- Replication value as a function of citation impact and sample size
- Peder Mortvedt Isager¹, Anna E. van 't Veer², & Daniël Lakens³
- ¹ Department of Psychology, Oslo New University College
- ² Methodology and Statistics unit, Institute of Psychology, Leiden University
- ³ Department of Industrial Engineering & Innovation Sciences, Eindhoven University of
- 6 Technology

- Submitted to Meta-Psychology. Participate in open peer review by commenting
- 9 through hypothes.is directly on this preprint. The full editorial process of all articles under
- 10 review at Meta-Psychology can be found following this link:
- 11 https://tinyurl.com/mp-submissions. You will find this preprint by searching for the first
- authors name. This work was funded by VIDI grant 452-17-013.
- The authors made the following contributions. Peder Mortvedt Isager:
- ¹⁴ Conceptualization, Investigation, Visualization, Writing Original Draft, Writing Review
- ¹⁵ & Editing; Anna E. van 't Veer: Supervision, Writing Review & Editing; Daniël Lakens:
- 16 Conceptualization, Funding acquisition, Investigation, Supervision, Writing Review &
- 17 Editing.
- 18 Correspondence concerning this article should be addressed to Peder Mortvedt Isager,
- Lovisenberggata 13, 0456 Oslo, Norway. E-mail: pederisager@gmail.com

20 Abstract

Researchers seeking to replicate original research often need to decide which of several 21 relevant candidates to select for replication. Several strategies for study selection have been 22 proposed, utilizing a variety of observed indicators as criteria for selection. However, few 23 strategies clearly specify the goal of study selection and how that goal is related to the 24 indicators that are utilized. We have previously formalized a decision model of replication 25 study selection in which the goal of study selection is to maximize the expected utility gain of the replication effort. We further define the concept of replication value as a proxy for 27 expected utility gain (Isager et al., 2020). In this article, we propose a quantitative operationalization of replication value. We first discuss how value and uncertainty - the two concepts used to determine replication value – could be estimated via information about citation count and sample size. Second, we propose an equation for combining these 31 indicators into an overall estimate of replication value, which we denote RV_{Cn} . Third, we 32 suggest how RV_{Cn} could be implemented as part of a broader study selection procedure. 33 Finally, we provide preliminary data suggesting that studies that were in fact selected for 34 replication tend to have relatively high RV_{Cn} estimates. The goal of this article is to 35 explain how RV_{Cn} is intended to work and, in doing so, demonstrate the many assumptions that should be explicit in any replication study selection strategy. 37

38 Keywords: replication value, replication, study selection, expected utility, RVcn

39 Word count: 10103

Replication value as a function of citation impact and sample size

Introduction

40

After close to a century of repeated reminders that replication studies play an 42 important role in establishing robust scientific knowledge, and following a number of 43 high-profile replications published in the last decade (Hagger et al., 2016; e.g., Open Science Collaboration, 2015; Ranehill et al., 2015; Ritchie, Wiseman, & French, 2012; Wagenmakers et al., 2016), researchers from many scientific disciplines are increasingly calling for a renewed focus on replication research (Blaszczynski & Gainsbury, 2019; Button et al., 2013; Heirene, 2021; Murphy, Mesquida, Caldwell, Earp, & Warne, 2021; Plucker & Makel, 2021; Sale & Mellor, 2018; Zwaan, Etz, Lucas, & Donnellan, 2018). Given the exponential growth of the empirical literature (Bornmann & Mutz, 2015; Parolo et al., 2015) and the low rates of replication up until this point (Makel, Plucker, & Hegarty, 2012; Mueller-Langer, Fecher, Harhoff, & Wagner, 2019), researchers interested in conducting replications of original research will often have to choose which of several replication targets to focus limited resources on. Similarly, funding bodies have to decide which of several proposed replication efforts to direct limited grant money towards, and journals that limit the number of articles they accept for publication might benefit from clearly communicating which replication studies will be accepted. Given that there will 57 often be a large number of replication targets to choose from, and assuming we can distinguish which targets would be most useful to replicate, study selection strategies are needed that will allow stakeholders to transparently discuss and compare the need for replication of candidate targets. 61

In response to this need, there has been an increasing interest in developing
quantitative, indicator-based strategies for estimating which targets are most in need of
replication (see Isager et al., 2020 for a review). Proposed quantitative strategies tend to
focus on different indicators. Field, Hoekstra, Bringmann, and Van Ravenzwaaij (2019)

base selection on Bayes Factors, Makel et al. (2012) formulate a selection criterion based on number of citations, Matiasz et al. (2018) devise a strategy based on the number of existing replications of the same target, and so on (see supplementary documents in Isager et al., 2020 for additional examples). Common to all such strategies is that they are all fundamentally measurement instruments. For a quantitative indicator-based strategy to make sense, there must exist some target attribute (e.g., "replication importance") that is related to our replication goals, and that can be quantified in a meaningful way.

Furthermore, we must assume that the observed indicator(s) of interest (Bayes factor, citation count, number of replications, etc.) is somehow a valid and accurate measure of the target attribute (Borsboom, Mellenbergh, & van Heerden, 2004).

So far there has been little formal analysis of the measurement assumptions
underlying proposed study selection indicators. For any given indicator it is often not made
explicit (1) what goal we are working towards, (2) how that goal is related to the target
attribute we are trying to measure, and (3) how well we are able to measure that target
attribute through the proposed indicator. This makes validation of proposed study
selection strategies challenging; it is difficult to say whether a strategy works as intended if
it is not first made clear how the strategy is intended to work. To examine the quality of a
given indicator-based strategy, stakeholders must therefore specify the goal of replication,
formalize how this goal can be achieved by selecting studies based on a particular target
attribute, and specify the measurement model that connects the target attribute to the
indicator(s) being measured.

The aim of this article is to address the need for measurement models by
demonstrating how important measurement assumptions could be worked out and reported
for a particular quantitative indicator. We begin by clearly stating the goal we want our
study selection strategy to achieve, and the target attribute relevant for achieving this goal.
We then propose a quantitative indicator to measure the target attribute. As part of this
process, we discuss the various assumptions that must be met for the indicator to work as

intended, and we show how these assumptions lead to observational predictions that let us test the validity of the proposed indicator. We are not advocating that the indicator we propose should be used for study selection without further validation of its usefulness, nor are we claiming that it is superior to already proposed study selection strategies.

The target attribute of replication study selection.

97

The first step in working out the measurement assumptions of a replication study 98 selection strategy is to clearly define the assumed goal of the replication effort. What are 99 we trying to achieve that should lead us to prefer one replication target over another? 100 Here, we build on the decision model of replication study selection proposed by Isager et al. (2020). This model proposes that the ultimate goal of replication study selection is to 102 select the replication target that, out of all targets considered, will yield the largest gain in 103 expected utility if replicated. Under certain additional assumptions, expected utility qain 104 can be approximated by replication value (RV; see Isager et al., 2020 for a formal definition 105 of the term and outline of important underlying assumptions). Replication value is a 106 function of the value of having accurate knowledge about a replication target claim, and 107 our uncertainty about the truth status of the target claim before it is replicated: 108

$$RV = f(value, uncertainty)$$
 (1)

If we further define value as "value of being correct about the truth status of a claim," and uncertainty as "the probability of being incorrect about the truth status of a claim," then we can express replication value more concretely as the product of value and uncertainty:

$$RV = value \times uncertainty \tag{2}$$

If we are not willing to define uncertainty in terms of probability, we would need a
different function to relate value and uncertainty. In the remainder of this article we will
assume the definitions of value and uncertainty stated above, and consequently assume
that equation (2) is an appropriate function for combining value and uncertainty into an
expected utility estimate. Expected utility gain is thus our ultimate target attribute, and
replication value is our proximate target attribute (under the assumption that expected
utility gain can be satisfyingly approximated by replication value as defined in equation 2).

Defining relevant measurement properties.

120

To create an indicator that is useful for replication study selection we must ensure 121 that the indicator is a valid, reliable and unbiased measure of replication value. Given 122 equation (2), this implies finding valid, reliable and unbiased operationalizations of value 123 and uncertainty. Following Isager (2020), we consider an indicator to be valid for 124 measuring replication value if (a) the attribute replication value exists and (b) replication 125 value is d-connected to the indicator (i.e., the two variables are joined by an unblocked 126 path in the causal model that describes their true relationship; see fine point 6.1 in Hernán 127 & Robins, 2020). 128

Following the APA Dictionary of Psychology, we consider an indicator to be a *reliable* measure of replication value if quantitative estimates of replication value are free of random error ("Reliability – APA Dictionary of Psychology," n.d.). For most measurement instruments, complete freedom from random error is unattainable, and reliability is thus a matter of degree. In practice, we will have to decide how reliable an indicator should be before we can use it for replication study selection.

We consider an indicator to be an *unbiased* measure of replication value if indicator rank-order estimates do not systematically diverge from the true rank-order difference in replication value between the claims being considered (Isager et al., 2020; "Test bias – APA"

Dictionary of Psychology," n.d.). It matters less whether absolute replication value is
over/under-estimated by the same amount for all candidates in a set of replication targets,
since only relative rank-order replication value within the set matters for answering "which
of the candidates in this set would I increase expected utility the most by replicating?"
Like reliability, bias is a matter of degree.

The validity, reliability and bias of an indicator combine to form the measurement quality of the indicator. Quality is often referred to as validity in practice, but validity refers to something more specific here (Borsboom et al., 2004; Isager, 2020). For an indicator of replication value to have high measurement quality it must be valid and reliable and unbiased¹.

Operationalizing an indicator of replication value.

148

Many replication value indicators have already been proposed. Of these, most focus 149 exclusively on either the value- or the uncertainty-side of equation 2. Some suggested 150 indicators take both value and uncertainty into account (Isager et al., 2020, supplementary 151 materials), but suffer from at least one of two problems. In some cases, indicators are 152 operationalized in obviously problematic ways. For example, if an indicator uses p-values 153 in their operationalization of uncertainty (e.g., https://osf.io/x73rk/) it is not possible to 154 differentiate between ambiguous statistical evidence (high uncertainty) and strong evidence 155 for the null (low uncertainty. See Field et al., 2019). In other cases, indicators depend on 156 information that is not easily available in practice, such as whether the claim studied is 157 considered "surprising" (https://osf.io/v8nkd/). Indicators that are difficult to calculate 158 will be difficult to use and to validate, since validation depends on our ability to collect 159 data about the indicator. For these reasons we here propose a novel quantitative indicator of replication value rather than analyze an existing indicator. However, we emphasize that 161

¹ Formally, we might say that a high-quality indicator implies high mutual information (DeDeo, 2018) between expected utility gain, replication value, and the indicator we use to estimate replication value.

working out the goals and assumptions of already proposed indicators would also be a highly worthwhile exercise.

We propose to operationalize the value of a claim as a function of the *citation count*of the original paper in which the claim is reported. We propose to operationalize the
uncertainty of the claim before replication as a function of the sample size of the replication
target study (or studies). The operationalized indicator of replication value thus becomes:

$$RV = value \times uncertainty$$

$$\approx f(citation\ count) \times f(sample\ size)$$
(3)

In what follows, we provide a rationale for the choice of these operationalizations, define the functions f(), and discuss known and potential issues related to validity, reliability and bias.

171 Citation count as an indicator of value

We define the value of a claim as the stakes involved in decision outcomes based on 172 the claim. Suppose a mining company is considering whether to establish a new mine. 173 Research suggests "there is gold in them there hills." If the claim is *correct* and the 174 company chooses to believe it, the new mine will turn a huge profit. If the claim is false 175 and the company chooses to believe it, the mine will be a complete waste of resources. 176 Thus, the value of the claim "there is gold in them there hills" depends on the interaction 177 between the decisions we make based on the claim and the truth status of the claim, which will determine the decision outcome. In other words, value is the utility of being correct 179 about the truth status of the claim (relative to the utility of being wrong). The value of a 180 claim is usually related to the impact of the claim in science and society. Scientific claims 181 can be used to build theories and research lines, transform clinical treatment, guide 182 education, inform public policy, etc. The potential benefits of such applications of scientific 183

claims will depend on the truth status of the claim. Consider the claim "naltrexone is an
effective treatment for drug craving." If true, this could have a huge impact on addiction
treatment policy. Because the stakes are high (we might drastically improve the life of drug
dependent patients if the claim is true, but we risk maltreating the same individuals if the
claim is false) the value of having accurate knowledge about the claim is high. Finding a
valid measure of value thus entails operationalizing one or more indicators of impact.

Here we propose to focus on scientific impact, and operationalize impact in terms of 190 citation count, which is generally considered a valid indicator of scientific impact (Aksnes, 191 Langfeldt, & Wouters, 2019). Scientific impact can be roughly defined as the impact of a 192 claim on future research decisions (such as whether to run a follow-up study, build on a 193 previous finding, etc.). If we can assume that researchers tend to cite articles as support for 194 important research decisions, then citation count will be sensitive to variation in value that 195 is caused by variation in scientific impact (figure 1). Based on comprehensive reviews of 196 the literature on citation behavior (Aksnes et al., 2019; Bornmann & Daniel, 2008) and 197 empirical studies of the relationship between citation count and perceived scientific impact 198 by researchers (Ioannidis, Boyack, Small, Sorensen, & Klavans, 2014; Radicchi, Weissman, 199 & Bollen, 2017), we consider this assumption to be plausible in many fields of science. By definition, it then follows that the number of citations of an article is a valid measure of the 201 perceived value of the claims in that article (Isager, 2020). Citation count also has several other desirable measurement properties. Its interpretation is relatively straight-forward. A 203 meticulous record of citation count is kept by many bibliometric sources, which means that 204 reliability of the count across sources can be studied (Martín-Martín, Orduna-Malea, 205 Thelwall, & Delgado López-Cózar, 2018). Bibliometric recordkeeping also means that 206 citation counts can be obtained with little effort in any scientific discipline, which is crucial 207 for being able to utilize citation count for study selection in practice. 208

However, citation count is clearly an imperfect indicator of value. First, citation count is not considered a very good measure of other sources of impact that we often have

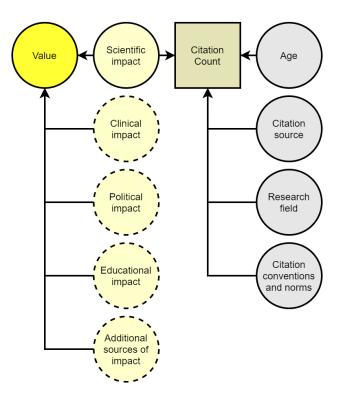


Figure 1: Proposed causal relationship between value and citation count (citation count is outlined by a square to signal it is observable). The figure can be interpreted as a directed acyclic graph model, where 'value' and 'citation count' are d-connected through 'scientific impact,' and all other variables act as independent noise factors on this relationship. Yellow-colored variables represent those we want information about. Dashed circles represent variables d-separated from 'citation count,' which we want to but do not have knowledge about. Since they are unmeasured causes of 'value' they also act as noise factors by distorting the relationship between value' and citation count. Grey-colored variables represent noise factors that influence citation count* but are d-separated from value. Additional sources of impact refer to any additional non-confounding causes of value that we could think of. Absence of a causal arrow between any two variables in the model implies the strong assumption that there is no causal relationship between these variables.

good reason to care about, such as societal or clinical impact. Impact on these dimensions 211 is less likely to generate citations from articles published in scientific outlets listed in major 212 bibliometric databases (Aksnes et al., 2019; Eck, Waltman, Raan, Klautz, & Peul, 2013). 213 Consequently, a traditional citation count metric will likely not be sensitive to variation in 214 value that is caused by variation in these dimensions of impact (see dashed circles in figure 215 1). When these other sources of impact are the most important to our overall definition of 216 value, citation count is going to be a poor measure of value (which, by extension, will lead 217 to poor estimates of replication value). 218

Second, even in cases where scientific impact is of primary interest, articles are cited 219 for a myriad different reasons that need not have anything to do with the scientific impact 220 of claims put forward in them. Non-relevant influences on citation count include (but are 221 not limited to) the time of publication, the bibliometric source that is doing the counting. 222 arbitrary citation conventions within scientific fields, the language in which an article is 223 written, reputation of article authors, preference for citing personal acquaintances, 224 bandwagon effects, self-citations to increase academic standing, etc. (Aksnes et al., 2019; 225 Bornmann & Daniel, 2008). To the extent that such factors are independent of article 226 impact, they act as random noise factors that reduce the reliability of citation count for measuring scientific impact and value (see grey circles in figure 1). 228

It is worthwhile to consider whether noise factors could be controlled to increase 220 reliability of the measurement. In some cases this will be challenging. For example, while it 230 might be possible to classify whether citations occurred due to arbitrary citation 231 conventions, citations would need to be manually classified for each replication target (though see Nicholson et al., 2021 for an example of innovations in citation classification). 233 However, three common and substantial sources of noise could likely be corrected for or held constant to improve the reliability of citation counts for measuring impact; the age of 235 the article (Bornmann & Daniel, 2008; D. Wang, Song, & Barabási, 2013), the source of 236 the citation count (Martín-Martín et al., 2018), and the research field the target claims are 237

238 part of (Waltman & van Eck, 2013).

Replication can only increase the expected utility of research decisions that have not 239 yet been made. Therefore, we specifically want to know the *future* scientific impact of a 240 replication target. If citation impact is taken as an indicator of simultaneous scientific 241 impact, it follows that what we are interested in is the future citation impact of a 242 replication target article. Current total citation count of an article is a measure of past 243 citation impact, but is likely predictive of future citation impact (M. Wang, Ren, Li, & 244 Chen, 2019). However, total citation count is not a useful predictor of future citations as 245 soon as we start comparing target articles across different publication years, as older 246 articles have had more time to be cited. As an example, imagine two articles, A and B, 247 that are both cited exactly 10 times every year. However, at the time of comparison study 248 A is 3 years old and has a total citation count of 30, while study B is 9 years old and thus 240 has a total citation count of 90. If we would use A and B's total citation count as a linear 250 predictor of their relative future citation count we will erroneously conclude that B will 251 receive three times as many citations as A. It is therefore sensible to adjust the two totals 252 for their respective ages and estimate the yearly citation rate (i.e. the slope, or the 253 derivative) of article A and B when determining their replication value.

Here we propose to adjust for publication age by simply dividing citation count by 255 the number of years since the article was published. That is, we operationalize the value of 256 a claim as the average citations per year of the paper that claim was reported in. In 257 practice this assumes that past average yearly citation rate is a good predictor of the future 258 trajectory of citations per year. This adjustment method works well when the future citation trajectory is relatively constant and similar to the past citation trajectory. Figure 2A displays the 50-year citation trajectory for an imagined article. The yellow line shows the predicted citation trajectory based on summing the citations from the first 25 years 262 (yellow bars) and dividing by the number of years. As expected, the prediction of the 263 actual citation trajectory (grey line) is very accurate. 264

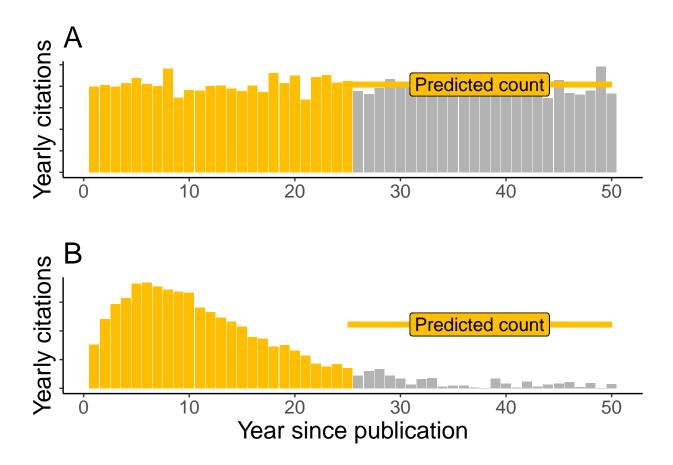


Figure 2: Simulated 50 year citation rate for two imagined articles. The yellow bars display the citation rate for the first 25 years in the article's history. The grey bars represent the future rate (year 26-50) that we want to predict. The yellow labeled line displays the predicted trajectory of future citations based on taking the sum of citations from the first 25 years and dividing by the number of years. (A) The true citation trajectory is defined as a constant citation count per year plus random error. The predicted trajectory is a very good approximation of the true citation rate. (B) The true citation trajectory is defined as a gamma function of the year since publication plus random error. The predicted trajectory substantially overestimates the true citation rate.

Adjustment by averaging over age works less well when the citation trajectory of an 265 article is not constant over time. Figure 2B displays a more realistic 50-year citation 266 trajectory (Parolo et al., 2015). In this context, predicting the future trajectory based on 267 past average yearly citation count leads to a considerable overestimate of future citations. 268 In fact, average yearly citation count will systematically over- or underestimate the future 260 citation trajectory whenever the trajectory is not converging towards the past average 270 yearly citation count. We therefore recommend that, whenever possible, prediction of 271 future citation impact should be based on more sophisticated prediction techniques that 272 utilize the entire distribution of past citation rate and other bibliometric features of the 273 target article (e.g., Chakraborty, Kumar, Goyal, Ganguly, & Mukherjee, 2014; Yuan, Tang, 274 Zhang, Wang, & Xiao, 2018)². However, these methods require advanced expertise to 275 understand and implement, and the information required to implement them may not always be available. For now, we focus on the simpler method of averaging over article age, 277 which provides a rough but useful method for adjusting citation impact estimates so replication targets of different ages can be compared.

In addition to controlling for age, we also need to consider the source that the 280 citation count estimate is derived from. Citation counts retrieved from different sources (such as Google Scholar, Crossref, Scopus or Web of Science) differ both in terms of their 282 reference coverage and in their exact citation counts for the same reference (Martín-Martín 283 et al., 2018). However, rank-order correlations between citation counts from different sources appear to be very high (Burgers, 2019; Martín-Martín et al., 2018). Thus, as long 285 as the same source is consistently used for all replication candidates under consideration, 286 the relative rank-order difference in citation count between two candidates should be highly 287 similar regardless of which source we use. 288

² Even detailed access to the past citation record of an article will not always allow for accurate prediction of future citation count however, due to phenomena such as "sleeping beauties" (Ke, Ferrara, Radicchi, & Flammini, 2015) and other hard-to-predict fluctuations in citation rate.

Finally, we need to consider the fact that article citation counts tend to 289 systematically vary between research fields for reasons that have nothing to do with the 290 value of claims (Bornmann & Daniel, 2008; e.g., Waltman & van Eck, 2013). Two fields 291 may differ in average article citation count due to differences in citation conventions, the 292 amount of journals or issues published per year, word limits and the maximum number of 293 references that is allowed for a specific article type, etc. A common approach to control for 294 such variation is to replace raw citation count with field-weighted citation impact (FWCI. 295 See Waltman & van Eck, 2019), in which citations are normalized against the average 296 citation count of articles from the same field of science. 297

One prominent difficulty with FWCI is determining the reference class to normalize 298 citation counts against. That is, when taking the average citation count of a research field, 299 which articles in the literature belong to that field? While it is common to use Web of 300 Science field categories for this delineation, these are considered too heterogeneous to 301 accurately control for variations in citation practices (Waltman & van Eck, 2013). Van 302 Eck, Waltman, van Raan, Klautz, and Peul (2013) have demonstrated that FWCI fails to 303 account for within-field variations in citations practices in medical science, and this can 304 lead to a severe underestimation of the impact of clinical research. Another challenge when 305 using FWCI to measure value is that some systematic differences in citation counts 306 between fields may genuinely be due to differences in the value society places on knowledge 307 produced in different fields. For example, according to Web of Science, the field of oncology 308 (cancer research) is nearly four times the size of dermatology (skin condition research) in 309 terms of sheer volume of articles published (1,802,676 vs 481,033 records, as of 2021-04-09) and at the time of writing, CA: A Cancer Journal for Clinicians is the highest impact 311 factor journal in the world with an impact factor of 509 as of 2022. If the research field of 312 oncology contains more researchers than dermatology - leading to more papers getting 313 published and more citations generated, on average, for each published paper - this could 314 reflect the fact that research in oncology is, on average, considered more valuable than 315

research in dermatology (to society, to clinicians, to the research community etc.). In this
case, FWCI becomes problematic because it partly suppresses the association between
citation impact and value. Researchers should think carefully about whether FWCI or raw
citation count is more appropriate for estimating value given the goals of their replication
effort, and should also consider whether some approaches to field-normalization are more
appropriate than others (Waltman & van Eck, 2019).

With the above-mentioned issues in mind, if we are still willing to assume that
scientific impact has a reasonably reliable causal effect on (age/source/field-corrected)
citation count, and if we are willing to assume that scientific impact has a reasonably
reliable causal effect on value, then citation count should be a useful measure of value.
Based on these assumptions, we propose the following operationalization of value for the
purposes of study selection:

$$value = \frac{w(C_S)}{Y+1} \tag{4}$$

where C stands for citation count of the article in which the claim in question is published, w() stands for the weighting function that should be applied to the citation count (a field normalization, a utility function, etc.³), S denotes the source the citation count is retrieved from, and Y denotes the age of the article in years. We add 1 to Y in order to prevent the equation from evaluating to infinity when the article was published less than a year ago and Y = 0.4 Thus, value is not identical but closely related to the

328

329

330

331

332

333

³ If raw citation scores are used then no weighting function is applied to C_S , $w(C_S) = C_S$, and w() can simply be removed from the equation.

⁴ Since all articles less than a year old will be adjusted by the same amount, an article published in January will be considered of the same age as an article published in December, even though the former has had many more months to acquire citations. This will likely matter less when both articles are several years old, but for young articles the monthly age difference may lead to substantial differences in their value estimates. If possible, instead of adding 1 to the citation estimate one could consider collecting the

average yearly (weighted) citation count of the article.

We could substitute equation (4) with a more intricate estimator and likely get a more 335 accurate impact estimate (Martin, 2011). However, there is an effort/accuracy tradeoff 336 involved in replication study selection. The benefits gained by more accurate estimates of 337 value need to be worth the additional efforts expended in collecting those estimates [e.g., in 338 reviewing documentation, interviewing key stakeholders, running case studies, etc; Klautzer et al. (2011), since all effort could be avoided by simply choosing a replication 340 candidate quickly based on random chance or personal interest, which would still lead to 341 some expected utility gain. Consequently, it may be preferable to use a less accurate but more easily derivable operationalization of value, so long as that measure is still accurate enough to yield study selection decisions that increase utility more than random selection.

Sample size as an indicator of uncertainty

Uncertainty, like value, is a multi-determined attribute. We may be uncertain about 346 a claim for a variety of reasons. The study design(s) used to test the claim may lack 347 internal or external validity, we may have a high prior that counteracts existing research, 348 statistical power might be too low to detect effect sizes of interest, original findings may 349 not have been independently replicated (or independent replication has failed to reproduce 350 the original study results), we may suspect that the evidence base is influenced by 351 publication bias, selective reporting, p-hacking, or fraud, and so on. Many factors related 352 to uncertainty could be quantified. For example, we can express our uncertainty about 353 parameter estimates using confidence intervals or other variability indices, and we can 354 express uncertainty about the relative likelihood of data given different hypotheses using 355 likelihood ratios or Bayes factors. We could incorporate uncertainty due to publication bias into Bayesian prior beliefs, which we could then combine with data into informed posterior exact publication date of each article, treat each day as 1/365th of a year, and simply divide citations less than a year old by 1/365th times the number of days since publication.

beliefs about claims. Alternatively, we could quantify uncertainty in terms of entropy using 358 principles from information theory (DeDeo, 2018). If our uncertainty depends on results 359 from multiple studies, we could combine their estimates through meta-analysis. The list 360 goes on. However, the information required to compute informative Bayesian posteriors, 361 Shannon entropy, etc., is often difficult to curate from published research reports, which 362 makes the task of estimating uncertainty based on these indicators difficult in practice. To 363 offer an indicator of uncertainty that is more easily determinable we here propose to 364 estimate uncertainty about a claim via the sample size of the existing study (or studies) 365 that investigate the claim. 366

Establishing the validity of sample size for measuring uncertainty requires a small 367 chain of measurement assumptions (figure 3). First, given a set of findings comprising the 368 original and the replication literature, we propose to approximate uncertainty as the 369 precision with which a parameter relevant to the claim has been estimated (e.g., for the 370 claim "stretching reduces the risk of athletic injury" we might be interested in the precision 371 of risk ratio estimates from relevant randomized controlled trials). Precision of parameter 372 estimates is obviously only one factor that makes up our overall uncertainty about a claim 373 (Isager et al., 2020). Even so, it seems reasonable to assume that (all else being equal) 374 increasing precision of a parameter estimate also decreases the uncertainty about claims 375 based on that parameter. Statistically we can define precision of the estimate as the 376 standard error of the parameter estimate⁵. 377

Second, when the standard deviation is not available we propose to make the
assumption that it is constant across all replication candidates, and to subsequently
estimate the precision of the estimate using only the sample size of the estimate. This
assumption will obviously always be false to some extent, which will lead to reduced
estimate reliability. However, it does capture the fact that, all else being equal, the higher

⁵ This definition of precision of the estimate will only make sense for quantitative parameters

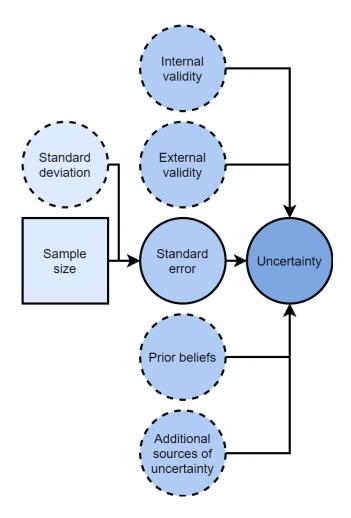


Figure 3: Proposed causal relationship between 'uncertainty' and 'sample size' ('sample size' is outlined by a square to signal it is observable). The figure can be interpreted as a directed acyclic graph model, where 'sample size' has a causal effect on 'uncertainty' by reducing the 'standard error' of relevant parameter estimates. Dashed circles represent variables d-separated from 'sample size,' which we want to but do not have knowledge about. Since they are unmeasured causes of 'uncertainty,' dashed circles also act as noise factors distorting the relationship between 'uncertainty' and 'sample size.' 'Additional sources of uncertainty' refer to any additional non-confounding causes of 'uncertainty' that we could think of. Absence of a causal arrow between any two variables in the model implies the strong assumption that there is no causal relationship between these variables.

the sample size the more precise the estimate. Based on the assumptions above, we propose the following equation for uncertainty:

$$uncertainty = \frac{\sigma}{\sqrt{n}} = \sigma \frac{1}{\sqrt{n}} \propto \frac{1}{\sqrt{n}}$$
 (5)

where σ is the standard deviation of the estimate, and n is the corresponding sample 385 size (i.e. the number of participants). When σ is constant, the equation is proportional to 386 $1 \div \sqrt{n}$. In other words, we assume that the standard deviation is the same for all 387 replication targets compared, and we assume that no other variance components but 388 participant variance are relevant to our standard error estimate (Westfall, Kenny, & Judd, 389 2014). When a claim depends on parameter estimates that have only been estimated in a single study, we can set n to the sample size of that study. When parameters have been 391 estimated in multiple studies, we should set n to the total sample size over all studies (see 392 supplementary material 1 for a rationale and relevant equations). Equation 4 behaves in line with the definition of uncertainty given by (Isager et al., 2020) whenever n>0.6

A limitation of this operationalization of uncertainty is that it ignores several 395 important factors that also contribute to uncertainty (dashed circles in figure 3). For example, we may be highly uncertain about a claim if the experiment used to support it is 397 poorly designed, poorly conducted, or if the claim suggests that experimental effects can be 398 generalized to a widely different context. This uncertainty may be completely independent 399 of the sample size of the study. In addition, there are instances where participant sample 400 size is not a strong determinant of the standard error. For example, participant sample size 401 is a poor estimator of the standard error in mixed model designs where the random effect 402 of participants is low compared with other random factors, such as stimulus (DeBruine & 403 Barr, 2021; Rouder & Haaf, 2018; Westfall et al., 2014). The same happens in repeated 404

⁶ When n = 1 it is not possible to estimate between-subject variance, so we should be maximally uncertain. When we are maximally uncertain, uncertainty = 1, which is what equation 4 yields for n = 1.

measures designs where the number of repeated samples per participant can matter more 405 than the number of participants per se, depending on the within-subject correlation (see 406 supplementary material 2 for a method of correcting the sample size estimate in such 407 cases). In general, because sample size is a cause of uncertainty, all determinants of 408 uncertainty that are independent of sample size will tend to reduce the reliability of sample 400 size as a measure of uncertainty. This problem is mitigated in more complex uncertainty 410 estimators such as Bayesian posteriors, which can incorporate several independent causes 411 of uncertainty into the estimate. 412

However, there are also advantages to using sample size as a measure of uncertainty. 413 Sample size contributes substantially to determining many of the previously mentioned factors that influence uncertainty, such as Bayesian posteriors, confidence/credible intervals, 415 etc., without invoking particular statistical inference philosophy. Sample size is also very 416 often readily available in published articles, regardless of the study design and statistical 417 analysis methods used (which is not the case for other potential uncertainty estimators 418 such as Bayesian posteriors). In summary, we believe sample size will be valid for 419 estimating overall uncertainty in most circumstances, but the reliability of these estimates 420 may be low due to the many independent factors influencing overall uncertainty. Reliability 421 could always be improved by substituting sample size with the full standard error in 422 equation 4, but this will require information that is not always available. To facilitate more 423 efficient estimates of uncertainty it is important that researchers begin to share statistical 424 information consistently and in machine-readable formats (Lakens & DeBruine, 2021). 425

Replication value as a function of citation count and sample size

If we operationalize value and uncertainty as specified in equations (4) and (5) and assume the structural equation for replication value defined in Isager et al. (2020)

(equation 2 in this article), we can construct a quantitative operationalization of replication value by multiplying equation (4) (our value estimator) with (5) (our uncertainty

estimator), which yields the following operational definition of replication value:

$$RV_{Cn} = value \times uncertainty = \frac{w(C_S)}{Y+1} \times \frac{1}{\sqrt{n}}$$
 (6)

where RV_{Cn} denotes a particular operationalization of replication value in terms of citation count C, and participant sample size n, w() stands for the weighting function that should be applied to the citation count, S denotes the source the citation count is retrieved from, and Y stands for the age of the article in years.

Figure 4 represents the measurement model for RV_{Cn} . It summarizes the causal assumptions that justify the use of RV_{Cn} as a measure of expected utility gain. Assuming that the casual relationships in the model holds, RV_{Cn} is d-connected with, and hence a valid measure of, expected utility gain. However, due to known unmeasured relationships (the U nodes in figure 4, which denote all the unmeasured variables in figures 1 and 3) RV_{Cn} can not be expected to be perfectly reliable. The exact reliability of RV_{Cn} is an important empirical question for future studies. Clearly, we must require a certain level of reliability in order to consider RV_{Cn} a useful measure of expected utility gain.

Figure 5 displays the distribution of RV_{Cn} over variations in the input parameters. RV_{Cn} estimates increase as the average yearly citation rate increases. Conversely, RV_{Cn} estimates decrease as the sample size increases. The axes in figure 5 cover a limited range of all possible input parameter values, but the distribution of RV_{Cn} remains similar for any range of input values.

In summary, RV_{Cn} is an appropriate operationalization of replication value when the following assumptions are met:

1. The goal of replication is to maximize expected utility gain for the claim(s) targeted for replication.

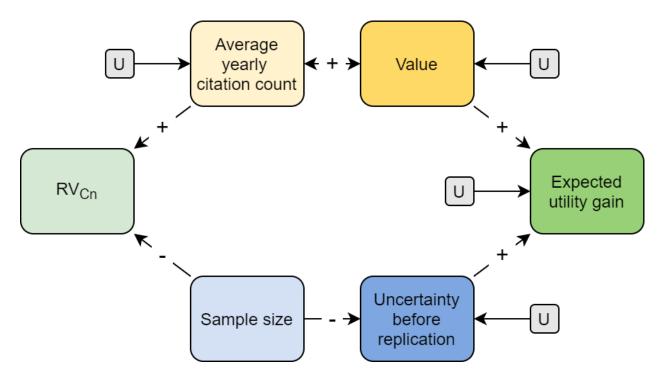


Figure 4: Measurement model justifying the validity of RV_{Cn} as a measure of expected utility gain. RV_{Cn} is d-connected with expected utility gain, as required for valid measurement by the d-connection definition of validity (Isager, 2020). The causal relations value \rightarrow expected utility gain and uncertainty before replication \rightarrow expected utility gain represent a simplified version of the structural causal model defined in Isager et al., (2020). All relations $U \rightarrow$ represent sources of noise due to unmeasured variables, such as the noise factors contained in figure 1 and figure 3. Absence of a causal arrow between any two variables in the model implies the strong assumption that there is no causal relationship between these variables.

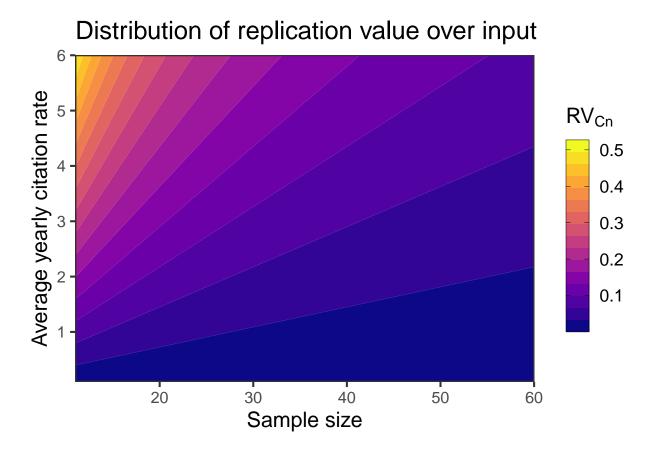


Figure 5: Distribution of RV_{Cn} for a range of input parameters.

453

454

457

458

459

460

- 2. Replication value, as defined in Isager et al. (2020) is a valid measure of the expected utility gain that a replication study would yield.
- 3. Equation (4) is valid and sufficiently positively associated with the true value of the to-be-replicated claim.
 - 4. Equation (5) is valid and sufficiently positively associated with the true uncertainty about the to-be-replicated claim.
 - 5. Equation (2) remains an appropriate specification of equation (1) when value is operationalized as equation (4) and uncertainty is operationalized as equation (5).
- In situations where any of these assumptions are violated, RV_{Cn} will cease to be a useful indicator of replication value.

General study selection strategy

463

Having provided the measurement rationale for RV_{Cn} , we must now specify how 464 RV_{Cn} can be implemented in a study selection strategy. In line with recommendations by 465 Field et al. (2019), we propose a procedure for study selection that combines RV_{Cn} with a 466 more comprehensive evaluation of expected utility gain. In-depth evaluation allows for 467 quality control and nuance during study selection. Given that multiple noise factors obscure the relationship between RV_{Cn} and expected utility, a certain amount of quality control will likely always be beneficial. However, detailed evaluation is also time-consuming and difficult to conduct in a systematic and unbiased way. To maximize the accuracy of 471 study selection while simultaneously minimizing the time spent on assessment, we propose a study selection procedure in which RV_{Cn} is used to narrow down the full set of replication candidates to a smaller subset of targets that are likely to be the most 474 worthwhile to replicate. A more detailed evaluation process can then be applied efficiently 475 to the highest RV_{Cn} candidates before a replication target is finally selected. Replication 476 study selection thus follows a general four-step procedure, outlined in figure 6. 477

First, a set of candidate replication targets is curated. This set should contain all 478 claims that are relevant to our interests and expertise. Targets that cannot be replicated 479 due to feasibility-constraints can be excluded from the initial candidate set for efficiency. 480 For example, suppose we want to form an initial candidate set of all experimental studies 481 on the psychoactive effects of cannabis in healthy human participants. Curation of this set 482 could begin by extracting all empirical articles on the psychoactive effects of cannabis from a bibliometric database. Subsequently, non-experimental studies, studies in non-human populations, studies in patient populations, etc., could be pruned away until the remaining 485 set of candidates matches our research interests and feasibility constraints. Second, RV_{Cn} is 486 calculated for all replication targets in the candidate set, so that replication targets can be 487 rank-ordered relative to each other based on their RV_{Cn} estimates. Third, a subset

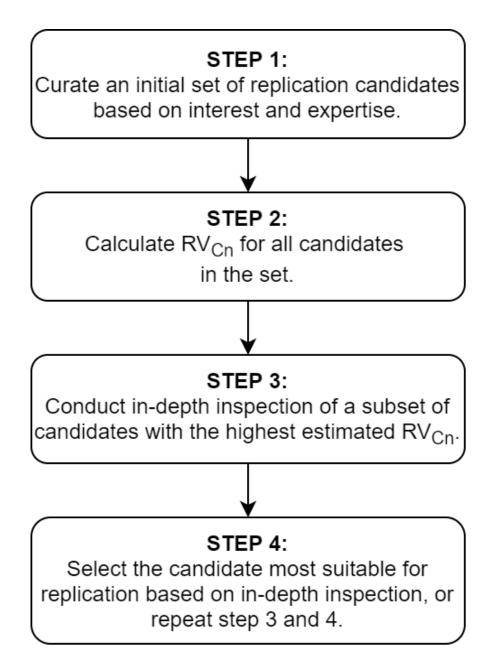


Figure 6: A general procedure for replication study selection. Quantitative estimation of replication value through RV_{Cn} (or through any other operationalization of replication value) can be implemented for study selection as part of this four-step procedure.

consisting of the highest RV_{Cn} targets in our candidate set is selected for closer inspection. 489 Assuming RV_{Cn} is a valid estimator of expected utility gain, this should increase the 490 probability that the studies evaluated further are those most likely to lead to high expected 491 utility gain if replicated. We can then invest a more substantial amount of time collecting 492 both qualitative and quantitative information about which studies in this subset would be 493 most worth replicating. How this evaluation should proceed is a topic of discussion (Field 494 et al., 2019; KNAW, 2018; Pittelkow, Hoekstra, Karsten, & van Ravenzwaaij, 2020), and 495 should probably be adapted depending on the goals of the stakeholder. However, the model 496 put forth in Isager et al. (2020) suggests that evaluation should always include some 497 consideration of (1) the perceived value of the research claim, (2) uncertainty about the 498 truth status of the claim prior to replication, (3) the ability of the planned replication study 499 design to reduce uncertainty about the claim, and (4) the estimated costs of conducting the replication effort. Empirical research to explore this issue in further detail is currently underway (Pittelkow, Field, Isager, Veer, & Ravenzwaaij, 2021). Finally, the replication 502 target that is considered the most promising after comprehensive evaluation is prioritized 503 for replication. Alternatively, if detailed evaluation suggests no study in the subset would 504 be worthwhile to replicate, step 3 can be repeated until a suitable target is found.

Preliminary validation of RV_{Cn} : assessing the replication value of replicated studies.

Whether RV_{Cn} is a valid and high-quality estimator of expected utility gain is an empirical question. Validation of replication value indicators will be challenging, since there does not exist an observable ground-truth measure of expected utility gain to benchmark indicators against. However, under some additional assumptions it is possible to make statements about what we expect to observe if RV_{Cn} works as intended. For example, if the true expected utility gain associated with a claim makes it more likely that researchers will select that claim for replication, then RV_{Cn} estimates and researchers'

selection preferences should be correlated under the causal model: $RV_{Cn} \leftarrow expected\ utility$ $gain \rightarrow selected\ by\ researchers\ for\ replication$. Consequently, studies that have been
replicated should have a higher RV_{Cn} on average than non-replicated empirical studies
from the same discipline. We here provide preliminary validation of this predicted
difference in a set of empirical articles from the field of psychology.

$_{20}$ Methods

Samples. We collected one dataset intended to represent the population of replicated studies in psychology, and one dataset intended to represent the general population of empirical studies in psychology.

The sample of replicated studies consisted of original study articles listed in the 524 Curate Science replication dataset (https://curatescience.org/app/replications). As of 525 2020-10-20, the Curate Science replication dataset contained information about 1127 526 replications of 202 original studies, primarily from the field of psychology. For these 527 original studies we estimated RV_{Cn} using sample size information available in the Curate 528 Science dataset, and using publication year and citation count information from Crossref 529 (https://www.crossref.org/). Citation counts were extracted from crossref 2020-10-20 using 530 the rcrossref package in R (Chamberlain, Zhu, Jahn, Boettiger, & Ram, 2020). Due to 531 missing DOI information, RV_{Cn} could not be calculated for 35 studies. The final sample of 532 replicated studies contained 167 studies from 145 articles. 533

The comparison sample consisted of 15104 empirical studies referenced in the tables
of meta-analyses published in Psychological Bulletin between the years 1914 and 2017.
Since all articles contained findings that have been referenced in meta-analysis tables in a
general-topic psychology journal, we assumed this sampling strategy would form a
reasonably representative sample of published empirical psychology studies. We also
assumed, given the generally low rate of replication in psychology (Makel et al., 2012), that

the comparison sample would consist largely of non-replicated original studies. We 540 estimated RV_{Cn} for each article using the sample size available in the meta-analysis tables, 541 and using publication year and citation count information from Crossref (citation counts 542 were extracted from crossref 2020-10-20). To retrieve citation count information, the DOI 543 of each study had to be retrieved, which was accomplished by applying a text-mining 544 procedure to the bibliometric information available in the meta-analysis tables and 545 reference lists. However, limited bibliometric information in the meta-analysis tables led to 546 many cases where the DOI could not be retrieved. Due to missing DOI information, RV_{Cn} could not be calculated for 11033 studies. The final sample of replicated studies contained 548 4071 studies from as many articles.

Statistical analyses. Our main hypothesis was that average RV_{Cn} would be higher 550 in the sample of replicated studies than in the comparison sample. In addition, we 551 analysed differences between samples in citation count, average citations per year, and 552 sample size, to better understand the causes of any potential differences in RV_{Cn} between 553 the two groups. Because all variables of interest were highly skewed, non-parametric 554 methods were used for all analyses. Median value and interquartile range were calculated 555 in both samples for each variable of interest. To compare differences between samples, we 556 calculated Vargha and Delaney's A for each variable of interest, which represents the 557 probability that a random observation from the sample of replicated studies has a higher 558 value than a random observation from the sample of non-replicated studies (Vargha & 550 Delaney, 2000). Bootstrapped 99% confidence intervals are reported for each effect size A. 560 Analyses were not preregistered. 561

The data files and analysis script used to generate all results reported below are openly available on OSF (https://osf.io/e35pu/).

Table 1: Summary statistics for variables of interest in the replicated (Curate Science) and comparison (Psychological Bulletin) samples.

variable	group	n	median	Q1	Q3	A
citation count	replicated	145	112.00	37.00	301.00	0.711 99%CI[0.652, 0.767]
citation count	comparison	4071	39.00	17.00	82.50	
citations per year	replicated	145	9.25	3.25	18.33	0.746 99%CI[0.686, 0.802]
citations per year	comparison	4071	2.67	1.12	5.63	
sample size	replicated	167	57.00	36.00	99.00	0.334 99%CI[0.294, 0.379]
sample size	comparison	4071	113.00	49.00	297.00	
replication value	replicated	167	1.01	0.46	2.70	0.797 99%CI[0.753, 0.841]
replication value	comparison	4071	0.22	0.08	0.53	

564 Results

Statistical results are presented in table 1. Cube-root transformed distributions of the variables of interest are presented in figure 7. RV_{Cn} was, on average, substantially greater in the sample of replicated studies than in the comparison sample (figure 7D). This difference seemed to be driven both by differences in citations per year, and by differences in sample size. Replicated studies received a greater number of citations (figure 7A), even after adjusting for article age (figure 7B). Conversely, replicated studies had substantially lower sample size, on average (figure 7C).

The patterns reported above are what we would expect to see if (1) researchers tend to follow the model of Isager et al. (2020) when selecting studies to replicate, (2) researchers can reliably predict the expected utility of replication efforts, and (3) RV_{Cn} is a valid but somewhat unreliable predictor of the expected utility of replication efforts. Under

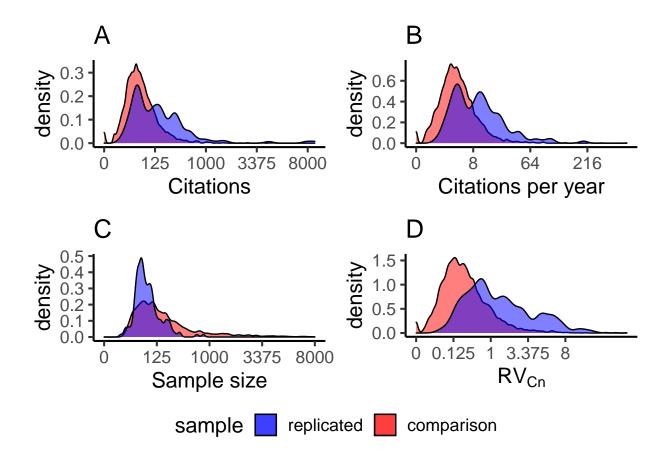


Figure 7: Distributions of various parameters in the comparison sample of psychological findings (red) and the sample of replicated findings in psychology (blue). The scale in all plots has been transformed by taking the cube root of the true values, which preserves the overall shape of the distribution but compresses the scale towards 1. (A) Distribution of citation counts. (B) Distribution of average citations per year. (C) Distribution of sample size. The x-axis limit is set to 8000, which excludes less than 1 percent of data points. (D) Distribution of RV_{Cn} replication value estimates.

these assumptions, a higher RV_{Cn} for replicated studies constitutes evidence for convergent validity between researchers' selection decisions and RV_{Cn} . Whether these assumptions truly hold, and whether there are other assumptions that would lead to the same predicted correlation without implying valid measurement, are still unresolved issues. Readers should therefore consider these results preliminary evidence for the validity of RV_{Cn} .

Our design is limited in a number of ways. The citation data used for these analyses 581 was collected in 2020 - years after the decisions to replicate were made. It is unclear what 582 effects this delay may have. In addition, our comparison sample consisted entirely of 583 empirical studies that are included in meta-analyses, which might differ from empirical 584 studies that are not in several aspects, including citation count and sample size. 585 Furthermore, RV_{Cn} could not be computed for a majority of the comparison sample due to 586 missing DOI information, and cases may not be missing at random. Thus, the 587 representativeness of the comparison sample could be questioned. An alternative method for deriving a comparison group would be to count citations of studies from the date each 589 study was selected for replication and select matched non-replicated controls from the same 590 year and research field. Finally, the external validity of the sample of replicated studies 591 may be reduced due to overrepresentation of a few large replication efforts. As one example, a large proportion of the studies included in the Curate Science dataset were replicated as part of the Reproducibility Project: Psychology (Open Science Collaboration, 594 2015). All studies replicated as part of this effort used the same general study selection 595 procedure (see https://osf.io/ezrsc/ for details about the sampling strategy). 596

A more fundamental limitation of this validation effort is the relatively low severity of the conclusions (Mayo, 2018). In effect, these data are weak falsifiers of the construct validity of RV_{Cn} . Consider what would have happened if the data showed no difference in replication value between the two samples. Would this have suggested that RV_{Cn} is a poor measure of replication value? Or would it have suggested that replication value is a poor measure of expected utility gain? Or would it have suggested that researchers' decisions of

what to replicate is a poor measure of expected utility gain? Perhaps RV_{Cn} is a better 603 measure of expected utility than researchers are? Conversely, a correlation between 604 researchers' study selection decisions and RV_{Cn} does not necessarily mean that either is a 605 valid measure of replication value. Perhaps both are consistently measuring a different 606 construct than the one we are interested in. The data presented here do not let us resolve 607 these issues, so they cannot definitely corroborate the validity of RV_{Cn} . Severity is also 608 lowered by the fact that these were exploratory analyses conducted by the same researchers 609 who proposed the model the data are trying to falsify. Preregistered independent 610 replication of these results in new data are therefore needed. 611

General discussion

612

To develop any strategy for efficient replication study selection, we need to (1) specify 613 the goal we are trying to achieve, (2) specify how this goal is realized by the strategy we 614 are proposing, and (3) empirically validate that the proposed strategy works as intended. 615 These three steps must be addressed in sequential order. We cannot specify how a given 616 strategy achieves a goal without first specifying the goal, and we cannot validate whether a 617 strategy works as intended before we can specify how the strategy is intended to work. 618 Isager et al. (2020) address the first of these steps. In this article, we have focused on the 619 second. Given the goal of maximizing expected utility gain (and the assumption that this 620 attribute can be approximated by replication value) we have formalized an observable 621 indicator - RV_{Cn} . We have formally specified the assumptions under which study selection 622 based on RV_{Cn} will maximize expected utility gain, and we have discussed a number of 623 obfuscating factors that limit the usefulness of RV_{Cn} . 624

To our knowledge, RV_{Cn} is the first replication study selection indicator that is explicitly treated as a measurement instrument. All indicators used for study selection are fundamentally measurement instruments, but RV_{Cn} is currently the only indicator for which a formal measurement model exists. By extension, RV_{Cn} is currently the only

indicator whose validity could be strictly falsified. RV_{Cn} is also the first indicator to 629 incorporate information about both value and uncertainty, which is important whenever 630 the goal of selection is to maximize expected utility gain. Should RV_{Cn} be a valid and 631 reliable measure of expected utility gain, it holds enormous potential for improving the 632 transparency and efficiency of resource allocation in replication research. The indicator 633 could be utilized by any researcher, funder, journal, or other stakeholder who wishes to 634 direct limited resources towards important replication targets. Because calculation only 635 requires information about citation impact and sample size RV_{Cn} should be applicable for 636 study selection across a variety of scientific disciplines. The strong measurement rationale 637 on which RV_{Cn} is built also makes it possible to improve and adapt the indicator to 638 different contexts. For example, in disciplines where sample size is a poor measure of 639 parameter estimate precision we can immediately see why RV_{Cn} estimates would be unreliable (assumptions underlying equation (5) are violated) and what remedy is required (replacing equation 5 with a valid measure of precision). 642

Three outstanding problems must be solved before we could confidently use RV_{Cn} for study selection in practice. These problems concern the feasibility, validity, and the overall quality of RV_{Cn} as a measurement instrument. We conclude this article with a brief sketch of each problem in turn.

First, in order to study the validity and usefulness of a study selection strategy based on RV_{Cn} , we need to know whether the strategy is feasible to apply in practice. Feasibility depends on a number of factors, including how difficult it is to assemble an initial pool of replication candidates and how difficult it will be to obtain the citation count and sample size of each candidate in the pool. Fortunately, obtaining citation counts will likely be a simple task regardless of the number of candidates, since records of such information are kept by multiple bibliometric services (e.g., Crossref, Scopus, and Web of Science), and can be automatically extracted for free through API interfaces such as the recossref package in R (Chamberlain et al., 2020). Collecting sample size information will likely be more

difficult. While sample size is almost always reported for empirical studies, it will likely
have to be obtained manually for each replication target, and issues related to participant
exclusion and multiple samples (e.g., when standard errors depend heavily on both the
number of participants and stimuli sampled) will have to be reckoned with. In a
forthcoming paper, we will explore the feasibility of applying a study selection procedure
based on RV_{Cn} within the context of fMRI research in social neuroscience. Similar studies
will likely be needed in other areas in which RV_{Cn} is to be applied.

Second, provided it is feasible to collect data about RV_{Cn} it must then be rigorously 663 validated. The lack of gold-standard indicators of expected utility gain prevents 664 straightforward benchmarking. As a substitute we could attempt to identify plausible 665 indicators of expected utility gain and compare RV_{Cn} to these, like is done in the study 666 reported above. However, interpretation of results will be more challenging in such studies 667 because we must add the auxiliary assumption that whatever attribute we compare RV_{Cn} 668 with is itself a good measure of expected utility gain (Meehl, 1997). Alternatively, RV_{Cn} 669 could be validated by corroborating the individual assumptions it is built on. The 670 constituent causal hypotheses in the chain that leads RV_{Cn} estimates to be a valid measure 671 of expected utility gain (figure 4) could be separated and tested in isolation. In principle, falsifying any single causal assumption would falsify the validity of RV_{Cn} overall, since valid 673 measurement requires that all the causal relationships hold. Finally, instead of examining 674 the validity of RV_{Cn} in isolation we could train machine-learning algorithms to predict 675 indicators of expected utility gain based on a broad range of bibliometric and statistical 676 information, as has recently been done to understand predictors of research replicability 677 (Altmejd et al., 2019; Yang, Wu, & Uzzi, 2020). This should yield a better understanding 678 of the predictive power of citation impact and sample size relative to other types of 679 information we could be considering. 680

Third, even if RV_{Cn} would be valid for measuring expected utility gain we still need to consider the broader issue of overall measurement quality (Borsboom et al., 2004). That

is, we need to empirically examine whether RV_{Cn} is sufficiently reliable and unbiased. If we 683 accept the causal hypotheses outlined in figure 4, then by definition replication RV_{Cn} must 684 be somewhat unreliable due to the influence of unmeasured noise factors in the 685 measurement process. The influence of the various noise factors will likely depend on the 686 diversity of the candidate set we are considering. Consider a set of studies, all from the 687 same subfield, all published within the span of a few years, and all utilizing similar study 688 designs. Now compare this to a set consisting of replication candidates from across a wide 689 time-span and many scientific fields, studying a wide range of topics with widely different 690 study designs. Variation in the noise factors included in figure 1 and figure 3 will likely be 691 much higher in the more diverse set of studies. Thus, the reliability of RV_{Cn} will likely 692 decrease as the diversity of replication targets increases, and there might even be a limit to 693 how broadly we can apply and compare RV_{Cn} estimates without compromising measurement reliability entirely.

 RV_{Cn} could also easily be a biased measure of expected utility gain. As an example, 696 consider the possibility that researchers prefer to cite articles that provide strong support 697 for the claims tested within (i.e., citation count is reduced by uncertainty about the 698 claims). In this scenario, RV_{Cn} would consistently underestimate expected utility gain since uncertainty ends up suppressing the positive effect of citation count on RV_{Cn} estimates (figure 8). Similarly, overestimation of expected utility gain can happen if research which 701 we are more certain will not replicate tends to get cited more often. Recent research 702 indicates the latter situation might often be the case (Serra-Garcia & Gneezy, 2021). Note 703 that we could in principle repeat this thought experiment for any two variables in figure 4 704 not already joined by a causal arrow. Whenever the causal relationship implied by such an 705 arrow seems plausible and causes systematic differences between the values of RV_{Cn} and 706 expected utility gain, we should be worried about bias in the measurement. 707

The key insight we wish to leave the reader with in this paper is that *any* strategy for replication study selection must be able to demonstrate usefulness, which entails assessing

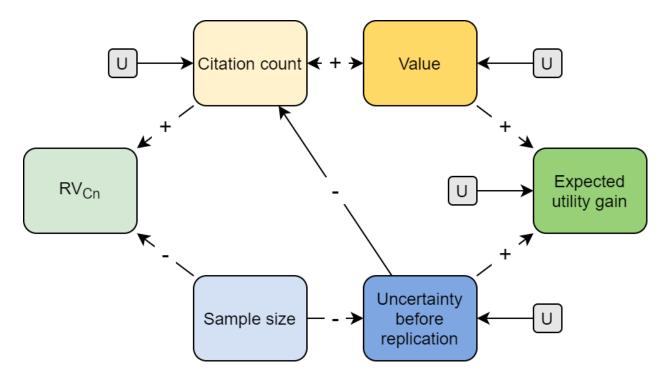


Figure 8: Causal scenario which will lead RV_{Cn} to become a biased measure of expected utility gain due to the influence of uncertainty before replication on citation count.

the feasibility and quality of the study selection criteria used. This assessment can only 710 happen if we first specify the goal of replication and formalize how the goal is achieved 711 through the selection strategy we are proposing. Only then can we evaluate whether the 712 strategy works as intended. For existing indicator-based strategies where the rationale 713 behind the indicator used remains unclear (e.g., Field et al., 2019; Isager, 2018; Isager et 714 al., 2020, supplementary materials; Makel et al., 2012), the measurement model should be 715 formalized so that validation and comparison of different strategies can take place. The 716 present article serves as a general example of how study selection strategies could explicitly be related to the goal we wish to accomplish when replicating, and how this in turn enables 718 us to examine whether a given strategy is likely to work as intended. In doing so we enable 719 a more transparent discussion of what is important to replicate, which will be crucial for 720 coordinating future replication efforts and direct resources to where they are needed the 721 most. Through such discussion and coordination we increase the overall impact of 722

replication studies on scientific discourse and strengthen the argument for why replications should be conducted and rewarded.

References

- Aksnes, D. W., Langfeldt, L., & Wouters, P. (2019). Citations, Citation Indicators, and
- Research Quality: An Overview of Basic Concepts and Theories. SAGE Open, 9(1),
- ⁷²⁸ 215824401982957. https://doi.org/10.1177/2158244019829575
- Altmejd, A., Dreber, A., Forsell, E., Huber, J., Imai, T., Johannesson, M., ... Camerer, C.
- (2019). Predicting the replicability of social science lab experiments. *PLOS ONE*,
- 14(12), e0225826. https://doi.org/10.1371/journal.pone.0225826
- Blaszczynski, A., & Gainsbury, S. M. (2019). Editor's note: Replication crisis in the social
- sciences. International Gambling Studies, 19(3), 359–361.
- https://doi.org/10.1080/14459795.2019.1673786
- Bornmann, L., & Daniel, H.-D. (2008). What do citation counts measure? A review of
- studies on citing behavior. Journal of Documentation, 64(1), 45-80.
- https://doi.org/10.1108/00220410810844150
- Bornmann, L., & Mutz, R. (2015). Growth rates of modern science: A bibliometric
- analysis based on the number of publications and cited references. Journal of the
- Association for Information Science and Technology, 66(11), 2215–2222.
- https://doi.org/10.1002/asi.23329
- Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The Concept of Validity.
- 743 Psychological Review, 111(4), 1061–1071.
- https://doi.org/10.1037/0033-295X.111.4.1061
- Burgers, J. (2019). Citation Counts as a Measure for Scientific Impact.
- 746 https://doi.org/10.31237/osf.io/y9r7c
- Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J.,
- ⁷⁴⁸ & Munafò, M. R. (2013). Power failure: Why small sample size undermines the

- reliability of neuroscience. Nature Reviews Neuroscience, 14(5), 365–376.
- 750 https://doi.org/10.1038/nrn3475
- Chakraborty, T., Kumar, S., Goyal, P., Ganguly, N., & Mukherjee, A. (2014). Towards a
- stratified learning approach to predict future citation counts. In IEEE/ACM Joint
- Conference on Digital Libraries (pp. 351–360).
- https://doi.org/10.1109/JCDL.2014.6970190
- Chamberlain, S., Zhu, H., Jahn, N., Boettiger, C., & Ram, K. (2020). Rcrossref.
- DeBruine, L. M., & Barr, D. J. (2021). Understanding Mixed-Effects Models Through
- Data Simulation. Advances in Methods and Practices in Psychological Science, 4(1),
- 758 2515245920965119. https://doi.org/10.1177/2515245920965119
- DeDeo, S. (2018). Information Theory for Intelligent People, 15.
- ⁷⁶⁰ Eck, N. J. van, Waltman, L., Raan, A. F. J. van, Klautz, R. J. M., & Peul, W. C. (2013).
- Citation Analysis May Severely Underestimate the Impact of Clinical Research as
- Compared to Basic Research. PLOS ONE, 8(4), e62395.
- 763 https://doi.org/10.1371/journal.pone.0062395
- 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.
- 766 https://doi.org/10.1525/collabra.218
- Hagger, M. S., Chatzisarantis, N. L. D., Alberts, H., Anggono, C. O., Batailler, C., Birt, A.
- R., ... Zwienenberg, M. (2016). A Multilab Preregistered Replication of the
- Ego-Depletion Effect. Perspectives on Psychological Science, 11(4), 546–573.
- https://doi.org/10.1177/1745691616652873
- 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
- 774 Hernán, M., & Robins, J. (2020). Causal Inference: What If. Boca Raton: Chapman &
- Hall/CRC.

- 776 Ioannidis, J. P. A., Boyack, K. W., Small, H., Sorensen, A. A., & Klavans, R. (2014).
- Bibliometrics: Is your most cited work your best? Nature News, 514 (7524), 561.
- https://doi.org/10.1038/514561a
- Isager, P. M. (2018). What To Replicate? Justifications Of Study Choice From 85
- Replication Studies. https://doi.org/10.5281/zenodo.1286715
- Isager, P. M. (2020). Test validity defined as d-connection between target and measured
- attribute: Expanding the causal definition of Borsboom et al. (2004). PsyArXiv.
- 783 https://doi.org/10.31234/osf.io/btgsr
- Isager, P. M., Aert, R. C. M. van, Bahník, Š., Brandt, M., DeSoto, K. A., Giner-Sorolla,
- R., ... Lakens, D. (2020). Deciding what to replicate: A formal definition of
- "replication value" and a decision model for replication study selection. MetaArXiv.
- 787 https://doi.org/10.31222/osf.io/2gurz
- Ke, Q., Ferrara, E., Radicchi, F., & Flammini, A. (2015). Defining and identifying Sleeping
- Beauties in science. Proceedings of the National Academy of Sciences, 112(24),
- 790 7426–7431. https://doi.org/10.1073/pnas.1424329112
- Klautzer, L., Hanney, S., Nason, E., Rubin, J., Grant, J., & Wooding, S. (2011). Assessing
- policy and practice impacts of social science research: The application of the Payback
- Framework to assess the Future of Work programme. Research Evaluation, 20(3),
- 794 201–209. https://doi.org/10.3152/095820211X13118583635675
- 795 KNAW. (2018). Replication studies Improving reproducibility in the empirical sciences.
- 796 Amsterdam: KNAW.
- 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), 2515245920970949.
- https://doi.org/10.1177/2515245920970949
- Makel, M. C., Plucker, J. A., & Hegarty, B. (2012). Replications in Psychology Research:
- How Often Do They Really Occur? Perspectives on Psychological Science, 7(6),

- 537–542. https://doi.org/10.1177/1745691612460688
- Martin, B. R. (2011). The Research Excellence Framework and the 'impact agenda': Are
- we creating a Frankenstein monster? Research Evaluation, 20(3), 247-254.
- https://doi.org/10.3152/095820211X13118583635693
- Martín-Martín, A., Orduna-Malea, E., Thelwall, M., & Delgado López-Cózar, E. (2018).
- Google Scholar, Web of Science, and Scopus: A systematic comparison of citations in
- 252 subject categories. Journal of Informetrics, 12(4), 1160–1177.
- https://doi.org/10.1016/j.joi.2018.09.002
- 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),
- e0195271. https://doi.org/10.1371/journal.pone.0195271
- Mayo, D. G. (2018). Statistical inference as severe testing. Cambridge: Cambridge
- 815 University Press.
- Meehl, P. E. (1997). The Problem Is Epistemology, Not Statistics: Replace Significance
- Tests by Confidence Intervals and Quantify Accuracy of Risky Numerical Predictions.
- In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), What If There Were No
- Significance Tests?
- Mueller-Langer, F., Fecher, B., Harhoff, D., & Wagner, G. G. (2019). Replication studies
- in economics—How many and which papers are chosen for replication, and why?
- 822 Research Policy, 48(1), 62-83. https://doi.org/10.1016/j.respol.2018.07.019
- Murphy, J., Mesquida, C., Caldwell, A. R., Earp, B. D., & Warne, J. (2021). Selection
- Protocol for Replication in Sports and Exercise Science. OSF Preprints.
- https://doi.org/10.31219/osf.io/v3wz4
- Nicholson, J. M., Mordaunt, M., Lopez, P., Uppala, A., Rosati, D., Rodrigues, N. P., ...
- Rife, S. C. (2021). Scite: A smart citation index that displays the context of citations
- and classifies their intent using deep learning. bioRxiv, 2021.03.15.435418.
- https://doi.org/10.1101/2021.03.15.435418

- Open Science Collaboration. (2015). Estimating the reproducibility of psychological
- science. Science, 349 (6251), aac4716-aac4716. https://doi.org/10.1126/science.aac4716
- Parolo, P. D. B., Pan, R. K., Ghosh, R., Huberman, B. A., Kaski, K., & Fortunato, S.
- (2015). Attention decay in science. Journal of Informetrics, 9(4), 734–745.
- https://doi.org/10.1016/j.joi.2015.07.006
- Pittelkow, M.-M., Field, S. M., Isager, P. M., Veer, A. van 't, & Ravenzwaaij, D. van.
- 836 (2021). The Process of Replication Target Selection: What to Consider?
- Pittelkow, M.-M., Hoekstra, R., Karsten, J., & van Ravenzwaaij, D. (2020). Replication
- 838 Crisis in Clinical Psychology: A Bayesian and Qualitative Re-evaluation (Preprint).
- PsyArXiv. https://doi.org/10.31234/osf.io/unezq
- Plucker, J. A., & Makel, M. C. (2021). Replication is important for educational
- psychology: Recent developments and key issues. Educational Psychologist, $\theta(0)$, 1–11.
- https://doi.org/10.1080/00461520.2021.1895796
- Radicchi, F., Weissman, A., & Bollen, J. (2017). Quantifying perceived impact of scientific
- publications. Journal of Informetrics, 11(3), 704–712.
- https://doi.org/10.1016/j.joi.2017.05.010
- Ranehill, E., Dreber, A., Johannesson, M., Leiberg, S., Sul, S., & Weber, R. A. (2015).
- Assessing the Robustness of Power Posing: No Effect on Hormones and Risk Tolerance
- in a Large Sample of Men and Women. Psychological Science, 26(5), 653–656.
- https://doi.org/10.1177/0956797614553946
- Reliability APA Dictionary of Psychology. (n.d.). https://dictionary.apa.org/reliability.
- Ritchie, S. J., Wiseman, R., & French, C. C. (2012). Failing the Future: Three
- Unsuccessful Attempts to Replicate Bem's 'Retroactive Facilitation of Recall' Effect.
- 853 PLoS ONE, 7(3), e33423. https://doi.org/10.1371/journal.pone.0033423
- Rouder, J. N., & Haaf, J. M. (2018). Power, Dominance, and Constraint: A Note on the
- Appeal of Different Design Traditions. Advances in Methods and Practices in
- Psychological Science, 1(1), 19–26. https://doi.org/10.1177/2515245917745058

- Sale, C., & Mellor, D. (2018). A call for replication studies in Nutrition and Health.
- Nutrition and Health, 24(4), 201–201. https://doi.org/10.1177/0260106018817675
- 859 Serra-Garcia, M., & Gneezy, U. (2021). Nonreplicable publications are cited more than
- replicable ones. Science Advances, 7(21), eabd1705.
- https://doi.org/10.1126/sciadv.abd1705
- Test bias APA Dictionary of Psychology. (n.d.). https://dictionary.apa.org/test-bias.
- van Eck, N. J., Waltman, L., van Raan, A. F. J., Klautz, R. J. M., & Peul, W. C. (2013).
- Citation Analysis May Severely Underestimate the Impact of Clinical Research as
- Compared to Basic Research. PLoS ONE, 8(4), e62395.
- https://doi.org/10.1371/journal.pone.0062395
- Vargha, A., & Delaney, H. D. (2000). A Critique and Improvement of the CL Common
- Language Effect Size Statistics of McGraw and Wong. Journal of Educational and
- Behavioral Statistics, 25(2), 101–132. https://doi.org/10.3102/10769986025002101
- Wagenmakers, E.-J., Beek, T., Dijkhoff, L., Gronau, Q. F., Acosta, A., Adams, R. B., ...
- Zwaan, R. A. (2016). Registered Replication Report: Strack, Martin, & Stepper (1988).
- Perspectives on Psychological Science, 11(6), 917–928.
- https://doi.org/10.1177/1745691616674458
- Waltman, L., & van Eck, N. J. (2013). A systematic empirical comparison of different
- approaches for normalizing citation impact indicators. Journal of Informetrics, 7(4),
- 833–849. https://doi.org/10.1016/j.joi.2013.08.002
- Waltman, L., & van Eck, N. J. (2019). Field Normalization of Scientometric Indicators. In
- W. Glänzel, H. F. Moed, U. Schmoch, & M. Thelwall (Eds.), Springer Handbook of
- Science and Technology Indicators (pp. 281–300). Cham: Springer International
- Publishing. https://doi.org/10.1007/978-3-030-02511-3 11
- Wang, D., Song, C., & Barabási, A.-L. (2013). Quantifying Long-Term Scientific Impact.
- Science, 342 (6154), 127–132. https://doi.org/10.1126/science.1237825
- Wang, M., Ren, J., Li, S., & Chen, G. (2019). Quantifying a Paper's Academic Impact by

Distinguishing the Unequal Intensities and Contributions of Citations. *IEEE Access*, 7,

- 96198–96214. https://doi.org/10.1109/ACCESS.2019.2927016
- 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
- Yang, Y., Wu, Y., & Uzzi, B. (2020). Estimating the deep replicability of scientific findings
- using human and artificial intelligence. Proceedings of the National Academy of
- Sciences, 117(20), 10762-10768.
- Yuan, S., Tang, J., Zhang, Y., Wang, Y., & Xiao, T. (2018). Modeling and Predicting
- ⁸⁹⁴ Citation Count via Recurrent Neural Network with Long Short-Term Memory.
- arXiv:1811.02129 [Physics]. Retrieved from https://arxiv.org/abs/1811.02129
- Zwaan, R. A., Etz, A., Lucas, R. E., & Donnellan, M. B. (2018). Making replication
- mainstream. Behavioral and Brain Sciences, 41.
- https://doi.org/10.1017/S0140525X17001972