AUTHOR's 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

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

final draft, January 15, 2013

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.

We thank the commentators for their thoughtful, and sometimes amusing, remarks, constructive criticism, and suggestions. We are delighted that most comments focus on concrete recommendations for improving the replicability of research findings, even by describing concrete actions in line with some of our recommendations (e. g., **Simpson**, **Spellman**). Thereby the peer commentary section and, hopefully, also our responses contribute to the translation of the current debate in psychology about the non-replicability of research findings toward concrete recommendations for improving replicability. To us, the most important, popular mindset to address was expressed best by **King** – that replication is akin to "hitting the lottery. Twice." In this response we hope to convince readers to alter their mindset that empirical research is a game of luck, and instead to remember that the goal of any empirical study is to learn something. The role of chance in research is in providing an indication of confidence in the result, not in determining whether we won the game.

What is Historically Different?

Commenters noted the historical context of recognizing challenges in replicability and of failing to take action or find correctives (see particularly **Westmeyer** and **King**). The intense

discussion at present could wither as well. However, we believe that it is different this time. First, prior cycles of this debate were somewhat isolated to different areas of psychology and other disciplines. This time, the discussion is an explicit, intense and widespread debate about the extent and the causes of non-replication. 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 channeled toward "getting it right", raises the stakes for a truly self-correcting movement in our science.

Second, the contributors to the debate recognized 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 like 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. Later, **Francis** continues to recommend meta-analysis as a solution for reducing uncertainty, and here we agree. But his

arguments draw 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 do not question the importance of replicability but they do question the usefulness of replication studies, with the argument that such studies are not needed if the original study was sufficiently powered. While we certainly agree with the call for greater power, it is not realistic to imagine that all studies will be sufficiently powered. The central challenge is resource allocation. Researchers that are 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 the idea has any merit. It is very difficult to justify placing a large bet at the outset of a research program. We agree this research strategy can lead to more false positives with lots of 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 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 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 non-generalization 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 in order to fully appreciate a phenomenon. 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 get lost in endless fluctuation between alternative generalization studies that add numerous boundary conditions but fail to advance theory.

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 view on the

importance of theory for determining the relevant facets of the 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 under-developed or completely missing (e. g., a theory of stimuli as a prerequisite of a contexualized 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 these issues. **Eid** noted that facets should not exclusively be considered as 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 defined as 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, as well. 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 on the facet of individuals, just as many cross-sectional studies are designs with no variation on 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 eyes 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 under-powered 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 on the pre-registration of hypotheses based on his assumption of low predictability. Instead, we consider pre-registration 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 pre-registration 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 plain 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 be submitted?

For example, in line with conventional rules, we may define as acceptable thresholds power at 0.80 with alpha at 0.05. 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 between-participants design (2 groups) or a correlation coefficient (1 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 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 trustable results given the expected or investigated effect, without considering the results reported in their study. If they don't do this, then the default rule of $n \geq 100$ would apply automatically, regardless of whether there are significant effects. Adoption of such rules would reduce the number of false positives and slow down the rate of publications, and 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 favorable toward 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 the non-replication of findings published earlier in the same journal. After a decade of non-replications of single gene and fMRI 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 behavioral 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, APA 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 the author's 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 pro-active encouragement of replication elsewhere would pay off much of our efforts. It would undoubtedly increase researchers' awareness of the importance of replicable findings and dampen the unhealthy tendency over the past decade to increasingly look out for "sexy" but unreliable findings that appeal to the mass media.

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 skeptical about **King**'s call for slowing down by concentrating on "significant" research questions. Although there are surely many non-significant 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 misperception. 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 curious to additionally 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 (see e.g., formatting standards on Wikipedia) standards for communication are more productively pursued as an emergent quality with existing data 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 an environment 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 skeptical about the specific suggestion of a bounty for non-replication. 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 replications would seem much more productive (e .g. by reserving a percentage of grant money for replication) than punishing non-replication. The best way of "changing hearts and minds" (**Lucas and Donnellan**) seems to us using incentives that enhance intrinsic motivation ("getting it better") and to incentives related to 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 feelings 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 even

more so in the near future.

References

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.