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. 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

* Correspondence to: Jens B. Asendorpf, Department of Psychology, Humboldt University, Unter den Linden 6, 10099 Berlin, Germany. E-mail: jens.asendorpf@online.de

European Journal of Personality, in press, November 5, 2012

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 toward 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.

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 easily 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.

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 over-developed culture of encouraging brief, eye-catching research publications that appeal to the media, and 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 toward 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 later section on sample sizes). The rate seems to be inflated by selective non-reporting of non-confirmations as well as *post hoc* invention of hypotheses and study designs that do not subject hypotheses to possibility of refutation. In contrast to the rosy picture presented by publications, in a recent worldwide poll of more than 1,000 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 2,000 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, details of structural equation models; Simmons, Nelson, & Simonsohn, 2011).

The question 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 pre-cognition 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 non-significant (see 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 non-replicability is the common thread that runs through most of the current debate, we address the 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 agencies contribute 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] In order to check reproducibility, Researcher B must have (a) the raw data, (b) the code book (variable names and labels, value labels, 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 (a) individuals (or dyads or groups), (b) situations (natural or experimental), (c) operationalizations (experimental manipulations, methods, 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 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 requires that researchers define not only the population of participantss 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 non-replication.

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 females, leaving it unclear to what extent results can be generalized to a population sample of males and females. 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 done in a specific historic context that is always changing). This

argument is often used to defend business as usual and avoid the problem of non-replication 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 (see below) 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 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 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).

1. 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 $\beta = .35$ (Bakker, Van Dijk, & Wicherts, in press). 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 $\beta = .59$, 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.

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; Fig. 2). Therefore, a focus on power suggests larger sample sizes for small effects and smaller ones for large effects compared to 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 confidence intervals 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, though studies labeled 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.

2. 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* confidence intervals 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.

3. 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.

4. 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 dependencies 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 (see Lee, 2012).

5. 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 (in press), 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.

6. Consider error introduced by multiple testing

When a study involves many variables and their interrelations, following the above recommendations becomes more complicated. As shown by Maxwell (2004), likelihood that some among multiple variables will show significant relations with another variable is higher with underpowered studies, although 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 non-significant) 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 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, though 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 (FDR, 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. Non-statistical solutions are required such as the explicit separation of *a priori* hypotheses pre-registered in a repository from exploratory *post hoc* hypotheses (see section on Implementation below).

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 an 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), $p = .051$ (second study). It can also be misleading if replication studies are underpowered, making non-replication 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, 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 $r = .25$ [.02, .76], which is significant at $p = .037$. A high-powered replication study of 1000 participants finds a correlation of $r = .05$ [-.01, .11], which besides being trivial is not significant ($p = .114$). 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, use of multiple inferential strategies along the lines suggested by Valentine et al. (2011, see 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.

1. 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. To make the research procedure 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) Pre-register research predictions. Where researchers have strong predictions these and the analysis plan for testing them should be registered prior to commencing the research (see section on Implementation below). Such pre-registered predictions should be labelled as such in the research reports and might be considered additional markers of quality. Pre-registration 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, 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 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 (see the section on Implementation below). Options are likely to vary in degree of technical sophistication.

2. 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, pre-publications can be posted in online repositories (see section on implementation below). The most prominent pre-print 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 above 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 creates 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) above.

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. In order to increase replicability, reviewers, editors, and journals should allow for and encourage the implementation of good research practices.

1. 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 non-replications 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 (see Schimmack, in press).

For example, if an article reports ten successful confirmations of an (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, in press).

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.

2. Pro-actively 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., pre-registration of hypotheses). As well, 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 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 analyzing the data from individual studies using null hypothesis significance testing. Undergraduate and even graduate statistical education is 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, has a checkered history, and other approaches are available (Gigerenzer, Swijnk, Porter, Daston, Beatty & Krüger, 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.

1. 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).

2. 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., confidence intervals) rather than outcomes of significance testing. Students should also learn to appreciate the value of non-significant 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 due to 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 (see later section on Implementation) and taught best practices in research (see Recommendations for Authors). To teach the value of replication of statistical analyses, students should reanalyze 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, analyzing, 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 they experience will reinforce its

importance.

3. Critical thinking

(a) Critical reading. Learning to see the advantages but 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 how to interpret effect sizes and confidence intervals correctly by means of examples. A didactical approach focused on multiple studies is well suited to explain 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 favor of or against a hypothesized effect.

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

(c) Critical evaluation of evidence (multi-study 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 above 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, submit and tolerate more null findings, focus more on calibrating estimation of effects rather than null hypothesis significance testing, and communicate the need for doing so to students, the culture will naturally accommodate the new values. That said, we are skeptical 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.

1. 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 non-replicable

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 above 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.

2. 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 endeavor 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 program 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 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 sub-disciplines. Some sub-disciplines 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 sub-disciplines with traditions of larger and more theoretically elaborated publications.

3. 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 (see 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 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) argues 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 (see 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 in order 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.

4. Change informal incentives

Finally, informal incentives within our guilds need to change in order 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 non-replicated 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 behavior 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 behavioral change. Ensuring success thus requires attention to the infrastructure and procedures required to implement the improvments.

The internet is a mechanism for sharing of materials and data that addresses some of the practical barriers. But its existence is not sufficient. A system is needed that (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 non-profit 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: CLARIN (Common Language Resources and Technology Infrastructure; http://www.clarin.eu/), financed by the European Seventh Framework programme, and DARIAH (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 programs 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 like 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 (SSRN; 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 trials 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 sub-disciplines. There are also groups organized around specific topics (e.g., on cognitive modeling, 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 options mentioned above within a single framework: The Open Science Framework (OSF; http://opensciencefframework.org/). The OSF 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 URL 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, toward improving scientific rigor.

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 confidence intervals 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 programs 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, like the recent fraud scandals and the publication of papers 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) is the reason 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. (in press). The rules of the game called psychological science. *Perspectives on Psychological Science*.

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

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.

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

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. *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.*

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

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, UK: 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 Med, 2,* e124.

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

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.

Nickerson, R.S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods*, *5*, 241–301.

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.*

Schimmack, U. (in press). The ironic effect of significant results on the credibility of multiple-study articles. *Psychological Methods*.

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*.

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., Kellam, S., Moscicki, E. K., & Schinke, S. P. (2011). Replication in prevention science. *Prevention Science*, *12*, 103-117.

Wachtel, P. L. (1980). Investigation and its discontents: Some constraints on progress in psychological research. *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

Author notes

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.

Footnotes

Footnote 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 in order to distinguish it from replicability.

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