Meta-Psychology, 2025, vol 9, MP.2024.4324
https://doi.org/10.15626/MP.2024.4324
Article type: Commentary
Published under the CC-BY4.0 license

Open data: Yes
Open materials: Not Applicable
Open code: Yes
Open and reproducible analysis: Yes
Open reviews and editorial process: Yes
Preregistration: No

Edited by: Felix D. Schönbrodt
Reviewed by: Paul Riesthuis, Merle-Marie Pittelkow
Analysis reproduced by: Lucija Batinović
All supplementary files can be accessed at OSF:
https://doi.org/10.17605/OSF.IO/6DEZM

Replication Value Usage and its Performance for Large Sample Sizes - Commentary on Isager et al. (2025)

Linda C. Bomm[1], Delaney Peterson[1], and Bert N. Bakker[1]
[1]Amsterdam School of Communication Research, University of Amsterdam

The Replication Value (RV_{Cn}) metric was introduced to help researchers prioritize studies for replication based on expected utility. While we welcome the introduction of this straightforward and systematic replication decision approach, we identify two limitations of the RV_{Cn}. First, when testing the "repeatability" of a study or systematically incorporating replication into a research workflow, the RV_{Cn} may not always be the most suitable metric to guide decisions. Use cases should consider the scope conditions of the metric. Second, the RV_{Cn} shows limited sensitivity in distinguishing between studies with large sample sizes. To address this, we propose a simple adjustment: a log transformation of the sample size component. This modification improves the metric's discriminatory power for high-N studies and better aligns the (RV_{Cn}) with its intended purpose: guiding efficient and meaningful replication efforts.

Keywords: Replication value, RV_{Cn}, replication, study selection, study comparison

Introduction

To help researchers determine which studies to replicate, Isager et al. (2025) introduced the Replication Value (RV_{Cn}), "a proxy for expected utility gain" (p. 1). While the RV_{Cn} is a promising tool for prioritizing replication efforts, its utility is limited in specific contexts. Our commentary addresses two key limitations of the RV_{Cn}. First, we identify situations where this metric is less appropriate to use (see, e.g., Freese & Peterson, 2017). Second, the RV_{Cn} loses discriminatory power when applied to studies with large sample sizes, which poses challenges for fields that rely heavily on such studies (see, e.g., Arel-Bundock et al., 2024; Sun et al., 2025). Addressing these issues is essential for ensuring the RV_{Cn} achieves its goal of guiding effective replication efforts.

Limitation 1: The RV_{Cn} is not applicable to all types of replications

The RV_{Cn} is, according to Isager et al. (2025), useful for "any researcher, funder, journal or other stakeholder who wishes to direct limited resources towards important replication targets" (p. 35). However, Isager et al. (2025) do not explicitly specify which *type* of replication is best suited for the RV_{Cn}. Here, we discuss two scenarios in which the RV_{Cn} may be less applicable.

Scenario 1: Replication to test a study's repeatability of a study

Repeatability is a type of replication where "researchers collect new data to determine whether key results of a study can be obtained using the original procedure" (Freese & Peterson, 2017, p. 152). In the social sciences, a researcher can, for instance, test the repeatability of a study across different contexts (e.g., different countries) or time periods. In such cases, the RV_{Cn} may not be the best metric to guide decisions. For example, some of us tested the repeatability of study findings by Soroka et al. (2019) in a different context. Using the same stimuli and procedures, we directly replicated and extended the study in the Netherlands (Dubèl et al., 2024). However, had we used the RV_{Cn} to rank-order studies as Isager et al. (2025) propose, the study by Soroka et al. (2019) would not have been selected. The RV_{Cn} for Soroka et al. (2019) is 1.67 (371 citations in 6 years and a sample size of $N = 1100$). This RV_{Cn} is lower than directly related studies (e.g., Lang et al., 1996; Soroka & McAdams, 2015), which have higher RV_{Cn} values of 5.66 (Soroka & McAdams, 2015) and 1.75 (Lang et al., 1996). Despite this, we chose to replicate Soroka et al. (2019) because of its agenda-setting nature and its impact across multiple fields.

Scenario 2: Systematic Replication in Research Workflows

Another scenario where the RV_{Cn} is less applicable is when a researcher systematically integrates replication into their workflow (Chambers, 2017). For instance, the starting point of a PhD dissertation could involve a direct replication of a study that is highly relevant to the dissertation's central research question. In such cases, a PhD student might replicate a study with relatively high uncertainty (i.e., low sample size) to increase confidence in the findings. Alternatively, a PhD student might choose to replicate a study with a lower RV_{Cn} because it aligns closest with the core idea of their dissertation. Moreover, some PhD students might choose to replicate studies with relatively low uncertainty (i.e., high sample size) to further strengthen confidence in an already robust result (Chambers, 2017).

These examples demonstrate that when testing the "repeatability" of a study or systematically incorporating replication into a research workflow, the RV_{Cn} may not always be the most suitable metric to guide decisions. While the RV_{Cn} provides a useful heuristic for prioritizing replication targets, it does not fully account for the diversity of replication objectives and contexts. Going forward, future refinements to the RV_{Cn} and its proposed use cases should consider the scope conditions of the metric.

Limitation 2: the RV_{Cn} does not Discriminate Well when Comparing Studies with Large Sample Sizes

The RV_{Cn} formula introduced by Isager et al. (2025) is a function of citation count and sample size. Citations and sample sizes, however, vary significantly across fields. The authors of the RV_{Cn} acknowledge that field-specific differences in citation practices influence the metric's behavior. Specifically, they note that "[...] article citation counts tend to systematically vary between research fields [...]" and propose addressing this with a "field-weighted citation impact" that normalizes citation counts against the average citation count within a specific field (p. 16, Isager et al., 2025). This adjustment ensures that the RV_{Cn} accounts for differences in citation norms across fields.

However, the RV_{Cn} does not account for systematic differences in sample size between fields, even though fields with large sample sizes also contend with uncertainty (Arel-Bundock et al., 2024; Sun et al., 2025). In disciplines like political science, communication science, and economics, large sample sizes are common (see, e.g., Amsalem & Zoizner, 2022; Huber et al., 2023; Kertzer, 2022, for meta analyses reporting large sample sizes in these fields). For instance, political science frequently relies on large-scale surveys such as the American National Election Studies (ANES) and the European Social Survey (ESS) to study public opinion. Similarly, economics utilizes datasets like the Panel Study of Income Dynamics (PSID) and the World Bank's Living Standards Measurement Study (LSMS). Communication research often involves datasets derived from social media platforms such as Twitter/X or Facebook. Ignoring systematic differences in sample size could reduce the RV_{Cn}'s discriminatory power across fields, negatively impacting its intended use "for study selection across a variety of scientific disciplines" (Isager et al., 2025, p. 35). This raises an important question: How well does the RV_{Cn} capture uncertainty when sample sizes are large? We investigate this question and propose a simple revision to the RV_{Cn} formula.

To illustrate the issue, let us consider two studies with relatively small sample sizes, Study A and Study B, which each receive 20 citations in their first year. Study A has a sample size of $N = 50$, resulting in a multiplicative term of 0.141 ($\frac{1}{\sqrt{50}}$). Study B, with a sample size of $N = 200$, results in a multiplicative term of 0.071 ($\frac{1}{\sqrt{200}}$). Consequently, the RV_{Cn} for Study A is 1.41 ($\frac{20}{1+1} \times 0.141$), while for Study B it is 0.71 ($\frac{20}{1+1} \times 0.071$). Study A therefore has a higher RV_{Cn} and would be prioritized for replication.

As sample sizes increase, however, the second part of the RV_{Cn} formula ($\frac{1}{\sqrt{n}}$) approaches zero. This means that for studies with large sample sizes, the RV_{Cn} becomes almost entirely dependent on citation count. To illustrate our point, we conducted simulations varying sample size and plotted the RV_{Cn} across a range of citation counts (x-axis) and years since publication (y-axis) – to access the code, see: https://osf.io/kqjpz/. In Figure 1, brighter yellow colors indicate higher RV_{Cn} values, while darker blue colors indicate lower values.

The top row of Figure 1 shows RV_{Cn} outcomes for sample sizes of $N = 50$ (left), $N = 500$ (middle), and $N = 3000$ (right). For smaller sample sizes ($N = 50$), the RV_{Cn} demonstrates good discriminatory power, with clear variation in values, especially as citation counts increase. However, as sample sizes grow ($N = 500$ and $N = 3000$), the RV_{Cn} values approach zero in most cases, as indicated by the predominantly dark blue panels. This demonstrates the reduced discriminatory power of the RV_{Cn} at large sample sizes.

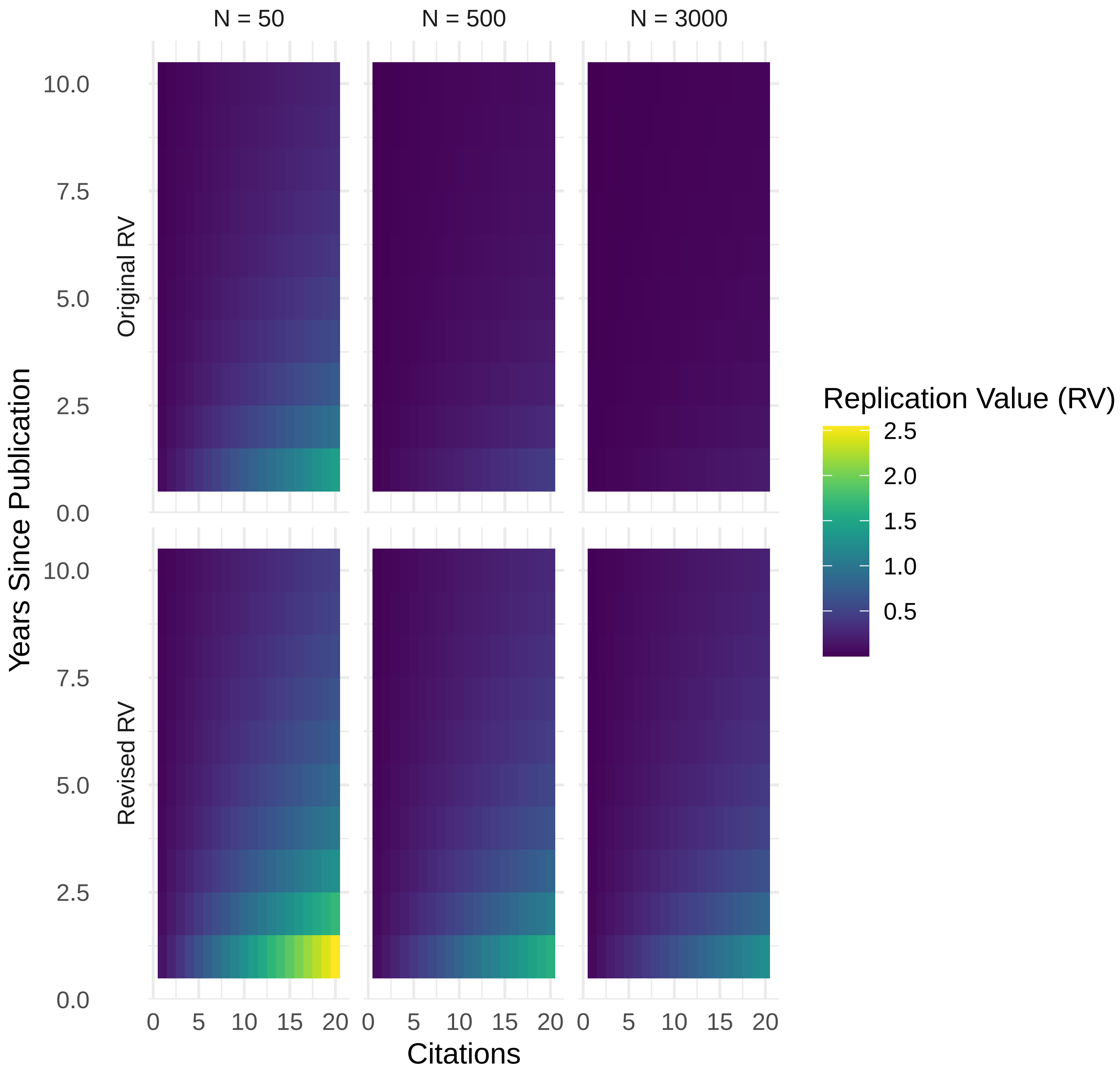
We propose a simple modification to the RV_{Cn} formula by taking the logarithm of the sample size to better account for large sample sizes:

$$RV_{Cn}\text{revised} = \left(\frac{\text{citations}}{\text{years\_since\_pub} + 1}\right) \times \left(\frac{1}{\log(n+1)}\right)$$

This simple adjustment reduces the influence of sample size as it grows, thereby improving the discriminatory power of the RV_{Cn} across studies with large samples. The bottom row of Figure 1 shows the revised RV_{Cn} outcomes for $N = 50$ (left), $N = 500$ (middle), and

Figure 1

Original and Revised Replication Values for Different Sample Sizes.



Note. Heatmaps of simulated replication values (top: original; bottom: revised) for three different sample sizes (left: N = 50; middle: N = 500; right: N = 3000). RV values are indicated by colour, with the colour spectrum ranging from dark blue (lowest RV) to yellow (highest RV).

$N = 3000$ (right). It demonstrates that at small samples ($N = 50$), the original and revised RV_{Cn} can be used as both have good discriminatory power. Yet, compared to the original RV_{Cn}, the revised version demonstrates greater variation in RV_{Cn} values for large sample sizes, as indicated by the broader range of colors in the middle and right panels.

Our simulations demonstrate that as sample sizes increase, the differences in the RV_{Cn} values in its current form diminish significantly, reducing its discriminatory power. Although the rank-order of studies remains unchanged, the RV_{Cn} is intended to represent replication value rather than mere rank. This distinction is critical because the magnitude of RV_{Cn} should accurately reflect meaningful differences, rather than being disproportionately influenced by field-specific variations in sample size. In fields with large sample sizes, stakeholders such as journal editors and funders, risk making decisions based on negligible differences at the second decimal place, while fields with smaller samples exhibit more substantial variation. This disparity becomes particularly problematic when the RV_{Cn} is used to prioritize replication targets across disciplines with systematically different norms for sample size.

Our revised RV_{Cn} formula offers a straightforward yet effective solution. By incorporating the logarithm of sample size, it mitigates the overweighting of citations and restores the metric's ability to discriminate between studies with large samples. This adjustment preserves the interpretability of the RV_{Cn} and the conciseness of its formula, while addressing its limitations, ensuring fairer and more accurate allocation of resources for replication efforts across diverse fields.

Conclusion

We applaud Isager et al. (2025) for the introduction of the RV_{Cn}. We hope that our comment supports the discussion about the use and functioning of the RV_{Cn} moving forward.

Author Contact

Corresponding author: Linda C. Bomm
ORCID: Linda C. Bomm: 0009-0009-5316-1071
ORCID: Delaney Peterson: 0009-0006-8825-5922
ORCID: Bert N. Bakker: 0000-0002-6491-5045

Conflict of Interest and Funding

This publication is part of the project "Under pressure: How citizens respond to threats and adopt the attitudes and behaviours to counter them" [project number VI.Vidi.211.055, awarded to Bert N. Bakker] of the NWO Talent Programme VIDI, which is financed by the Dutch Research Council (NWO).
The authors have no conflicts of interests related to this publication.

Author Contributions

We base this author contribution statement on the CRedit norms, see https://www.elsevier.com/researcher/author/policies-and-guidelines/credit-author-statement. Within each credit category, the authors' initials are listed in no particular order.
Conceptualization: LCB, DP, BNB; Data Curation: BNB; Formal Analysis: LCB, BNB; Funding Acquisition: BNB; Investigation: BNB; Methodology: LCB, DP, BNB; Project Administration: BNB; Resources: BNB; Supervision: BNB; Validation: LCB, DP, BNB; Visualization: LCB; Writing – original draft: LCB, DP, BNB; Writing – review & editing: LCB, DP, BNB.

Open Science Practices

This article earned the Open Data and Open Code badge making the data, and code openly available. It has been verified that the analysis reproduced the results presented in the article. The entire editorial process, including the open reviews, is published in the online supplement.

References

Amsalem, E., & Zoizner, A. (2022). Real, but limited: A meta-analytic assessment of framing effects in the political domain. *British Journal of Political Science*, *52*(1), 221–237.

Arel-Bundock, V., Briggs, R. C., Doucouliagos, H., Mendoza Aviña, M., & Stanley, T. D. (2024). Quantitative political science research is greatly underpowered. *The Journal of Politics*.

Chambers, C. (2017). *The seven deadly sins of psychology: A manifesto for reforming the culture of scientific practice*. Princeton University Press.

Dubèl, R., Schumacher, G., Homan, M. D., Peterson, D., & Bakker, B. N. (2024). Replicating and extending soroka, fournier, and nir: Negative news increases arousal and negative affect. *Media and Communication*, *12*.

Freese, J., & Peterson, D. (2017). Replication in social science. *Annual Review of Sociology*, *43*(1), 147–165.

Huber, C., Dreber, A., Huber, J., Johannesson, M., Kirchler, M., Weitzel, U., Abellán, M., Adayeva, X., Ay, F. C., Barron, K., et al. (2023). Competition and moral behavior: A meta-analysis of forty-five crowd-sourced experimental designs. *Proceedings of the National Academy of Sciences*, *120*(23), e2215572120.

Isager, P., van 't Veer, A., & Lakens, D. (2025). Replication value as a function of citation impact and sample size. *Meta-Psychology*, *9*. https://doi.org/10.15626/MP.2022.3300

Kertzer, J. D. (2022). Re-assessing elite-public gaps in political behavior. *American Journal of Political Science*, *66*(3), 539–553.

Lang, A., Newhagen, J., & Reeves, B. (1996). Negative video as structure: Emotion, attention, capacity, and memory. *Journal of Broadcasting & Electronic Media*, *40*(4), 460–477.

Soroka, S., Fournier, P., & Nir, L. (2019). Cross-national evidence of a negativity bias in psychophysiological reactions to news. *Proceedings of the National Academy of Sciences*, *116*(38), 18888–18892.

Soroka, S., & McAdams, S. (2015). News, politics, and negativity. *Political communication*, *32*(1), 1–22.

Sun, Y., Shen, L., Pan, Z., & Qian, S. (2025). Toward a more powerful experimental communication science: An assessment of two decades' research (2001–2023). *Communication Research*, 00936502241308599.