Meta-Psychology, 2025, vol 9, MP.2024.4323
https://doi.org/10.15626/MP.2024.4323
Article type: Commentary

Open data: Yes
Open materials: Yes
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, Adrien Fillon
Analysis reproduced by: Lucija Batinović
All supplementary files can be accessed at OSF:
https://doi.org/10.17605/OSF.IO/4TDF3

Thinking beyond RV_{CN}: Adressing the complexity of replication target selection.

Merle-Marie Pittelkow[1], Sarahanne M. Field[2], and Don van Ravenzwaaij[3]
[1]QUEST Center for Responsible Research, Berlin Institute of Health at Charité - Universitätsmedizin Berlin, Berlin, Germany
[2]Department of Pedagogy, Faculty of Behavioral and Social Sciences, University of Groningen, Groningen, The Netherlands
[3]Department of Psychometrics and Statistics, Faculty of Behavioral and Social Sciences, University of Groningen, Groningen, The Netherlands

Isager, van 't Veer and Lakens (2025) proposed a quantitative operationalization of replication value (denoted RV_{CN}), using average yearly citation count and sample size as proxies for value and uncertainty, respectively. In this commentary, we suggest that the approach of Isager et al., while a good theoretical departure point, oversimplifies the complex decision-making process that underpins replication target selection in practice. We present what we view as some issues with RV_{CN}, notably the use of citation count and ambiguity as to whether RV_{CN} is prescriptive or descriptive. We also present preliminary empirical evidence that RV_{CN} diverges on its performance as a replication target selection method, compared with existing selection methods (such as those published by us in the past). We conclude with the recommendation that going forward, approaches should emphasize the multifaceted nature of replication target selection to maximize their practical utility.

Keywords: replication target selection, qualitative frameworks

Isager, van 't Veer, and Lakens (2025 [IVL21]) put forth a proposal to quantify replication value. Replication value, denoted as RV_{CN}, is calculated as the product of the average yearly citation count of a given article in which the original effect was reported (value) and the sample size used to investigate the original study (uncertainty). The authors further suggest that replication targets should be identified using a four-step selection procedure including: (1) curation of an initial set of candidate studies, (2) calculation of RV_{CN} for the candidate studies, (3) in-depth evaluation of the subset of studies with the highest RV_{CN}, and (4) selection of the most suitable candidate based on RV_{CN} and the in-depth evaluation.

We agree with the notion that replication value is best thought of as a function of value and uncertainty. However, we view the proposal of IVL21 as a starting point for a much-needed expert discussion about the conceptualization and potential quantification of replication value as well as other assessment strategies for replication targets. In this commentary, we identify what we believe to be some issues with RV_{CN} before comparing it to our own selection procedures (Field et al., 2019; Pittelkow et al., 2021, 2023).

RV_{CN} fails to capture the complexity of replication target selection in practice

Our primary point of critique is that the conceptualization of what makes up value and uncertainty, operationalized by just one measure each, oversimplifies the multifaceted approach typically employed in practice for choosing replication targets. Results from surveying researchers who have conducted, or plan to conduct a replication suggest that they typically consider additional aspects such as feasibility and methodology when deciding what to replicate (Pittelkow et al., 2023). Feasibility, while not related to the concepts of value and uncertainty, is considered very important by replicating authors because replication is only possible when the necessary resources (e.g., money, time, staff, expertise) are available. (In-)appropriate methodology informs both value and uncertainty. Although IVL21 refer to additional qualitative assessments for studies with the largest RV_{CN}, we argue that it is premature to speak of the formula as assessing replication value without including these important additional concepts.

It is unclear whether RV_{CN} is prescriptive or descriptive

An overarching question central to the utility of measures such as RV_{CN} is whether it is meant to be prescriptive (what ought to be selected) or descriptive (what is selected in practice). The first half of the article by IVL21 suggests RV_{CN} is prescriptive, as the aim of the article is to develop a measurement model of a study selection metric aiming to maximize the expected utility gain of a single replication. From the section *"Preliminary validation [...] studies"* on, the distinction between RV_{CN} as a prescriptive and descriptive measure becomes murkier. In this section, the authors propose validating the RV_{CN} metric by correlating actual study selection behaviour with what RV_{CN} suggests researchers should have selected. In our assessment, this correlation only makes sense when evaluating RV_{CN} as a descriptive measure. If researchers already select studies with the largest utility gain for replication, there would be no need for a measure like RV_{CN}. Our own work suggests that replication study selection in practice is often content-driven (e.g., based on interest) and not utility-based (Pittelkow et al., 2023). Thus, we argue that the preliminary validation strategy of correlating RV_{CN} to actual replication behaviour is unlikely to be particularly informative.

Indicators depend on their specification

Citation count is an imperfect indicator of scientific value. As the authors mention in Figure 1, there are many reasons why articles might get cited that have little to do with the scientific value of an article, such as studies which have had a lot of social media attention for being novel or controversial (and not necessarily reliable or valid). Examples include scandals surrounding the authors, methodological flaws, or citation practices (self-citations by authors or by journals, and citation bias). Citation counts also vary considerably between scientific subfields (Patience et al., 2017) making RV_{CN} unsuitable for assessments or comparisons across fields.

Every metric has faults, but accepting the premise that citation count might possess some utility as a metric for scientific value, we turn to the specific version the authors use in their preliminary validation, which closely follows the predicted citation count of articles as described in Figure 2A of IVL21 (i.e., a uniform distribution). The authors acknowledge that the actual trajectory of citations through time might be better approximated by Figure 2B of IVL21 (i.e., a gamma distribution). We agree with this observation and to assess robustness of the chosen distribution for predicting citation count, we have redone the preliminary validation of IVL21, using their parametrization of the gamma distribution instead[1]. To this end, we adjusted the citation counts by assuming x is gamma distributed with a shape parameter of 2 and a rate parameter of 0.8. Parameter x is defined as years since publication + 1 divided by 5 (see IVL21). This distribution produces a corrected citation count accounting for the expected increase in citations over time thus normalizing the citation count in relation to the age of the publication (see Figure 1B). This impacted the values of RV_{CN} leading to less overlap between RV_{CN} values for replicated and non-replicated studies (see Figure 1D). This demonstrates the dependency of RV_{CN} on the distributional assumptions for citation count and underscores the need for follow-up work on the best way to quantify 'value'.

Comparison of RV_{CN} to previous suggestions

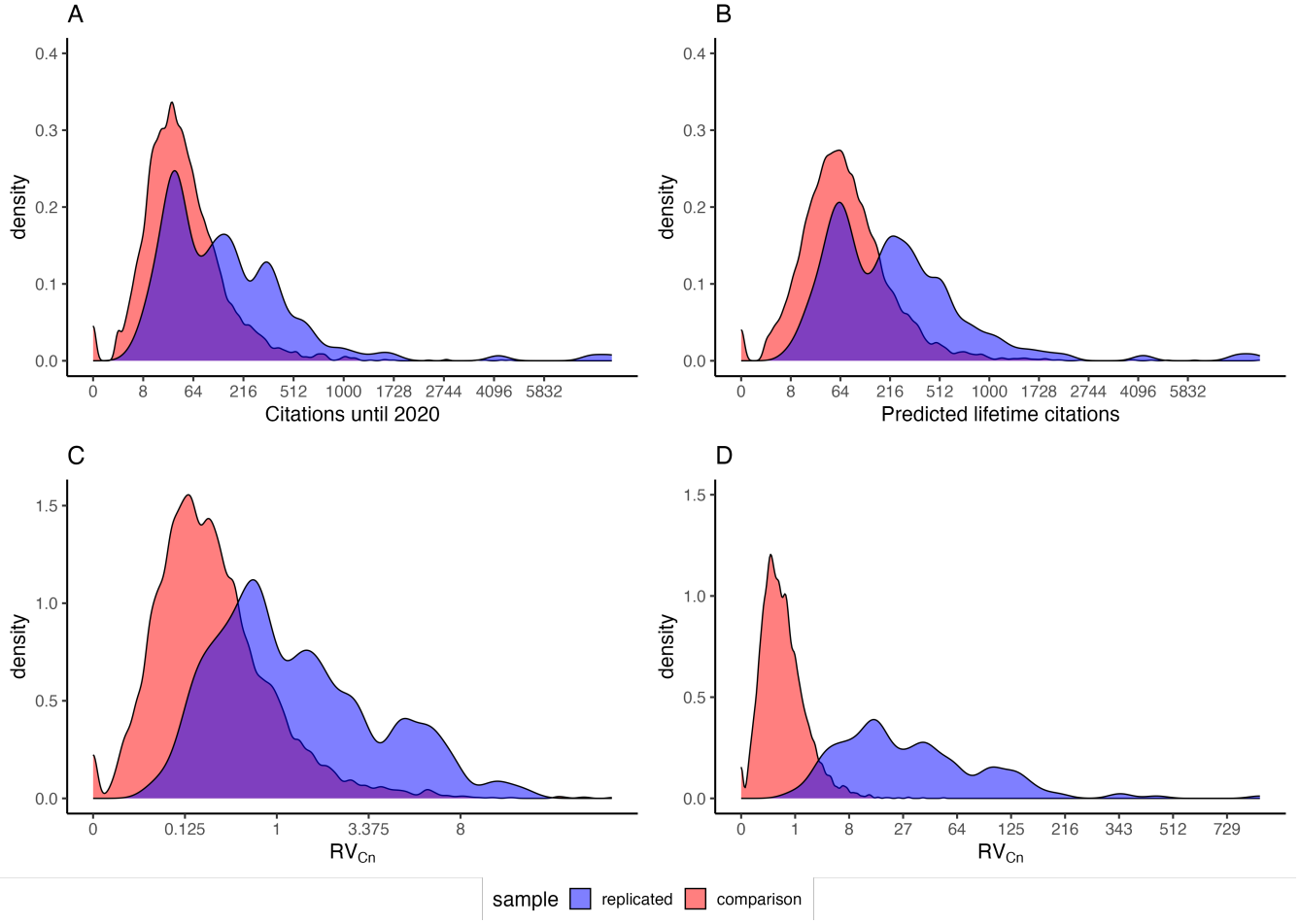
In this section, we compare RV_{CN} to replication target selection procedures outlined in our earlier publications (Field et al., 2019; Pittelkow et al., 2021). A full description of these procedures is beyond the scope of this commentary. Briefly, we proposed that studies with uncertain evidence are in greater need of corroboration than studies with strong evidence either for or against a particular effect. To quantify this uncertainty, we advocated for the use of Bayes Factors (BFs), which assess the relative strength of evidence supporting two competing hypotheses, thereby providing a quantitative and continuous measure of uncertainty. Studies identified as having ambiguous evidence were subsequently assessed based on qualitative criteria, as illustrated in Table 1 and recommendations for possible replication targets were made. Following these projects, we conducted a large-scale initiative involving a survey and a Delphi consensus method to develop a comprehensive set of qualitative criteria (see Table 1) for replication target selection in practice (Pittelkow et al., 2023).

Our assessment of the proposal by IVL21 is that RV_{CN} acts as a filter to make a pre-selection based on a set of candidate studies. As such it resembles the function of the Bayes Factor (BF) as described in our earlier publications (Field et al., 2019; Pittelkow et al., 2021). We therefore decided to compare the performance of RV_{CN} to the BF and the results of Field et al. (2019) and Pittelkow et al. (2021). Field et al. (2019) extracted data for 57 results from 30 Psychological Science articles published between 2015-2016, that reported significant statistical tests (one–sample, paired, and independent t-tests), associated with primary research questions. Pittelkow et al. (2021) extracted data for 97 results from

[1]While this exact parametrization is to some extent arbitrary, we agree with the authors that "*Figure 2B displays a more realistic 50-year citation trajectory (Parolo et al., 2015)*".

Figure 1

Distribution of overall citation count and RV_{CN} in the comparison sample of psychological findings (red) and the sample of replicated findings in psychology (blue). Following IVL21, the scale in all plots has been transformed by taking the cube root of the true values, which preserves the overall shape of the distribution but compresses the scale towards 1. (A) Overall citation count as reported in IVL21. (B) Overall citation count as estimated based on the gamma distribution. (C) RV_{CN} as reported by IVL21 (D) RV_{CN} based on total predicted citation count based on the gamma parameterization as specified by IVL21 in Figure 2.



78 articles published in the Journal of Consulting and Clinical Psychology published between 2012 and 2016, which supported their primary outcome by a statistically significant *t*-test. We had previously extracted the sample size. To calculate RV_{CN} we collected additional data on citation count from Crossref on January 16th 2025 following the method proposed by IVL21. Importantly, our data includes information from additional qualitative analyses assessing the value of the replication targets. If RV_{CN} captures 'replication value', the studies identified by qualitative assessment should also present relatively larger RV_{CN} values. Figure 2 presents the results of our re-analysis.

There was no clear relationship between RV_{CN} and BFs (r_{Field} = -0.09, $r_{Pittelkow}$ = -0.12), suggesting that the two measures capture distinct aspects of replication value. Furthermore, RV_{CN} scores did not differ between studies identified as replication targets through qualitative assessment and those that did not (see Table 2). For the dataset from Field et al. (2019), RV_{CN} values were even lower for studies suggested as replication targets compared to those not identified as replication targets. This implies that potentially important replication targets might be prematurely excluded from consideration based solely on RV_{CN} scores. However, this does not mean that these studies are unsuitable replication tar-

Table 1

Qualitative criteria proposed in previous work

Field et al. (2019)	**Pittelkow et al. (2021)**	**Pittelkow et al. (2023)**
Relevance	Relevance	Clinical relevance
		Interventional study
Interest	Relevance of the OS for current research	
	Clinical sample	Doubt
	Severity of the condition	Current strength of evidence in favor of the OS
		(Un)clarity and (Un)replicability of the original protocol
Insufficient investigation	Scientific relevance	Evidence base
Theoretical importance	Quality	Theory
	Scientific background sound	Theoretical relevance of the OS
	Clear rationale	Implications of the OS (e.g., for practice, policy, or clinical work)
Methodology	Methodology	CONSORT criteria
	Statistical method appropriate	Flaws of the OS
	Interpretation	Interpretation and conclusion follow logically
	Generalizability	Concerns about questionable research practices
	Robustness	Generalizability of the OS
	Feasibility	Resources available for replicating the OS
		The replicating team's experience or expertise with the OS

Note. This table does not provide a row-by-row comparison but lists different criteria. OS = original study, equivalent to the original claim.

gets. Every selection procedure has its strengths and limitations, but the minimal overlap between our qualitative suggestions and RV_{CN} -based selections, particularly in the Field et al., 2019 dataset, begs the question of why two different selection procedures come to such different conclusions.

Concluding remarks

We hope that this commentary can provide some nuance and multidimensionality to the perspective of IVL21. While their approach to assessing replication values facilitates the evaluation of large sets of potential targets, it risks oversimplifying a complex reality. However, we believe the IVL21 approach provides a systematic starting point that is preferable to having no structured methodology at all. Simplification by quantitative approximation does carry risks, but these need to be weighed against the practical benefits of transparent and systematic approximations.

Still, we argue that failing to emphasize the complexity of replication target selection may compromise the utility of their method. Although we have previously proposed partially quantitative methods ourselves, our recent work has shifted towards more qualitative and open approaches for describing replication value, exemplified by the checklist for transparent replication target selection (for more details please review Pittelkow et al., 2023). We contend that such flexible methodolo-

Figure 2

Scatterplots plotting replication value (RV_{CN}) against Bayes Factors (BF). BFs are presented on a reversed log 10 scale. The dotted lines indicate $BF = 1$ and $BF = 3$. We excluded very large BFs and RV_{CN} ($BF > 10^{10}$; $RV_{CN} > 10$).

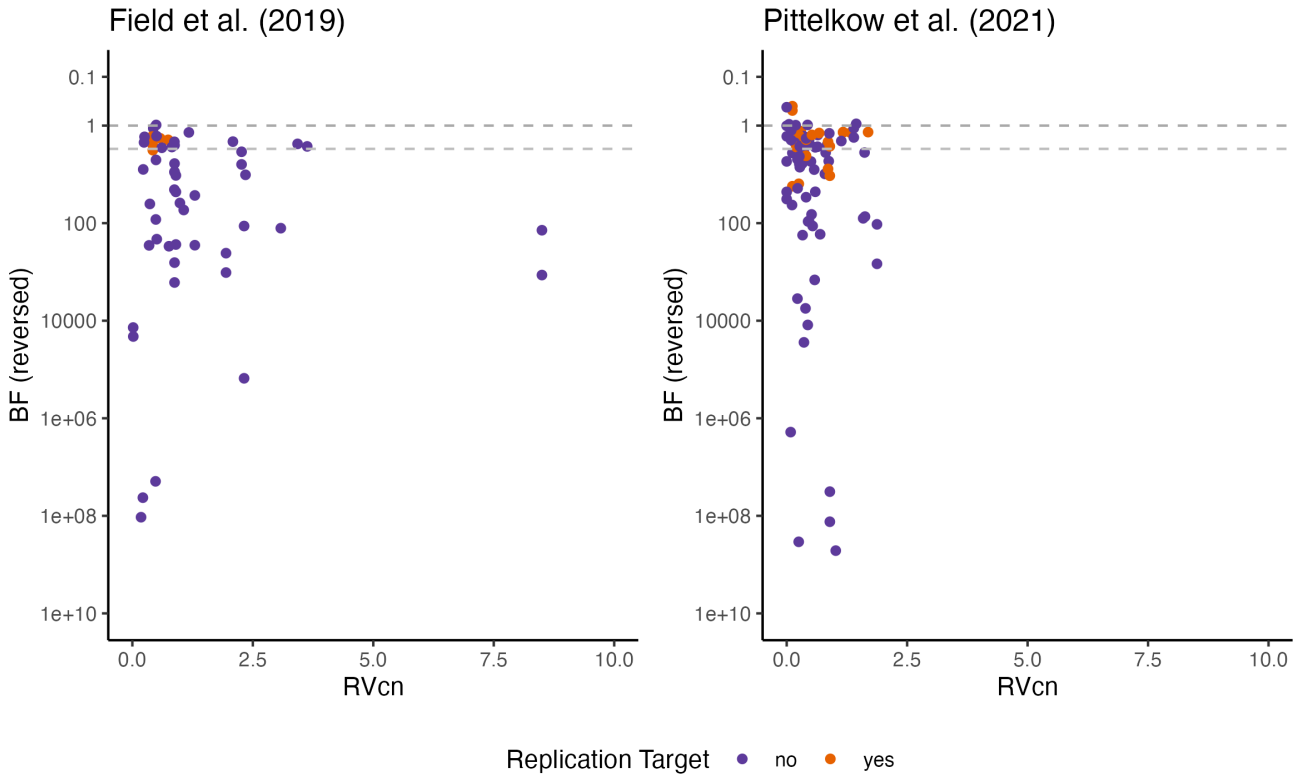


Table 2

Average RV_{CN} scores for studies suggested and not suggested as replication targets based on previous analyses

	RT	N	M	Mdn	IQR
Field et al.	Yes	8	0.48	0.42	[0.42, 0.46]
	No	49	1.38	0.88	[0.49, 1.94]
Pittelkow et al.	Yes	21	0.61	0.40	[0.25, 0.89]
	No	73	0.53	0.40	[0.23, 0.65]

Note. RV_{CN} = Replication Value Citation Number. M = Mean, Mdn = Median, IQR = Interquartile Range, RT = Replication Target.

gies offer greater practical utility in the nuanced landscape of replication research. At the same time, we caution against definitive conclusions about the superiority of any one method in capturing replication value. By continuing to assess these approaches critically, we can refine strategies for replication target selection.

Author Contact

Merle-Marie Pittelkow, merle-marie.pittelkow@bih-charite.de, Anna-Louisa-Karsch-Straße 2, 10178 Berlin. ORCID: https://orcid.org/0000-0002-7487-7898 Sarahanne M. Field ORCID: https://orcid.org/0000-0001-7874-1261 Don van Ravenzwaaij: ORCID: https://orcid.org/0000-0002-5030-4091

Data Availability

The data and code that support the findings of this study are openly available on the Open Science Framework (OSF) at https://osf.io/hnsry/; DOI: 10.17605/OSF.IO/HNSRY.

Conflict of Interest and Funding

The authors declare no conflict of interests.

Author Contributions

M.-M.P.: Conceptualization, Data curation, Formal analysis, Project administration, Visualization, and Writing - original draft. **S.M.F.**: Conceptualization, Data curation, and Writing - review & editing. **D.v.R.**: Conceptualization, Formal analysis, and Writing - original draft.

Open Science Practices

This article earned the Open Data, Open Materials, and Open Code badge for making the data, materials, 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

Field, S. M., Hoekstra, R., Bringmann, L., & van Ravenzwaaij, D. (2019). When and Why to Replicate: As Easy as 1, 2, 3? *Collabra: Psychology*, *5*(1), 46. https://doi.org/10.1525/collabra.218

Isager, P. M., van't Veer, A. E., & 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

Patience, G. S., Patience, C. A., Blais, B., & Bertrand, F. (2017). Citation analysis of scientific categories. *Heliyon*, *3*(5), e00300. https://doi.org/10.1016/j.heliyon.2017.e00300

Pittelkow, M.-M., Field, S. M., Isager, P. M., van't Veer, A. E., Anderson, T., Cole, S. N., Dominik, T., Giner-Sorolla, R., Gok, S., Heyman, T., Jekel, M., Luke, T. J., Mitchell, D. B., Peels, R., Pendrous, R., Sarrazin, S., Schauer, J. M., Specker, E., Tran, U. S., ... van Ravenzwaaij, D. (2023). The process of replication target selection in psychology: What to consider? [Publisher: Royal Society]. *Royal Society Open Science*, *10*(2), 210586. https://doi.org/10.1098/rsos.210586

Pittelkow, M.-M., Hoekstra, R., Karsten, J., & van Ravenzwaaij, D. (2021). Replication target selection in clinical psychology: A Bayesian and qualitative reevaluation. *Clinical Psychology: Science and Practice*, *28*(2), 210–221. https://doi.org/10.1037/CPS0000013