, without proposing practical means of achieving it, amount to support for 'motherhood and apple pie', MAPPLE for short. The proverbial warning 'Be careful of what you wish for, you might get it' is also relevant. Present practices used to evaluate scientific contributions evolved for reasons. Changing these practices to

DOI: 10.1002/per

achieve one goal may have unintended consequences that influence other goals.

Transparency: The solvable problem

Asendorpf et al. say that studies are transparent when data and procedures are accessible and limits on conclusions are stated. Accessibility requires data archiving at reasonable cost. Just saying 'Keep good lab notes' is MAPPLE. There must be standards for record keeping. The issue is not simple. Psychological studies range from laboratory experiments to analyses of government records. Confidentiality and the proprietary nature of some data must be considered. In some cases, there are issues of security. Recall the debate over whether or not the genomes for pandemic influenzas should be reported. Psychology has similar, less dramatic, cases.

Improving record keeping would do more than improve replicability. Asendorpf et al. and others are concerned about pressures on authors to rush towards publication. Clear records aid an investigator in thinking about just how strongly a claim can be made, especially when the investigator realizes that the data will be available for examination. Similarly, record keeping will not prevent fraud, but it will make it somewhat more difficult. Good record keeping is also one of the best defences against unjustified charges of fraud. A lack of transparent records has been a factor in several such allegations, including the famous Cyril Burt case.

The professional societies, such as the Association for Psychological Science, are the logical agencies to be responsible for both establishing standards and maintaining the archives. The project is substantial but doable. Creating archives before record-keeping standards are established puts the cart before the horse.

Modifying behaviour

Asendorpf et al. recommend changes in the behaviour of researchers and in reviewing practices, both for manuscripts and for professional advancement. The recommendations for researchers to 'accelerate scientific progress' and to 'engage in scientific debate' are pure MAPPLE. The changes in reviewing and personnel evaluation practices, although eminently reasonable (almost to the point of MAPPLE), may have unintended consequences. The devil is once again in the details.

The current reviewing system is already overwhelmed. There is an inevitable conflict between the desire for rapid reviewing and careful reviewing. As for rewards, it is highly likely that reviewing will remain a virtuous activity. The rewards for virtue are well known.

Of course, evaluation committees should look at quality rather than quantity. MAPPLE! But quantity is objective, whereas judgments of quality are often subjective. This does not make evaluation of quality by 'learned judges' invalid. It does make decisions difficult to defend when review systems are held accountable for fairness, including unconscious biases, and productivity.

Statistical solutions

Asendorpf et al. propose changes in statistical practice and training that are good in themselves but that suffer from two problems: unintended consequences and conceptual limitation.

The statistical training curriculum in psychology is already overcrowded. A call to add something to the curriculum, however good that 'something' is in isolation, is MAPPLE, unless the call is accompanied by suggestions for dropping topics currently taught.

The conceptual limitation is more subtle.

Many discussions, including those of Asendorpf et al., seem to assume that a psychological study can be regarded as a point sampled from a space of possible studies. For example, they suggest that independent variables be analysed as random effects. This model works for laboratory studies but does not fit many studies outside the laboratory. Longitudinal studies take place at a particular place and time. And what about studies of major social trends, such as increases in intelligence test scores throughout the 20th century or the social and psychological effects of, say, the increase in urbanization throughout the world? Such studies can be extremely important to the social sciences, issues of transparency are highly relevant, but the relevance of models of statistical inference based on sampling is questionable.

In such cases, statistical models and significance tests provide useful descriptive statements because they contrast the results to ones that might have been observed in (unobtainable) ideal conditions. The statistics of a unique study cannot be used to support inferences about parameters in some nonexistent population of possible studies. Generalization should be based on a careful analysis of the study and the nature of the generalization desired. Statistical analysis is part of this process but often not the most important part.

So what to do?

Costs must be weighed against benefits. Increasing transparency is a low-cost, high-benefit action. It should be taken now.

The replicability problem is more complicated because costs are often indirect and because the remedies conflict with other legitimate goals of the scientific and academic systems. However, there is an unfortunate characteristic of the social and behavioural scientific literature that makes the issue more manageable.

Eighty-eight per cent of the 1745 articles in the 2005 Social Science Citation index received less than two citations (Bensman, 2008). Only four had more than 10. Highly cited studies should be replicated. The ever-more frequent publication of meta-analyses, including tests for 'file drawer' issues, shows that in fact this is being performed. Otherwise, meta-analysis would not be possible. Why bother to replicate the virtually uncited studies?

DOI: 10.1002/per

Rejoice! In Replication

HANS IJZERMAN, MARK J. BRANDT, AND JOB VAN WOLFEREN

Tilburg University
h.ijzerman@tilburguniversity.edu

Abstract

$$\frac{\text{solid science}\\ \text{theoretical advances}\\ \text{teaching opportunities} +}{\text{Rejoice!}}$$

We found the target article one of the most enlightening contributions to the 'replicability debate' through many cogent and nuanced recommendations for improving research standards. Contributions such as this quickly aid in remedying sloppy science ('slodderwetenschap') and enabling solid science (KNAW, 2012). The primary contribution of our commentary is the following equation:

$$\frac{\text{solid science}\\ \text{theoretical advances}\\ \text{teaching opportunities} +}{\text{Rejoice!}}$$

The case for replication as a part of solid science was made in the target article. We thus focus on the latter last two pieces of our equation.

Replications are theoretically consequential

Generally, we appreciate the recent contributions to the discussions of verifiability of research findings (e.g., Ferguson & Heene, 2012; Fiedler, Kutzner, & Krueger, 2012; Simmons et al., 2011). Conducting replications is a dirty business, and to date, few researchers have been motivated to do it. This may be mostly because, as the target article points out, researchers believe success of 'direct replications' to be unlikely (Fuchs, Jenny, and Fiedler, 2012).

This latter point is important because it shows one way that replications can help advance theory. High-powered failures to replicate, in our eyes, have two (or more) potential reasons (OSC, 2012), assuming that the replication study has adequate statistical power and the researcher the ability to replicate the study. First, failures to replicate can be interpreted as indications that the original effect is context dependent. Psychological findings are often influenced by many environmental factors, from culture (Henrich, Heine, and Norenzayan, 2010) to specific subpopulations in a culture (Henry, 2008), and even minor variations in the same laboratory (e.g., research on priming and social tuning; Sinclair, Lowery, Hardin, and Colangelo, 2005). Replications, thus, involve reproducing a variety of factors that are rarely recorded (and of which the original researchers may not be aware) and that may be as trivial as temperature or lighting (IJzerman & Koole, 2011).

This suggests that the ideal of 'direct replication' may be harder to achieve than expected and that *any* replication is a conceptual replication, with some being *more or less direct* than others. But fear not! Rejoice! Variations in replications can be to our theoretical *advantage* as they may illustrate which factors facilitated an effect in the original study and which factors prevented the effect from being observed in a replication attempt. *More* direct replications of a study's methods provide us with information regarding the stability of the effect and its contextual moderators. As suggested by the target article, when the effect size across replications is heterogeneous, moderators of the variation can provide valuable theoretical insights.

A second reason an effect may fail to replicate is that the effect size is small (and potentially zero) and thus more difficult to uncover than expected. In our experience, this is typically the assumed cause of a failure to replicate. Researchers thus consider, rightfully so, the possibility that initial findings result from type I errors. However, a failure to replicate is as convincing as the initial study (assuming similar power), and failures to replicate may actually increase one's confidence in an effect because they suggest there is not a vast hidden file drawer (Bakker et al., 2012; Schimmack, 2012). Presuming that an effect is due to a type I error after a single replication attempt is as problematic as committing that initial type I error (Doyen et al., 2012). However, multiple replication attempt effect sizes that are homogenous around zero (without reasons for the original effect to differ) suggest that the original effect was a fluke.

One direct implication is that replications require many attempts across multiple contexts to provide valid inferences. Only after systematic replications can we infer how robust and how veracious an effect is. Despite additional efforts, we believe that we should rejoice in replications as they lend credibility to research and help us make theoretical progress. Replications can thus help solve not only the 'replicability crisis' but also the 'theory crisis' (cf. Kruglanski, 2001). The true size of the effect, predictors of effect size variation, and knowledge of whether an effect is 'true' or not all advance understanding of human psychology.

Facilitating solid science: Walking the talk

Systematic replication attempts can be more easily achieved by facilitating transparency of published research and by systematically contributing to replication studies. To facilitate replications (and solid science more generally), we first examined our research practices. We determined that for other researchers to effectively

DOI: 10.1002/per

replicate *our* work, it is essential to trace *all* steps from raw data (participants' behaviour) to its final form (the published research paper). We upload all files to a national server, interface with Dataverse to provide a digital object identifier, and link them to the published paper. This should make it feasible for others to replicate the crucial aspects of our work. Our own detailed document can be found online (Tilburg Data Sharing Committee, 2012), including exceptions for researchers with sensitive data.

Provided that raw materials of research are easily available, replication becomes astonishingly easy to integrate into researchers' scholarly habits and teaching (Frank & Saxe, 2012). Recently, we have implemented replication projects with our bachelor students. With the current sample ($N = 3$), we can attest to how fun it is, how well students pick up on power analyses, and how easy it is to use this to let students first learn how to crawl before they walk in 'research land'. Thus, we should rejoice in replications because they solidify our science, facilitate theoretical advancement, and serve as valuable teaching tools.

Finally, although many researchers (including us) have pointed to the necessity of replication, without innovation, there is no replication. A research culture of pure replication is just as harmful for the future of the study of human psychology as a research culture of pure innovation and exploration. Taking seriously the idea of systematic replication attempts, in our eyes, forces us to go beyond *weird* samples and *odd* research methods (Rai & Fiske, 2011). As psychologists work through the current 'crisis', we urge researchers to both rejoice in replication and be enlightened in exploration.

Let Us Slow Down

LAURA A. KING

University of Missouri, Columbia
kingla@missouri.edu

Abstract: Although I largely agree with the proposals in the target article, I note that many of these suggestions echo past calls for changes in the field that have fallen on deaf ears. I present two examples that suggest some modest changes are underway. Finally, I suggest that one reason for the alarmingly high productivity in the field is that the research questions we are addressing are not particularly significant. Copyright © 2013 John Wiley & Sons, Ltd.

Although, generally speaking, I applaud the suggestions made by the esteemed authors of the target article, I cannot help but note that their voices join a chorus of similar calls that have been made not only in response to recent events but also historically. We have been lectured for decades about the problems inherent in null-hypothesis significance testing; the wonders of confidence intervals, effect sizes, and Bayesian analyses; the value of replication; and the importance of large samples. The necessities of reliable measurement and critical thinking are *de rigueur* in introductory psychology. Certainly, with regard to practice, the authors add some good and useful new ideas based on innovations in the field and the world, but the spirit of this call is not qualitatively different from calls we have been ignoring for years. The field has continued to rely on problematic practices and, if anything, has exacerbated them with increasing pressure to publish more and more (and shorter and shorter) articles, and to do so as quickly as possible. As a result, criticizing our research practices has become its own cottage industry. Will anything ever change? Here, I offer two bits of anecdotal evidence that the times might be a-changing. The first involves my own editorial consciousness raising and the second an inspiring tenure review.

I do not believe that top journals will (or could, or even *should*) begin to publish replications as stand-alone contributions. However, as an editor who reads these critiques, I have tried, in admittedly small ways, to institute greater respect (and occasional demand) for replications. For example, in its initial submission, one paper, currently 'in press' in the *Journal of Personality and Social Psychology*, presented several studies. The final study, the one that truly tested the main predictions of the package, was underpowered. The results were 'barely' significant and looked weird, as results from small sample studies often do. With the echoes of critiques of the field ringing in my ears, I drew a line in the sand and requested a *direct* replication with a much larger sample. I took this step with misgivings: Was it unfair? Should this work be held to such a standard? Was this 'revise and resubmit', in fact, its death knell in disguise? When the revision arrived, I fully expected a cover letter explaining why the replication was unnecessary. To my surprise, the direct replication was presented, and the predictions were strongly supported. Good news all around, but the experience gave me pause. True confession: I felt like I was demanding that those authors *hit the lottery. Twice*.

In our world of *p*-values, it is sometimes hard to remember that producing good science is not about hitting the lottery. Nor is it about taking whatever the data show and declaring that one *has* won the lottery. Importantly, within the editorial process, the 'preregistration' of predictions (that the authors of the target article suggest) often happens, inevitably. When new analyses or new studies are requested, as in the earlier case, authors' hands are tied. I realize that, typically, *JPSP* is considered a slow and stodgy animal in the publication world. Such lingering 'conversations' over papers would seem to be rather exceptional. If speed is the utmost value, journals are less likely to request new data than to simply reject a paper. If we could all slow down just a bit, it might help. A thoughtful and sometimes slow process has its advantages.

DOI: 10.1002/per

The nuance and breadth of the target article's treatment of institutional factors (especially in terms of tenure) warrant high praise. I believe that the problem of determining the quality of scholarship is just as complex as the authors suggest. Recently, I wrote a tenure letter for an apparently wise institution, with the following criterion noted:

> If you were to compile a list of the most significant books or articles to appear recently in this field, would any of the candidate's publications be on your list? Which ones? Why?

Such a criterion represents the kind of principles that ought to motivate our science, more generally. If we had this criterion in mind, what sorts of science might we produce?

Aside from data-faking scoundrels, we work very hard, as is evidenced in the astronomical number of articles published in our field. The target article suggests that at least some of this work is probably neither replicable nor particularly sound. Surely, changing our practices would improve all of this science.

But the gist of this critique, as well as others, is that in some ways, the sheer *amount* of research itself is problematic. And doing this science even very, very well would not necessarily reduce this enormous corpus of research. And here I come to the thought that suggested itself to me as I read and ruminated over the target article and that I hesitate to share. I do not mean to sound overly harsh or dismissive of the hard work we do. But, is it possible that we are able to produce so much because what we are producing is potentially trivial, relatively easy, and preoccupied with effects rather than phenomena? It seems to me that our energies are disproportionately dedicated to developing amazingly clever research designs rather than amazingly important *research questions*. Perhaps not only the practices but also the *content* of our scholarship requires rethinking. Yes, let us slow down and do our science right. But let us also slow down and remember that ours is the science of human behaviour. Too often, we limit ourselves to problems we have invented in the lab, to topics and variables that implicate very little in terms of human behaviour. Consider a paraphrase of the aforementioned criterion:

> If you were to compile a list of the most significant articles to appear in this field, would **any recent publications** be on your list?

With this criterion in mind, what sorts of *research questions* might we ask?

Stop Me If You Think You Have Heard This One Before: The Challenges of Implementing Methodological Reforms

RICHARD E. LUCAS AND M. BRENT DONNELLAN

Michigan State University
lucasri@msu.edu

Abstract: Asendorpf et al. offer important recommendations that will improve the replicability of psychological research. Many of these recommendations echo previous calls for methodological reform dating back decades. In this comment, we identify practical steps that can be taken to persuade researchers to share data, to follow appropriate research practices, or to conduct replications. Copyright © 2013 John Wiley & Sons, Ltd.

The target article offers several recommendations that will improve psychological research. These suggestions are based on sound methodological arguments and a common sense approach for building a cumulative science. Nothing the authors recommend strikes us as unreasonable or overly burdensome. Yet, their recommendations echo previous calls for more replication studies (e.g., Smith, 1970), greater statistical power (Cohen, 1962; Rossi, 1990), and increased transparency and data sharing (Johnson, 1964; Lykken, 1991; Wollins, 1962). These prior calls have gone unheeded, and thus, if there is to be any hope of lasting methodological reforms, the field must confront the obstacles that have prevented such reforms from being implemented in the past.

Although many psychologists will agree in principle with the suggestions made in the target article, we suspect there will also be vocal opposition to specific recommendations. Bakker et al. (2012) showed that the most successful strategy for finding a statistically significant result is to go against the recommendations of the target article and to run a large number of flexibly analysed studies with very small samples. Thus, in the current system, questionable research practices can produce the raw materials for a multistudy article that will be publishable in the most prestigious journals in our field. It will be difficult to convince those for whom this approach has been successful to change their behaviours.

In terms of implementation, the target article mainly focuses on making desirable research practices less burdensome. As one example, the authors highlight available resources for archiving data. However, it will be important to acknowledge that some researchers will object to specific policy changes for reasons that go beyond researcher burden. For instance, existing research shows that few authors are currently willing to provide data even to specific requests from other individual researchers (Wicherts, Borsboom, Kats, & Molenaar, 2006); we suspect that ease of sharing data is not the primary reason for refusal.

Currently, there are few consequences for researchers who fail to adhere to optimal research practices, such as sharing their data. Highlighting the problems with such policies to funding representatives may be fruitful, especially given the emphasis on accountability that accompanies funding. It is also disconcerting that journals published by the American Psychological Association and the Association for Psychological Science—journals that

DOI: 10.1002/per

are typically quite prestigious and could therefore afford a slight drop in submissions—have no stated penalties for researchers who go against guidelines and refuse to share data. One option is to make it standard journal policy that papers are retracted when authors refuse to share data from recently published papers unless there are compelling mitigating circumstances that prevent sharing. Any inability to share data with interested psychologists should be disclosed to the editor at the time of submission (Kashy, Donnellan, Ackerman, & Russell, 2009).

What are other ways that data sharing can be encouraged? One possibility is simply to make data sharing more normative. If you are interested in someone's data, you should request it and make sure you can replicate their results. In fact, it is probably not a bad idea to ask our close colleagues for their data, just to make the process more commonplace and less adversarial. As anyone who has been asked to share data knows, it only takes 1- or 2-day-long scrambles to compile and annotate existing messy data before you develop better procedures to prevent future occurrences.

In addition to targeting recommendations to those who have leverage, it is also worthwhile considering which recommendations have the largest 'bang for the buck'. It should be clear that many (if not most) studies in psychology are underpowered. The small sample sizes that plague our field have serious consequences in terms of imprecise parameter estimates and reduced credibility. Fortunately, this problem is easy to fix by demanding larger sample sizes. Editors and reviewers should simply require that authors start with the assumption that their effects will be no larger than what is typical for the field unless there is solid evidence that the specific effect under investigation will be larger. Thus, we suggest that power and precision be used as explicit grounds for a desk rejection.

Similarly, replication studies are easy to conduct and will have great benefit for the field. It is less important whether such replications are conducted by students or senior researchers or whether they are published in online repositories or special sections of existing journals. The real issue is making sure that the results are made available and that those who conduct independent replications are given credit for their efforts. Any reader who agrees with the recommendations provided in the target article can make an immediate contribution to the field by committing to conduct regular replications of their own and others' work and to make sure that the results are made accessible. In addition, concerned researchers should consider refusing to support journals that do not publish replications as a matter of policy.

The fact that so much has been written about methodological reform in the last 2 years is both encouraging and depressing. It is encouraging because these articles could be a harbinger of major changes in how psychological science is conducted. Such articles can also be depressing because the current discussions have an eerie similarity to those from the past decades. As it stands, many of the discussions about methodological reform operate on the assumption that there is basic agreement about the ultimate point of psychological research, which is to gain a clearer understanding of reality. However, it might be worth questioning this basic assumption. What if some researchers believe that the point of psychological science is simply to amass evidence for a particular theoretical proposition? Those with such a worldview might find the recommendations provided by the target article to be unnecessary roadblocks that limit their productivity. If so, then methodological reform needs to confront the reality that improving psychological research must involve changing hearts and minds as well as encouraging more concrete changes in behaviours.

Put Your Money Where Your Mouth Is: Incentivizing the Truth by Making Nonreplicability Costly

CORY A. RIETH[1], STEVEN T. PIANTADOSI[2], KEVIN A. SMITH[1], EDWARD VUL[1]

[1]*University of California, San Diego*
[2]*Univesity of Rochester*
edwardvul@gmail.com

Abstract: We argue that every published result should be backed by an author-issued 'nonreplication bounty': an amount of money the author is willing to pay if their result fails to replicate. We contrast the virtuous incentives and signals that arise in such a system with the confluence of factors that provide incentives to streamline publication of the low-confidence results that have precipitated the current replicability crisis in psychology. © 2013 The Authors. European Journal of Personality

A major part of the replicability 'crisis' in psychology is that commonly reported statistics often do not reflect the authors' confidence in their findings. Moreover, there is little incentive to attempt direct replications, as they are difficult, if not impossible, to publish. We propose a solution to both problems: For each result, authors must name a one-time nonreplication 'bounty' specifying the amount they would be willing to pay if the result did not replicate (e.g. $t(30) = 2.40$, $p < .05$, nonreplication bounty: \$1000). Thus, when you report a finding, you are effectively making a one-sided bet: if it replicates, you gain nothing, but if it fails to replicate, you pay the bounty using personal income. The bounty should be proportional to your confidence—if you are unsure, it could be \$1; if you know the results replicate, it could be a huge sum. This bounty measures the authors' subjective confidence on a scale that is universally interpretable, penalizes authors for overconfidence, and provides direct incentives for replication. Tabling the implementation details, consider the benefits:

(1) *Author confidence is clearly reported*

DOI: 10.1002/per

Ultimately, only the authors know exactly how their study was conducted and how they analysed their results. Their confidence is the best available signal of the robustness of their results, and a nonreplication bounty offers a clear signal of this confidence. This clear signal offers naïve readers an effortless assessment of the soundness of a result, as well as a quantitative metric to evaluate authors and journals. Thus, instead of rewarding raw publication and citation counts and encouraging the frequent publication of surprising, low-confidence results—one systemic problem contributing to the replicability crisis (Ledgerwood & Sherman, 2012)—sound results could be rewarded for both authors and journals.

(2) *Authors have incentive to provide an accurate signal*

The nonreplication bounty is not only a clear signal of confidence but also costly to fake. A low-confidence result offers authors two choices: overestimate their own confidence and suffer a considerable risk, or publish a result with low confidence, which readers will know should be ignored. Neither of these will be appealing, so authors will be altogether less eager to publish low-confidence results. If authors systematically overstate their own confidence, intentionally or not, they will face high costs and will either calibrate or leave the field.

(3) *Replications are directly encouraged*

Replication attempts receive direct incentives: Nonreplications pay a bounty. Moreover, replication attempts would be targeted towards the same results that naïve readers of the literature would have most confidence in: The higher the bounty, the more seriously the result will be taken, and the greater is the incentive for replications. Furthermore, such a system necessitates publication of replication successes and failures, adding further replication incentives.

We believe that many of the other proposed solutions to the replicability crisis ultimately will not work because they fail to provide appropriate incentive to authors (Nosek, Spies, & Motyl, 2012). For instance, the literature has suggested a number of metrics offering more reliable objective signals of result soundness: use of confidence intervals (Cumming & Finch, 2005), effect sizes (Cohen, 1994), Bayesian posterior intervals (Burton, Gurrin, & Campbell, 1998; Kruschke, Aguinis, & Joo, 2012), Bayes factors (Wagenmakers, Wetzels, Borsboom, & Van der Maas, 2011), and various disclaimers pertaining to the analysis procedures (Simmons, Nelson, & Simonsohn, 2012). Although these are useful statistical tools and policies, none is so sound as to avoid the possibility of being gamed, as they do not make errors costly to the authors. Running many low-powered studies, *post hoc* selection of independent or dependent variables, and other p-hackery (Simmons et al., 2011) would all yield nice results under these metrics. We believe that a remedy to these ailments must provide incentives to authors to offer clear, unbiased estimates of the soundness of their results, in place of the current incentives for authors to directly or indirectly overstate their confidence in and the reliability of their data.

Similarly, many proposals for remedying the replicability crisis (such as the target article) have focused on rules that publication gatekeepers (reviewers and editors) should enforce so as to increase the soundness of results. In contrast, nonreplication bounties would provide a clear and reliable signal that would alleviate some of the burden on volunteer reviewers and editors, rather than increase it. Authors would no longer receive incentives to sneak low-confidence results past reviewers, and reviewers could take on more thoughtful roles in trying to assess the validity of the measures and manipulations: Does the empirical result really have the theoretical and practical implications that the authors claim? Furthermore, as long as we have a reliable confidence signal associated with each result, there need not be an argument about whether type I or type II errors are more worrisome (Fiedler et al., 2012): Journal editors can choose to publish exciting, but speculative, findings or to publish only high-confidence results.

As proposed (Asendorpf et al., this issue; Koole & Lakens, 2012), encouraging replication attempts and the publicity of their outcomes is certainly beneficial. However, without quantitative metrics of result soundness, there is little incentive for journals to publish replications as impact factor only rewards short-term citations, which largely reflect the novelty and noteworthiness of a result.

The status quo indirectly provides incentives for rapid publication of low-confidence outcomes and their misrepresentation as high-confidence results: a practice that appears to be undermining the legitimacy of our science. We believe that local changes that do not restructure authors' incentives are only stopgaps for a deep-seated problem. Under our scheme, authors would have incentives to offer the most calibrated, precise estimates of the soundness of their available results.

Our position is best summarized by Alex Tabarrok (2012): 'I am for betting because I am against bullshit. Bullshit is polluting our discourse and drowning the facts. A bet costs the bullshitter more than the non-bullshitter so the willingness to bet signals honest belief. A bet is a tax on bullshit; and it is a just tax, tribute paid by the bullshitters to those with genuine knowledge'.

Increasing Replicability Requires Reallocating Research Resources

ULRICH SCHIMMACK AND GIUSEPPINA DINOLFO

University of Toronto Mississauga
uli.schimmack@utoronto.ca

Abstract: We strongly support the recommendation to increase sample sizes. We recommend that researchers, editors, and granting agencies take statistical power more seriously. Researchers need to realize that multiple studies, including exact replication studies, increase the chances of type II errors and reduce total power. As a result, they have to either publish inconclusive null results or use questionable research methods to report false-positive results. Given limited resources, researchers should use their resources to conduct fewer original studies with high power rather than use precious resources for exact replication studies. Copyright © 2013 John Wiley & Sons, Ltd.

DOI: 10.1002/per

Psychology has had a replicability problem for decades (Sterling, 1959), but it seems as if the time has finally come to improve scientific practices in psychology (Schimmack, 2012). Asendorpf et al. (this issue) make numerous recommendations that deserve careful consideration, and we can only comment on a few of them. We fully concur with the recommendation to increase sample sizes, but past attempts to raise sample sizes have failed (Cohen, 1990; Maxwell, 2004; Schimmack, 2012). One potential reason for the persistent status quo is that larger samples reduce research output (number of significant *p*-values). This is a disadvantage in a game (reinforcement schedule) that rewards quantity of publications. If sample size is ignored in evaluations of manuscripts, it is rational for researchers to conduct many studies with small samples. Thus, it is paramount to reward costly studies with adequate power in the review process. As most manuscripts contain more than one statistical test, it is also important to take the number of statistical tests into account (Maxwell, 2004; Schimmack, 2012). Even if a single statistical test has adequate power, total power (i.e. the power to obtain significant results for several tests) decreases exponentially with the number of statistical tests (Schimmack, 2012). As a result, holding other criteria constant, a manuscript with one study, one hypothesis, and a large sample is likely to produce more replicable results than a manuscript with many studies, multiple hypotheses, and small samples. We can only hope that editors will no longer reject manuscripts because they report only a single study because a greater number of studies is actually a negative predictor of replicability. Instead, editors should focus on total power and reward manuscripts that report studies with high statistical power because statistical power is essential for avoiding type I and II errors (Asendorf et al., in press; Maxwell, 2004).

We disagree with the recommendation that researchers should conduct (exact) replication studies because this recommendation is antithetical to the recommendation to increase sample sizes. Demanding a replication study is tantamount to asking researchers to split their original sample into two random halves and to demonstrate the effect twice. For example, if the original study and the replication study have 80% power, total power is only 64%, meaning every third set of studies produces at least one type II error. In contrast, combining the two samples produces one study with 98% power.

We agree with the recommendation to focus on research quality over quantity. We believe the main reason for the focus on quantity is that it is an objective indicator and easier to measure than subjective indicators of research quality. We think it is useful to complement number of publications with other objective indicators of quality such as number of citations and the h-index. Another useful indicator could be the incredibility index (Schimmack, 2012). A low incredibility index suggests that a researcher conducted studies with adequate power and was willing to publish null findings. In contrast, a high incredibility index suggests that a researcher used questionable research practices to publish results with a lower chance of replication.

We agree that funding agencies have the most power to change current research practices, but we do not agree that funding agencies should allocate resources to exact replication studies. It would be more beneficial for funding agencies to enforce good research practices so that original studies produce replicable results. Funding agencies already request power analyses in grant applications, but there is no indication that this requirement has increased power of published studies. A simple way to increase replicability would be to instruct review panels to pay more attention to total power and to fund research programmes that have a high probability to produce replicable results.

Finally, we agree with the recommendation to change the informal incentives in the field. Ideally, psychologists have a common goal of working together to obtain a better understanding of human nature. However, limited resources create conflict among psychologists. One way to decrease conflict would be to encourage collaboration. For example, granting agencies could reward applications by teams of researchers that pool resources to conduct studies that cannot be examined by a single researcher in one lab. It would also be helpful if researchers would be less attached to their theories or prior findings. Science is a process, and to see one's work as a contribution to a process makes it easier to accept that future work will improve and qualify earlier findings and conclusions. In this spirit, we hope that the authors of the target article see our comments as an attempt to contribute to a common goal to improve psychological science.

In Defence of Short and Sexy

DANIEL J. SIMONS

Department of Psychology, University of Illinois
dsimons@illinois.edu

Abstract: Proposals to remedy bad practices in psychology invariably highlight the 'problem' of brief empirical reports of 'sexy' findings. Even if such papers are disproportionately represented among the disputed findings in our field, discouraging brevity or interestingness is the wrong way to cure what ails us. We should encourage publication of reliable research and discourage underpowered or unreliable studies, regardless of their length or sexiness. Improved guidelines may help, but researchers will police themselves only if we change the incentive structure by systematically publishing direct replications. © 2013 The Authors. European Journal of Personality

DOI: 10.1002/per

Revelations of statistical bad practices, methodological shortcuts, and outright fraud in psychology have almost invariably led to criticism of novel, 'sexy' findings that seem disproportionately well represented among the contested claims in our field. Here I use the term "sexy" in the same manner that the target article does, to refer to a finding or claim that is provocative and exciting. True, many of the problematic findings in the literature take the form of sexy short reports. But they are problematic not by virtue of being brief or interesting, but by being wrong. Sexy findings need not be wrong, though, and dampening enthusiasm because a finding happens to be interesting or a paper brief will not cure what ails the field. We should encourage publication of highly powered, replicable, and interesting findings that people will want to read. We should dampen enthusiasm for (and, ideally, publication of) underpowered and unreliable studies, regardless of their length or the appeal of the topic.

Many of the changes proposed in the target paper will improve the collective quality of our publications. Journal articles are the currency of our field, and improved reporting requirements would increase their value. I applaud the new statistical and method standards adopted by the Psychonomic Society for all of its journals, the initiatives under consideration at the Association for Psychological Science to improve the state of our science, and the call in the target article for funding agencies to expect more rigorous statistical and methodological practices. But these changes are not enough because they will not address the real problem afflicting our field: the lack of incentives for individual researchers to publish *replicable* results.

Bad practices in psychology are prevalent in part because there is little public cost to being wrong and little direct benefit for being right. At present, the impact of a finding both for its author and for the field is largely unrelated to its correctness. Imagine conducting an underpowered study and finding a sexy result at $p < .05$. There are many incentives to publish that sexy result in a top journal immediately, including that the journal likely would want to publish it. Journal editors justifiably want to publish important, sexy findings that people will want to read. What incentive do you have to make sure your own work will replicate with a larger sample before publishing it? If you conducted the larger study, the extra effort might incrementally increase the chances of publication success, but probably not enough to justify the costs. The finding would garner the same visibility with or without the larger replication. A larger-scale replication would allow you to take pride in yourself as a good scientist who verifies before publishing, but most of the p-hackers among us already think of themselves as good scientists.

Once published, your sexy result will become the standard by which future studies are judged, you likely will serve as a reviewer for any findings that challenge yours (or build on it), and the literature will forever give more attention to your paper than any challenges to it (e.g., Bangerter & Heath, 2004). There is little danger that it will be corrected even if it is false (e.g., Ioannidis, 2012). Consequently, there are no incentives working against the immediate publication of sexy but underpowered studies.

The only incentives that would induce consistent changes in publishing practices are those that work for or against the interests of the individual researcher. We must provide incentives for publication of replicable findings and introduce consequences for publishing iffy ones. One initiative would have that effect: encouraging systematic publication of replications. Specifically, journals should encourage direct replications, conducted by multiple labs using the original protocol, and published regardless of outcome. The primary goal would be a cumulative estimate of the true effect size, but a secondary benefit would be a change to the publication incentives.

Imagine you have a new, sexy finding that just barely reached $p < .05$ with a small sample. Would you publish it right away if there were a sizable chance that multiple other labs would try to replicate it and their replication attempts would be published? The risk of embarrassment for being publicly wrong and the accompanying hit to your scientific credibility would provide a large incentive to make sure you are right before publishing, particularly if the result is sexy. To the extent that sexy findings challenge well-established evidence, they merit greater scrutiny: Extraordinary claims require extraordinary evidence. The sexier the claim, the more likely that other labs would want to replicate it, and the greater the incentive for the original researcher to make sure the result is solid before publishing. The end result might be fewer sexy findings in our top journals, but that outcome would emerge not by discouraging interesting results but by providing incentives for publication of reliable ones.

Better design, analysis, and reporting standards of the sort proposed in the target article are essential if we hope to improve the reliability and replicability of published psychology research, but only by changing the incentives for individual researchers can the field move away from publishing underpowered sexy findings and towards publication of well-powered, robust, and reliable sexy findings. With the right incentives in place, researchers will verify before publishing, and some initially promising results will vanish as a result. But sexy findings that withstand replication are the ones we want in our journals and the ones our journals should want to publish.

Replicability Without Stifling Redundancy

JEFFRY A. SIMPSON

University of Minnesota
simps108@umn.edu

Abstract: Many papers in psychology are written in a hypothesis-confirming mode, perhaps because authors believe that if some of their a priori predictions fail, their papers will be rejected. This practice must change. However, we must achieve replicability without stifling redundancy, which cannot occur unless scholars use (or develop) strong theoretical principles to derive, frame, and test their predictions. We also need to establish reasonable expectations about data sharing and data providing that are sensitive to the investment required to generate and maintain different kinds of datasets. © 2013 The Authors. European Journal of Personality

DOI: 10.1002/per

It is hard to disagree with most of what is said and recommended in this well-written and well-argued target article. Psychology has clearly reached a crossroads, and the time has come to focus much more attention on 'getting findings right'. As the authors note, portions of the blueprint for change already exist in other fields (e.g., medicine), but larger institutional values, priorities, and practices need to shift within psychology at the level of authors, editors and reviewers, departments and universities, and granting agencies.

Most papers in psychology are written in a hypothesis-confirming mode, which may partially explain why the current confirmation rate in psychology is 92% and has increased sharply during the last 20 years. Many authors implicitly believe that if even a few of their *a priori* predictions fail to work as planned, their papers will suffer in the review process. Some scholars (e.g., Bem, 1987) have actually advocated writing introductions so that they provide a coherent story that funnels readers towards *a priori* predictions that are predominately supported by the reported data. This practice has been harshly criticized by Kerr (1998) and others, and it needs to change. As the authors note, editors and reviewers can both play important roles in facilitating this change. However, we need to achieve replicability without stifling redundancy, which cannot occur unless scholars use strong theories to derive and test their predictions and to guide them when prior results consistently fail to replicate.

There is little if any argument that we, as a field, need to increase the size of our samples, improve the reliability (and validity) of our measures, ensure that our studies have sensitive designs, conduct proper statistical analyses, avoid reporting underpowered studies, and think more carefully about the error introduced when multiple statistical tests are performed. There is also little if any argument that authors should provide comprehensive literature reviews in their introductions, report their sample size decision making within papers, be much clearer about what their 'strong' *a priori* predictions actually are, and archive their research materials (and data, when realistic) so other investigators can evaluate what they have carried out. Authors also need to communicate more frequently, directly, and openly with colleagues who are conducting similar research, and not only individual investigators but also different *teams* of researchers located in different labs should routinely replicate each other's work when feasible.

Editors and reviewers also need to alter some of their expectations and practices. From my vantage point as the current editor of the *Journal of Personality and Social Psychology: Interpersonal Relations and Group Processes (JPSP: IRGP)*, I believe that we *cannot* view the 'perfectly confirmatory paper' as the gold standard for acceptance and that editors should be willing to publish well-conducted, sufficiently powered studies that fail to replicate important predictions and hypotheses. This is occurring at *JPSP*. The *Journal of Personality and Social Psychology: Attitudes and Social Cognition* section, for example, just published a set of studies that failed to replicate Bem's (2011) retroactive facilitation of recall effects (Galak, LeBoeuf, Nelson, & Simmons, 2012). *JPSP: IRGP* recently accepted a paper showing that the findings of several previous candidate gene studies do not replicate in the *National Institute of Child Health and Human Development* Study of Early Child Care and Youth Development dataset (Fraley, Roisman, Booth-LaForce, Owen, & Holland, in press).

Although the target article is exemplary in many ways, it does not address two sets of considerations relevant to the successful implementation of the recommendations offered. First, the article says relatively little about the essential roles that good theory and careful theorizing need to assume to make future findings in our field more replicable. The authors are correct in emphasizing that facets of studies vary in terms of individuals/dyads/groups (the observed units), situations (natural or experimental), operationalizations (manipulations, methods, and measures), and time points. They also acknowledge that 'Which [facet] dimensions are relevant depends on the relevant theory' (pp. XX of the target article). However, many researchers do not derive, frame, or test their hypotheses from the foundation of 'strong' theories that make specific predictions about the following: (i) which individuals should (and should not) show a specific effect; (ii) the situations or contexts in which the effect should (and should not) emerge; (iii) the manipulations, methods, or measures that should (and should not) produce the effect in certain people exposed to certain situations; and (iv) when the effect should be stronger and weaker (i.e. its time course). Some theories do offer reasonably good precision on some of these dimensions (e.g. certain diathesis–stress models; Simpson & Rholes, 2012), but more careful and detailed theorizing must be performed 'upfront' if future investigators are going to have a chance to replicate certain effects. Cast another way, we must do a better job of thinking theoretically to pin down how the most critical facets associated with different research designs should operate.

Second, the target article does not address the complications that may arise when data sharing extends beyond easier-to-collect cross-sectional experiments or self-report studies. Some research projects are extremely intensive in terms of time, effort, and cost, such as large *N* social interaction studies that may require years of behavioural coding, and major longitudinal projects that follow the same people over many years while collecting hundreds or sometimes thousands of measures. Scholars who work on these projects often devote most of their careers to these highly intensive data collection efforts, which can produce exactly what the authors call for—very high-quality data that can generate reliable, valuable, and very difficult-to-obtain findings. Unless data-sharing expectations and rules are carefully crafted, future investigators who might be interested in devoting their careers to collecting these high-investment datasets may be disinclined to do so, which would have a very negative impact on our field. Thus, there must be clear and reasonable expectations about both data sharing and data providing that are sensitive to the amount of investment required to generate and maintain different types of datasets.

DOI: 10.1002/per

There Is No Such Thing as Replication, but We Should Do It Anyway

BARBARA A. SPELLMAN

University of Virginia
spellman@virginia.edu

Abstract: Despite the fact that 'exact' replications are impossible to perform, it is important for a science to make close attempts and to systematically collect such attempts. It is also important, however, not to treat data as only relevant to the hypothesis being tested. Good data should be respected; they can tell us about things we have not yet thought. Plans for publishing systematic attempts at replication (regardless of outcome) are being developed. Copyright © 2013 John Wiley & Sons, Ltd.

Claiming 'there is no such thing as replication' may sound odd coming from a scientist, but the authors of the target article (Asendorpf et al., this issue) also make that point. More accurately, the statement should be, 'There is no such thing as *exact* replication'. Each study is different—different subjects, materials, time of day, time in history, and so on. The fact that each one is different is true not only in psychological science but also in other sciences. It is different atoms, bacteria, fossils, plants, and stars. The success of a replication depends on, among other things, the variability of the relevant features within the population studied. Often a scientist does not know the variability; and often a scientist does not even know the relevant features.

Neither one failure to replicate, nor one successful replication, tells us much. But a pattern of failures and successes does. One study is an existence proof—such a thing can happen. But having multiple studies, each slightly different, each varying on one or more dimensions, gives us information about robustness, about generalizability, about boundary conditions, about the features that matter, and, therefore, about our scientific theory.

In this comment, I discuss three points for psychological scientists. First, we should do more to respect our data. Second, we should do more to recognize that our data speak to more than one theory. Third, we should do more to amass our data to help us understand the robustness and generalizability of what we (think we) know. Finally, I report on a project in progress involving *Perspectives on Psychological Science* that should help with that last goal.

'The data are…'

A wonderful graduate school professor of mine, the late Tom Wickens, would often say, 'The data are…'. What he was doing was correcting our grammar, making sure that we knew that 'data' is the plural of 'datum' (and thereby reminding us that we were reporting more than one observation). But when I imagine Tom's long-ago admonition, I often think of an additional interpretation: We should give more respect to our data because 'the data are'. The data exist, and they are trying to tell us something. We should listen closely.

Scientists work hard to collect data, but sometimes we carelessly toss them away. That is fine if we discovered something truly 'bad' about the data—such as typos that changed the meaning of the stimuli or the measurement scales, a glitch in a randomization procedure, a manipulation check that reveals that subjects did not understand the instructions, the fact that your graduate student says to you 'I just finished running condition 1; I will start condition 2 tomorrow' (true story).

But that we collect 'good' data and then toss them away or bury them in a (virtual) file drawer because the study did not 'work'—that is, because it did not confirm our hypothesis or replicate a previous research finding (and that therefore we have no 'use' for it and no place to publish it)—well, that is sad for science. Good data are good data, and we should respect them.

Bull's eye

There is a story that is told in many guises but one version I like is this: 'A woman is driving through the countryside and sees a barn on which many huge targets are painted. Smack in the middle of the bull's eye in each target is a bullet hole. The woman stops and talks to the farmer. "You are such an amazing shot", she says, "a bull's eye every time." "Oh no", he says, "first I shoot at the barn, then I paint the targets around the holes."'

This feat is analogous to HARKing in science (hypothesizing after the results are known; Kerr, 1998). We have a hypothesis, we design a (good) study, we collect some data, but the study 'doesn't work' to confirm our hypothesis. However, after many analyses have been carried out, something in the study turns out to be significant, and we write an article as if those data answered the hypothesis we were asking all along; that is, we paint the target around the data rather than where we were aiming in the first place. Many current calls for reforming how we do science, including the target article, suggest that researchers register their hypotheses before testing them to avoid HARKing and to avoid the antics so nicely illustrated by Simmons et al. (2011) that caused subjects to age before their eyes.

Registering hypotheses is a fine idea but it seems equally important that we tell both what we were aiming for and what we hit. Listen to the data; it is informative about more than one hypothesis (Fiedler et al., 2012). As Nelson Goodman (1955, in his discussion of 'grue') said (more or less): 'Every piece of data is consistent with an infinite number of hypotheses; it is also inconsistent with an infinite number of hypotheses'.

DOI: 10.1002/per

Or, as I like to say: 'One scientist's .25 is another scientist's .01'.

What is to be done?

The target article makes some recommendations to increase not simply 'replicability' as their title says, but really our knowledge of which findings are robust and generalizable. Again, I agree. We need to not only save and publish more of our data but also better amass our results. We need better ways to connect our findings—not just knowing who has cited whom but what they have cited each other for (Spellman, 2012). We do not need to publish single replications or single failures to replicate; rather, we need systematic attempts at replication and meta-analysis that do not suffer from massive file drawer problems.

Among the suggestions of the target article is that journals be willing to 'go even further by launching calls to replicate important but controversial findings' with a guarantee of publication 'provided that there is agreement on method before the study is conducted'. In fact, *Perspectives on Psychological Science* has plans to do that in the works. I do not know if it will be in place by the time this comment is published, but readers can check for updates and instructions at http://morepops.wordpress.com.

The Significance Test Controversy Reloaded?

HANS WESTMEYER

Free University of Berlin
hans.westmeyer@fu-berlin.de

Abstract: Asendorpf et al. addressed the currently much-discussed problem of poor replicability of scientific findings in many areas of psychological research and recommended several reasonable measures to improve the situation. The current debate rekindles issues that have a long history in psychology and other social and behavioural sciences. In this comment, I will focus on some precursors of the current debate. Copyright © 2013 John Wiley & Sons, Ltd.

The target paper is a very important and highly welcome contribution to our current research practice. The replication of scientific findings is a neglected topic in many areas of psychology, and the recommendations for increasing replicability are well founded and worthy of adoption by researchers, editors, reviewers, teachers, employers, and granting agencies. The topic of replicability has a long history in our discipline and, at least in certain areas of psychology, has been with us all the time.

The target paper reminds me of a book entitled *The significance test controversy* edited by Morrison and Henkel (1970a). Their book is 'a reader representing the major issues in the continuing debate about the problems, pitfalls, and ultimate value of one of the most important tools of contemporary research' (text on the front cover). In one of the reprinted papers in this book, Sterling (1970, originally published in 1959) discussed 'publication decisions and their possible effects on inferences drawn from tests of significance—or vice versa'. He presented a table (p. 296) of the significance test outcomes performed in all contributions to four renowned psychology research journals published in 1955 (three journals) and in 1956 (one journal). The total number of published research reports was 362; 294 of these used significance tests; in 286 contributions, the null hypothesis had been rejected (alpha $\leq .05$); only in 8 out of 294 research reports (2.72%) had the null hypothesis not been rejected; not a single study was a replication of a previously published experiment. The target paper shows that the situation has not changed much within the last 50 years.

In a comment on the Sterling paper, Tullock (1970, originally published in 1959) drew the following conclusion: 'The tradition of independent repetition of experiments should be transferred from physics and chemistry to the areas where it is now a rarity. It should be realized that repeating an experiment, although not necessarily showing great originality of mind, is nevertheless an important function. Journals should make space for brief reports of such repetitions, and foundations should undertake their support. Academics in the social sciences should learn to feel no more embarrassment in repeating someone else's experiment than their colleagues in the physics and chemistry departments do now' (p. 302). That is not far from an admittedly very brief version of the recommendations given in the target paper.

There is one important disagreement between the editors of the aforementioned book, Morrison and Henkel, and the authors of the target article. Morrison and Henkel (1970b) came to very sceptical conclusions concerning the significance of significance tests in scientific research and briefly addressed the question, 'What do we do without significance tests?' (p. 310f), whereas the authors of the target article do not explicitly question the application of significance tests. At least they mention, as an alternative approach, parameter estimation and the computation of confidence intervals. This approach had also been addressed in the Morrison and Henkel book by a contribution by Rozeboom (1970, originally published in 1960).

One reason for drawing sceptical conclusions concerning the significance of significance tests in psychological research is the requirement of random samples drawn from specified populations (cf. Morrison & Henkel, 1970b, p. 305f). The authors of the target article emphasize this point: 'Brunswikian replicability requires that researchers define not only the population of participants, but also the universe of situations, operationalizations, and time points relevant to their designs'. This reminds me of the structuralist or

DOI: 10.1002/per

nonstatement view of scientific theories that requires determination of the set of intended applications as an indispensable part of any proper formulation of a scientific theory (or hypothesis; cf. Westmeyer, 1989, 1992). Let us remain more modest and be satisfied with studies conducted on random samples drawn from *defined populations of participants*. But are most psychological studies of this kind? I doubt it. Many of our studies are conducted on groups of students, quite often from our own department, without any previous population specification. These groups are not random samples; even the term 'convenience sample' is hardly appropriate. For something to be a sample, there has to be to a targeted population. What would the population for a group of students be? The population of all persons worldwide? The population of all students worldwide? The population of all students in a certain country? The population of students from a certain university? The population of students from a certain department? And what about the time points? Is a specification of the time points necessary, or do the respective populations also comprise future (and former) students? If we take the requirement of (random) sampling from prespecified populations seriously, a remarkable change in our research practice and the way we formulate our hypotheses is the consequence. That change would greatly facilitate the replicability of our findings. Differential psychology and psychological assessment are among the few areas of psychology in which many studies already satisfy the discussed requirement (e.g. when properly constructing tests).

It is regrettable that the target article does not refer to previous explications of the terms 'replication' and 'replicability'. These explications have been with us for a long time. Take, for example, differentiation of replication into *direct and systematic replications* by Sidman (1960) and further differentiation of direct replication into *intergroup or intersubject* and *intragroup or intrasubject replications*, not to mention still-further differentiations of systematic replication. For Sidman, replicability is one of the most important evaluation criteria for scientific findings, although there is no place for significance tests in his methodology. And take, for example, Lykken (1970, originally published in 1968), who introduced three kinds of replication: literal replication, operational replication, and constructive replication. Sidman's differentiations, in particular, would enrich the terminology proposed in the target article, which does not even mention experimental single-case studies as a possible alternative to the study of large samples (cf. Kazdin, 2010).

These omissions in no way decrease the importance and merits of the recommendations made in the target article. I really hope that the new debate will have long-lasting consequences.

AUTHORS' RESPONSE

Replication is More than Hitting the Lottery Twice

JENS B. ASENDORPF[1*], MARK CONNER[2], FILIP DE FRUYT[3], JAN DE HOUWER[4], JAAP J. A. DENISSEN[5], KLAUS FIEDLER[6], SUSANN FIEDLER[7], DAVID C. FUNDER[8], REINHOLD KLIEGL[9], BRIAN A. NOSEK[10], MARCO PERUGINI[11], BRENT W. ROBERTS[12], MANFRED SCHMITT[13], MARCEL A. G. VAN AKEN[14], HANNELORE WEBER[15], JELTE M. WICHERTS[5]

[1]*Department of Psychology, Humboldt University Berlin, Germany*
[2]*Institute of Psychological Sciences, University of Leeds, UK*
[3]*Department of Developmental, Personality and Social Psychology, Ghent University, Belgium*
[4]*Department of Experimental Clinical and Health Psychology, Ghent University, Belgium*
[5]*School of Social and Behavioral Sciences, Tilburg University, The Netherlands*
[6]*Department of Psychology, University of Heidelberg, Germany*
[7]*Max Planck Institute for Research on Collective Goods, Bonn, Germany*
[8]*Department of Psychology, University of California at Riverside, USA*
[9]*Department of Psychology, University of Potsdam, Germany*
[10]*Department of Psychology, University of Virginia, USA*
[11]*Department of Psychology, University of Milano-Bicocca, Italy*
[12]*Department of Psychology, University of Illinois, USA*
[13]*Department of Psychology, University of Koblenz-Landau, Germany*
[14]*Department of Psychology, Utrecht University, The Netherlands*
[15]*Department of Psychology, University of Greifswald, Germany*
jens.asendorpf@online.de

Abstract: The main goal of our target article was to provide concrete recommendations for improving the replicability of research findings. Most of the comments focus on this point. In addition, a few comments were concerned with the distinction between replicability and generalizability and the role of theory in replication. We address all comments within the conceptual structure of the target article and hope to convince readers that replication in psychological science amounts to much more than hitting the lottery twice. Copyright © 2013 John Wiley & Sons, Ltd.

DOI: 10.1002/per

We thank the commentators for their thoughtful, and sometimes amusing, remarks, constructive criticisms, and suggestions. We are delighted that most comments focused on concrete recommendations for improving the replicability of research findings, even describing concrete actions in line with some of our recommendations (e.g., **Simpson** and **Spellman**). Thereby, the peer commentary section and, we hope, our response contribute to the current debate in psychology about the poor replicability of research findings and how to improve it. To us, the most important and commonly expressed mindset to address was stated best by **King**—that replication is akin to 'hitting the lottery. Twice.' In this response, we hope to convince readers that empirical research is more than a game of luck and to keep in mind that the goal of any empirical study is to learn something. The role of chance in research is to provide an indication of confidence in the result, not to determine whether we won the game.

WHAT IS HISTORICALLY DIFFERENT THIS TIME?

Commenters noted the historical cycles of recognizing challenges in replicability and failing to take action or find correctives (see particularly **Westmeyer** and **King**). The current intense discussion could wither as well. However, we believe that it is different this time. First, prior cycles of this debate were somewhat isolated to specific areas of psychology and other disciplines. This time, the discussion is an explicit, intense, and widespread debate about the extent and the causes of nonreplication. The issue is dominating discussion across the sciences and includes all major stakeholders—societies, journals, funders, and scientists themselves. This gives the debate a stronger impetus than ever before, which, if wisely channelled towards 'getting it right', increases the chances for a truly self-correcting movement in our science.

Second, contributors to the debate recognize that the issue is systemic—not isolated to a particular practice, discipline, or part of the research process. Our target article acknowledges this by recommending actions at multiple levels. Third, there exists an infrastructure—the Internet—that can enable solutions such as data sharing on a scale that was simply not conceivable in previous epochs. Now, the barriers are not technical, they are social. Therefore, we are more optimistic than some of the commentators that the current debate offers opportunity for real reform and improvement.

NEED FOR REPLICATION

Two commentators questioned the need for conducting replication studies. **Francis** questioned replicability as a core requirement for psychological findings by drawing a distinction between physics and chemistry on the one hand and psychology on the other because psychological findings are more 'uncertain'. But, as quantum physics teaches us, uncertainty is inherent in many physical phenomena, and the role of statistics is to solve problems of probabilistic relations, whether in physics, chemistry, or psychology. **Francis** recommended meta-analysis as a solution for reducing uncertainty, and here we agree. But his arguments drew a false distinction between replication and meta-analysis. Replication is the stuff that makes meta-analysis possible (see also our section in the target article on 'small' meta-analyses for evaluating the replicability of an effect size).

Schimmack and Dinolfo did not question the importance of replicability, but they did question the usefulness of replication studies, with the argument that such studies are not needed if the original study was sufficiently powered. Although we certainly agree with the implied call for greater power, it is not realistic to imagine that all studies will be sufficiently powered. The central challenge is resource allocation. Researchers pushing the boundaries of knowledge take risks and venture into the unknown. In these cases, it is easy to justify placing a small bet to see if an idea has any merit. It is very difficult to justify placing a large bet at the outset of a research programme. We agree that this research strategy can lead to false positives resulting from many small bets, but it is also a means of reducing false negatives. If we can only place large bets, then we will take very few risks and miss perhaps the most important opportunities to learn something. So, what is the solution? Replication. When one finds some initial evidence, then a larger bet is justifiable. Our suggestion is that it is not only justifiable; it is essential. We believe that this strategy recognizes the conflicting challenges facing the pursuit of innovation and confirmation in knowledge accumulation.

Although it is true that one well-powered study is better than two, each with half the sample size (see also our section in the target article on the dangers of multiple underpowered studies), the argument ignores the point, reiterated by many other commentators, that exact replication is never possible; even studies designed as direct replications will inevitably vary some more or less subtle features of the original study. Thus, replication studies have merits even in an ideal Schimmack and Dinolfo world where only well-powered studies are conducted, by making sure that the design described by the original authors and copied by the replicators sufficiently describes all causally relevant features. In many areas of current psychology, well-powered replication attempts of equally well-powered original studies will sometimes fail, turning the replication studies into assessments of the limits of generalizability.

FROM REPLICABILITY TO GENERALIZABILITY

We view direct replicability as one extreme pole of a continuous dimension extending to broad generalizability at the other pole, ranging across multiple, theoretically relevant facets of study design. **Cacioppo and Caccioppo** called direct replication 'minimal replication' and linked inability to generalize to fruitful theoretical challenges. We fully endorse this view (see also **IJzerman et al**.). When replication fails, it can provide an opportunity for condition seeking—what are the boundary conditions for the effect?—that can stimulate theory advancement. We also like the argument by **Cacioppo and Caccioppo** that the multiple determination of virtually all psychological phenomena requires generalization rather than replication studies to appreciate a

DOI: 10.1002/per

phenomenon fully. Nevertheless, we insist that replicability is a necessary condition for further generalization and thus indispensable for building solid starting points for theoretical development. Without such starting points, research may become lost in endless fluctuation between alternative generalization studies that add numerous boundary conditions but fail to advance theory about why these boundary conditions exist.

ROLE OF THEORY

We agree that our recommendations could have done more to emphasize the role of theory. As **Simpson** correctly noted, we only briefly cited theory as a means of guiding the selection or construction of relevant design facets. The main reason is that our focus was on replication, not on generalization. In any case, we fully endorse **Simpson**'s and **Eid**'s views on the importance of theory for determining the relevant facets of an experimental design, for operationalizing them such that they fit the underlying theory, and for generating a design that is best suited to study the expected effects. Also, we like **Eid**'s discussion of the importance of deciding what should be considered measurement error and what should be considered substantive variation on theoretical grounds and his reminder that in many areas of psychology theories for important facets are underdeveloped or completely missing (e.g., a theory of stimuli as a prerequisite of a contextualized theory of perception or a theory of situations as a prerequisite of a contextualized theory of personality). We only insist that replication studies have their own virtue by providing solid starting points for generalization (see also the preceding section).

STUDY DESIGN AND DATA ANALYSIS

Only two comments focused directly on study design and data analysis. **Eid** noted that facets should not exclusively be considered random; whether they should be considered random or fixed is a theoretical issue. Actually, we did not propose in the target article that all facets should be considered random; instead, we proposed that researchers should at least *consider* that a facet might be better considered random rather than fixed. Whereas individuals are routinely treated as random factors, stimuli or situations are routinely considered fixed in most studies even though there are often good reasons for treating them as random. Related was **Westmeyer**'s remark that we discussed only designs including samples of individuals, ignoring single-case studies. We agree that we should have noted that our facet approach does include single-case studies as designs with no variation in the facet of individuals, just as many cross-sectional studies are designs with no variation in the facet of developmental time.

PUBLICATION PROCESS

Many comments concerned our recommendations for reforming the publication process on the part of reviewers, editors, and journals. We were most curious to read the comments by **Fanelli** because of his bird's-eye view on psychological publications in the context of publications in other areas of science and by the editors of flagship journals, **King**, **Simpson**, and **Spellman**, because we were quite critical about the current policies of many such journals that discourage direct replications and encourage sequences of underpowered studies.

Fanelli's remark about an equal citation rate of negative and positive results in psychological publications took us by surprise, because in the target article, we discussed confirmation bias of authors and publication bias of journal policies but not citation bias. Also, it seems to us that **Fanelli** underestimated the ability to predict study outcomes in at least some areas of psychology. To cite examples from personality psychology, the effect size of certain gender differences, the agreement between self and others on reliable measures of the Big Five factors of personality, and the longitudinal stability of such measures across a specified retest interval starting at a particular age can be predicted quite well. Psychology is not astrophysics, to be sure, but it offers much better predictions than astrology.

Therefore, we disagree with **Fanelli**'s negative view of the preregistration of hypotheses, based as it appears to be on his assumption of low predictability. Instead, we consider preregistration to be one of the most promising means for confirmatory testing. When the researcher has a strong *a priori* hypothesis, the best way to affirm the *p*-value's uncertainty estimation is to register the analysis plan in advance. Without it, flexibility in analysis strategies and motivated reasoning can lead to inflation of false positives and reduction of replicability in the process (see also the section on multiple hypothesis testing in the target article and **King**'s remarks on preregistration during longer review processes).

We fully agree with **Fanelli**'s view on the merits of purely exploratory research, but if and only if the research process and the results are fully and transparently reported. Such transparency requires standards for reporting, and we consider **Fanelli**'s suggestions for more specific reporting guidelines to be adopted by major journals a welcome addition to our own recommendations.

King's call for 'slowing down', by pressing authors for additional work invested in conducting additional studies or ruling out alternative explanations, is well taken in the current mad rush for quick-and-many publications. We would only add that instead of responding to a low-powered study by desk rejection as recommended by **Lucas and Donnellan**, a more constructive slowing-down response might be to ask for additional data to achieve sufficient power. An even better approach would be to take Cohen's call for sufficiently powered research seriously, just as many journals finally are beginning to take his call for reporting effect sizes seriously. Why do journals not adopt explicit rules that only studies with sufficient power to address their main research questions should be submitted?

For example, in line with conventional rules, we may define as acceptable thresholds power at .80 with alpha at .05. Given that recent meta-analyses converge in indicating that the average effect size in published psychological research is around $d = 0.50$, an approximate power calculation would result in $n = 100$ for a one-tail hypothesis for a simple

DOI: 10.1002/per

between-participants design (two groups) or a correlation coefficient (one group). Of course, there are many exceptions; within-participants designs have much more power, several effects are greater than $d = 0.50$, and so on. Therefore, this guideline should be flexible and adjustable to the conditions of specific studies.

The adoption of such a simple but flexible guideline would provide a clear incentive to authors to make a case, if needed, why in their specific study a different effect size should be expected given previous relevant studies and reasonable arguments. Thus, the authors should be able to justify why their specific sample size should give reliable results given the expected or investigated effect, without considering the results they obtained. If they did not do this, then the default rule of $n > 100$ would apply automatically, regardless of whether there were significant effects. Adoption of such rules would reduce the number of false positives and slow down the rate of publication. Slow publication in this sense may eventually become an indicator of quality similar to slow food.

For reasons spelled out in detail in the target article, we strongly disagree with *Journal of Personality and Social Psychology: Personality Processes and Individual Differences* editor **King**'s statement that replication studies should not be published in top journals. Interestingly, *Journal of Personality and Social Psychology: Interpersonal Relations and Group Processes* editor **Simpson** seems more favourable towards replication studies, at least if they present solid evidence that a seemingly established finding is not valid. We applaud **Simpson**'s view and would only ask that it should particularly be applied to failures to replicate findings published earlier in the same journal. After a decade of nonreplications of single-gene and functional magnetic resonance imaging results published in top biomedical journals, we are confident that such a policy would increase rather than decrease the reputation of any psychology journal that followed it.

We also share **Simpson**'s view that transparency, data archiving, and data sharing are particularly important for costly longitudinal and behavioural observation studies. Many funding agencies now require these for large projects, and journals could join the bandwagon by requiring them too, as long as confidentiality concerns or legal rights are not violated. In fact, the American Psychological Association publication guideline 8.14 requires data sharing on request of competent peers 'provided that the confidentiality of the participants can be protected and unless legal rights concerning proprietary data preclude their release', but it seems that this guideline is not taken seriously by authors and editors (Wicherts, Bakker & Molenaar, 2011). Retraction of an article because of violation of this guideline (as suggested by **Lucas and Donnellan**) should be a last resort, but a letter from the editor reminding an author of the commitment he or she has already signed may help to increase willingness to share data with peers.

We were particularly pleased by **Spellman**'s announcement that *Perspectives on Psychological Science* (*PPS*) will soon take up our suggestion of launching calls to replicate important but controversial findings with a guarantee of publication, provided that there is agreement on method before the study is conducted. In this action, *PPS* converges with the *European Journal of Personality*, which encourages such activities as well as articles concerned with replication issues. Spreading similar proactive encouragement of replication elsewhere would benefit much of our efforts. It would undoubtedly increase researchers' awareness of the importance of replicable findings and dampen the increasing unhealthy tendency over the past decade to look for 'sexy' findings that appeal to the mass media but later prove unreliable.

In his comment on this issue, **Simons** correctly pointed out that the 'sexiness' of a publication should not be a criterion for its quality, and we do not consider 'sexiness' as necessarily bad either. However, **Simons**' conclusion that '. . .sexy findings that withstand replication are the ones that we want in our journals' could be interpreted as 'sexy replicable findings are better than non-sexy replicable findings', which would run against the independence of 'sexiness' and scientific quality.

In a similar vein, we are sceptical about **King**'s call for slowing down by concentrating on 'significant' research questions. Although there are surely many nonsignificant questions around, what is viewed as significant may depend on what issues are currently mainstream and the flux and flow of fashions. Trying to steer science by significant questions may be as short-sighted as steering science by application questions. The history of science is full of examples where answers to questions that seemed awkward or trivial at the time later became critically important in a different and unforeseen context.

TEACHING

The enthusiastic comment by **IJzerman et al**. on the joys of teaching the importance of replication somewhat compensates for the fact that these joys were based on $N = 3$ students. **Hunt**'s perception that we are recommending more teaching of methodology and statistics, probably the most unpopular subjects for most psychology students at most departments, is a misinterpretation. We do not recommend *more* methodology and statistics; we recommend certain shifts of focus within the teaching of methodology and statistics (e.g., from null hypothesis testing in single studies to replication of effect sizes in multiple studies).

INSTITUTIONAL INCENTIVES

After many of us used Google to learn about **Hunt**'s usage of 'motherhood and apple pie' (it is always enchanting to learn new phrases of local dialect), we were additionally curious to learn what concrete recommendations he might offer that would differ from our own. We found two but disagree with both. First, we disagree with 'Creating archives before record-keeping standards are established puts the cart before the horse'. Standardization for documentation (within limits) is certainly a worthwhile goal, but waiting for standards is a good way to guarantee that archives will never happen. As the Internet age has demonstrated (e.g., formatting standards on Wikipedia), standards for communication are more productively pursued as an emergent quality with existing data

DOI: 10.1002/per

rather than developed in the abstract and then applied *en masse*. Waiting until professional societies agree on standards would be counterproductive—both for increasing sharing and for developing the standards.

Second, we disagree with **Hunt**'s suggestion that impact should be the sole criterion for launching replication studies. Relevance to scientific theory and opportunities to resolve controversy seem more important to us, and these are not always the same as impact. But we do agree with **Bakker et al**. that highly cited textbook findings need to be shown to be replicable; 'textbook-proof' is not sufficient, and we are pleased to see initiatives such as Open Science Framework (http://openscienceframework.org/) and PsychFileDrawer (http://psychfiledrawer.org/) providing environments for uploading and discussing the results of such replication studies.

Rieth et al.'s call for clearer signals of authors' confidence is not without merits, but we are more than sceptical about the specific suggestion of a nonreplication bounty. Assuming that the suggestion is serious and not satirical, such a measure would be misguided for two reasons. First, it would contribute to unhealthy tendencies to focus only on scientists' extrinsic motivation. As motivational psychology tells us, intrinsic motivations such as striving for discovery and truth can be corrupted by monetary reward and punishment. Second, if one wants to use money as an incentive, rewarding successful replications would seem much more productive (e.g., by reserving a percentage of grant money for replication) than punishing inability to replicate. The best way of 'changing hearts and minds' (**Lucas and Donnellan**) seems to us to be to use incentives that enhance intrinsic scientific motivation ('getting it better') and concern with peer reputation, as spelled out in some detail in the target article.

Conclusion

Taken as a package, we hope that our and the commentators' recommendations will counteract beliefs of some colleagues that successful replication amounts to hitting the lottery twice. We are convinced that psychological science can do much better than that now, and better still in the near future.

REFERENCES

Allport, G. W. (1968). The historical background of modern social psychology. In G. Lindzey, & E. Aronson (Eds), *The handbook of social psychology* (Vol. *1*, pp. 1–80). Reading, MA: Addison-Wesley.

Anisfeld, M. (1991). Neonatal imitation. *Developmental Review*, *11*, 60–97. doi: 10.1016/0273-2297.

Appley, M. H. (1990). Time for reintegration? *Science Agenda*, *3*, 12–13.

Augoustinos, M., Walker, I., & Donaghue, N. (2006). *Social cognition: An integrated introduction* (2nd ed). London, UK: Sage Publications Ltd.

Bakker, M., van Dijk, A., & Wicherts, J. M. (2012). The rules of the game called psychological science. *Perspectives on Psychological Science*, *7*, 543–554.

Bangerter, A., & Heath, C. (2004). The Mozart effect: Tracking the evolution of scientific legend. *British Journal of Social Psychology*, *43*, 605–623.

Bargh, J. A., Chen, M., & Burrows, L. (1996). Automaticity of social behavior: Direct effects of trait construct and stereotype activation on action. *Journal of Personality and Social Psychology*, *71*, 230–244. doi: 10.1037/0022-3514.71.2.230

Bem, D. J. (1987). Writing the empirical journal article. In M. Zanna, & J. Darley (Eds), *The compleat academic: A practical guide for the beginning social scientist* (pp. 171–201). New York: Random House.

Bem, D. J. (2011). Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology*, *100*, 407–425. doi: 10.1037/a0021524.

Bensman, S. J. (2008), Distributional differences of the impact factor in the sciences Versus the social sciences: An analysis of the probabilistic structure of the 2005 Journal Citation Reports. *Journal of the American society for information science and technology*, *59*, 1366–1382.

Berk, L. E. (2013). *Child development* (9th ed). Boston: Pearson.

Bless, H., Fiedler, K., & Strack, F. (2004). *Social cognition: How individuals construct reality*. East Sussex, UK: Psychology Press.

Burton, P., Gurrin, L., & Campbell, M. (1998). Clinical significance not statistical significance: A simple Bayesian alternative to *p* values. *Journal of Epidemiology and Community Health*, *52*(5), 318–323. doi:10.1136/jech.52.5.318.

Cacioppo, J. T., & Berntson, G. G. (1992). The principles of multiple, nonadditive, and reciprocal determinism: Implications for social psychological research and levels of analysis. In D. Ruble, P. Costanzo, & M. Oliveri (Eds), *The social psychology of mental health: Basic mechanisms and applications* (pp. 328–349). New York: Guilford Press.

Campbell, D. (1997). *In United States Patent and Trademark Office* (Ed.), *The Mozart effect*. US Patent 75094728.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait–multimethod matrix. *Psychological Bulletin*, *56*, 81–105.

Cesario, J., Plaks, J. E., & Higgins, E. T. (2006). Automatic social behavior as motivated preparation to interact. *Journal of Personality and Social Psychology*, *90*, 893. doi: 10.1037/0022-3514.90.6.893.

Cohen, J. (1962). Statistical power of abnormal–social psychological research: A review. *Journal of Abnormal and Social Psychology*, *65*, 145–153.

Cohen, J. (1990). Things I have learned (so far). *American Psychologist*, *45*(12), 1304–1312. doi:10.1037/0003-066X.45.12.1304.

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist*, *49*(12), 997–1003. doi:10.1037/0003-066X.49.12.997.

Cromie, W. J. (1999). *Mozart effect hits sour notes.* Retrieved 12/10, 2012, from http://news.harvard.edu/gazette/1999/09.16/mozart.html

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York: Wiley.

Cumming, G., & Finch, S. (2005). Inference by eye: confidence intervals and how to read pictures of data. *American Psychologist*, *60*(2), 170–80. doi:10.1037/0003-066X.60.2.170.

Doyen, S., Klein, O., Pichon, C.-L., & Cleeremans, A. (2012). Behavioral priming: It's all in the mind, but whose mind? *PLoS ONE*, *7*, e29081.

Eid, M., Geiser, C. & Nussbeck, F. W. (2009). Multitrait-multimethod analysis in psychotherapy research: New methodological approaches. *Psychotherapy Research*, *19*, 390–396.

Eid, M., Nussbeck, F., Geiser, C., Cole, D., Gollwitzer, M. & Lischetzke, T. (2008). Structural equation modeling of multitrait-multimethod data: Different models for different types of methods. *Psychological Methods*, *13*, 230–253.

Fanelli, D. (2010). "Positive" Results Increase Down the Hierarchy of the Sciences. *PLoS One*, *5*(3). doi: 10.1371/journal.pone.0010068.

Fanelli, D. (2012a). Positive results receive more citations, but only in some disciplines. *Scientometrics*, 1–9. doi: 10.1007/s11192-012-0757-y.

Fanelli, D. (2012b). *Project for a Scientific System Based on Transparency*. Paper presented at the EQUATOR Network Scientific Symposium Freiburg, Germany. http://www.equator-network.org/index.aspx?o=5605.

Fiedler, K., Kutzner, F., & Krueger, J. I. (2012). The long way from α-error control to validity proper: Problems with a short-sighted false-positive debate. *Perspectives on Psychological Science*, *7*, 661–669.

Fraley, R. C., Roisman, G. I., LaForce, C., Owen, M. T., & Holland, A. S. (in press). Interpersonal and genetic origins of adult attachment styles: A longitudinal study from infancy to early adulthood. *Journal of Personality and Social Psychology*.

Frank, M. C., & Saxe, R. (2012). Teaching replication. *Perspectives on Psychological Science*, *7*, 600–604

Fuchs, H., Jenny, M., & Fiedler, S. (2012). Psychologists are open to change, yet wary of rules. *Perspectives on Psychological Science*, *7*, 639–642.

Galak, J., LeBoeuf, R. A., Nelson, L. D., & Simmons, J. P. (2012). Correcting the past: Failures to replicate Psi. *Journal of Personality and Social Psychology*, *103*, 933–948. doi: 10.1037/a0029709.

Goodman, N. (1955). *Fact, fiction, and forecast*. Cambridge: Harvard University Press.

Hayes, L. A., & Watson, J. S (1981). Neonatal imitation: Fact or artifact? *Developmental Psychology*, *17*, 655–660. doi: 10.1037/0012-1649.17.5.655.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, *33*, 61–135.

Henry, P. J. (2008). College sophomores in the laboratory redux: Influences of a narrow data base on social psychology's view of the nature of prejudice. *Psychological Inquiry*, *19*, 49–71.

Hewstone, M., Stroebe, W., & Jonas, K. (2012). *An introduction to social psychology*. West Sussex, UK: Wiley-Blackwell.

Hull, J. G., Slone, L. B., Meteyer, K. B., & Matthews, A. R. (2002). The nonconsciousness of self-consciousness. *Journal of Personality and Social Psychology*, *83*, 406. doi: 10.1037/0022-3514.83.2.406.

IJzerman, H., & Koole, S. L. (2011). From perceptual rags to metaphoric riches: Bodily, social, and cultural constraints on socio-cognitive metaphors. *Psychological Bulletin*, *137*, 355–361.

Ioannidis, J. P. A. (2012). Why science is not necessarily self-correcting. *Perspectives in Psychological Science*, *7*, 645–654.

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

Johnson, R. W. (1964). Retain the original data! *American Psychologist*, *19*, 350–351.

Kashy, D. A., Donnellan, M. B., Ackerman, R. A., & Russell, D. W. (2009). Reporting and interpreting research in PSPB: Practices, principles, and pragmatics. *Personality and Social Psychology Bulletin*, *35*, 1131–1142.

Kazdin, A. E. (2010). *Single case research designs: Methods for clinical and applied settings* (2nd ed). New York: Oxford University Press.

Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review*, *2*, 196–217.

Koepke, J. E., Hamm, M., Legerstee, M., & Russell, M. (1983). Neonatal imitation: Two failures to replicate. *Infant behavior and development*, *6*, 97–102. doi: 10.1016/S0163-6383(83)80012-5.

Koninklijke Nederlandse Academie voor de Wetenschappen (KNAW, 2012). *Zorgvuldig en integer omgaan met wetenschappelijke onderzoeksgegevens [Handling scientific data with care and integrity]*. Retrieved December 2012 from http://www.knaw.nl/Content/Internet_KNAW/publicaties/pdf/20121004.pdf.

Koole, S. L., & Lakens, D. (2012). Rewarding Replications: A Sure and Simple Way to Improve Psychological Science. *Perspectives on Psychological Science*, *7*(6), 608–614. doi:10.1177/1745691612462586.

Kruglanski, A. W. (2001). That "vision thing": The state of theory in social and personality psychology at the edge of the new millennium. *Journal of Personality and Social Psychology*, *80*, 871–875.

Kruschke, J. K., Aguinis, H., & Joo, H. (2012). The Time Has Come: Bayesian Methods for Data Analysis in the Organizational Sciences. *Organizational Research Methods*, *15*(4), 722–752. doi:10.1177/1094428112457829.

LeBel, E. P., & Peters, K. R. (2011). Fearing the Future of Empirical Psychology: Bem's (2011) Evidence of Psi as a Case Study of Deficiencies in Modal Research Practice. *Review of General Psychology*, *15*(4), 371–379. doi: 10.1037/a0025172.

Ledgerwood, A., & Sherman, J. (2012). Short, sweet, and problematic? The rise of the short report in psychological science. *Perspectives on Psychological Science*, *7*(1), 60–66. doi:10.1177/1745691611427304.

Leman, P., Bremner, A., Parke, R. D., & Gauvain, M. (2012). *Developmental Psychology*. London: McGraw Hill.

Levelt Committee, Noort Committee, & Drenth Committee. (2012). *Flawed science: The fraudulent research practices of social psychologist Diederik Stapel*.

Lykken, D.T. (1968). Statistical significance in psychological research. *Psychological Bulletin*, *70*, 151–159. Reprinted in Morrison & Henkel (1970, pp. 267-279).

Lykken, D. T. (1991). What's wrong with psychology anyway? In D. Cichhetti, & W. M. Grove (Eds), *Thinking Clearly about Psychology. Volume I: Matters of Public Interest* (pp. 3–39). Minneapolis: MN: University of Minnesota Press.

Makel, M. C., Plucker, J. A., & Hegarty, B. (2012). Replications in psychology research: How often do they really occur? *Perspectives on Psychological Science*, *7*, 537–542. doi: 10.1177/1745691612460688.

Maxwell, S. E. (2004). The persistence of underpowered studies in psychological research: Causes, consequences, and remedies. [Article]. *Psychological Methods*, *9*(2), 147–163. doi:10.1037/1082-989X.9.2.147.

McCall, R. B., & Carriger, M. S. (1993). A meta-analysis of infant habituation and recognition memory performance as predictors of later IQ. *Child Development*, *64*, 57–79. doi: 10.1111/j.1467-8624.1993.tb02895.x.

Meltzoff, A. N., & Moore, M. K. (1977). Imitation of facial and manual gestures by human neonates. *Science*, *198*(4312), 75–78. Retrieved from http://www.jstor.org/stable/1744187

Morrison, D.E., & Henkel, R.E. (Eds). (1970a). *The significance test controversy - A reader*. Chicago: Aldine.

Morrison, D.E., & Henkel, R.E. (1970b). Significance tests in behavioral research: Skeptical conclusions and beyond. In D.E. Morrison & R.E. Henkel (Eds), *The significance test controversy - A reader* (pp. 305–311). Chicago: Aldine.

Newman, J., Rosenbach, J. H., Burns, K. L., Latimer, B. C., Matocha, H. R., & Rosenthal Vogt, E. (1995). An experimental test of the Mozart effect': Does listening to his music improve spatial ability? *Perceptual and Motor Skills*, *81*, 1379–1387. doi: 10.2466/pms.1995.81.3f.1379.

Nicholson, J. M., & Ioannidis, J. P. A. (2012). Research grants: Confirm and be funded. *Nature*, *492*, 34–36.

Nosek, B. A., Spies, J. R., & Motyl, M. (2012). Scientific Utopia: II. Restructuring Incentives and Practices to Promote Truth Over Publishability. *Perspectives on Psychological Science*, *7*(6), 615–631. doi:10.1177/1745691612459058.

Open Science Collaboration. (2012). An open, large-scale, collaborative effort to estimate the reproducibility of psychological science. *Perspectives on Psychological Science*, *7*, 652–655.

Pashler, H., Harris, C., & Coburn, N. *Elderly-Related Words Prime Slow Walking*. (2011, September 15). Retrieved 06:15, December 12, 2012 from http://www.PsychFileDrawer.org/replication.php?attempt=MTU%3D.

DOI: 10.1002/per

Petty, R. E., & Cacioppo, J. T. (1981). *Attitudes and persuasion: Classic and contemporary approaches*. Dubuque, Iowa: Wm. C. Brown.

Petty, R. E., & Cacioppo, J. T. (1986). *Communication and persuasion: Central and peripheral routes to attitude change*. New York: Springer-Verlag.

Pietschnig, J., Voracek, M., & Formann, A. K. (2010). Mozart effect–Shmozart effect: A meta-analysis. *Intelligence*, *38*, 314–323. doi:10.1016/j.intell.2010.03.001.

Rai, T. S., & Fiske, A. P. (2010). Psychological research methods are ODD (observation and description deprived). *Brain and Behavioral Science*, *33*, 106–107.

Rauscher, F. H., Shaw, G. L., & Ky, C. N. (1993). Music and spatial task performance. *Nature*, *365*, 611. doi: 10.1038/365611a0

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, *86*, 638–641. doi: 10.1037/0033-2909.86.3.638.

Rossi, J. S. (1990). Statistical power of psychological research: What have we gained in 20 years? *Journal of Consulting and Clinical Psychology*, *58*, 646–656.

Rozeboom, W. W. (1960). The fallacy of the null hypothesis significance test. *Psychological Bulletin*, *57*, 416–428. Reprinted in Morrison & Henkel (1970, pp. 216-230).

Schachter, S., Christenfeld, N., Ravina, B., & Bilous, F. (1991). Speech disfluency and the structure of knowledge. *Journal of Personality and Social Psychology*, *60*, 362–367.

Schimmack, U. (2012, August 27). The ironic effect of significant results on the credibility of multiple-study articles. *Psychological Methods*. Advance online publication. doi: 10.1037/a0029487.

Sechrest, L., Davis, M., Stickle, T., & McKnight, P. (2000). Understanding "method" variance. In L. Bickman (Ed.), *Research design: Donald Campbell's legacy* (pp. 63–87). Thousand Oaks, CA: Sage.

Sidman, M. (1960). *Tactics of scientific research. Evaluating experimental data in psychology*. New York: Basic Books.

Shaffer, D. R., & Kipp, K. (2009). *Developmental psychology: Childhood and adolescence* (8th ed.). Belmont, CA: Wadsworth.

Siegler, R. S., DeLoache, J. S., & Eisenberg, N. (2011). *How children develop* (3th ed.). New York: Worth publishers.

Simmons, J. P., Nelson, L. D., Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*, 1359–1366.

Simmons, J., Nelson, L., & Simonsohn, U. (2012). *A 21 Word Solution*. *Available at SSRN*: http://ssrn.com/abstract=2160588.

Simonton, D. K. (2004). Psychology's status as a scientific discipline: Its empirical placement within an implicit hierarchy of the sciences. *Review of General Psychology*, *8*(1), 59–67. doi: 10.1037/1089-2680.8.1.59.

Simpson, J. A., & Rholes, W. S. (2012). *Adult attachment orientations, stress, and romantic relationships*. In P. G. Devine, A. Plant, J. Olson, & M. Zanna (Eds.), *Advances in Experimental Social Psychology, 45*, 279-328. doi: 10.1016/B978-0-12-394286-9.00006-8.

Sinclair, S., Lowery, B. S., Hardin, C. D., & Colangelo, A. (2005). Social tuning of automatic racial attitudes: the role of affiliative motivation. *Journal of Personality and Social Psychology*, *89*, 583–592.

Smith, N.C, Jr. (1970). Replication studies: A neglected aspect of psychological research. *American Psychologist*, *25*, 970–975.

Spellman, B. A. (2012). Scientific utopia. . . or too much information? Comment on Nosek and Bar-Anan. *Psychological Inquiry*, *23*, 303–304.

Staats, A. W. (1989). Unificationism: Philosophy for the modern disunified science of psychology. *Philosophical Psychology*, *2*, 143–164.

Sterling, T. D. (1959). Publication decisions and their possible effects on inferences drawn from tests of significance: Or vice versa. *Journal of the American Statistical Association*, *54*, 30–34. doi:10.2307/2282137.

Tabarrok, A. (2012, Nov 2). *A Bet is a Tax on Bullshit*. Marginal Revolution. Retrieved from http://marginalrevolution.com/marginalrevolution/2012/11/a-bet-is-a-tax-on-bullshit.html.

Tilburg Data Sharing Committee (2012). *Manual for data-sharing*. Retrieved December 2012 from http://www.academia.edu/2233260/Manual_for_Data_Sharing_—_Tilburg_University.

Tullock, G. (1959). Publication decisions and tests of significance: A comment. *Journal of the American Statistical Association*, *54*, 593. Reprinted in Morrison & Henkel (1970, pp. 301-302).

Wagenmakers, E., Wetzels, R., Borsboom, D., & Van der Maas, H. (2011). Why psychologists must change the way they analyze their data: the case of psi: Comment on Bem (2011). *Journal of Personality and Social Psychology*, *100*(3), 426–432. doi:10.1037/a0022790.

Wagenmakers, E. J., Wetzels, R., Borsboom, D., van der Maas, H. L. J., & Kievit, R. A. (2012). An agenda for purely confirmatory research. *Perspectives on Psychological Science*, *7*, 632–638. doi:10.1177/1745691612463078.

Weisburd, D., & Piquero, A. R. (2008). *How well do criminologists explain crime? Statistical modeling in published studies Crime and Justice: a Review of Research*, (Vol. *37*, pp. 453–502). Chicago: Univ Chicago Press.

Westmeyer, H. (Ed.) (1989). *Psychological theories* from a structuralist point of view. New York: Springer-Verlag.

Westmeyer, H. (Ed.) (1992). *The structuralist program in psychology: Foundations and applications*. Toronto: Hogrefe & Huber Publishers.

Wicherts, J. M., Borsboom, D., Kats, J., Molenaar, D. (2006). The poor availability of psychological research data for reanalysis. *American Psychologist*, *61*, 726–728.

Wicherts, J. M., Bakker, M., & Molenaar, D. (2011). Willingness to share research data is related to the strength of the evidence and the quality of reporting of statistical results. *PLoS One*, *6*, e26828.

Wollins, L. (1962). Responsibility for raw data. *American Psychologist*, *17*, 657–658.

Yong, E. (2012). Nobel laureate challenges psychologists to clean up their act: Social-priming research needs "daisy chain" of replication. *Nature*, *485*(7398), 298–300.

DOI: 10.1002/per