



UvA-DARE (Digital Academic Repository)

Analysis Blinding as a Potential Means to Foster a Productive Collaboration Between Original Authors and Replicators

Sarafoglou, Alexandra; Hoogeveen, Suzanne

DOI

[10.1525/collabra.136869](https://doi.org/10.1525/collabra.136869)

Publication date

2025

Document Version

Final published version

Published in

Collabra: Psychology

License

CC BY

[Link to publication](#)

Citation for published version (APA):

Sarafoglou, A., & Hoogeveen, S. (2025). Analysis Blinding as a Potential Means to Foster a Productive Collaboration Between Original Authors and Replicators. *Collabra: Psychology*, 11(1), Article 136869. <https://doi.org/10.1525/collabra.136869>

General rights

It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations

If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: <https://uba.uva.nl/en/contact>, or a letter to: Library of the University of Amsterdam, Secretariat, P.O. Box 19185, 1000 GD Amsterdam, The Netherlands. You will be contacted as soon as possible.

UvA-DARE is a service provided by the library of the University of Amsterdam (<https://dare.uva.nl>)

Methodology and Research Practice

Analysis Blinding as a Potential Means to Foster a Productive Collaboration Between Original Authors and Replicators

Alexandra Sarafoglou¹^a, Suzanne Hoogveen²

¹ Department of Psychology, University of Amsterdam, Amsterdam, The Netherlands, ² Department of Methodology and Statistics, Utrecht University, Utrecht, The Netherlands

Keywords: Replication, Analytic Flexibility, Bias Control, Analysis Blinding

<https://doi.org/10.1525/collabra.136869>

Collabra: Psychology

Recent awareness of the importance of rigor and robustness have deemed replication efforts vital for scientific advance. Yet the value of replication projects may often be undermined by post-hoc disputes with the original authors about the replication outcomes, for instance, concerning data quality, unanticipated deviations from the data collection protocol, or diverging implementations of the analysis strategy. In this comment, we reflect on the tension between replicators and original authors and advocate for analysis blinding as a means to prevent such unproductive post-hoc discussions. Analysis blinding involves the alteration of data to remove the key effect of interest while preserving all other aspects. This methodology allows for an assessment of important properties of the data (manipulation checks, outliers, data quality) without introducing bias or risking the perception of attempting to manipulate the results. We discuss three replication studies we were responsible for in the Holzmeister et al. (2025) project and demonstrate how to effectively blind data for each of them. We argue that analysis blinding has the potential to prevent fruitless discussions and tension between original authors and replication teams in replication projects while preserving a healthy scientific debate.

Introduction

Replication studies are crucial to establish the credibility of scientific claims. If the evidence supporting these claims cannot be reestablished in repeated studies, it may indicate that the purportedly discovered effects or relations are mere noise or chance findings, or that they are not as robust as initially assumed. However, replication studies are costly—both financially and in terms of researcher time. Compared to novel research, replication projects often involve an additional layer of effort: communication with original authors to leverage their expert experience, optimize the replication study design, and prevent post-hoc disputes over the interpretation of the replication results. However, while we fully support consultation with original authors, we noticed that in practice this interaction can be time-consuming and ineffective, as post-hoc disputes are still common. In this comment, we share insights and advice for future replication projects regarding the interaction with original authors after replication results are known. In particular, we recommend *analysis blinding* as a methodology that offers a chance to give an unbiased evaluation of the replication data and analysis (Dutilh et al., 2019; MacCoun & Perlmutter, 2015). Our suggestions are based on

and illustrated by our experience as one of the replication teams involved in the Holzmeister et al. (2025) project.

The aim of the Holzmeister et al. (2025) project was to replicate social science studies that were published in the *Proceedings of the National Academy of Sciences (PNAS)* between 2015–2018 and that were conducted on the online data collection platform Amazon Mechanical Turk (MTurk). The replications were high-powered, preregistered, and kept as similar as possible to the original studies. The team leaders' strict emphasis on direct replication ensured that the sampling platform, instructions, materials, and analysis remained identical to those of the original study. This was done to protect the project from concerns that hidden moderators might influence the results of the replications. Deviations from the original protocol were not allowed, with changes implemented only in exceptional cases (e.g., if an error was discovered in the original analysis). The original authors were involved in providing feedback and approving the replication protocols. These discussions took place before data collection for all replication attempts, with original authors explicitly asked to confirm the correspondence of the replication protocol to the original study. When original authors raised concerns, their arguments were considered before the lead team made a final decision. If the

^a Correspondence concerning this article should be addressed to: Alexandra Sarafoglou, Nieuwe Achtergracht 129B, 1001 NK Amsterdam, The Netherlands, E-mail: alexandra.sarafoglou@gmail.com.

original authors did not respond after multiple follow-up emails, the lead team allowed the replication teams to proceed without permission. After the replication was completed and the data were analyzed, the original authors were contacted again to review and approve the final replication report summarizing the results. When requested, the original authors received access to the replication data and the analysis code, and had the possibility to submit a commentary on the project's OSF page (<https://osf.io/sk82q/>) that was published alongside the replication reports.

In the Appendix, we provide a detailed account of our interactions with the original authors for the Holzmeister et al. (2025) replication project, including their specific concerns and our attempts to address those. We would like to emphasize that all interactions we had with the original authors as well as the project lead team were constructive, pleasant, and respectful in tone. Yet for any issues raised after data collection, it was impossible to consider the original authors' arguments in an objective manner, as both parties were already aware of the replication's success or failure. The post-hoc problems identified by the original authors mainly concerned data quality and coding errors, both in their original scripts and in our replication scripts. Problems with data quality manifested in high participant dropouts in the replication attempts and a high percentage of failed comprehension check questions. These patterns raised questions about validity and echoed general quality concerns regarding the MTurk platform (Chmielewski & Kucker, 2020; Douglas et al., 2023; Peer et al., 2022, 2023; Stagnaro et al., 2024). Coding errors were identified both in the original studies (e.g., in the statistical model) and in the replication analyses (e.g., in the preprocessing code). While clear mistakes were rectified in the final analyses, it is not guaranteed that the detection of errors is independent from the outcomes. That is, people tend to evaluate evidence less critically when it aligns with their preferences (Ditto & Lopez, 1992) or beliefs (Lord et al., 1979) than when it contradicts them, a tendency referred to as biased debugging (see e.g., Poldrack, 2013; Stokes, 2013).

Potential Risks When Disputing Replication Results

Our experiences in the Holzmeister et al. (2025) project provide a telling illustration of the tensions between replication teams and original authors (see also Chatard et al., 2020; Simons, 2014; Stroebe & Strack, 2014). On the one hand, the conversations between original authors and replicators attest to a healthy scientific discourse. On the other hand, many issues were raised by the original authors when they reviewed the final replication report – that is, *after* they had seen the replication results. This timing, in combination with the project setup of conducting direct, pre-approved replications, severely limited our flexibility in incorporating original authors' perspectives, however reasonable given the situation at hand. We believe the tension is fed by implicit assumptions from both sides, which are often perceived as having conflicting interests.

In the worst case scenario, the replication team is accused of attempting a replication effort of insufficient qual-

ity, perhaps because they lack the necessary expertise, investment, or even have an interest in discrediting the original findings. These potential risks can be mitigated by preregistering the design and analysis plan and inviting the original authors to provide feedback before data collection – this was the approach used in the PNAS replication project. However, a pre-approved preregistration does not safeguard against disputes arising from unanticipated events (e.g., data collection needs to be repeated, changes in the code are required). Conversely, the original authors may be suspected of post-hoc explanations, biased debugging, or even (unconscious) *p*-hacking in order to preserve their effects. Critically, when replications fail, the original authors have no opportunity to raise concerns and prove their presence independently of the outcomes. These constraints can make the post-analysis discussions incredibly unproductive and hinder the core scientific goal of assessing whether or not an effect exists.

The core problem is not differing opinions on what constitutes a valid replication (e.g., whether changes in the subject pool or the “state of the world” affect comparability), but rather the challenge of addressing data-dependent concerns in an unbiased way. We believe that even in the context of strict direct replications, some flexibility in protocol deviations should be maintained – *if* justified independently of the key statistical outcomes. For instance, if comprehension rates show a significant decline from the original study, a decision how to proceed (e.g., revise inclusion criteria) should be considered without knowledge of the replication results.

To reduce tensions and encourage productive discourse, original authors should have opportunities to raise concerns, audit code, and propose data-based checks in a fair and unbiased manner – before seeing the replication results. At the same time, we believe the original authors should have access to the replication data to identify and address potential peculiarities, such as substantial quality concerns.

Implementing Analysis Blinding in Replication Projects

We believe that analysis blinding (Dutilh et al., 2019; MacCoun & Perlmutter, 2015) can have great merit in replication projects, in addition to preregistration. The idea behind analysis blinding is that the analysis is conducted in two steps and managed by two independent teams: the data manager and the analyst. The data manager receives the real data and alters it in a way that any potential effect of interest is removed by design (e.g., be shuffling values in the outcome variable, or adding noise to the data). While the effect of interest disappears in these blinded data, all other information, such as responses to comprehension questions or manipulation checks, stay intact. The blinded data are given to the analyst who can then proceed to develop the entire analysis pipeline. When the analysis pipeline is finalized, the pipeline is registered, and the analyst receives the real data. The resulting analysis is fair because the analysis pipeline is shielded against potential biases driven by the eventual results. At the same time, the

resulting analysis is also flexible in the sense that the analytic strategy can be adapted to peculiarities even after data collection. This flexibility offers a key benefit as Sarafoglou, Hoogeveen, and Wagenmakers (2023) found: analytic strategies using analysis blinding lead to fewer deviations than those relying on preregistered plans.

In the context of replication projects we propose to use preregistration and analysis blinding in combination. Specifically, a feedback round for the original authors could be implemented after data collection but before revealing the final data. The original authors would then receive the blinded replication data and the finalized analysis pipeline, but not the replication outcome. Based on these, the original authors can express any concerns that can still be addressed in an unbiased manner. This would increase the credibility of the submitted feedback. For instance, quality concerns based on comprehension questions may receive more weight when raised while blind to the results than while the original authors know their effect did not replicate. In the end, this procedure benefits both the replication team and the original authors. The replication team can collaborate and accept reasonable changes from the preregistered protocol suggested by the original authors. The original authors are given a fair chance to criticize the study and give an unbiased evaluation of the data and analysis.

Analysis blinding can be easily implemented (see e.g., Dutilh et al., 2019, for guidance on how to effectively blind data in common experimental designs in the social sciences). In replication projects, the replication team could serve as data managers and be responsible for the analysis blinding. In the PNAS replication project, we took the following general analysis blinding strategy across all studies: we shuffled the dependent variable while leaving all other variables unaffected; for within-subjects designs, we shuffled within subjects, for between-subjects designs, we shuffled across subjects. If the dependent variable consisted of multiple items, these were kept together per subject, allowing, for instance, for the assessment of reliability. We opted to alter the outcome instead of the experimental condition assignment, as most manipulation checks and comprehension questions are based on the independent rather than the dependent variable (e.g., did the manipulation of stereotypicality of targets indeed result in difference in perceived stereotypicality). Finally, shuffling the data, rather than adding noise or replacing the actual data with random values, has the benefit that the overall distribution of the data remains intact and can hence be evaluated (e.g., baseline rates of sharing in economic games). Details on the exact blinding procedure and its results for each individual replication study can be found in the Appendix. In summary, applying a simple shuffling strategy, taking into account the peculiarities of each individual design, resulted in effectively removing the focal effect of interest while allowing for meaningful manipulation and data quality checks.

Concluding Remarks

Replication projects form a crucial empirical tool to evaluate scientific reliability and robustness. Yet even in direct replication projects, the exact implementation and analysis are often not clear-cut. Post-hoc disputes between original authors and replicators are common, either as part of the project itself as in Holzmeister et al. (2025), or publicly after the replication results are shared (see for instance the academic (preprint) interaction about the ManyLabs 4 replication of the mortality salience effect; (Chatard et al., 2020; Hoogeveen et al., 2023; Klein et al., 2022)). We believe analysis blinding presents a powerful tool to prevent fruitless disputes and tension between involved teams. Sharing the blinded data from the replication study as an intermediate step between data collection and the release of final results allows original authors to provide feedback on unanticipated peculiarities in the data or issues with the analysis code, while remaining shielded from bias arising from knowledge of the replication outcomes.

Analysis blinding is in most cases straightforward to implement; as illustrated on the case studies here, shuffling the dependent variable is often sufficient to remove the key effect of interest yet retain information needed for quality control and manipulation checks. As the exact blinding strategy requires some knowledge on the structure of the data and design (e.g., within- or between-subjects), we would recommend the data managers to be either the experts within the replication team themselves or be provided with detailed instructions or a blinding script based on dummy data.

Analysis blinding is not a panacea – it may not, for instance, resolve fundamental differences in perspective between replicators and original authors on what qualifies as a good (direct) replication. In the Holzmeister et al. (2025) project, the lead team considered the data collection platform (i.e., MTurk) an essential part of the replication. However, some original authors argued that changes in the subject pool on MTurk rendered the newly collected data incomparable to the original data. Such divergent opinions should ideally be resolved beforehand, or else accepted as inherent disagreements. Nevertheless, analysis blinding could still add an opportunity to assess the data quality and subsequently make adjustments that both replicators and original authors agree upon.

We believe any additional time and effort related to blinding the data is more than compensated for by the time and effort it may save on unproductive discussions between the original authors and the replicators. It gives the original authors a fair chance to voice concerns without the risk of being perceived as trying to manipulate the results in their favor. Above all, in our view, analysis blinding fosters a more collaborative environment and hence may lead to a more satisfying experience for all parties involved in replication projects.

.....

Conflict of Interest

The authors declare that there were no conflicts of interest with respect to the authorship or the publication of this article.

Ethical approval

For this perspective piece, no ethical approval was needed as it did not involve testing of human participants. In the Holzmeister et al. (2025) replication project, all par-

ticipants were treated in accordance with the Declaration of Helsinki.

Funding information

A.S. was supported by a 2024 Ammodo Science Award.

Submitted: December 11, 2024 PDT. Accepted: March 28, 2025 PDT.



This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CCBY-4.0). View this license's legal deed at <http://creativecommons.org/licenses/by/4.0> and legal code at <http://creativecommons.org/licenses/by/4.0/legalcode> for more information.

References

- Chatard, A., Hirschberger, G., & Pyszczynski, T. (2020). *A word of caution about Many Labs 4: If you fail to follow your preregistered plan, you may fail to find a real effect* [Preprint]. PsyArXiv. <https://doi.org/10.31234/osf.io/ejubn>
- Chmielewski, M., & Kucker, S. C. (2020). An MTurk crisis? Shifts in data quality and the impact on study results. *Social Psychological and Personality Science*, *11*, 464–473. <https://doi.org/10.1177/1948550619875149>
- Ditto, P. H., & Lopez, D. F. (1992). Motivated skepticism: Use of differential decision criteria for preferred and nonpreferred conclusions. *Journal of Personality and Social Psychology*, *63*, 568–584. <https://doi.org/10.1037/0022-3514.63.4.568>
- Douglas, B. D., Ewell, P. J., & Brauer, M. (2023). Data quality in online human-subjects research: Comparisons between MTurk, prolific, CloudResearch, qualtrics, and SONA. *Plos One*, *18*, e0279720. <https://doi.org/10.1371/journal.pone.0279720>
- Dutilh, G., Sarafoglou, A., & Wagenmakers, E.-J. (2019). Flexible yet fair: Blinding analyses in experimental psychology. *Synthese*, *198*, S5745–S5772. <https://doi.org/10.1007/s11229-019-02456-7>
- Holzmeister, F., Johannesson, M., Camerer, C. F., Chen, Y., Ho, T.-H., Hoogeveen, S., Huber, J., Imai, T., Jin, L., Kirchner, M., Ly, A., Mandl, B., Manfredi, D., Nave, G., Nosek, B. A., Pfeiffer, T., Sarafoglou, A., Schwaiger, R., Wagenmakers, E. J., ... Dreber, A. (2025). Examining the replicability of online experiments selected by a decision market. *Nature Human Behaviour*, *9*(2), 316–300. <https://doi.org/10.1038/s41562-024-02062-9>
- Hoogeveen, S., Berkhout, S. W., Gronau, Q. F., Wagenmakers, E.-J., & Haaf, J. M. (2023). Improving statistical analysis in team science: The case of a Bayesian multiverse of Many Labs 4. *Advances in Methods and Practices in Psychological Science*, *6*(3), 25152459231182318. <https://doi.org/10.1177/25152459231182318>
- John, L. K., Barasz, K., & Norton, M. I. (2016). Hiding personal information reveals the worst. *Proceedings of the National Academy of Sciences*, *113*(4), 954–959. <https://doi.org/10.1073/pnas.1516868113>
- Jordan, J. J., Hoffman, M., Nowak, M. A., & Rand, D. G. (2016). Uncalculating cooperation is used to signal trustworthiness. *Proceedings of the National Academy of Sciences*, *113*(31), 8658–8663. <https://doi.org/10.1073/pnas.1601280113>
- Klein, R. A., Cook, C. L., Ebersole, C. R., Vitiello, C., Nosek, B. A., Hilgard, J., Ahn, P. H., Brady, A. J., Chartier, C. R., Christopherson, C. D., Clay, S., Collisson, B., Crawford, J. T., Cromar, R., Gardiner, G., Gosnell, C. L., Grahe, J., Hall, C., Howard, I., ... Ratliff, K. A. (2022). Many labs 4: Failure to replicate mortality salience effect with and without original author involvement. *Collabra: Psychology*, *8*(1), 35271. <https://doi.org/10.1525/collabra.35271>
- Lord, C. G., Ross, L., & Lepper, M. R. (1979). Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of Personality and Social Psychology*, *37*, 2098–2109. <https://doi.org/10.1037/0022-3514.37.11.2098>
- MacCoun, R., & Perlmutter, S. (2015). Blind analysis: Hide results to seek the truth. *Nature*, *526*, 187–190. <https://doi.org/10.1038/526187a>
- Morris, A., MacGlashan, J., Littman, M. L., & Cushman, F. (2017). Evolution of flexibility and rigidity in retaliatory punishment. *Proceedings of the National Academy of Sciences*, *114*(39), 10396–10401. <https://doi.org/10.1073/pnas.1704032114>
- Peer, E., Rothschild, D., & Gordon, A. (2023). Platform over procedure: Online platforms that pre-vet participants yield higher data quality without sacrificing diversity. *PsyArXiv*, 1–10. <https://doi.org/10.31234/osf.io/buzwn>
- Peer, E., Rothschild, D., Gordon, A., Evernden, Z., & Damer, E. (2022). Data quality of platforms and panels for online behavioral research. *Behavior Research Methods*, 1643–1662. <https://doi.org/10.3758/s13428-021-01694-3>
- Poldrack, R. (2013). *Anatomy of a coding error* [Scientific Blog]. <https://poldrack.github.io/blog/posts/anatomy-of-coding-error/index.html>
- Reeck, C., Wall, D., & Johnson, E. J. (2017). Search predicts and changes patience in intertemporal choice. *Proceedings of the National Academy of Sciences*, *114*, 11890–11895. <https://doi.org/10.1073/pnas.1707040114>
- Sarafoglou, A., Hoogeveen, S., & Wagenmakers, E.-J. (2023). Comparing analysis blinding with preregistration in the many-analysts religion project. *Advances in Methods and Practices in Psychological Science*, *6*(1), 25152459221128319.
- Simons, D. J. (2014). The value of direct replication. *Perspectives on Psychological Science*, *1*, 76–80. <https://doi.org/10.1177/1745691613514755>
- Stagnaro, M. N., Druckman, J., Berinsky, A., Arechar, A. A., Willer, R., & Rand, D. (2024). *Representativeness versus attentiveness: A comparison across nine online survey samples*. PsyArXiv. <https://doi.org/10.31234/osf.io/h9j2d>
- Stern, C., West, T. V., & Rule, N. O. (2015). Conservatives negatively evaluate counterstereotypical people to maintain a sense of certainty. *Proceedings of the National Academy of Sciences*, *112*(50), 15337–15342. <https://doi.org/10.1073/pnas.1517662112>
- Stokes, M. (2013). *Biased debugging* [Scientific Blog]. <http://the-brain-box.blogspot.com/2013/02/biased-debugging.html>
- Stroebe, W., & Strack, F. (2014). The alleged crisis and the illusion of exact replication. *Perspectives on Psychological Science*, *9*(1), 59–71.

Williams, K. E. G., Sng, O., & Neuberg, S. L. (2016).
Ecology-driven stereotypes override race stereotypes.
Proceedings of the National Academy of Sciences,
113(2), 310–315. [https://doi.org/10.1073/
pnas.1519401113](https://doi.org/10.1073/pnas.1519401113)

Appendix. Illustration: the Holzmeister et al. (2025) replication studies

In the winter of 2018, the lead team of the replication project (Holzmeister et al., 2025) recruited five research teams to prepare (and potentially execute) 41 replication protocols for online social science studies conducted on MTurk and published in the *Proceedings of the National Academy of Sciences (PNAS)*. The study designs were high-powered and kept as similar as possible to the original ones. Materials and code from the original studies were used whenever available and changes in the protocol were implemented only when absolutely necessary (e.g., in cases of inaccessible original materials). All replication protocols were sent to the original authors and preregistered upon approval. The 41 target replications then entered the online decision markets, in which social scientists ($n = 162$) traded on the replication outcomes. The top 12 studies with the highest and the bottom 12 studies with the lowest final market prices, plus two randomly chosen studies, were selected for the actual replications. After the data for these 26 studies were collected, the replication teams conducted the preregistered analyses that were cross-validated by the lead team. The primary indicator of replication success was a statistically significant effect in the hypothesized direction with $\alpha = .05$. The final replication reports were then sent to the original authors for approval.

Our team was responsible for six of the 26 conducted replications; initially, one of which was qualified as a successful replication (i.e., showed a significant effect in the hypothesized direction) and five of which were qualified as failed replications. For three of the five initially unsuccessful replications, that is, the studies by Morris et al. (2017), Jordan et al. (2016), and Reeck, Wall, and Johnson (2017), we had intensive exchange with the original authors after we sent them the replication reports. In the following we will briefly describe these three case studies in which problems were raised by the original authors after the data were analyzed and the non-significant results were communicated.

Disclosure: Data availability and code

Readers can access the R code to blind the data and conduct all analyses (including all figures), in our OSF folder at: <https://osf.io/h2feq/>. The replication data and the replication reports can be found on the project page by Holzmeister et al. (2025) at: <https://osf.io/sk82q/>.

Discussions with original authors

The original authors from the Morris et al. (2017) study expressed concerns regarding the quality of the data. These concerns had already been raised prior to data collection,

when the authors were invited to review the preregistration, and were raised again when reviewing the final replication report. The general policy for safeguarding the quality of the data in the replications involved applying the same inclusion/exclusion criteria as the original study (e.g., based on attention or comprehension checks) plus an additional IP-address check to filter out potential bots completing the surveys. According to the lead team, this additional IP-check should ensure sufficient data quality, potentially higher than in the original studies that may have included invalid bot responses.

However, the authors also encountered other issues while reviewing the report. Upon scrutinizing the analysis script, the original authors discovered a coding error in their initial analysis. Although this error did not alter the conclusions of the original study, it led to a change in conclusions for the replication attempt. That is, the initially non-significant effect ($p = .305$) now became significant ($p = .019$); thus, the effect replicated. The error was subsequently confirmed by the lead team of the replication project and resulted in adjustments to the analysis code in the replication attempt.¹ Furthermore, the expressed concerns about the data quality of the replication study prompted the original authors to rerun another replication on an alternative platform with stricter quality control. Here again, the effect replicated.

Due to a misunderstanding, data for the Jordan et al. (2016) replication study was collected twice. We disregarded the first data batch without looking at the results and only analyzed data from the second batch. After reading the replication report, Jordan et al. requested to see both data batches. They pointed out that the data quality in the second and final batch was much lower than in the first batch and the original study, as indicated by the comprehension rates (final batch: 11% answered 7/7 correctly, $M = 3.45$; first batch: 17% answered 7/7 correctly, $M = 4.25$; original study: 29%, $M = 5.44$). Importantly, the comprehension questions were *not* used as an exclusion criterion in the original study, a strategy we copied for the replication attempt. The original authors, however, argued that only reporting the second batch was misleading, especially given the poor quality and the fact that the first batch was perfectly valid from a scientific perspective, which they had already indicated prior to knowing the results. They also showed that the focal replication effect was significant in the first batch including all subjects ($p = .026$), in the first batch including only full comprehenders ($p = .010$), in the second (final) batch including only full comprehenders ($p = .024$), in a meta-analysis including both batches with all subjects ($p = .016$) and in a meta-analysis with only full comprehenders ($p = .001$), though not in their original study including only full comprehenders ($p = .217$). We note that as reported in Holzmeister et al. (2025), the effect also replicated when applying different indicators for replication success besides the primary criterion of statistical

¹ Importantly, the original authors also submitted a corrigendum to their original work to rectify the error.

significance (i.e., the small telescope approach and meta-analytic effect sizes, but not Bayes factors and prediction intervals).²

Finally, perhaps the most severe case was the replication attempt of the study by Reeck, Wall, and Johnson (2017). In this study, we (i.e., the replication team) preregistered that all analyses were based on the code shared by the original authors. However, we later realized that only the code for the main analysis had been shared by the original authors, while the code for preprocessing the data for the main analysis and conducting the manipulation check had not been shared. What made this case particularly intricate was that the data structure was highly complex and necessitated sophisticated preprocessing. After collecting the data, we implemented the preprocessing to the best of our knowledge. However, upon sharing it with the original authors, some severe coding mistakes were revealed. As a result, the original authors wrote an independent script to preprocess the data, and we sought consultation from an independent analyst (who was part of the lead team for the replication project) to help us rectify the errors and to cross-validate the code we had written. In the end, the coding errors could be resolved but major disagreements persisted between the preprocessing protocol from the replication team and the one from the original authors (exacerbated by the ambiguity in the original paper regarding the exact preprocessing steps).

Points of disagreement included (1) whether or not catch trials should be included in the final analysis, (2) whether participants should be excluded who failed one or both catch trials, and (3) whether trial exclusions relevant for the manipulation check should also be implemented in the main analysis. The original authors' preferred preprocessing approach led to an increase in the exclusion of observations, resulting in a sample size of $n = 857$, which not only raised concerns about the validity of the results due to falling below the preregistered target sample size but also about the data quality, given that this constituted nearly half of all participants (49.6%). Our preferred preprocessing approach led to a sample size of 1044, which matched the target sample size, with a drop out rate of (38.6%). Notably, the replication result was non-significant with both approaches.

Study descriptions and analysis blinding

In the following, we describe how analysis blinding could have been implemented in these particular studies. Specifically, in this counterfactual world, we could have given the authors from the Reeck et al. study the blinded data and ask them to audit the preprocessing code while ensuring that ambiguities are not (consciously or subconsciously) used as loopholes to exploit researchers degrees of freedom. When Jordan et al. had received the blinded data from both data batches, they could have inspected the data quality and

based on quality alone make suggestions on whether the disregarded batch 1 data should be used in the analysis and/or whether the comprehension questions should be applied as inclusion criteria. Morris et al. could have audited the code and discovered the erroneous analysis without any suspicion of p -hacking.

Morris et al.: reacting to rigid thieves and punishers

In the within-subjects experiment by Morris et al. (2017) participants played repeated steal/punish games. In each round, participants were randomly assigned to play either as "thieves" or as "victims". The opponent adhered rigidly to their stealing or punishing behavior, consistently following a strategy of always stealing or always punishing theft. Participants had full knowledge of their opponent's previous actions and could choose their own actions accordingly. The replication study investigated the hypothesis that in repeated games against rigid opponents, participants acting as thieves demonstrate relatively flexible stealing behavior, whereas those in the role of victims exhibit relatively rigid punishment behavior.

The main outcome variable was the participants' choice of either stealing or punishing in their roles as thieves or victims. The critical aspects that required blinding were the relationship between participants' choices and their assigned roles, as well as between participants' choices and the trial number. To blind the data, we shuffled the main outcome variable within participants, thereby removing both the main effects of condition and trial number, as well as the interaction effect between them. This way, the analysis code (and thus the coding error), could still be reviewed.

Figure A1 illustrates the original and blinded response patterns in the Morris et al. (2017) replication data. The interaction between participant role and trial number was tested using a likelihood ratio test against a model without the interaction term and showed a significant result ($\chi^2(1) = 5.495, p = 0.019$). Conducting the analysis on the blinded data yielded a non-significant result ($\chi^2(1) = 0.554, p = 0.457$).

Jordan et al.: uncalculated cooperation to signal trustworthiness

In the between-subject experiment by Jordan et al. (2016), participants played a two-stage economic game: a Helping Game which they play by themselves and a Trust Game which they play with another participant. In the Helping Game, the participant received a certain monetary amount and had to decide whether they wanted to pay a cost to help out another participant. The decision could be made either after looking at the precise cost for helping or without looking at the cost. Participants in the Helping Game are randomized to one of two groups. In one group, participants are told that their looking decisions and their

² Similar to the original authors from Morris et al. (2017), the authors additionally conducted their own replication on a different platform, again finding a significant overall effect (though not when including only full comprehenders).

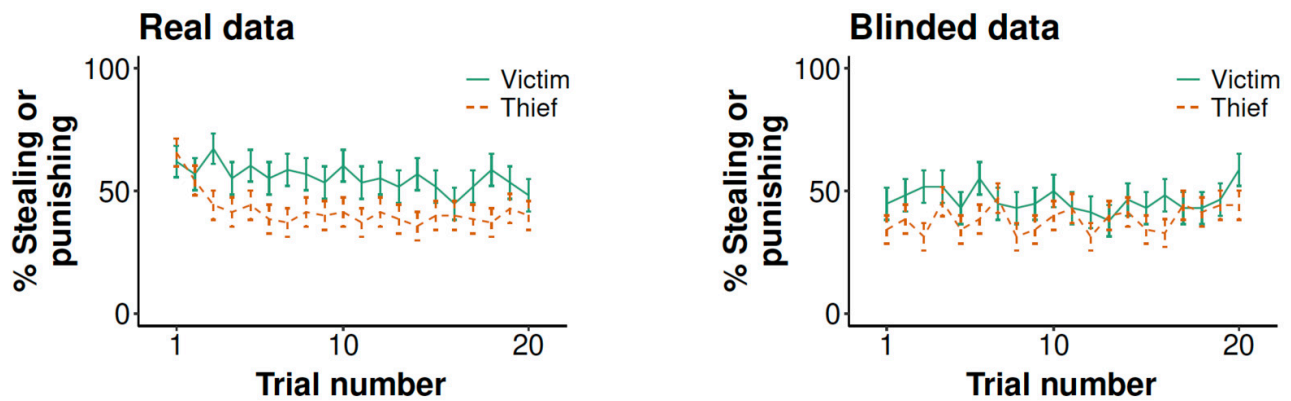


Figure A1. Response patterns for the Morris et al. replication data. The left panel illustrates the replication data, which indicates a weak but significant interaction between participants' roles ("thief" or "victim") and trial numbers on their choice of either stealing or punishing ($p = 0.019$). The right panel displays the blinded version of these data, where the interaction has been removed ($p = 0.457$).

helping decisions can be monitored by the subject they will later play the Trust Game with (process observable condition). In the other group, the players are told that only their helping decisions - not whether they look at the price of helping - can be monitored by the subject they later play the Trust Game with (process hidden condition). After deciding whether or not to look at the cost of helping and whether or not to help, participants play the Trust Game in the role of the receiver. The replication study investigated the hypothesis that participants are more likely to behave in an uncalculating manner (i.e., not looking at the cost of helping) in the process observable condition (when their reputation is at stake) than in the process hidden condition (when their reputation is not at stake).

The main outcome variable was the binary decision whether or not to check the cost of helping. To blind the data, we shuffled the main outcome variable across participants. This way, the comprehension check questions, which were dependent on the assigned experimental condition, could still be evaluated.

The hypothesis was evaluated in a logistic regression with process observability as the independent variable and looking decision as the dependent variable. The effect of process observability was statistically non-significant and negative ($b = -0.129$, $se = 0.122$, $z = -1.062$, $p = 0.288$, $n = 1826$). Applying the analysis to the blinded data likewise yielded a non-significant result ($b = 0.070$, $se = 0.122$, $z = 0.580$, $p = 0.562$, $n = 1826$).

Reeck et al.: the effect of search strategies on patience

In the between-subjects experiment by Reeck, Wall, and Johnson (2017) participants were randomly assigned to one of two search strategy conditions: the easy comparative condition or the easy integrative condition, where the ease of search strategies was manipulated. During each trial, participants had to choose between receiving a smaller monetary reward sooner or a larger monetary award later. The replication study investigated the hypothesis that par-

ticipants in the easy comparative condition would demonstrate greater patience (i.e., opting more often for the larger monetary reward delivered later) compared to those in the easy integrative condition.

The main outcome variable was the participants' choice: either participants chose the smaller, sooner reward or the larger, later one. To blind the data, we shuffled the main outcome variable across participants to leave most aspects of this complex dataset intact. This way, the manipulation could still be evaluated, which depended on participants' searching behavior (but not their choices), and the preprocessing could be implemented in an unbiased manner.

While shuffling the choice data across participants, we needed to ensure that certain features of the data remained intact to facilitate an informed assessment of data quality and a correct implementation of the preprocessing steps. Specifically, we preserved the choices from catch trials and implemented some initial preprocessing before blinding the data. That involved excluding participants who did not complete the experiment, along with those who completed it on mobile devices or tablets. This step aimed to prevent the mixing of choices from invalid participants with those from valid ones during the data blinding process which can cause higher exclusion rates and lower data quality. Additionally, we assigned a unique identifier to each trial in the experiment.

The focal hypothesis was tested using a hierarchical logistic regression. The coefficient estimate for the easy integrative condition was statistically non-significant ($\beta = -0.051$, $se = 0.047$, $z = -1.084$, $p = 0.278$, $n = 38378$, clustered on 1044 participants). Conducting the analysis on the blinded data likewise yielded a non-significant result ($\beta = -0.006$, $se = 0.01$, $z = -0.57$, $p = 0.569$, $n = 38378$, clustered on 1044 participants).

Analysis blinding for non-disputed replication studies

Next, we additionally demonstrate how data can be blinded for the three non-disputed replication stud-

ies–Stern, West, and Rule (2015), Williams, Sng, and Neuberg (2016), and John, Barasz, and Norton (2016)—we analyzed.

Stern et al.: money allocation to counterstereotypical targets by conservatives

In the within-subject experiment by Stern, West, and Rule (2015), participants read an excerpt of a story describing two fictitious groups, “Niffites” and “Luupites”. Participants were then shown four photographs of men, two of which belonged to each group. In both groups, one target confirmed the stereotype from the story of having / not having facial moles, and the other target deviated from the stereotype. Afterwards, participants were asked to allocate money (as gifts) to the four targets and rate their own political ideology. The replication study investigated the hypothesis that conservatives allocate less money to targets who deviate from stereotypes (i.e., counterstereotypical targets) than to targets who confirm the stereotype.

The main outcome variable was the percentage of money allocated to the four targets. To blind the data, we shuffled the division of allocated money to the four targets across participants. This way, the checks to validate the perceived stereotypicality of the targets could still be assessed.

The hypothesis was evaluated in a generalized estimating equations regression analysis with allocation as the dependent variable and stereotypicality, participants’ ideology (recentered on conservatives), and the interaction thereof as predictors. The main effect of stereotypicality was the focal test of interest. In the replication data, this effect was statistically non-significant ($b = 0.028$, $se = 0.0110$, $z = 0.252$, $p = 0.801$, $n = 503$). In the blinded data, the result was also statistically non-significant ($b = 0.022$, $se = 0.0110$, $z = 0.213$, $p = 0.831$, $n = 503$).

Williams et al.: the effect of ecology on perceived life history strategies

In the between-subject experiment by Williams, Sng, and Neuberg (2016), participants evaluated perceived life history strategies related to ecology stereotypes of a pictured target (a 24 year old man). The target was described as wealthy, randomized to be either from a hopeful or desperate ecology, as indicated by a picture of his neighborhood. Participants then answered how they perceived the target in terms of five ecology stereotype constructs (i.e., sexual unrestrictedness, impulsivity, opportunistic behavior, investment in own education, and investment in children). The replication study investigated the hypothesis that high-wealth individuals from desperate ecologies are stereotyped as possessing faster life history strategies (e.g., act impulsively, have more children than can be financially supported) than high-wealth individuals from hopeful ecologies.

The main outcome variable was the composite score of 18 indicators reflecting the five life history strategies. To blind the data, we shuffled the responses on the 18 items across participants, keeping the responses per participant together. This allowed us to still investigate the reliability of the five strategies as well as the overall construct in the blinded data.

The hypothesis was evaluated in a independent samples t -test with the composite score of fast life history strategies as the dependent variable and ecology condition as the independent variable. In the replication data, the effect of ecology was statistically significant ($M_{desperate} = 3.68$, $M_{hopeful} = 3.07$, difference = 0.607, $se = 0.185$, $t(110) = 3.286$, $p = 0.001$). In the blinded data, this effect disappeared ($M_{desperate} = 3.26$, $M_{hopeful} = 3.48$, difference = -0.219 , $se = 0.192$, $t(110) = -1.138$, $p = 0.257$).

John et al.: the effect of hiding personal information on dating interest

In the between-subject experiment by John, Barasz, and Norton (2016), participants viewed a completed questionnaire for a hypothetical dating prospect who had indicated the frequency with which he/she had engaged in a series of desirable behaviors. All participants saw three questions answered with a mix of “Sometimes” and “Frequently”. For the focal comparison, participants were randomized to either the Hider condition (with two additional questions that were answered as “Choose not to answer”) or the Inadvertent Nondiscloser condition (with two additional questions with a red X icon alongside each response option for these questions, as if due to a technical failure). Participants were then asked how interested they would be in dating this man/woman on a scale from 1 to 10. The replication study investigated the hypothesis that people are more interested in potential dates who inadvertently did not answer all questions on their desirable behaviors than potential dates who deliberately do not provide answers to all questions.

The main outcome variable was the dating interest on a 1-10 scale. To blind the data, we shuffled the interest score across participants.

The hypothesis was evaluated in a independent samples t -test with the dating interest as the dependent variable and disclosing condition as the independent variable. In the replication data, the effect of disclosing condition was statistically non-significant ($M_{nondisclosure} = 7.134$, $M_{hider} = 7.057$, difference = 0.077, $se = 0.112$, $t(1222) = 0.688$, $p = 0.492$). In the blinded data, similarly, the effect was non-significant ($M_{nondisclosure} = 7.102$, $M_{hider} = 7.088$, difference = 0.014, $se = 0.112$, $t(1222) = 0.69$, $p = 0.492$).

Supplementary Materials

Peer Review Communication

Download: https://collabra.scholasticahq.com/article/136869-analysis-blinding-as-a-potential-means-to-foster-a-productive-collaboration-between-original-authors-and-replicators/attachment/279286.docx?auth_token=jLlxsr9_B6u2_xywRf8T
