Exploring a formal approach to selecting studies for replication: a feasibility study in social neuroscience

Isager, P.M.; Lakens, D.; Leeuwen, T. van; Veer, A.E. van 't

CORTEX 171 (2024) 330–346

Available online at www.sciencedirect.com

ScienceDirect

Journal homepage: www.elsevier.com/locate/cortex

Special Issue "Strengthening Derivation Chains in Cognitive Neuroscience": Exploratory Report

Exploring a formal approach to selecting studies for replication: A feasibility study in social neuroscience

Peder M. Isager [a], *Daniël Lakens* [b], *Thed van Leeuwen* [c] *and Anna E. van 't Veer* [d,*]

[a] *Department of Psychology, Oslo New University College, Norway*
[b] *Department of Industrial Engineering & Innovation Sciences, Eindhoven University of Technology, the Netherlands*
[c] *Centre for Science and Technology Studies, Leiden University, the Netherlands*
[d] *Methodology and Statistics Unit, Institute of Psychology, Leiden University, the Netherlands*

ARTICLE INFO

Article history:
Received 28 February 2023
Reviewed 30 May 2023
Revised 12 September 2023
Accepted 2 October 2023
Action editor Robert D. McIntosh
Published online 7 November 2023

Keywords:
Replication value
Replication
Social neuroscience
Bibliometric analysis
Expected utility
Exploratory report

ABSTRACT

Replication of published results is crucial for ensuring the robustness and self-correction of research, yet replications are scarce in many fields. Replicating researchers will therefore often have to decide which of several relevant candidates to target for replication. Formal strategies for efficient study selection have been proposed, but none have been explored for practical feasibility – a prerequisite for validation. Here we move one step closer to efficient replication study selection by exploring the feasibility of a particular selection strategy that estimates *replication value* as a function of citation impact and sample size (Isager, van 't Veer, & Lakens, 2021). We tested our strategy on a sample of fMRI studies in social neuroscience. We first report our efforts to generate a representative candidate set of replication targets. We then explore the feasibility and reliability of estimating replication value for the targets in our set, resulting in a dataset of 1358 studies ranked on their value of prioritising them for replication. In addition, we carefully examine possible measures, test auxiliary assumptions, and identify boundary conditions of measuring value and uncertainty. We end our report by discussing how future validation studies might be designed. Our study demonstrates the importance of investigating how to implement study selection strategies in practice. Our sample and study design can be extended to explore the feasibility of other formal study selection strategies that have been proposed.

* *Corresponding author*. Leiden University, Institute of Psychology, Methodology and Statistics Unit, Wassenaarseweg 52, 2333 AK Leiden, the Netherlands.
E-mail address: a.e.van.t.veer@fsw.leidenuniv.nl (A.E. van 't Veer).
https://doi.org/10.1016/j.cortex.2023.10.012

1. Introduction

Close replication of original research results is essential to increase our confidence that empirical findings are reliable (LeBel, McCarthy, Earp, Elson, & Vanpaemel, 2018; Schmidt, 2009). When a study procedure is repeated in a new sample with high similarity to the original (preferably by a novel research team), spurious data patterns can be detected and discarded. Close replication reduces research waste by preventing researchers from building on seemingly promising findings that are not true. Various formal strategies for finding the studies that need replication the most have been developed in recent years (Field, Hoekstra, Bringmann, & Van Ravenzwaaij, 2019; Isager et al., 2023; Matiasz et al., 2018). If effective, such strategies have great potential to increase the transparency and efficiency of replication study selection. When criteria for study selection are made transparent, it becomes easier to discuss which replication studies are most important to fund, conduct, and publish. Additionally, formal strategies allow researchers who agree on criteria to more easily identify and coordinate replication of high-value studies in their field. By increasing the efficiency of coordination and resource spending in replication research, formal study selection strategies present a major step forward towards the important goal of making replication part of mainstream research practice.

This is especially important in human behavioral neuroscience. Research in these fields are often vulnerable to inflated false positive rates and overestimation of effect sizes due to a combination of (1) low statistical power (Szucs & Ioannidis, 2017), (2) substantial researcher degrees of freedom that inflate the Type 1 error rate (Botvinik-Nezer et al., 2020; Carp, 2012), and (3) incentives to publish statistically significant results (Button et al., 2013). Unsurprisingly, rates of successful replications are low (Boekel et al., 2015). In spite of this, close replications of original studies are not common practice (Ashar et al., 2021; Huber, Potter, & Huszar, 2019; Poldrack et al., 2017). While the evidence cited deals primarily with cognitive neuroscience, we believe these issues generalize to most areas of neuroscience that utilize imaging techniques to study the neural correlates of human behavior. At the same time, the cost of data collection in such research is high (Poldrack et al., 2017). This leads to a conundrum. On the one hand, high data collection costs make it all the more important to conduct close replications and prevent costly studies from being built on spurious findings. On the other hand, high costs limit how often replication studies can be conducted. With limited resources and many non-replicated studies to choose from, researchers in social and cognitive neuroscience should consider which studies in the published literature would be the most important to replicate, so that resources are directed towards replication can be spent optimally.

However, no formal study selection strategies have been tested for application in human behavioral neuroscience. To be applicable, a strategy must meet two basic conditions. First, it must be feasible to apply the strategy in practice. That is, the information needed to execute the strategy must be possible to obtain given reasonable time and resource constraints. Most formal study selection strategies are based on a combination of statistical, bibliometric, and substantive information about the candidate studies, which is often not easy to access (Federer et al., 2018; Furukawa, Barbui, Cipriani, Brambilla, & Watanabe, 2006; Glasziou, Meats, Heneghan, & Shepperd, 2008; Sullivan & Feinn, 2012; e.g., Tay, Kramer, & Waltman, 2020). The feasibility of existing strategies for application in any particular area of research is therefore uncertain. Second, provided that the strategy is feasible to apply we must validate that the strategy is actually helping us reach prespecified research goals. All feasible selection strategies lead to *some* prioritization of studies, but whether this prioritization has any validity and practical utility is an empirical question.

In this article we explore how feasible it is to apply a particular replication study selection strategy to fMRI research in social neuroscience, hence no part of the study procedures or analysis plan was preregistered prior to the study being conducted. We focus on a strategy previously developed by the first, second, and last author (Isager, van 't Veer, & Lakens, 2021). This main advantage of this strategy over potential alternatives is that it is (in theory) easy to apply, because it only requires information about the sample size and article citation count of each study that is considered for replication. It should therefore also be possible to apply the strategy to large bodies of research, such as all fMRI studies in social neuroscience. However, the strategy has never been utilized in practice, leaving many questions of practical application open. The goal of this article is therefore to apply the strategy proposed by Isager et al. (2021) in practice, to explore important implementation questions and identify real-world challenges and limitations that are so often overlooked in theoretical analyses.

We focus on fMRI research in social neuroscience because replication studies in this field are both scarce and costly, and because of all methods and areas within the neurosciences, this is what the first and last author are the most familiar with. In other words, the field of social fMRI was chosen because we believed it would provide a sensible test context for the study selection strategy. It is not our aim to study the relative need for replication studies in social neuroscience versus other areas of neuroscience. The goal is simply to provide a case study for testing the feasibility of our selection strategy within the realm of human behavioral neuroscience. We reflect on the generalizability of our conclusions to other research areas in the general discussion.

2. A four step approach to select studies for replication

The concept *replication value* is defined in the formal decision model for replication study selection proposed by Isager et al. (2023) that has been developed to select which empirical claims in the scientific literature to replicate. According to this model, the goal of a replication effort is to maximize the *expected utility* of knowledge gained. *Expected utility gain* can be approximated by the *replication value* of the target claim we want to replicate. In this model replication value is a function of the *value* (or importance) of having accurate knowledge about the target claim, and our *uncertainty* about the truth

status of the claim based on available evidence prior to replicating. Research claims that are highly valuable or important, and about which we are highly uncertain, will have a high replication value, and should be prioritized for replication in order to maximize expected utility gain.

Isager et al. (2021), propose a quantitative method for estimating replication value in which value is operationalized as the average yearly citation impact of the article in which a claim is reported, and in which uncertainty is operationalized as the sample size used to investigate the claim. Replication value is then operationalized as the indicator RV_{Cn}:

$$RV_{Cn} = \text{value} \times \text{uncertainty} = \frac{w(C_S)}{Y+1} \times \frac{1}{\sqrt{n}}$$

where RV_{Cn} denotes a particular operationalization of replication value, C stands for citation impact, n stands for the total number of participants included in the study, $w(C_S)$ stands for the weighting function that should be applied to the citation impact (such as removing self-citations or not), s denotes the source the citation data is retrieved from, and Y stands for the age of the article in years. The equation assumes that average yearly citation impact is causally influenced by scientific impact, and that scientific impact partly determines the value of a claim. It also assumes that sample size partly determines the standard error of estimates in a study, which in turn partly determines the uncertainty about claims studied. Although both the average citation per year and the sample size are imperfect measures of value and uncertainty, our auxiliary assumption is that they are sufficiently correlated with value and uncertainty to generate a useful initial rank order of replication value.

RV_{Cn} is embedded in a four-step procedure for replication study selection based on RV_{Cn} (see Fig. 1). In the first step a set of candidate studies is identified based on the research interests and resource constraints of the replicating researcher. As with every systematic review of the literature, the scope needs to be broad enough to encompass all claims of interest to the researchers, but narrow enough so that the review process remains feasible. In the second step RV_{Cn} is calculated for each study included in the set to create an initial estimate of rank-order expected utility gain. In the third step a subset of the studies with the highest RV_{Cn} is inspected in-depth by reading the article. This step functions as an additional check of the RV_{Cn} estimates, and has as the primary goal to evaluate additional factors relevant to replication value (e.g., Field et al., 2019; Heirene, 2021; KNAW, 2018). In this step researchers can also evaluate the feasibility of a replication study given the resources they have available, and the extent to which a replication study will be able to reduce uncertainty about the effect (Isager et al., 2023). Finally, in this step researchers can check if the article is cited for its empirical claim, and remove replication candidates if the article is cited for other reasons (e.g., the use of a new method, or proposing a new theoretical idea). In the fourth step the candidate deemed most worthwhile to replicate is selected. Alternatively, if the researcher thinks the subset of studies that has been inspected contains no candidate that is worth replicating or feasible to replicate, step 3 and 4 can be repeated for a second subset of studies. We recommend that researchers register their literature search (e.g., using PROSPERO), as well as the replication value formula they will use, and specify as well as they are able any selection criteria for the manual screening phase. This should prevent concerns about the opportunistic use of inclusion criteria to end up with a desired set of studies with a high replication value.

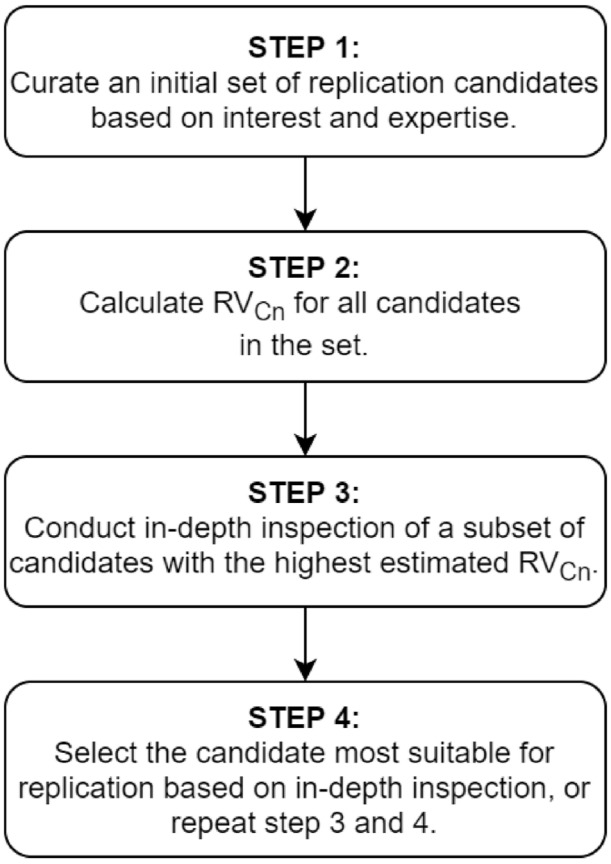

Fig. 1 – General study selection procedure in which the RV_{Cn} indicator is implemented.

3. The current study – exploring the feasibility of using RV_{Cn} for study selection in social fMRI research

RV_{Cn} was developed to enable more efficient coordination of replication efforts. However, it is not clear whether it is feasible to use RV_{Cn} in practice for study selection in a research area such as social fMRI research. Our exploration focuses on the first two steps of the four-step procedure listed in Fig. 1. We report the results of our attempt to implement these steps in practice, including our method for collecting a sample set of replication candidates (step 1), and more importantly, our method for collecting the citation impact and sample size data necessary to calculate RV_{Cn}, the reliability of our methods for generating accurate measures of citation counts and sample sizes, and the distribution of RV_{Cn} for our set of candidates (step 2). In supplementary materials we also summarize our unsuccessful pilot efforts to collect additional quantitative information related to the *main finding* for each candidate

studies in our set. Where a main finding could often be identified based on the abstract, it proved too difficult to identify which statistical test was the basis of this main finding. Finally, we also provide a brief qualitative evaluation of the recommendations produced by RV_{Cn} to better understand what sort of studies are being recommended, what the boundary conditions of this study selection strategy are, and to understand the factors one might want to evaluate when implementing step 3 and 4. We report how we determined our sample size, all data exclusions (if any), all inclusion/exclusion criteria, whether inclusion/exclusion criteria were established prior to data analysis, all manipulations (there were none), and all measures in the study. We conclude the article by generating hypotheses for studies that could be undertaken to test the validity of RV_{Cn}.

3.1. *Step 1 – determining an initial set of candidate studies*

3.1.1. *Eligibility criteria*

To test the feasibility of calculating RV_{Cn} we first set out to determine a suitable set of candidate articles. This step is similar to any systematic literature review (e.g., a meta-analysis). We restricted our search for studies to fMRI research within social neuroscience between 2009 and 2019 at the time this decision was made. Although there is no need to restrict study selection to a specific time period, we reasoned that researchers might be especially interested in conducting replications of studies within a relatively recent time window to prevent unproductive follow-up research (when the original research is non-replicable) or stimulate follow-up research (when the original research is replicable).

3.1.2. *Search strategy*

We used the Web of Science (WoS; www.webofknowledge.com) database to construct our candidate dataset. WoS does not have a predefined field category for social neuroscience. To identify articles related to social neuroscience, we implemented a two-pronged search strategy on 2019-02-21. We first identified four journals in the WoS database as social neuroscience journals (Social Cognitive and Affective Neuroscience; Social Neuroscience; Behavioral Neuroscience; and Socioaffective Neuroscience Psychology). Empirical articles published in these journals were identified by submitting the following search term to WoS:

(SO=(social neuroscience OR social cognitive and affective neuroscience OR behavioral neuroscience OR socioaffective neuroscience psychology) AND PY=(2019 OR 2009 OR 2018 OR 2017 OR 2016 OR 2015 OR 2014 OR 2013 OR 2012 OR 2011 OR 2010)) AND DOCUMENT TYPES: (Article) Timespan: 2009–2019. Indexes: SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, ESCI.

To identify social neuroscience articles in general topic journals we searched the entire WoS database for articles containing the keywords "social" and "fMRI" in all fields. Empirical articles containing the relevant keyword information were identified by submitting the following search term to WoS:

ALL FIELDS: (fmri AND social) Refined by: DOCUMENT TYPES: (ARTICLE) Timespan: 2009–2019. Indexes: SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, ESCI.

3.1.3. *Selection process*

The two search strategies yielded overlapping results. After removing duplicate records, the two search strategies yielded 7413 unique empirical articles in total (see Fig. 2). Basic bibliometric information about each article, including author-provided keywords, were downloaded for all articles.

Authors PMI and AvtV reviewed the initial set of articles and excluded articles they did not believe would be feasible to replicate given their expertise and available resources, which meant excluding animal model research, highly invasive study designs, imaging methods outside our area of expertise, research on patient groups, and other keywords signaling the study would require highly specific samples, procedures, or technologies to perform. At this stage, exclusion criteria were not predetermined, but were exploratorily derived through inspecting keyword information in our initial candidate set. Note that if future replicators want to apply this step of selecting a candidate set it can be done in a number of different ways depending on their specific interest or expertise. For our decision rationale for each excluded keyword, a written record is made openly available on OSF (https://osf.io/mtx72/). Our final set of candidates contained 2268 empirical articles.

3.1.4. *Exploration of sample representativeness*

Once the final set of candidate records was determined, we explored the available bibliographic information to ensure that the sample indeed consisted of the field of studies using fMRI in social neuroscience. The full dataset, including all bibliometric variables and a variable codebook, are available on OSF (https://osf.io/f7zdq/). The articles included in our dataset were published in 329 unique journals, consistent with our expectation that social neuroscience is a broad and loosely connected discipline of researchers from many subfields, who publish in a variety of specialty- and general-topic journals. Table 1 displays the name and frequency of the 20 journals most frequently published in (70.99% of all articles in the set were published in these 20 journals).

We used the statistical visualization software VOSviewer (van Eck & Waltman, 2010) to extract commonly mentioned terms from the titles and abstracts of all studies. Additional analyses of keywords retrieved from the Centre for Science and Technology Studies (CWTS, https://www.cwts.nl/) are reported in supplementary material SM1. All data included in the initial candidate set were subjected to analysis in VOSviewer (co-occurrence map with parameters set to binary counting, minimum number of occurrences set to 15, maximum number of keywords set to 200. Age-related and generic terms were excluded. The list of excluded keywords and map files to recreate the reported co-occurrence map can be found on OSF: https://osf.io/f7zdq/). Fig. 3 displays the co-occurrence map between commonly mentioned keywords in our dataset.

The VOSviewer co-occurrence map corroborates that themes commonly studied in social neuroscience frequently co-occur in the titles and abstracts of articles in our data. Further, the analysis shows that individual topics could be organized into larger categories based on keyword co-occurrence clusters [represented as keyword colors in Fig. 3; van Eck and Waltman (2014)]. As expected from a set of

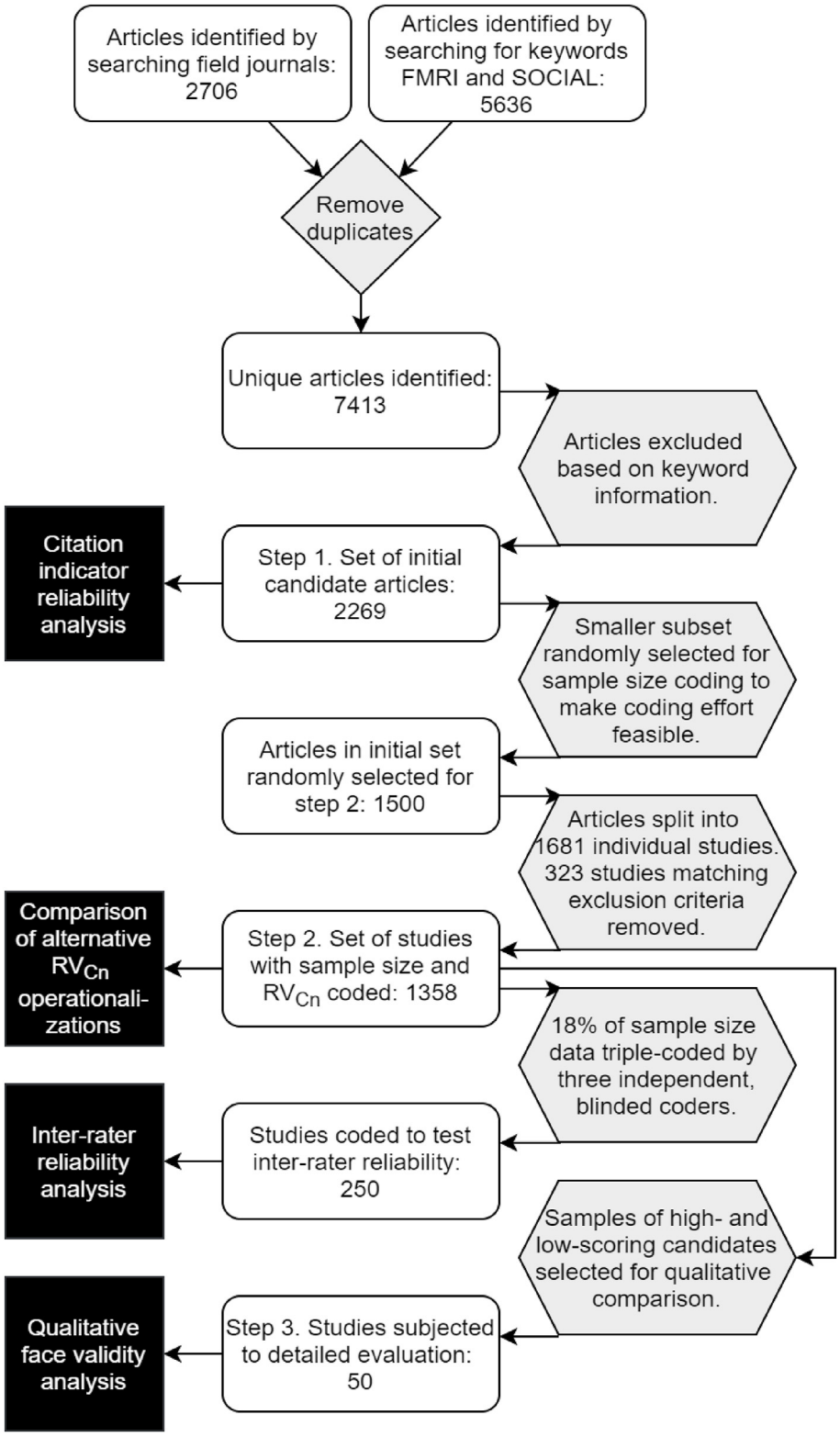

Fig. 2 – Overview of candidate selection process and data points available for each respective analysis reported below.

articles sampled from social neuroscience, these categories center around themes such as face perception (purple cluster), judgment and decision-making (green cluster), language (red cluster), and social pain/ostracism/exclusion (blue cluster). The default mode network (yellow cluster) also has clear ties to social neuroscience research (Li, Mai, & Liu, 2014).

Converging lines of evidence suggest that our search strategy and selection process was successful in curating a

Table 1 – Journals which the articles in our initial candidate set were most frequently published in.

Journal	Frequency
Social Cognitive and Affective Neuroscience	324
Neuroimage	236
Frontiers in Human Neuroscience	115
PLOS One	112
Human Brain Mapping	109
Social Neuroscience	109
Journal of Neuroscience	80
Journal of Cognitive Neuroscience	78
Neuropsychologia	77
Cerebral Cortex	63
Scientific Reports	63
Frontiers in Psychology	51
PNAS	34
Cognitive Affective & Behavioral Neuroscience	30
Cortex	25
Frontiers in Behavioral Neuroscience	23
Brain Research	22
Experimental Brain Research	22
Brain and Language	19
Developmental Cognitive Neuroscience	18

dataset both representative of, and exclusive to our target population of healthy human social fMRI research. Note that our sampling and selection process was largely constructed to overcome the problem that social fMRI is not a well-defined bibliometric category. Determining an initial set of candidates will likely be more straightforward when the field of interest aligns more closely with a well-defined bibliometric category (e.g., a WoS field category) or search terms related to more narrowly defined researcher interest and/or expertise.

We subsequently set out to quantitatively estimate the replication value for each study in this set (see Fig. 1, step 2). Following Isager et al. (2021) we chose RV_{Cn} as our operationalization of replication value (see equation under Section 2). To quantify the replication value, researchers need to specify what function w should be used to weigh the citations, which type of citation impact C is used, as well as source S of that citation impact, if multiple sources are available. In the sections below we explain how citation impact and sample size data were collected in practice, and we explore the reliability of the collected data.

3.2. *Step 2 - calculating RV*

3.2.1. *Operationalizing value as citation impact*

To explore the impact of choosing one specification over another, we studied the reliability of citation impact estimates across a range of impact types C, sources S, and functions w. Although changes to these values will immediately impact the absolute replication value that is calculated, we are mainly interested in their impact on the relative ranking of studies in terms of replication value. Two qualitatively different types of citation impact C were collected; traditional academic citation indexes and Altmetric attention scores. Altmetric attention scores were collected using the *rAltmetric* package in R [Ram (2017); download date: 2020-10-30]. Altmetric attention scores are a weighted count of news- and social-media attention an article has received. For traditional citation impact, we collected data from multiple sources, including WoS (collected 2020-11-07 using the WoS web interface), Crossref [collected 2020-10-30 using the rCrossref package in R; Chamberlain, Zhu, Jahn, Boettiger, & Ram (2020)], Scopus [collected 2020-10-30 using the rScopus package in R;

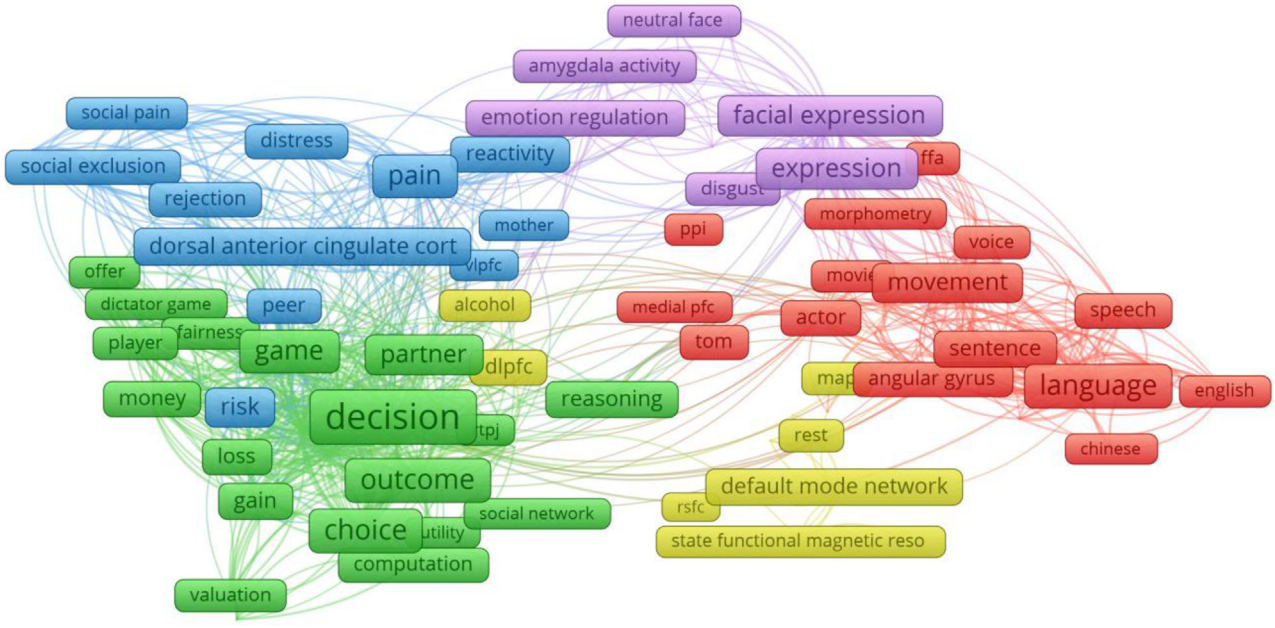

Fig. 3 – VOSviewer co-occurrence map of substantive keywords retrieved from the title and abstract of articles in our dataset. Colors represent VOSviewer-defined clusters of closely related keywords. See van Eck and Waltman (2014) for further details on clustering in VOSviewer. Online interactive version of the figure: https://bit.ly/3yDPMup.

Muschelli (2019)], CWTS (collected 2020-10-28 from the CWTS database by author TvL), and scite™ (www.scite.ai; obtained 2021-08-23 by scite™ staff on request). WoS, Crossref, Scopus and scite™ citation counts are all unweighted raw counts of incoming citations of an article. CWTS citation counts consist only of incoming citations that are not self-citations. We also collected field- and age-normalized citation counts from the CWTS database. This normalization process corrects for differences between subfields in how often papers are cited on average, with the aim to treat publications from different fields equally (for details about the normalization procedure, and a discussion of the use of arithmetic averages in skewed distributions, see Waltman, van Eck, van Leeuwen, Visser, & van Raan, 2011). The score represents how many more times the article is cited relative to the average citation count of an article in its field from the same year. Thus, our data contained three different functions w of traditional citation impact (raw count, self-citations subtracted, and field/age-normalized). Publication year data Y was collected from the WoS database.

3.2.2. *Operationalizing uncertainty as sample size*

Following the rationale of Isager et al. (2021) we operationalized the uncertainty about a claim before replication in terms of the standard error of effects supporting the claim. The standard error can be computed based on the standard deviation and the sample size, which is a combination of the number of participants and the number of observations per participant. We originally aimed to collect multiple sources of information that are relevant to quantifying the uncertainty such as information about the statistical test and the test results (e.g., the standard deviation of the dependent variable), the experimental design (e.g., the number of trials), the number of existing replications, etc., as such information can be used to compute and evaluate alternative operationalizations of replication value. This information would allow us to compare estimates from the RV_{Cn} indicator with other proposed indicators of replication value (e.g., Field et al., 2019, which requires information about bayes factors). We performed two pilot studies to 1) identify additional information that could be coded to quantify uncertainty, and 2) examine if this information could be efficiently coded (see supplementary materials SM2 and SM3, respectively). From these pilot studies we concluded that it was possible to identify the main claim in a paper (which was often feasible based on the abstract), but that it was not feasible to identify the results of the statistical test that provided empirical support for the main claim. It was often not possible to identify which of many statistical tests authors reported were the basis of the main claim. The main reason for this difficulty was the fact that the verbally stated hypotheses were often too ambiguously connected to the reported statistical tests, making it difficult to identify which statistical results would corroborate or falsify the main claim of the paper. This problem is frequently experienced when coding claims and the corresponding tests from the scientific literature (Edelsbrunner & Thurn, 2020; e.g., Scheel, 2022). Furthermore, statistical results were often not reported in sufficient detail to extract information (e.g., about the standard deviation of the measure). We concluded that it would not be feasible to collect additional information related to the uncertainty of the claim on a large scale from the social fMRI literature. In the end, the number of participants was the only operationalization of uncertainty we were able to move forward with in this study. This is an approximation of uncertainty that ignores variation in standard deviations, and the number of trials in a study. In addition, we did not identify replication studies of the studies in our candidate set. It is reasonable to assume that some studies in our set have been replicated. Replication studies should normally reduce our uncertainty about a claim, and methods for incorporating replication information in the RV_{Cn} estimate have been developed (Isager et al., 2021, Supplementary material 1). However, because original and replication studies are not systematically connected in the bibliometric record, and because we believe replication studies to be quite rare within social fMRI research anyway, we elected not to code such information in this study.

3.2.3. *Collecting and inspecting the reliability of* RV_{Cn} *input*

3.2.3.1. Reliability of citation impact across sources. To better understand the relationship between different variables related to the citation impact C across sources S, we explored the strength of the association between a variety of citation metrics (Table 2).

All metrics were retrieved within a time span of two weeks to prevent differences due to a time-lag. Fig. 4 displays the distributions of all citation metrics. All metrics are heavily right skewed. The distributions of raw citation counts are highly overlapping across sources (Fig. 4A). CWTS citation counts are more heavily skewed towards zero than raw counts from other metrics, likely due to the fact that CWTS subtracts self-citations from the total citation count.

To examine how strongly WoS, Crossref, Scopus, CWTS, and scite™ were correlated measures of the same underlying construct - the raw academic citation impact of an article – we subjected the citation data from these sources to an intraclass correlation analysis [model = two-way fixed effects, type = single rater, definition = consistency; Koo and Li (2016)] using the ICC function in the R package *psych* [Revelle (2021); ICC3 output reported]. Because all citation metrics have a skewed distribution, and because we are primarily concerned with the rank-ordering of studies we retrieved citation metrics for (Isager et al., 2021) Spearman's rho correlation was used to assess the strength of association.

Fig. 5 displays the rank-order correlations between various citation metrics. The correlation between raw citation counts from any two sources was very high (always >.94). The inter-rater reliability between these metrics was similarly high, ICC = .97, CI 95% [.96, .97]. When self-citations are subtracted, as is done in the CWTS citation counts, correlations are only slightly lower compared to intercorrelations between the other sources, suggesting that self-citations will not have a large impact on the computation of a replication value.

As expected based on the prior literature (Costas, Zahedi, & Wouters, 2015) the correlations between Altmetric scores and all other metrics were consistently low. The correlation between normalized and non-normalized citation counts was consistently high across sources, though substantially lower than the inter-correlation between different raw citation counts. As will be discussed in more detail below, this

Table 2 – Frequency of various citation metrics available for our data. Web of Science citation counts were originally available for all articles, but some could not be retrieved when the citation count data was updated in 2020.

Citation metric	Description	N
WoS	Web of Science Core Collection Times Cited Count	2105
Crossref	Crossref citation counts	2253
Scopus	Scopus citation counts	2238
CWTS	CWTS citation counts – excluding self-citations	2220
CWTS normalized	Total Field-Normalized Citation Score. CWTS citation impact of article relative to the primary field to which the article belongs.	2220
scite	The total scite citation count from publication until year 2020.	2091
Altmetric	Altmetric score	1874
Total	Number of articles for which all citation metrics were available	1590

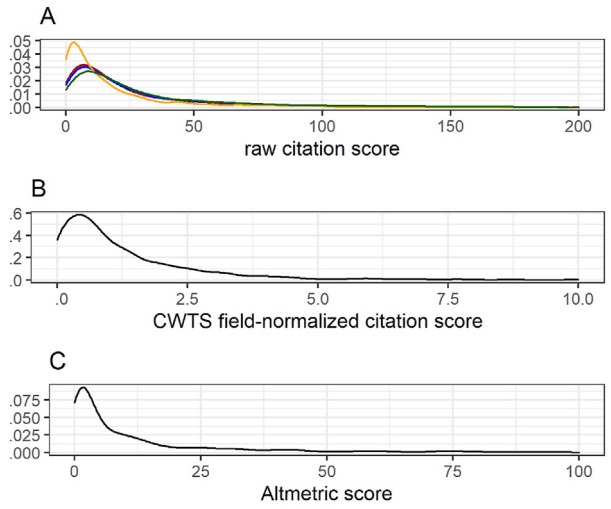

Fig. 4 – Density distribution of citation metrics up to 200 citations. A) The distribution of raw citation counts from Web of Science (black), Crossref (red), Scopus (blue) and CWTS (orange). B) The distribution of CWTS citation impact up to a score of 10, normalized by research field/cluster. C) The distribution of Altmetric attention scores up to 100.

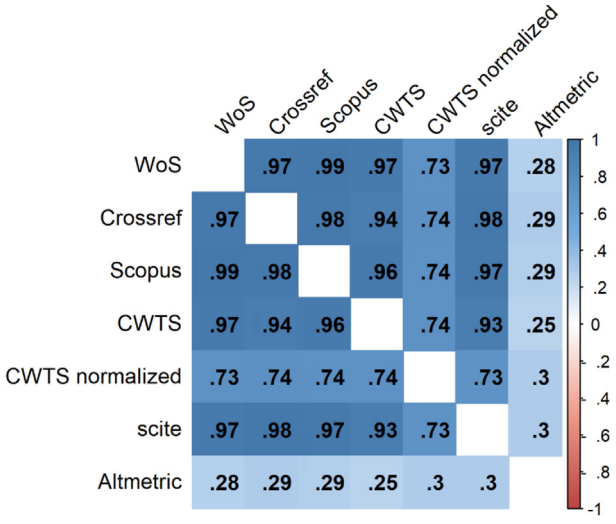

Fig. 5 – Matrix of bi-variate correlations between the citation metrics available for the articles in our dataset.

suggests that it matters little for RV_{Cn} estimates which source S is used, but it does matter whether one chooses raw or field-normalized citation count as the operationalization of $w(C)$, and it would matter substantially whether one chooses to use traditional citation count or news/social-media impact as the operationalization of C. The reliability of Altmetric attention scores as estimates of news/social-media impact remains unclear, as we had no other metrics for this kind of impact to compare against. We will examine the consequences of using Altmetric scores or field-normalized citation counts on the computation of replication value scores below.

3.2.3.2. Accuracy and unbiasedness of average yearly citation count. The ideal citation metric of RV_{Cn} is the number of *future* citations an article will receive (Isager et al., 2021). Total citation count is not a useful estimator of future citation impact because citations accumulate over time. As an article gets older it will tend to get a higher total citation count. This could mean that a 50 year old article cited once per year has the same total citations as an article published last year that has been cited 50 times, even though we should expect the latter to have much more impact on the field in the future. To prevent age from impacting the replication value of articles, RV_{Cn} uses the average yearly citation count instead of the total count as an operational measure of value.

To examine how well average yearly citation count predicts future citation count we obtained the yearly citation rate for each year separately from scite™, including the citation counts for 2020. Then, with the average yearly citation count of each article from all years until 2019, we predicted the citation rate of each article in our data for year 2020 (the last complete year in the data from scite™). To examine whether average yearly citation count is a sufficient approach to correct for the effect of age on citation counts we examined the correlation between age and average yearly citation count. In addition, we explored the relationship between age-averaged citation count and age/field-normalized CWTS citation count, which are age-adjusted using the superior method of normalizing the citation count against all articles from the same year. If age-averaging is an effective method for age adjustment, age-averaged citation count should correlate more strongly with CWTS normalized scores than raw citation count. Finally, we also examined the effect of age-averaging on Altmetric attention scores. Our goal in examining the relationship between these variables is to gain a better

understanding of which data should be used to quantify the value of a published study.

We focus on scite™ citation count data in these analyses since it was the only source from which we could obtain data on yearly citation rate. However, the reported pattern of results is highly similar regardless of which citation source is used (see supplementary material SM4).

3.2.3.2.1. Predictive accuracy. Fig. 6A displays the scite™ citation rate trajectory for all articles in our data. Fig. 6B displays the same trajectories on a log+1 scale with box plots summarizing the distribution for each year since publication, which gives a better sense of the overall trend. On average, most articles seem to be cited at an increasing rate for about the first two years since publication. Then the citation rate stabilizes, possibly increasing slightly around year ten. Given this general trend, our auxiliary assumption that average yearly citation count is on average a useful predictor of future citation impact is supported. Including citations from the two first years seems to lead to an underestimation of the citation rate in later years, but this might not directly affect any rank-order of citation counts.

Fig. 6C displays the accuracy of average yearly citation count (using data until 2019) to predict the "future" citation count in 2020. Predictive accuracy is quite good, but far from perfect, $\rho = .75$, CI 95%[.72, .77]. As noted above, average yearly citation count consistently underestimates how many citations are obtained in 2020. The two first years since publication are included in the average yearly citation count, which tends to drag down the average. Also as expected, underestimation of citations in 2020 seem to be particularly severe for more recently published articles (more yellow dots above the line). The younger the article, the more its average yearly citation count is influenced by the relatively lower number of yearly citations in the two first years since publication. Because total citation counts obtained from scite™ were highly correlated with total citation count obtained from other sources, we believe the results reported here likely generalize to citations from WoS, Crossref, Scopus, and CWTS. The results suggest that the predictive accuracy of RV_{Cn} could be improved by excluding citations from the first two years since publication. Alternatively, accuracy could be improved through more accurate modeling of each article's citation trend. Such improvements require data on citations per year, which is not easily accessible to most researchers [the information was provided to us by scite™ (www.scite.ai)].

3.2.3.2.2. Predictive unbiasedness. Article age was very weakly correlated with the number of scite™ citations an article received from january to december of the year 2020, $\rho =$ .07, CI 95%[.02, .11], suggesting article age is not a determinant of future citation impact and can safely be corrected for. To examine how well age-averaging corrects citation estimates for age, we computed pairwise spearman correlations between publication age, scite™ citation count, Altmetric scores, scite™ citation count divided by years since publication, Altmetric scores divided by years since publication, and CWTS normalized citation count.

Fig. 7 displays the correlation coefficients between all variables of interest. Not surprisingly, there was a strong correlation between age and raw scite™ citation count, $\rho = .54$, CI 95%[.51, .57]. The correlation between citations and age dropped substantially when citation count was divided by years since publication. However, a small residual correlation

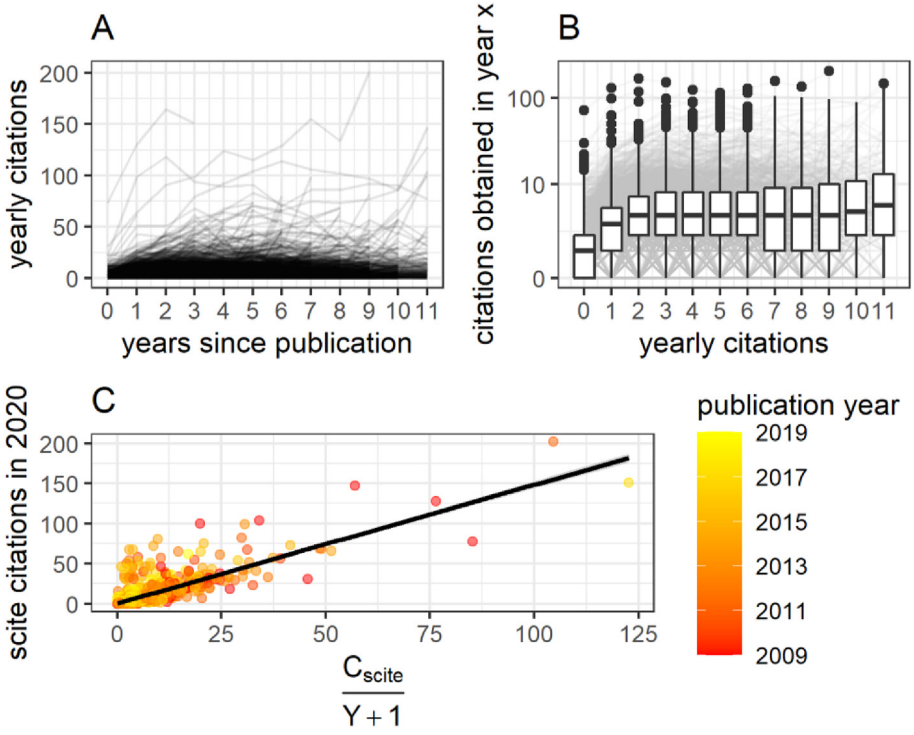

Fig. 6 – A) Citation trajectories for all articles in the dataset. B) Log citation trajectories, with box plot summaries for each year. C) Citations obtained in 2020 predicted by the average yearly citation count from the articles publication year until 2019.

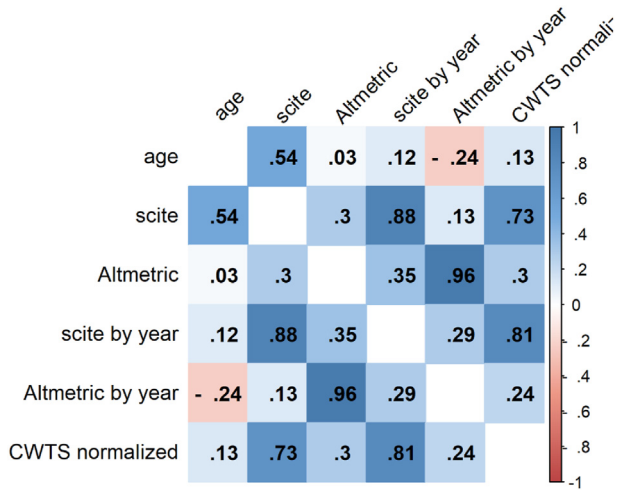

Fig. 7 – Matrix of bi-variate correlations between age and citation indices.

between average yearly citation rate and publication age remains, $\rho = .12$, CI 95%[.07, .16]. This suggests that dividing total citation count by the number of years since publication is an imperfect age adjustment method, but the correction substantially reduces the correlation between age and citation count, and is therefore a substantially better measure than total citation counts. Averaging over age works best if citation time accumulates at a constant rate, but this rate is quite variable for most articles (Fig. 6A). Encouragingly, however, averaging citation count by age does increase the correlation between citation count and CWTS normalized scores, whose method of age correction is superior as it corrects for the average number of citations of all publications published in the same field and in the same year. Interestingly, even CWTS scores are weakly positively correlated with age, suggesting that perfectly adjusting for article age is challenging. In summary, taking the average yearly citation count seems to be an imperfect but efficient method for age adjustment in traditional citation metrics.

3.2.3.3. Coding number of participants. The number of participants for each study in our dataset was coded manually. Manually coding the number of participants for all studies in the full set of 2268 candidate articles was assumed to be costly and time consuming from the outset. In practice, we expect most researchers to have more narrow inclusion criteria when computing the replication value for a set of replication targets. For feasibility reasons, we aimed at coding 1000 articles at random from the full set of 2268 articles and began the process of splitting these into individual studies for coding the number of participants. While coding, it became clear that many studies did not meet our inclusion criteria. To ensure we would end up with at least 1000 articles we oversampled with an additional 500 articles drawn at random from the full set. The exact code used to draw the sample is available on OSF (https://osf.io/rxukq/). After removing articles that matched our initial exclusion criteria (e.g., single non-fMRI studies from multi-study articles, such as De Vries, Fennis, Bijmolt, Ter Horst, & Marsman, 2018, study 4) the number of participants was coded for each fMRI study in the article.

Coding was performed by a team of three undergraduate research assistants. For each article we identified the number of studies reported in the article. For each study we recorded the number of participants who contributed any fMRI data to analyses reported in the study (even if their data were excluded from some analyses). For further details about how coders were instructed to proceed with coding the number of participants, see the supplementary coding instructions (https://osf.io/j3pxf/).

The 1500 articles contained 1681 individual studies, of which 323 matched our exclusion criteria. The final dataset contained 1358 individual studies from 1283 unique articles. Coding time was a few minutes when the number of participants and exclusion criteria were clearly summarized in either the study abstract or the "participants" subsection of the methods section, but could take longer if reporting was less structured. In order to ensure that the number of participants was reliably coded, a subset of 250 studies, randomly selected from the larger set of 1358, were double-coded by independent coders and subjected to an inter-rater reliability analysis. Two additional coders (one additional undergraduate student, the undergraduate coder, and the first author, the PhD coder) re-coded the number of participants for each study in this subset. While coding, all coders were blind to the number of participants provided by other coders. To study inter-rater reliability, we subsequently calculated the percentage agreement between each of the coders, and we calculated the intraclass correlation coefficient between coders (model = one-way fixed effects, type = single rater, definition = absolute agreement) using ICC function in the R package *psych* (ICC3 output reported). Overall, there was a high but imperfect agreement between the three coders (percentage exact agreement = .77). The intraclass correlation coefficient between raters was high, ICC = .82, CI 95%[.79, .86]. Fig. 8 displays the variation in sample size between the coders, plotted on log scale.

Coders disagreed in 57 cases. All disagreements between coders were resolved by the PhD coder after inspecting comments by the other coders. In addition to the cases of disagreements identified in the data used for inter-rater reliability analysis, one additional sample size coding error in the full set of 1358 studies was detected and corrected at a later time during the analyses. Fig. 9 displays the distribution of sample size in our data after resolving coder disagreements (mode = 20, median = 24, frequency of $n <= 10 = 37$, $11–20 = 479$, $21–30 = 365$, $31–40 = 184$, $41–50 = 97$, $51–60 = 60$, $61–70 = 27$, $71–80 = 25$, $81–90 = 10$, $91–100 = 10$, $n > 100 = 64$).

3.2.4. *Calculating and comparing alternative operationalizations of* RV_{Cn}

Having established that sufficiently accurate citation counts and the number of participants can be collected, we proceeded with the calculation of RV_{Cn}. Because replicating researchers may end up relying on any of several citation metrics to estimate value, we decided to compare the results of several alternative operationalizations of replication value; one indicator measured value via the WoS citation count of the articles (RV_{WoS}), one via the Scopus citation count (RV_{Scopus}), one via the field-normalized citation counts (RV_{tncs}), one via the RV_{scite} and one indicator measured value via

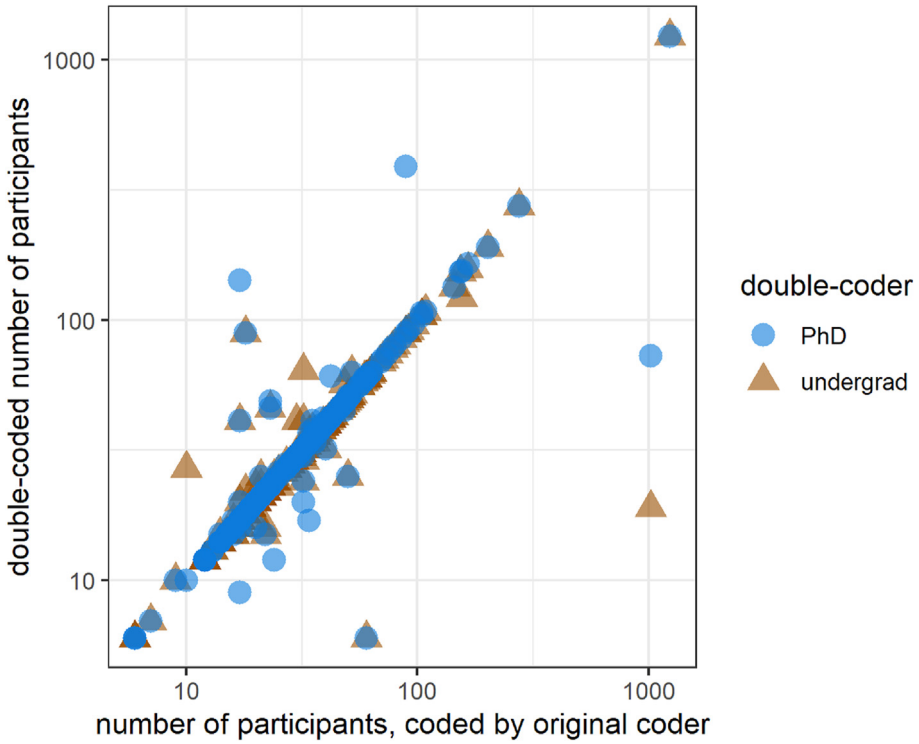

Fig. 8 – Variation in sample size between coders. Sample size is plotted on log scale. The original sample size coded is represented on the x-axis. Double-coded sample size values are represented on the y-axis. Blue circles represent values from the PhD-student coder. Brown triangles represent values from the undergraduate student coder.

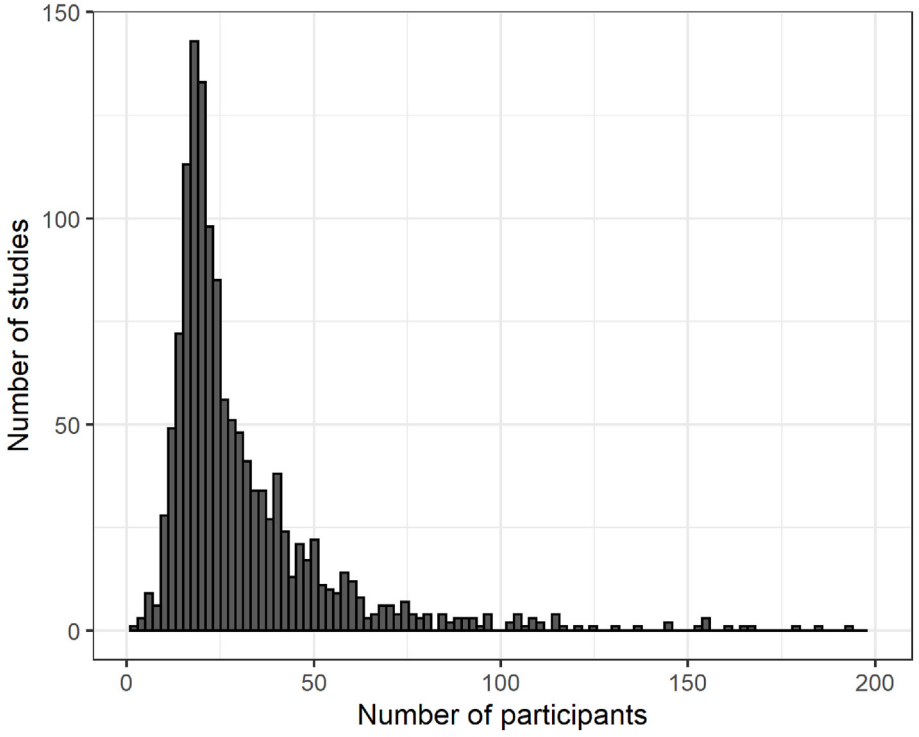

Fig. 9 – Distribution of sample sizes in the dataset. For visualization purposes, the x-axis limit is set to $n = 200$.

Altmetric score of the articles (RV_{Alt}). All indicators used sample size as a measure of uncertainty.

RV_{WoS} was based on the equations derived by Isager et al. (2021), and calculated in the following way:

$$RV_{Wos} = \frac{C_{Wos}}{Y+1} \times \frac{1}{\sqrt{n}}$$

where C_{WoS} denotes the WoS citation count of the article a study is reported in, Y denotes the article age in years, and n denotes the sample size of the study after exclusion. The three measures using Scopus, scite™, and cluster-normalized citation scores were computed in the same way as RV_{Wos}.

RV_{Alt} was calculated in the following way:

$$RV_{Alt} = C_{Alt} \times \frac{1}{\sqrt{n}}$$

where C_{Alt} denotes the Altmetric attention score of the article, and n denotes the sample size of the study after exclusion. Because the analyses above revealed that Altmetric attention scores are not strongly correlated with article age in our data, we did not average C_{Alt} over publication year in this replication value indicator. Many articles are not mentioned in any sources that are tracked by Altmetric, and therefore have a score of 0. In our dataset C_{Alt} could only be calculated for 1156 of 1358 studies.

Importantly, we calculated all replication values under the assumption that no study in our candidate set is a replication of another study in the set, implying that no studies should be combined in the estimate of n. Because lack of replication research in fMRI research (Poldrack et al., 2017) implies that only very few articles in our dataset would be replications of one another, we found it acceptable to proceed with calculation under the assumption that there were no replications in the data. Where direct replication studies have been performed, it would have been more appropriate to combine the sample size from the original study and it's replications (Isager et al., 2021, supplementary material 1). However, there are no databases that store information about direct replication in social neuroscience. Whenever researchers compute the replication value for a more specific population, information about direct replications might be more readily available, or it can be manually searched and coded in step 3.

The distribution of replication value from all indicators was visually inspected, and estimates from indicators were correlated to study their similarity. Spearman's rho was used since the rank-order correlation between different indicators is of primary interest. 95% bootstrap confidence intervals were calculated for the correlation estimate using the spearman.ci function of the RVAideMemoire package in R (Hervé, 2021).

Fig. 10 displays the distribution of RV_{WoS}, RV_{Alt}, RV_{Scopus}, RV_{tncs} (field-normalized citation scores), RV_{scite}, and their associations with RV_{WoS}. Overall, all distributions are highly skewed with most scores distributed around low values, which is expected given that the number of participants, citation counts, and Altmetric attention scores are all highly skewed as well (see Figs. 4 and 9). Overall rank-order correlations were high for different citation sources (WoS, Scopus, scite), lower for field-normalized citation counts, and low for

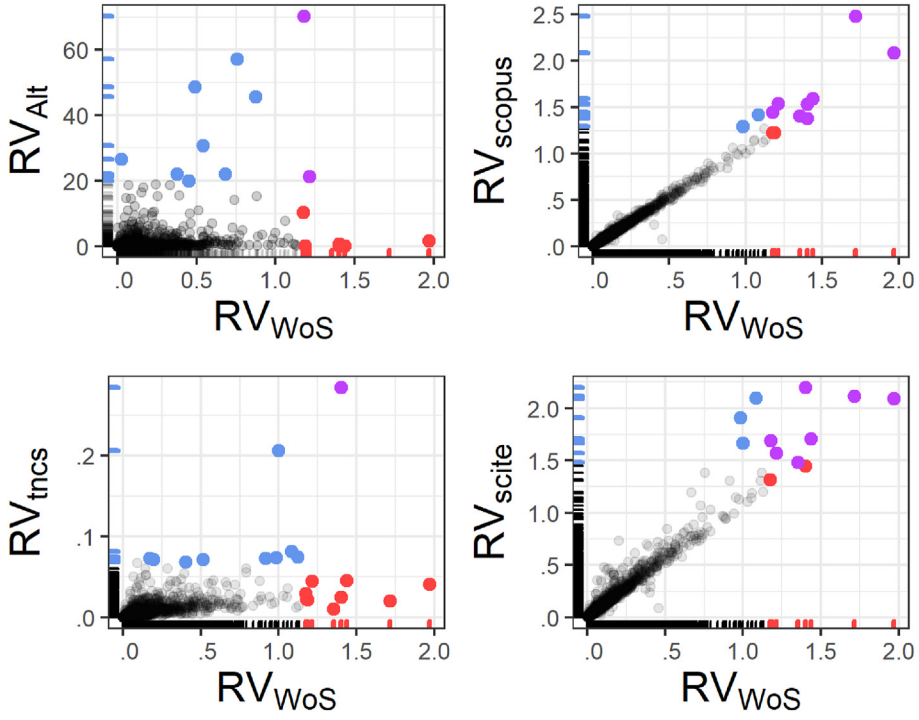

Fig. 10 – Scatter plot visualizing the relationship between RV_{WoS} and RV_{Scopus}, RV_{tncs}, RV_{scite}, and RV_{Alt}. Distribution of RV_{WoS} indicators are visualized as bars on the x-axis. Distribution of the other replication values are visualized as bars on the y-axis. Blue bars (and dots) represent the 10 highest scores on the y-axis. Red bars (and dots) represent the 10 highest RV_{WoS} scores. Purple dots represent scores that are among the 10 highest scores on both estimators. Two of the ten studies with the highest RV_{WoS} scores are not included in the scatter plot with the RV_{Alt} scores, as the RV_{Alt} could not be computed due to missing Altmetric attention scores.

Altmetric scores (see Fig. 11). As a consequence, only two studies (Kassam, Markey, Cherkassky, Loewenstein, & Just, 2013; Tamir & Mitchell, 2012) were ranked among the top ten in both WoS and Altmetric rank-orderings (purple-colored points in Fig. 10). The same was true for field-normalized citation scores, where the overlap between top-ranked studies using WoS citation scores and field-normalized citation scores was very low (despite the relatively high correlations between the two measures). Traditional citation impact and altmetric attention scores are generally thought to measure different aspects of impact and are known to be weakly associated. It is clear field-normalized citation scores also measure impact in a substantially different manner than raw citation counts. The overlap between citation counts from different sources such as scite™ or Scopus does not lead to substantially different selections, even though even there some variation in the last one or two studies included when selecting the X highest ranked studies (e.g., the 9th and 10th study included in a Top 10) should be expected to vary.

To conclude, quantitative recommendations for which studies to replicate will vary substantially based on whether traditional, field-normalized, or altmetric citation impact is used to estimate replication value, because these impact metrics measure non-overlapping aspects of scientific impact. Different stakeholders may prefer either operationalization, depending on what aspects of impact they find most relevant. Altmetric attention scores are only weakly correlated with traditional citation counts, which has a substantial impact on RV_{Cn} estimates.

3.3. *Step 3 – In depth review of recommended candidates*

The next step when selecting a replication target is an in-depth inspection of studies with a high replication value. For our exploratory purposes, we expect such an in-depth review to reveal certain boundary conditions of when the number of participants and/or the citation count do not accurately reflect the value and impact of a study. We subjected the 10 studies with the highest and lowest replication value on either RV_{WoS} or RV_{Alt} to an in-depth inspection. In addition, we included the 10 lowest non-zero estimates from the RV_{WoS} distribution, because RV_{WoS} scores of 0 often simply reflect a paper too young to have picked up citations yet. In total, 44 unique studies were included in our face validity review (6 studies were among the highest or lowest scores for both indicators).

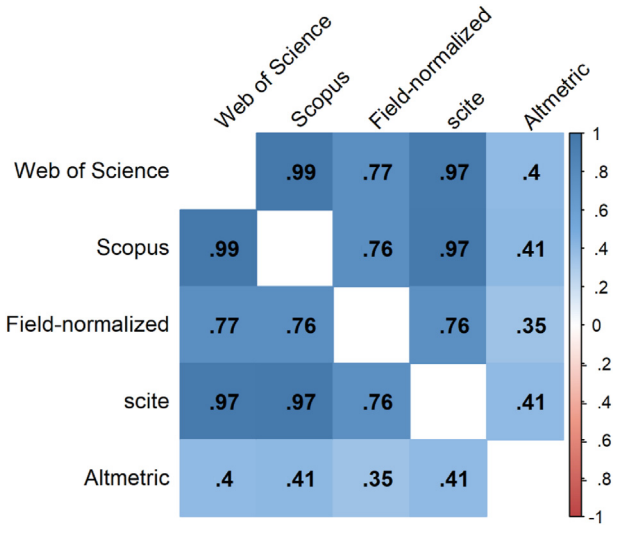

Fig. 11 – Matrix of bi-variate correlations between replication value indices computed based on different operationalizations of value through citations or Altmetrics.

We wanted to see whether quantitative replication value estimates would conform to our own intuitions about replication value, and identify factors that would lead to a high replication value using a formula-based approach, without actually warranting a replication. Such boundary conditions are likely present in other sets of replication targets as well, and identifying such factors will help researchers during the in-depth inspection in step 3. For example, an article may be highly cited for reasons other than the empirical studies it reports, which would lead to a highly cited paper while the study in the article is not worth replicating. As such, the goal is to identify potential issues with validity, reliability and measurement error that future validation studies of RV_{Cn} may want to follow up on.

Authors PMI and AvtV read the title and abstracts of all studies included in the review, consulted the article text intermittently for clarifications, and reviewed quantitative information related to the replication value estimates of these studies (i.e., reviewers were not blinded to a record's rank position). Both reviewers first made notes for each study in private, focusing on their intuitive validity judgment of the replication value estimate and on potential sources of error and bias. Notes were then discussed by PMI, AvtV, and DL in two meetings to distill the most central outcomes of the review effort. The full set of notes is available on OSF for author PMI (https://osf.io/vwpqs/) and AvtV (https://osf.io/953rh/).

3.4. *Central outcomes of the review process*

The in-depth review yielded several insights. A detailed inspection of quantitative replication value estimates is important for quality control. In two studies, coders had erroneously coded an incorrect number of participants (due to a transcription error, and overlooking data exclusions). Eight articles turned out not to be connected to social neuroscience, and one study did not utilize fMRI for imaging. Finally, in one case we had incorrectly labeled a single two-session repeated measures study as two separate studies. Together, these studies make up one quarter of the entire sample selected for review. This clearly indicates that, in this particular context, RV_{Cn} is a noisy measure of replication value, and finding the studies most in need of replication is highly dependent on the third step of the procedure.

There was not always an intuitive correspondence between the RV_{Cn} rank order and our intuitions about the replication value of the claims purely based on the title and information in the abstract. One reason for this lack of correspondence may have been that reviewers were not blind to the replication value ranking, and had access to the citation count and number of participants, which were so salient they were difficult to not take into account. Another reason was that without other

explicit criteria to determine the value of a replication study, there was substantial subjectivity in the value of each study as judged by both reviewers. This is not unexpected, as peer evaluations of the value of a study are variable, and not strongly related to eventual citation scores (Gottfredson, 1978). A final reason for the low perceived correspondence between indicator estimates and reviewer intuitions were a number of boundary conditions where the RV_{Cn} estimates did not accurately reflect the value and uncertainty of the studies.

The first boundary condition was that many studies used within-subject designs, where the number of participants does not fully capture the uncertainty, as it ignores the number of measurements per participant. The use of within-subject designs seemed to be common among the highest ranked studies, as such designs require less participants for high statistical power, and therefore get a higher replication value when uncertainty is based only on the number of participants. This is clearly an important limitation, especially when the number of trials in each study varies substantially between studies (as was the case in the set of studies we examined). In future applications of RV_{Cn}-based study selection we therefore recommend that uncertainty is quantified during step 2 based on both the number of participants, and the number of observations per participant. If this is unfeasible (which is likely given how unsystematically this information is reported in the literature), the number of observations should be taken into account during step 3 (see supplementary material 2 in Isager et al., 2021 for technical details on such a correction method). Alternatively, selecting a narrower set of candidates with homogeneous study designs in step 1 will alleviate this limitation. Another boundary condition concerned a study that already had been replicated in the literature. Although rare, when replication studies already exist, the replication value should be computed based on the uncertainty remaining after all replication studies (Isager et al., 2021).

Other boundary conditions concerned the reason why an article was highly cited. One article containing both a literature review and an empirical study seemed to be cited primarily due to the literature review (Dimoka, Pavlou, & Davis, 2011). Another study on human navigation appeared to receive a large Altmetric score primarily due to speculative news reports claiming that GPS use can "turn the brain off" – even though this conclusion did not follow from the study (Javadi et al., 2017). A replication of the study results would do little to avert such speculations, since the speculations are not grounded in the actual study results. The boundary conditions identified so far seem general enough to incorporate in the in-depth review process of replication targets by default. Future research should give us a better understanding of which additional factors to consider during in-depth review of replication candidates (e.g., Pittelkow et al., 2023).

4. General discussion

The overall aim of this exploratory study was to test the feasibility of implementing the four-step replication study selection procedure based on RV_{Cn} proposed by Isager et al. (2021) in a large body of social fMRI research. The current exploratory report shows the importance of testing the feasibility of proposed selection strategies, as well as carefully examining possible measures, auxiliary assumptions, and boundary conditions. We show it is possible to calculate RV_{Cn} for a large candidate set of studies identified based on bibliometric information. We were able to reliably code the total number of participants and retrieve citation count data for each study in order to calculate RV_{Cn} (step 2 in Fig. 1). However, we were only able to code uncertainty coarsely with 'number of participants in study', omitting the number of trials per participant, which also determines the standard error of the estimate (Westfall, Kenny, & Judd, 2014).

Traditional citation count metrics were highly rank-order correlated, meaning there is little difference in which source S is used in the calculation of RV_{Cn}. Field-normalized citation counts provide a somewhat different measure of citation impact, and lead to less overlap in the final rank-order than non-normalized citation scores, especially in an interdisciplinary research topic such as social neuroscience, where publications appear across scientific fields, which leads to different articles being normalized against different citation cluster averages. Altmetric attention scores are weakly correlated with traditional citation impact, and represent a qualitatively different approach to measuring value. Whichever measure is preferred, both Altmetric scores and traditional citation counts could easily be extracted using free and open source applications (Chamberlain et al., 2020; e.g., Ram, 2017), where field-normalized citation counts or citation counts per year are not publicly available.

Finally, in-depth review of the highest ranking indicator estimates from step 2 appears to be an important method of quality control before a candidate is selected for replication. This review revealed important boundary conditions of using citation counts and the total number of participants as measures of value and uncertainty. Auxiliary hypotheses that we explored, such that past citation counts predict future citation counts, that the source of the citation counts do not substantially affect citation rank-order, and that we can control for the age of the article, were all supported.

Overall, however, we do not think our implementation of RV_{Cn} in the social fMRI literature was successful. Modifications to either the selection procedure or scope are needed for future application in this research area. While it was feasible to reliably code sample size and citation count for over one thousand studies, several challenges hindered efficient implementation. First, the topic boundaries of a research area like "social fMRI research" are fuzzy. Social neuroscience clearly does not include volcanology studies, but it is not trivial (and perhaps not even possible) to define the borders between social neuroscience and related neuroscientific disciplines. This made it very difficult to execute step 1 of the strategy, and in spite of our best efforts to develop reliable inclusion and exclusion procedures, in every review step we discovered a substantial number of studies that should not have been included given our exclusion criteria. Second, it is difficult to say whether "number of participants" is a meaningful indicator of general uncertainty in a candidate set that contains such a wide range of study designs. While it is possible to correct for study design in theory (Isager et al., 2021), this is not possible in practice for such a large set of studies with widely varying within-subject structures. This

reduces the usefulness of step 2, which is the very core of the RV_{Cn} strategy. Third, we believe that, in step 3, more expertise with the study topics under review may be required in order to provide adequate face validation of the candidates ranked highly in step 2. In this study, we wanted a large candidate set, as a primary aim was to test the feasibility of applying step 2 to a large set of studies. However, future researchers aiming to use these or similar steps in selecting a candidate for replication may already start with a more narrow candidate set in step 1, based on their research interest and expertise.

Whether these challenges generalize to application of RV_{Cn} in other disciplines is an open question which will need to be empirically examined. The use of RV_{Cn} might be more straightforward in more homogenous literatures, especially if these mainly rely on between-participant designs. It may also be more feasible to adapt or modify RV_{Cn} to account for variations in study design (Isager et al., 2021, supplementary material 2) in fields where such information is easier to curate from the articles.

The current report provides insights into how RV_{Cn} can be applied in practice, and how its feasibility can be evaluated. However, it doesn't yet provide insight into the validity of RV_{Cn} as a measure of replication value. Future research could attempt to provide criterion validation of RV_{Cn} by investigating whether RV_{Cn} is associated with other operational measures that are hypothesized to predict expected utility gain. For example, we would expect RV_{Cn} to predict which studies are chosen for replication in practice under the assumption that both RV_{Cn} and the selection criteria used by researchers who perform replication studies are caused by the expected utility of the replication effort (Isager et al., 2021). It might also be possible to validate RV_{Cn} by examining the extent to which RV_{Cn} predicts subjective estimates of the relative replication value of a set of studies. Future studies could also aim to increase the understanding of which factors researchers usually consider when selecting a study for replication. Recently, Pittelkow et al. (2023) identified a number of criteria such as interest, doubt, impact, methodology, and feasibility. How feasible it is to include such factors in a formal study selection strategy remains an open question.

We end this article with some recommendations for researchers looking to apply replication study selection strategies. For researchers specifically interested in using RV_{Cn} to identify important-to-replicate fMRI studies in social neuroscience, our study provides some important insights. First, focusing on a relatively well-defined subject within social neuroscience literature, rather than all studies in the discipline, seems wise. For a recent successful implementation, see Zaragoza-Jimenez et al. (2023). Although this will restrict how broadly one can search for replication candidates, it will likely make it much easier to curate a candidate set of studies that includes only studies relevant to one's interests and expertise. Second, since within-subject designs are very common in fMRI studies, the RV_{Cn} uncertainty estimate should ideally be based on the standard deviation in this field. If one elects to use sample size, it should be corrected for the design used (Isager et al., 2021, supplement 2). Be aware, however, that by using the standard deviation to estimate uncertainty one is forced to identify the effect of interest for each study in the candidate set, which will add additional work to the procedure. Taken together, while RV_{Cn} itself can reliably and efficiently be computed for hundreds of studies, the general selection procedure (Fig. 1) seems more suited to a smaller, more homogenous set of studies than what we aimed for in this study.

It may of course also be valuable to study whether other potential selection strategies would work better than RV_{Cn} in social fMRI research. We encourage interested researchers to conduct additional feasibility studies for other proposed strategies.

Finally, some general recommendations can be given to facilitate more efficient replication research in *any* discipline. First, it is important to conduct feasibility studies of a range of study selection strategies in more disciplines. As our study demonstrates, it is not enough to show that a study selection strategy works in theory or in toy examples. If we want replication study selection to be more strategic and efficient, replicating researchers will need clear guidelines for how to implement and adapt strategies in practice. Feasibility studies are needed to develop such practical guidelines. Second, this work again highlights the need to standardize the reporting of study design and statistical uncertainty as much as possible. The task of evaluating the uncertainty in scientific claims becomes easier if researchers adhered to reporting standards, and when the relationship between statistical tests and scientific claims are more clearly specified in the article (Appelbaum et al., 2018; Lakens & DeBruine, 2021). Third, in any replication study, we recommend explicitly stating why the original study was selected for replication (e.g., Pittelkow et al., 2023). By exploring and documenting the wealth of information relevant to replication study selection, we can increase the ability of researchers to make well-informed decisions about which original research would be the most important to replicate.

Funding

This work was funded by VIDI grant 452-17-013 from the Netherlands Organisation for Scientific Research. We want to thank all student assistants from Leiden University and Eindhoven University of Technology who contributed to data collection for this project at one point or another. All data to reproduce this manuscript can be found at https://github.com/pederisager/NeuroRep_RV.

Author contributions

Peder Mortvedt Isager: Conceptualization, Investigation, Visualization, Writing – Original Draft, Writing – Review & Editing; Daniël Lakens: Funding acquisition, Supervision, Writing – Review & Editing; Thed van Leeuwen: Data curation, Investigation, Resources; Anna E. van 't Veer: Conceptualization, Supervision, Investigation, Writing – Review & Editing.

Open practices

The study in this article earned Open Data and Open Material badges for transparent practices. The data and materials used

in this study are available at: https://github.com/pederisager/NeuroRep_RV.

REFERENCES

Appelbaum, M., Cooper, H., Kline, R. B., Mayo-Wilson, E., Nezu, A. M., & Rao, S. M. (2018). Journal article reporting standards for quantitative research in psychology: The APA Publications and Communications Board task force report. *American Psychologist, 73*(1), 3–25. https://doi.org/10.1037/amp0000191

Ashar, Y. K., Clark, J., Gunning, F. M., Goldin, P., Gross, J. J., & Wager, T. D. (2021). Brain markers predicting response to cognitive-behavioral therapy for social anxiety disorder: An independent replication of Whitfield-Gabrieli et al. 2015. *Translational Psychiatry, 11*(1), 260. https://doi.org/10.1038/s41398-021-01366-y

Boekel, W., Wagenmakers, E.-J., Belay, L., Verhagen, J., Brown, S., & Forstmann, B. U. (2015). A purely confirmatory replication study of structural brain-behavior correlations. *Cortex; a Journal Devoted to the Study of the Nervous System and Behavior, 66*, 115–133. https://doi.org/10.1016/j.cortex.2014.11.019

Botvinik-Nezer, R., Holzmeister, F., Camerer, C. F., Dreber, A., Huber, J., Johannesson, M., … Schonberg, T. (2020). Variability in the analysis of a single neuroimaging dataset by many teams. *Nature, 582*(7810), 84–88. https://doi.org/10.1038/s41586-020-2314-9

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience, 14*(5), 365–376. https://doi.org/10.1038/nrn3475

Carp, J. (2012). On the plurality of (methodological) worlds: Estimating the analytic flexibility of FMRI experiments. *Frontiers in Neuroscience, 6*(OCT). https://doi.org/10.3389/fnins.2012.00149, 149–149.

Chamberlain, S., Zhu, H., Jahn, N., Boettiger, C., & Ram, K. (2020). Rcrossref: Client for Various 'Crossref' 'APIs' (1.1.0) [Computer software]. https://cran.r-project.org/web/packages/rcrossref/index.html

Costas, R., Zahedi, Z., & Wouters, P. (2015). Do "altmetrics" correlate with citations? Extensive comparison of altmetric indicators with citations from a multidisciplinary perspective. *Journal of the Association for Information Science and Technology, 66*(10), 2003–2019. https://doi.org/10.1002/asi.23309

De Vries, E. L. E., Fennis, B. M., Bijmolt, T. H. A., Ter Horst, G. J., & Marsman, J.-B. C. (2018). Friends with benefits: Behavioral and fMRI studies on the effect of friendship reminders on self-control for compulsive and non-compulsive buyers. *International Journal of Research in Marketing, 35*(2), 336–358. https://doi.org/10.1016/j.ijresmar.2017.12.004

Dimoka, A., Pavlou, P. A., & Davis, F. D. (2011). NeuroIS: The potential of cognitive neuroscience for information systems research. *Information Systems Research, 22*(4), 687–702. https://doi.org/10.1287/isre.1100.0284

Edelsbrunner, P. A., & Thurn, C. (2020). Improving the utility of non-significant results for educational research [preprint]. *PsyArXiv*. https://doi.org/10.31234/osf.io/j93a2

Federer, L. M., Belter, C. W., Joubert, D. J., Livinski, A., Lu, Y.-L., Snyders, L. N., & Thompson, H. (2018). Data sharing in PLOS ONE: An analysis of data availability statements. *PLoS One, 13*(5), Article e0194768. https://doi.org/10.1371/journal.pone.0194768

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

Furukawa, T. A., Barbui, C., Cipriani, A., Brambilla, P., & Watanabe, N. (2006). Imputing missing standard deviations in meta-analyses can provide accurate results. *Journal of Clinical Epidemiology, 59*(1), 7–10. https://doi.org/10.1016/j.jclinepi.2005.06.006

Glasziou, P., Meats, E., Heneghan, C., & Shepperd, S. (2008). What is missing from descriptions of treatment in trials and reviews? *Bmj: British Medical Journal, 336*(7659), 1472–1474. https://doi.org/10.1136/bmj.39590.732037.47

Gottfredson, S. D. (1978). Evaluating psychological research reports: Dimensions, reliability, and correlates of quality judgments. *American Psychologist, 33*(10), 920–934. https://doi.org/10.1037/0003-066X.33.10.920

Heirene, R. M. (2021). A call for replications of addiction research: Which studies should we replicate and what constitutes a "successful" replication? *Addiction Research & Theory, 29*(2), 89–97. https://doi.org/10.1080/16066359.2020.1751130

Hervé, M. (2021). *RVAideMemoire: Testing and plotting procedures for biostatistics*.

Huber, D. E., Potter, K. W., & Huszar, L. D. (2019). Less "story" and more "reliability" in cognitive neuroscience. *Cortex; a Journal Devoted to the Study of the Nervous System and Behavior, 113*, 347–349. https://doi.org/10.1016/j.cortex.2018.10.030

Isager, P. M., van Aert, R. C. M., Bahník, Š., Brandt, M. J., DeSoto, K. A., Giner-Sorolla, R., … 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. M., van 't Veer, A. E., & Lakens, D. (2021). Replication value as a function of citation impact and sample size. *MetaArXiv*. https://doi.org/10.31222/osf.io/knjea

Javadi, A.-H., Emo, B., Howard, L. R., Zisch, F. E., Yu, Y., Knight, R., … Spiers, H. J. (2017). Hippocampal and prefrontal processing of network topology to simulate the future. *Nature Communications, 8*(1), Article 14652. https://doi.org/10.1038/ncomms14652

Kassam, K. S., Markey, A. R., Cherkassky, V. L., Loewenstein, G., & Just, M. A. (2013). Identifying emotions on the basis of neural activation. *PLoS One, 8*(6), Article e66032. https://doi.org/10.1371/journal.pone.0066032

KNAW. (2018). *Replication studies improving reproducibility in the empirical sciences*. Amsterdam: KNAW.

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine, 15*(2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012

Lakens, D., & DeBruine, L. M. (2021). Improving transparency, falsifiability, and rigor by making hypothesis tests machine-readable. *Advances in Methods and Practices in Psychological Science, 4*(2), Article 2515245920970949. https://doi.org/10.1177/2515245920970949

LeBel, E. P., McCarthy, R. J., Earp, B. D., Elson, M., & Vanpaemel, W. (2018). A unified framework to quantify the credibility of scientific findings. *Advances in Methods and Practices in Psychological Science, 1*(3), 389–402. https://doi.org/10.1177/2515245918787489

Li, W., Mai, X., & Liu, C. (2014). The default mode network and social understanding of others: What do brain connectivity studies tell us. *Frontiers in Human Neuroscience*. https://doi.org/10.3389/fnhum.2014.00074, 0.

Matiasz, N. J., Wood, J., Doshi, P., Speier, W., Beckemeyer, B., Wang, W., … Silva, A. J. (2018). ResearchMaps.org for integrating and planning research. *PLoS One, 13*(5), Article e0195271. https://doi.org/10.1371/journal.pone.0195271

Muschelli, J. (2019). *Rscopus: Scopus database 'API' interface*.

Pittelkow, M.-M., Field, S. M., Isager, P. M., van 't Veer, A. E., Anderson, T., Cole, S. N., … van Ravenzwaaij, D. (2023). The process of replication target selection in psychology: What to

consider? *Royal Society Open Science, 10*(2), Article 210586. https://doi.org/10.1098/rsos.210586

Poldrack, R. A., Baker, C. I., Durnez, J., Gorgolewski, K. J., Matthews, P. M., Munafò, M. R., … Yarkoni, T. (2017). Scanning the horizon: Towards transparent and reproducible neuroimaging research. *Nature Reviews Neuroscience, 18*(2), 115–126. https://doi.org/10.1038/nrn.2016.167

Ram, K. (2017). *rAltmetric: Retrieves altmerics data for any published paper from 'Altmetric.com'*.

Revelle, W. (2021). *Psych: Procedures for psychological, psychometric, and personality research*.

Scheel, A. M. (2022). Why most psychological research findings are not even wrong. *Infant and Child Development, 31*(1). https://doi.org/10.1002/icd.2295

Schmidt, S. (2009). Shall we really do it again? The powerful concept of replication is neglected in the social sciences. *Review of General Psychology, 13*(2), 90–100. https://doi.org/10.1037/a0015108

Sullivan, G. M., & Feinn, R. (2012). Using effect sizeor why the P value is not enough. *Journal of Graduate Medical Education, 4*(3), 279–282. https://doi.org/10.4300/JGME-D-12-00156.1

Szucs, D., & Ioannidis, J. P. A. (2017). Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. *PLoS Biology, 15*(3), Article e2000797. https://doi.org/10.1371/journal.pbio.2000797

Tamir, D. I., & Mitchell, J. P. (2012). Disclosing information about the self is intrinsically rewarding. *Proceedings of the National Academy of Sciences, 109*(21), 8038–8043. https://doi.org/10.1073/pnas.1202129109

Tay, A., Kramer, B., & Waltman, L. (2020). *Why openly available abstracts are important – overview of the current state of affairs*. https://www.leidenmadtrics.nl/articles/why-openly-available-abstracts-are-important-overview-of-the-current-state-of-affairs.

van Eck, N. J., & Waltman, L. (2010). Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics, 2*(84), 523–538. https://doi.org/10.1007/s11192-009-0146-3

van Eck, N. J., & Waltman, L. (2014). Visualizing bibliometric networks. In Y. Ding, R. Rousseau, & D. Wolfram (Eds.), *Measuring scholarly impact: Methods and practice* (pp. 285–320). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-10377-8_13.

Waltman, L., van Eck, N. J., van Leeuwen, T. N., Visser, M. S., & van Raan, A. F. J. (2011). Towards a new crown indicator: Some theoretical considerations. *Journal of Informetrics, 5*(1), 37–47. https://doi.org/10.1016/j.joi.2010.08.001

Westfall, J., Kenny, D. A., & Judd, C. M. (2014). Statistical power and optimal design in experiments in which samples of participants respond to samples of stimuli. *Journal of Experimental Psychology: General, 143*(5), 2020–2045. https://doi.org/10.1037/xge0000014

Zaragoza-Jimenez, N., Niehaus, H., Thome, I., Vogelbacher, C., Ende, G., Kamp-Becker, I., … Jansen, A. (2023). Modeling face recognition in the predictive coding framework: A combined computational modeling and functional imaging study. *Cortex; a Journal Devoted To the Study of the Nervous System and Behavior*, Article S0010945223001648. https://doi.org/10.1016/j.cortex.2023.05.021