Taylor & Francis
Taylor & Francis Group

Accountability in Research

Policies and Quality Assurance

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/gacr20

Replication and trustworthiness

Rik Peels & Lex Bouter

To cite this article: Rik Peels & Lex Bouter (2021): Replication and trustworthiness, Accountability in Research, DOI: 10.1080/08989621.2021.1963708

To link to this article: https://doi.org/10.1080/08989621.2021.1963708

Published online: 08 Aug 2021.

ACCOUNTABILITY IN RESEARCH, 2021
https://doi.org/10.1080/08989621.2021.1963708

Taylor & Francis
Taylor & Francis Group

Replication and trustworthiness

Rik Peels[a] and Lex Bouter[b]

[a]Philosophy Departmentand Faculty of Religion and Theology, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands; [b]Department Of Epidemiology And Data Science, Amsterdam University Medical Centers, Amsterdam, The Netherlands

ABSTRACT

This paper explores various relations that exist between replication and trustworthiness. After defining "trust", "trustworthiness", "replicability", "replication study", and "successful replication", we consider, respectively, how trustworthiness relates to each of the three main kinds of replication: reproductions, direct replications, and conceptual replications. Subsequently, we explore how trustworthiness relates to the intentionality of a replication. After that, we discuss whether the trustworthiness of research findings depends merely on evidential considerations or also on what is at stake. We conclude by adding replication to the other issues that should be considered in assessing the trustworthiness of research findings: (1) the likelihood of the findings before the primary study was done (that is, the prior probability of the findings), (2) the study size and the methodological quality of the primary study, (3) the number of replications that were performed and the quality and consistency of their aggregated findings, and (4) what is at stake.

KEYWORDS

Replication; trustworthiness; trust; replicability; reproducibility

1. Introduction

Some have dubbed the replication crisis a "trust crisis" (Hendriks, Kienhues, and Bromme 2020). It seems indeed plausible that successful replication can increase the trustworthiness of the findings of the study that was replicated (Bouter and Ter Riet, 2021). However, this claim, though intuitively plausible, raises many questions. What is object of trust: the findings of the primary study, the quality of the primary study protocol, the combined findings of the primary study and its replication, only the main findings or all findings reported, or yet something else? Is replication *required* for trustworthiness? Does a study's being replicable without the actual attempt having been made already make it more trustworthy? Does a successful replication always increase trustworthiness? Does a failure to replicate always come with a decrease in trustworthiness? Should a replication be intentional in order to increase trustworthiness? How often should a study be

CONTACT Rik Peels Philosophy Department and Faculty of Religion and Theology, Vrije Universiteit Amsterdam, de Boelelaan 1105, Amsterdam, 1081 The Netherlands mail@rikpeels.nl

successfully replicated to confidently be able to trust its findings? Is the trustworthiness of research findings also dependent on what is at stake when they are put into practice? This paper will try to answer these questions by exploring the relation between trustworthiness and replication.

2. Trustworthiness, replicability, and replication

The notion of *trust* has recently received much conceptual attention in ethics, epistemology, and social epistemology (O'Neill 2002; Faulkner and Simpson, 2017; McLeod 2015). A widely adopted view is that one trusts that some proposition p is true if and only if (i) one takes the truth of p for granted in one's practical and theoretical reasoning, and (ii) one has some sort of positive attitude toward p. Thus, trust in study findings would be something like: taking it for granted that they are reliably produced, based on solid methods and sound data, well founded on what we already know, and so on.

Let us now turn to trust*worthiness*, that is, something's being *worthy of* trust. The notion of trustworthiness has a normative dimension: it implies that it is proper, justified or rational to trust the study findings in question. If one trusts the findings of a study, but merely because it was performed by someone who has such a high status in the field that no one dares to question her work, the study and its findings are *trusted* but not necessarily *trustworthy*. Findings of a study are trustworthy if there is good reason to think that they are true or at least sufficiently close to the truth and that they are based on sufficient high-quality evidence. For those who, in the wake of Karl Popper, hesitate to use the concept of truth for research findings, we can say that findings are trustworthy if they have been subjected to thorough testing and have not yet been falsified. It is, of course, a matter of debate exactly when there is *good reason* to think that a study's findings are trustworthy and we will not try to settle that debate here. Our point is merely that we should not confuse *trust in* research findings with the *trustworthiness of* these findings.

Trustworthiness is always trustworthiness *for someone*. Obviously, findings may be trustworthy to some, but not to others. We have in mind trustworthiness *for the academic community* rather than for the non-expert public or for policy makers. The non-expert public will often not have the relevant background knowledge to assess research findings and policy makers have to consider numerous practical and tactical considerations that go far beyond epistemic ones. We take it, therefore, that findings are trustworthy if the relevant academic community can properly trust these findings.

It is also important to be explicit on how we use "replicability", "replication", and "successful replication", as these words are understood quite differently in different academic fields and have gone through many changes in meaning and interpretation over time (Vachon et al. 2020). For our

exploration of the relation between replication and trustworthiness we selected a taxonomy of three types of replication that was also used by others (Bouter and Ter Riet 2021; KNAW: Royal Dutch Academy of Arts and Sciences 2018), where we take "research protocol" to refer to the combination of a study protocol and a data-analysis plan:

(1) Replication with *existing data* and the *same research protocol*: re-analysis of the data from the primary study with the primary research question. This we refer to as a "reproduction". It can be argued that a publication should first be checked for numerical inconsistencies in the data and in the statistical parameters. Then a re-analysis of the existing data with the same data-analysis plan comes in view, potentially followed by analyses of the existing data with one or more alternative data-analyses plans. When these tests are passed satisfactorily it can be concluded that the findings of the primary study are robust (Nuijten 2021).[1]
(2) Replication with a *new data* collection and with the *same research protocol* as the primary study. This we refer to as a "direct replication".
(3) Replication with *new data* and with a *somewhat different research protocol* aimed at answering the same research question. This we refer to as a "conceptual replication".

Of course, this threefold distinction is somewhat arbitrary – for a slightly different way to carve things up, see, for instance, NAS 2019. One of us has defended this three-fold distinction in more detail elsewhere (Bouter and Ter Riet, 2021), so we will not repeat the argument here.[2] It suffices to say that many different distinctions can be made with regard to kinds of replication and that it is perhaps best just to be as explicit as possible on how one understands each of these notions.

We should also note that only a successful replication increases the trustworthiness of the findings of the primary study. In order to keep this in mind, let us distinguish between replicability and successful replication:

> *Replicability* means a study's having certain features such that a replication study of it could be carried out.
>
> A *Successful Replication* implies that a replication has successfully been carried out, producing findings that agree sufficiently with the primary study.

What is the object of trust when one says that replication enlarges a study's trustworthiness? We suggest that it is *the main findings of the primary study* rather than the primary study as a whole or its research protocol.[3] By "findings" we mean the combination of the data or results obtained and the conclusions that are drawn from them. A direct replication can help to rule out chance findings (due to random error)[4] while conceptual

replications can additionally detect biases (due to systematic error). While we focus on replication of findings it can be argued that at least in some disciplines the issue might be more one of a "theory crisis" than a "replication crisis" (Oberauer and Lewandowsky 2019). If that is indeed the case priority should be given to the formulation of strong theories and sharp hypotheses for which perhaps conceptual replication studies could be instrumental. Additionally, strong theories may point the way to findings that are unlikely to occur and thus avoid be led by highly improbable assumptions (Ulrich and Miller 2020).

When are findings *successfully replicated*? There is no straightforward answer to this question (as also pointed out by Nosek and Errington 2017, 2020), for at least four reasons. First, it is often not a dichotomous matter whether a replication is successful or not. Successful replication does not require that the replication study obtains *the exact same* findings. Rather, it is a matter of obtaining a finding *that points in the same direction and has an effect size and precision that are similar enough* (Bouter and Ter Riet 2021). Because the rules for what is "similar enough" can vary, they need to be specified when the conclusion is drawn that the replication was successful. As Brian Nosek and Timothy Errington rightly point out, asking such questions as the following will be helpful in determining whether findings are similar enough: "Does the replication produce a statistically significant effect in the same direction as the original? Is the effect size in the replication similar to the effect size in the original? Does the original effect size fall within the confidence or prediction interval of the replication (and vice versa)? Does a meta-analytic combination of results from the original experiment and the replication yield a statistically significant effect? And do the results of the original experiment and the replication appear to be consistent?" (Nosek and Errington 2017, 3) Second, most studies report multiple findings that may differ in similarity in a replication. Third, one will have to choose what to take into consideration in comparing various findings. An extensive evaluation of 100 experimental and correlational replication studies in three psychology journals by the Open Science Foundation, for instance, considered significance and P values, effect sizes, subjective assessments of replication teams, and meta-analyses of effect-sizes (Open Science Foundation 2015). Fourth, in qualitative studies, things are even more complicated. In interviews on post-traumatic stress syndrome, for instance, one will have to interpret subjects' words and actions, confer meaning to them, and compare them while there are differences between interviewees in use of words, cultural background factors, and so on. Of course, replication works differently for qualitative studies from replication in quantitative studies (Aguinis and Solarino 2019), yet our notions of replicability, replication study, and

successful replication, as well as our focus on the replicability of findings, seem to apply here as well.

All this is not to deny that successful replications that increase trustworthiness are possible. It *is* to stress that the notion is ambiguous and needs a lot of qualification.

Let us also be explicit that replication studies that were not successful can be useful in a wide variety of ways. They often give rise to new ideas, they can suggest alternations in methods, they can call for new data, and sometimes even point in the direction of an entirely new theory. That only a successful replication is valuable from the standpoint of the trustworthiness of the primary study findings, then, does not mean that if a replication is not successful it does not have any value.

So far, we have spoken of a "primary study" and a "replication study", as is common in the literature. The whole notion of a "primary study" is somewhat problematic though. As we explain below, one can replicate a study without being aware of the primary study; in what sense is the other study primary in such a case? If we nonetheless insist on speaking of a "primary study", which moment determines whether a study is primary in comparison with another: when the idea was born, submission of the research proposal or its being accepted, preregistration of the research protocol, approval by an ethics committee, the start of the data collection phase, posting of a preprint, the submission for publication, the acceptance for publication, or the actual publication? Yet, usually a replication study is clearly an attempt to replicate a known earlier study. In that case, it makes sense to speak of a primary study and a replication study.

Finally, we should note that mere replicability does not increase the trustworthiness of a study's findings. Although replicability makes it easier or even plain possible in the first place to replicate a study, it does not tell us anything about whether such replications will actually be performed or whether they will be successful. The availability of a detailed study protocol and data-analysis plan, however, *does* enable a replication and thereby opens up the possibility to increase its trustworthiness. That being said, even if replication studies are not carried out, the requirement of replicability may actually provide a strong incentive to perform the study well. The demand to be replicable comes with a risk that flaws and errors will be identified by others. That awareness in itself may make the results of a study more trustworthy due to a more careful execution.

3. Trustworthiness and kinds of replication

Above, we noted that there are three kinds of replication: a reproduction, a direct replication, and a conceptual replication. There is a difference between these kinds of replications when it comes to trustworthiness.

A *reproduction* increases the trustworthiness of the findings and it reduces the likelihood that errors have been made in the initial data analysis, because the same data are analyzed again using the same data analysis plan. Additionally, the primary data can be analyzed with an adapted data-analysis plan when the primary analysis is suspected of being inadequate or to check whether alternative approaches will lead to similar findings. If this turns out to be the case this indicates that the findings are robust and thus increases trustworthiness. Of course, there are important differences between *robustness* (different analyses lead to the same findings), evidence for *accuracy* (successful direct replication), and evidence for *validity* (successful conceptual replication), but they all increase the trustworthiness of the primary findings.

A successful direct replication often renders the primary findings more trustworthy than a reproduction because it concerns the collection of *new data*. The larger the evidential base for a particular finding, the more trustworthy it is.

In other words, enlarging the evidential base by gathering new data adds *more* to the trustworthiness of the primary findings than merely re-analyzing the primary data. This is because it demonstrates that sampling variability has no strong influence on the findings and thus lowers the likelihood of chance findings. Of course, all this is true only on certain constraints. Data collected in a different time period, e.g., after rather than before an election, can sometimes not properly be compared with the original data and then do not bear on the trustworthiness of the primary findings.

Finally, a successful conceptual replication renders the primary findings even more trustworthy, because it employs a somewhat different research protocol and demonstrates that the methodological choices that were varied do not influence the findings.[5] The point is: conceptual replications usually add most to the trustworthiness of the primary findings, because they not only help to rule out chance findings but also can strengthen trust in the validity of the findings.

That being said, we should note that even here things are complicated: a successful conceptual replication renders the primary findings more trustworthy than a direct replication but *only when it is clear that it can truly count as a replication rather than a new study on a related but different research question*. An example of this is the famous Marshmallow experiment (Shoda, Mischel, and Peake 1990). Many have taken it to show that strong bivariate correlations exist between a child's ability to delay gratification just before entering school on the one hand and adolescent achievement and socioemotional behavior on the other hand. An attempt at conceptual replication by Watts, Duncan, and Quan (2018), though, yielded a bivariate correlation of only half the size of that reported in the original study. In fact, it was reduced by two thirds after adjustment for factors like family

background, early cognitive ability, and home environment. Others, however, have argued that given the differences in study design and data-analysis, the original findings do actually hold up rather well (e.g., Collins 2018). Our point about conceptual replication and trustworthiness, then, only applies in those cases in which it is sufficiently clear that *it actually aims to answer the same research question*.

Of course, conceptual replications require usually more work, time, and financial resources – we say "usually" because in some cases a new method may be more efficient and cheaper. When they fail it is difficult to decide what the reason for it was because both a different data set and a somewhat different research protocol were used. It is not uncommon, then, to first carry out a direct replication and only then engage in a conceptual replication. Our view is that it is usually best to start with a reproduction. Only when the findings of a primary study have been shown to be robust by checking for numerical inconsistencies and a re-analysis of the data with the same or an alternative data-analysis plan leads to the same findings, it makes sense to devote time and energy to a direct replication (Nuijten 2021). If that is successful as well, a conceptual replication might be considered.

4. Trustworthiness and unintentional replication

Replications are usually intentional. Replication *studies*, after all, are purposely set-up as systematic replications of a primary study. It is important to note, though, that replications need not be intentional. It is possible, for instance, that at the literature review stage, a "primary study" is overlooked, for instance, because of language barriers or publication in a non-indexed journal. It is also possible that the publication or preprint of the findings of the primary study has not yet appeared at the time the study that is an unintended replication is designed. The notion of "multiple independent discoveries" has received quite a bit of attention in the sociology and philosophy of science, particularly in the work of Robert K. Merton (e.g., Merton 1963). Many of such discoveries will also qualify as unintentional replications.

Such unintentional replications tend to be *conceptual*, that is, a replication that uses a somewhat different research protocol to answer the same study question with new data. One could argue that when the differences in research protocols are too large it is not a replication but just another primary study on the same study question. A better term would then be *triangulation* which indicates that the same study question is approached with different methods. It is, however, also possible to unintentionally use the *same* research protocol with new data (a so-called *direct* replication), for instance, when it comes to a relatively straightforward research question,

such as comparing the mortality for a new pharmaceutical drug with that for a placebo.

What is the relation between (un)intentionality in replication and trustworthiness? An *unintentional* replication is not less valuable than an *intentional* replication: if successful, it contributes as much to the trustworthiness of the primary findings as an intentional replication. Ceteris paribus – its being set up rigorously, its taking potential biases into account (for conceptual replications), and so on – an unintentional successful publication is as worthwhile, as it adduces further evidence for the same research question.[6]

5. Trustworthiness and stakes

So far, we have gone along with a widely but tacitly accepted understanding of trustworthiness, namely that findings of a primary study are trustworthy when they are sufficiently warranted by the evidence. It seems to us, though, that things are somewhat more complicated than that: whether findings are trustworthy is not merely a matter of evidential or epistemic considerations, but also of practical or pragmatic considerations. In other words: whether findings are trustworthy depends on *what is at stake*. Let us explain.

An important position in the theory of knowledge nowadays is that whether or not one knows that a particular proposition p is true depends not only on the quality of one's evidence or reasons, but also on what is at stake. This view is called "pragmatic encroachment" (Stanley 2005). Sometimes, the costs of making an error are higher than at other times. We need lower error rates if the costs are higher. Whether or not pragmatic encroachment is correct for knowledge, it certainly seems true for trust: if there is only a 1% chance that the gun in my hand is loaded and the purpose is to shoot in the air (e.g., for celebrating a victory), I can trust that it is empty. If there is only a 1% chance that the gun is loaded, but I am putting the gun to my head, I *cannot* rationally trust that it is empty.

The pragmatic dimension of trust also seems to apply to trustworthiness of research findings: whether primary findings are trustworthy of course partially depends on the quality and robustness of the evidence but also on what is at stake. Thus, a study about how rehearsal studies of Shakespeare can inform and enrich our understanding of historical performance genres needs little replication in order to be trustworthy, but a study result about the intended and unintended effects of a new Covid-19 vaccine that is meant to be distributed to many millions of people needs to be replicated more thoroughly, given the enormous stakes involved.

Therefore, trustworthiness is not only a matter of the evidence adduced in the primary study and its replications, but also a matter of what is at stake.

6. How to assess trustworthiness

Increasingly, replication is perceived and accepted as part and parcel of academic research (thus also Lakens 2020; Zwaan, Etz, and Donnellan 2018). That is a development we wholeheartedly embrace. Yet, little attention has been paid to the relation between replication and trustworthiness. We suggest, in conclusion, that the assessment of trustworthiness of study findings should be extended to include not only widely accepted points like:

(1) What was the likelihood of the findings before the primary study was done? This is often referred to as the prior probability of the findings.
(2) What was the study size and the methodological quality of the primary study? In other words: was the primary study valid and precise enough?
(3) How many reproductions, direct replications and conceptual replications were performed? And what was their study size and methodological quality? Which proportion of them was successful?
(4) What are the stakes at issue? In other words: is another replication indicated before we can act upon the aggregated findings?[7]

Of course, much more work is needed to operationalize or even more precisely define some of these concepts. For instance, how should we understand the "methodological quality" of a study and how can that be operationalized for the different disciplinary fields and study types? The same holds for "prior probability", "stakes", and so on. Systematic reviews, meta-analyses, Bayesian inferences and decision theory can help to answer these four questions. A particularly interesting approach concerns calculation of the level of "acceptable regret", which is the likelihood of making a wrong decision on the research hypothesis we deem acceptable (Djulbegovic and Hozo 2007). This, however, would go far beyond the point of the present paper. Our point is merely that these things need to be taken into account – and then, of course, further analyzed and operationalized – in assessing the trustworthiness of the findings of a study.

Notes

1. This is sometimes called "reproducibility", that is, computational reproducibility of analyses following identical scrips, data artifacts, hardware, and so on.
2. Recently, some have critiqued the idea that conceptual replications are truly replications. See, for instance, Machery (2019). For the sake of argument, we will here assume the mainstream position which says that there are three kinds of replications: reproductions, direct replications, and conceptual replications. However, we admit that when the differences between the research protocol of the primary study and its replication become too large it makes little sense to label it as a replication.

3. Among other things, when it comes to the trustworthiness of a study protocol or even an entire study, various elements that these are composed of may come in various degrees of trustworthiness. For instance, one of two methods used may be more trustworthy than the other, one set of data used may be more trustworthy than the other ones, or a laboratory instrument may be more trustworthy than another one.
4. When the objective of a study is estimation (e.g., of the effect size) as opposed to hypothesis testing, direct and conceptual replication will increase the precision of the estimates calculated from the aggregated data of the primary study and its replications. Consequently, all replications will increase trust in the accuracy of the study findings when estimation is at issue.
5. It follows from how we have defined the three kinds of replication that replication is not identical to validity. Validity concerns whether a study really measures what it is supposed to measure in the sense of being acceptably free of bias (lack of systematic errors). Only conceptual replications provide information on validity but reproductions and direct replications only provide an insight in the accuracy (lack of random error) of the findings. If a study design is biased a reproduction or a direct replication will only lead to a more precise estimate of the wrong answer.
6. Studies can also indirectly replicate earlier findings even though that was not the purpose of that study. For instance, by looking at the relation between education and income in a series of studies that happened to collect data of both these variables. In order to make things not unnecessarily complicated, we have left such indirect replications aside here.
7. In assessing the trustworthiness of findings, replication, then, plays a role among many other factors.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by the Templeton World Charity Foundation [TWCF0436].

References

Aguinis, H., and A. M. Solarino. 2019. "Transparency and Replicability in Qualitative Research: The Case of Interviews with Elite Informants." *Strategic Management Journal* 40 (8): 1291–1315.

Brigitte Vachon, Janet A. Curran, Sathya Karunananthan, Jamie Brehaut, Ian D. Graham, David Moher, Anne E. Sales, Sharon E. Straus, Michele Fiander, P. Alison Paprica, Jeremy M. Grimshaw, 2020. "A Concept Analysis and Meta-narrative Review Established A Comprehensive Theoretical Definition of Replication Research to Improve Its Use." Journal of Clinical Epidemiology, 129: 176–187. doi:10.1016/j.jclinepi.2020.07.006

Collins, J. 2018. "The Marshmallow Test Held Up OK." https://jasoncollins.blog/the-marshmallow-test-held-up-ok/

Djulbegovic, B., and I. Hozo. 2007. "When Should Potentially False Research Findings Be Considered Acceptable?" *PLoS Medicine* 4 (2): e26. doi:10.1371/journal.pmed.0040026.

Faulkner, P., and T. Simpson, eds. 2017. *The Philosophy of Trust*. Oxford: Oxford University Press.

Hendriks, F., D. Kienhues, and R. Bromme. 2020. "Replication Crisis = Trust Crisis? The Effect of Successful Vs Failed Replications on Laypeople's Trust in Researchers and Research." *Public Understanding of Science* 29 (3): 270–288. doi:10.1177/0963662520902383.

KNAW: Royal Dutch Academy of Arts and Sciences. 2018. "Replication Studies: Improving Reproducibility in the Empirical Sciences (Amsterdam)." Accessed 19 November 2020. https://knaw.nl/en/news/publications/replication-studies

Lakens, D. 2020. "The 20% Statistician: A Blog on Statistics, Methods, Philosophy of Science, and Open Science." Accessed 29 November 2020. http://daniellakens.blogspot.com/2020/11/why-i-care-about-replication-studies.html

Machery, E. 2019. "What Is a Replication?" *Philosophy of Science* 87 (4): 545–567. doi:10.1086/709701.

McLeod, C. 2015. "Trust." In *The Stanford Encyclopedia of Philosophy*, edited by N. Z. Edward. Accessed 29 November 2020. https://plato.stanford.edu/entries/trust/

Merton, R. K. 1963. "Resistance to the Systematic Study of Multiple Discoveries in Science." *European Journal of Sociology* 4 (2): 237–282. doi:10.1017/S0003975600000801.

National Academies of Sciences, Engineering, & Medicine. 2019. *Reproducibility and Replicability in Science*. Washington, DC: National Academies Press.

Nosek, B. A., and T. M. Errington. 2017. "Making Sense of Replications." *eLife* 6 (e23383): 19. doi:10.7554/eLife.23383.

Nosek, B. A., and T. M. Errington. 2020. "What Is Replication?" *PLOS Biology* 18 (3): e3000691. doi:10.1371/journal.pbio.3000691.

Nuijten, M. B. 2021. "Assessing and Improving Robustness of Psychological Research Findings in Four Steps." *PsyArXiv*, 8 April. doi:10.31234/osf.io/a4bu2.

O'Neill, O. 2002. *A Question of Trust: The BBC Reith Lectures*. Cambridge: Cambridge University Press.

Oberauer, K., and S. Lewandowsky. 2019. "Addressing the Theory Crisis in Psychology." *Psychonomic Bulletin & Review* 26 (5): 1596–1618. doi:10.3758/s13423-019-01645-2.

Open Science Foundation. 2015. "Estimating the Reproducibility of Psychological Science." *Science* 349 (6251): aac4716. doi:10.1126/science.aac4716.

Shoda, Y., W. Mischel, and P. K. Peake. 1990. "Predicting Adolescent Cognitive and Self-regulatory Competencies from Preschool Delay of Gratification: Identifying Diagnostic Conditions." *Developmental Psychology* 26 (6): 978–986. doi:10.1037/0012-1649.26.6.978.

Stanley, J. 2005. *Knowledge and Practical Interests*. New York: Oxford University Press.

Ulrich, R., and J. Miller. 2020. "Questionable Research Practices May Have Little Effect on Replicability." *eLife* 9: e58237. doi:10.7554/eLife.58237.

Watts, T. W., G. J. Duncan, and H. Quan. 2018. "Revisiting the Marshmallow Test: A Conceptual Replication Investigating Links between Early Delay of Gratification and Later Outcomes." *Psychological Science* 29 (7): 1159–1177. doi:10.1177/0956797618761661.

Zwaan, R. A., R. L. Etz, and M. Donnellan. 2018. "Making Replication Mainstream." *Behavioral and Brain Sciences* 41 (41): e120. doi:10.1017/S0140525X17001972.