Meta-Psychology, 2025, vol 9, MP.2024.4210
https://doi.org/10.15626/MP.2024.4210
Article type: Commentary

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

Edited by: Felix D. Schönbrodt
Reviewed by: Rene Bekkers, Steven Verheyen
Analysis reproduced by: Not Applicable
All supplementary files can be accessed at OSF:
https://doi.org/10.17605/OSF.IO/VA28K

Valuing replication value

Kaito Takashima and Yuki Yamada
Kyushu University

This is a commentary piece on the proposal by Isager et al. (2025) for a new metric, RV_{Cn}, designed to evaluate the replication value of psychological studies. We discuss the hope of utilizing the RV_{Cn} metric in undergraduate education and possible improvements on using some elements other than original sample size to derive this metric.

Keywords: replication value, uncertainty, mentoring, sample size

The reproducibility problem in psychology has rendered uncertain which of the daily explosion of psychological findings can be trustworthy, affecting all researchers, especially early career researchers (ECRs). In the past, countless ECRs have left academia only to find that the previous studies on which they relied were completely unreproducible. Replication is the necessary process for verifying the reproducibility of studies. Yet, until now, which of the myriad of previous studies should be replicated has been determined based on gut feelings rather than systematic criteria, but this has obvious limits.

Thus, P. Isager et al. (2025) proposed employing RV_{Cn}, a metric of replication value derived from the value of a study as a function of citation count and from the uncertainty of the study as a function of its original sample size. Based on P. M. Isager (2020) and P. M. Isager et al. (2023), this metric aims to maximize utility from replication efforts and offers a straightforward way to prioritize studies. However, as the original authors note, this metric is not intended to function as a definitive or sole criterion. Instead, here we will discuss that it is good to consider other factors—such as study quality, effect size, and the existence of previous replications as well—when determining the priority of replication targets.

This is a challenge for any researcher interested in the reproducibility problem in psychology, as they should choose what to replicate amidst limited time, labor, and financial resources. With such a metric, replication studies could become a lower hurdle, more efficient, and more effective for the self-correction of psychology[1] Here, we discuss both the potential benefits and problems of this metric toward its more effective practical application in the future.

May aid in mentoring many students

RV_{Cn} could be helpful as a criterion for undergraduate and graduate students when selecting studies for replication, such as in experimental exercises or for their theses. Based on our observation, several principal investigators (PIs) with many students cannot always be intimately familiar with every student's research plan or its academic context (this is a major problem in itself...). Some PIs find great satisfaction in writing papers with undergraduates, yet they also experience a shortage of time to dedicate to the process (Howell et al., 2024). When advising on replication proposals (including conceptual replications) brought by students and making the "GO" decision, PIs often find it challenging to judge. In such cases, quantifying the worthiness of replication could aid decision-making without requiring in-depth expertise in all those topics.

Our observation is that it is relatively common, especially at private universities in Japan, for a single PI to in fact mentor a number of undergraduate research projects beyond coverage within the capacity of a single PI to handle. We conducted a brief survey within the Psychological Science Accelerator, a community of psychologists worldwide, at their Slack and within a Japanese community at X (N = 97 in total) (https://osf.io/ch397). As a result, we found that 27.29% of PIs mentor more than 10 students per year, and surprisingly, nearly 10% of PIs mentor more than 15 students per year. While the specific characteristics that lead to high-capacity labs remain unclear, it is evident that there is indeed a non-trivial proportion of PIs who annually manage and handle more than 10 theses. Imagine that one day, 15 students, who are all studying different topics, come to you and each consult with you about selecting one study for replication out of 10 previous studies. Would you be able to immediately give them all an appropriate answer? In such a situation, it is likely that busy PIs can receive the benefit of the RV_{Cn} and fo-

[1] This example represents only one context in which the metric may be applied, and there are likely many other scenarios in which it could be similarly beneficial.

cus their mentoring on the more specific aspects of the study. Following this initial screening, PIs can devote their limited time to evaluating other important factors. For example, they may assess whether the chosen experiment is feasible given the laboratory's available resources, determine how best to allocate those resources, and prepare the research environment.

Original sample size as the indicator of uncertainty

Sample size is a reasonable metric for replication decisions, but larger samples do not necessarily ensure higher quality. This is especially true in fields like visual science, where strong effects can be observed with a minimal sample. In such cases, the sample size may not serve as a reliable indicator of uncertainty. For example, consider visual illusions, a common phenomenon or subject in visual science. Research on visual illusions often reports strong effects with a few participants (sometimes only the authors themselves serve) (e.g., Anstis, 2022; Balas, 2021; Sugihara, 2023). In these situations, there is little to no uncertainty, and the sample size does not also reflect the strength or reliability of the effect. However, the calculation method for RV_{Cn} could lead to significantly different evaluations of the replication value of studies with an equal citation count solely based on the difference in sample size. As a result, a larger-sample study with a smaller but more reliable effect may appear less valuable than a smaller-sample study with an anticipated larger effect. This bias, in turn, could affect how limited research resources are allocated. While we acknowledge that comparing such cases can be challenging, it is nonetheless important to consider these factors to ensure that limited research resources are allocated effectively. Such considerations highlight the need to consider factors other than sample size, such as effect size and research quality when determining the replication value. As we will note later, the original authors also seem to share similar concerns and do not intend to rely solely on sample size.

Additionally, large-scale surveys, common in social sciences and psychology, often exceed 1,000 participants. However, a large sample does not necessarily indicate high quality. Online crowdsourcing allows rapid data collection, but data quality depends on well-defined exclusion criteria and satisficing detection.

Another important consideration in selecting studies for replication is adjusting for the presence of already conducted replications. If there are already many replication efforts for an original study, the replication value of the research would naturally decrease due to the diminishing new insights or value and reduced uncertainty that additional replications would provide. As also noted by the original authors, it is reasonable to include the total sample size of existing replications in the selection strategy for replication studies. As a tentative approach, we suggest taking the square root of each study's sample size first and then summing them, rather than summing *n*s and taking the square root afterward. This approach may help avoid a potential underestimation of the standard deviation (*SD*) when dealing with large sample sizes. This arises from the nonlinear relationship between sample size and *SD*. These approaches could lower the priority of extensively replicated studies, allowing resources to be concentrated on studies replicated only a few or zero times. This suggestion aligns with the original authors' vision of integrating the sample sizes of previously conducted replications to collectively assess replication value.

Moreover, the characteristics of the original study's sample are crucial points in selecting replication studies. To enhance the generalizability of research findings, replications targeting different demographic or geographical groups may be particularly valuable if the original research is biased toward a specific population. For example, using the cultural distance between the sample of the planned replication and the sample of the original study as part of the value metric might be beneficial. Effectively quantifying cultural distance (e.g., Correa da Cunha et al., 2022) would be ideal, but even a simple constant to indicate whether the sample is the same as the original study or not could be still meaningful (e.g., same sample: 1, different sample: 2). An interesting aspect of this idea is that the replication value could also depend on the relationship with the replication plan. Of course, cultural distance is only one dimension. Others might include historical contexts, languages, or technological factors. Integrating these considerations could provide a more holistic picture of generalizability. Nevertheless, we recognize that this approach adds complexity and may go beyond the original authors' primary focus. Therefore, we present this idea as a tentative, illustrative extension rather than a definitive addition to the replication value metric.

Even if it is for the initial quick screening, relying solely on sample size in selecting replication studies may risk overlooking the true value and impact of the research. We understand that the original authors recognize sample size as one important, yet not exclusive, criterion (see Figure 3 in P. Isager et al. (2025) for detail). Building on this foundation, more accurately assessing replication value benefits from incorporating additional measurable variables, where available. For example, weighted variables that consider the effect size of the original research, whether the study was preregistered, the openness of the analysis code and data, and the number of hypotheses tested could play a criti-

cal role in enhancing the evaluation process. Reaffirming the importance of these factors in the replication study selection process and developing a more comprehensive evaluation framework is necessary. The difficult thing is probably to reconcile that with the brevity and computability of the calculation.

Future with RV_{Cn}

Finally, consider an academic world where RV_{Cn} is widely implemented. Imagine an undergraduate student, X, who plans to replicate an experiment from similar studies A, B, and C. After calculating RV_{Cn} for each, they find A > B > C, indicating A has the highest value. However, if X is most interested in C, how could they justify its replication? Prioritizing studies this way may make it harder to defend lower-ranked replications, raising concerns about publication bias. While the metric can help researchers decide how to allocate resources among possible replications, relying too heavily on this metric for evaluating replication studies is undesirable. Hopefully, guidelines for the use of this metric will be developed in the future.

Author Contact

Correspondence concerning this article should be addressed to Yuki Yamada. Email: yamadayuk@gmail.com
ORCID - Yuki yamada: 0000-0003-1431-568X
ORCID - Kaito Takashima: 0000-0002-7489-5682

Conflict of Interest and Funding

The authors declare no conflict of interest. K.T. was funded by Support for Pioneering Research Initiated by the Next Generation: SPRING (grant number: JPMJSP2136). However, views and opinions expressed are those of the authors only and do not necessarily reflect those of the SPRING. The granting authority can not be held responsible for them.

Author Contributions

K.T.: Conceptualization, Funding acquisition, Writing - original draft, and Writing - review & editing.
Y.Y.: Conceptualization, Project administration, Supervision, Writing - original draft, and Writing - review & editing.

Open Science Practices

This article earned the Open Data badge for making the data openly available. The entire editorial process, including the open reviews, is published in the online supplement.

References

Anstis, S. (2022). A pink illusion. *Journal of Illusion*, *3*. https://doi.org/10.47691/joi.v3.8786

Balas, B. (2021). Faces behind bars: Illusory eye movements induced by gratings. *Journal of Illusion*, *2*. https://doi.org/10.47691/joi.v2.8005

Correa da Cunha, H., Farrell, C., Andersson, S., Amal, M., & Floriani, D. E. (2022). Toward a more in-depth measurement of cultural distance: A re-evaluation of the underlying assumptions. *International Journal of Cross Cultural Management*, *22*(1), 157–188. https://doi.org/10.1177/14705958221089192

Howell, J. L., Giuliano, T. A., & Hebl, W. I. (2024). Successfully publishing with undergraduate coauthors in psychology: Insights from faculty with top track records. *Collabra: Psychology*, *10*(1). https://doi.org/10.1525/collabra.94261

Isager, P. M. (2020). Test validity defined as d-connection between target and measured attribute: Expanding the causal definition of borsboom et al. (2004). https://doi.org/10.31234/osf.io/btgsr

Isager, P. M., van Aert, R. C. M., Bahník, S., Brandt, M. J., DeSoto, K. A., Giner-Sorolla, R., Krueger, J. I., Perugini, M., Ropovik, I., van 't Veer, A. E., Vranka, M., & Lakens, D. (2023). Deciding what to replicate: A decision model for replication study selection under resource and knowledge constraints. *Psychological Methods*, *28*(2), 438–451. https://doi.org/10.1037/met0000438

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

Sugihara, K. (2023). Five types of anomalous perceptions created by the same mirror-reflection process. *Journal of Illusion*, *4*. https://doi.org/10.47691/joi.v4.8993