Deciding what to replicate: a formal definition of “replication value” and a decision model for replication study selection

Isager, P.M.; Van Aert, R.C.M.; Bahník, S.; Brandt, M.J.; DeSoto, K.A.; Giner-Sorolla, R. Krueger, J.; ... ; Lakens, D.

Deciding what to replicate: A decision model for replication study selection under resource and knowledge constraints.

Peder Mortvedt Isager[1], Robbie C.M. van Aert[2], Štěpán Bahník[3], Mark J. Brandt[4, 13], K. Andrew DeSoto[5], Roger Giner-Sorolla[6], Joachim I. Krueger[7], Marco Perugini[8], Ivan Ropovik[9, 10], Anna E. van 't Veer[11], Marek Vranka[12], & Daniël Lakens[1]

[1] Department of Industrial Engineering & Innovation Sciences, Eindhoven University of Technology
[2] Department of Methodology and Statistics, Tilburg University
[3] Faculty of Business Administration, Prague University of Economics and Business
[4] Department of Social Psychology, Tilburg University
[5] Association for Psychological Science
[6] School of Psychology, University of Kent
[7] Department of Cognitive, Linguistic & Psychological Sciences, Brown University
[8] Dipartimento di Psicologia, University of Milano-Bicocca
[9] Faculty of Education, Charles University
[10] Faculty of Education, University of Presov
[11] Methodology and Statistics unit, Institute of Psychology, Leiden University
[12] Faculty of Social Sciences, Charles University
[13] Department of Psychology, Michigan State University

Robust scientific knowledge is contingent upon replication of original findings. However, replicating researchers are constrained by resources, and will almost always have to choose one replication effort to focus on from a set of potential candidates. To select a candidate efficiently in these cases, we need methods for deciding which out of all candidates considered would be the most useful to replicate, given some overall goal researchers wish to achieve. In this article we assume that the overall goal researchers wish to achieve is to maximize the utility gained by conducting the replication study. We then propose a general rule for study selection in replication research based on the *replication value* of the set of claims considered for replication. The *replication value* of a claim is defined as the maximum expected utility we could gain by conducting a replication of the claim, and is a function of (1) the value of being certain about the claim, and (2) uncertainty about the claim based on current evidence. We formalize this definition in terms of a causal decision model, utilizing concepts from decision theory and causal graph modeling. We discuss the validity of using *replication value* as a measure of expected utility gain, and we suggest approaches for deriving quantitative estimates of *replication value*. Our goal in this article is not to define concrete guidelines for study selection, but to provide the necessary theoretical foundations on which such concrete guidelines could be built.

Keywords: expected utility, replication, replication value, study selection
Word count: 9107

1. Introduction

The goal of science is the advancement of knowledge (Kitcher, 1995). To achieve this goal, scientists need to generate novel claims[1] about the world, and they need to ensure that these claims represent true and robust knowledge. An important first step in ensuring the robustness of many scientific claims is to test whether the observations that support the claim are replicable. Non-replicable observational claims are unlikely to represent true and robust knowledge, so it is important to differentiate replicable from spurious claims - preferably before the latter have an unwarranted impact on scientific theories or collective beliefs in society. This concern is amplified by evidence that (a) researchers overestimate the replicabil-

[1]Throughout this article we will use the term 'claim' to refer to the target property of a replication study (i.e., the phenomenon being replicated), unless we refer directly to previous work that uses another term. Many terms could be used to refer to the replication target; a result, a study, a finding, an effect, a procedure used to generate an effect, etc. There is at present no consensus on which of these terms is the most appropriate to use. Preferred terms vary across articles, and many authors use different terms interchangeably within the same articles (Brandt et al., 2014; Coles, Tiokhin, Scheel, Isager, & Lakens, 2018; Field, Hoekstra, Bringmann, & Van Ravenzwaaij, 2019; Hardwicke, Tessler, Peloquin, & Frank, 2018; Heirene, 2020; Kuehberger & Schulte-Mecklenbeck, 2018; LeBel, McCarthy, Earp, Elson, & Vanpaemel, 2018; Mackey, 2012; Schmidt, 2009; Zwaan, Etz, Lucas, & Donnellan, 2018).

ity of significant claims (Tversky & Kahneman, 1971), (b) published articles report an implausibly high rate of positive claims (Fanelli, 2010, 2012; Scheel, Schijen, & Lakens, 2019), (c) there are many scientific practices that can increase the false-positive rate in published reports (e.g., Simmons, Nelson, & Simonsohn, 2011), and (d) such practices may be relatively common (Agnoli, Wicherts, Veldkamp, Albiero, & Cubelli, 2017; Banks et al., 2016; Fiedler & Schwarz, 2016; John, Loewenstein, & Prelec, 2012; LeBel et al., 2013).

The definition of what constitutes a replication is a topic under constant debate, on which many authors have weighed in over the decades (for summaries, see Schmidt, 2009; Zwaan et al., 2018; or Machery, 2020). In this article we start from the definition of replication by Nosek and Errington (2020): "to be a replication, [two] things must be true: outcomes consistent with a prior claim would increase confidence in the claim, and outcomes inconsistent with a prior claim would decrease confidence in the claim". We believe this definition provides sufficient clarity about what is meant by replication throughout this article. However, it is unlikely to be the final say in the definition debate, and we urge the reader to consider whether the arguments that follow here would make sense under other definitions of replication as well.

Previously, many scientific literatures have favored *conceptual* replication; extending an already-tested claim by testing it in a new method or context. This replication scheme is effective for testing boundary conditions and generalizability of replicable claims. However, in this scheme it is not straight-forward to adjust confidence in the original study's claim based on replication results, because any inconsistent result might be due to variations in context rather than to the original finding being a false positive (LeBel & Peters, 2011; Nosek & Errington, 2020). More recently, there have been increasing calls to conduct and publish replication studies that follow as faithfully as possible the methods and conditions of previously published research, in order to test the robustness of the reported claims. Throughout this article the term 'replication' is used to refer to studies that are 'close' (Brandt et al., 2014; LeBel et al., 2018) or 'true' (Moonesinghe, Khoury, & Janssens, 2007) to the original study, often also referred to as direct replications (Schmidt, 2009).

In the last decade, a number of failed (close) replications of prominent claims from the published literature (e.g., Doyen, Klein, Pichon, & Cleeremans, 2012; Hagger et al., 2016; Nosek, Spies, & Motyl, 2012; Open Science Collaboration, 2015; Ranehill et al., 2015; Ritchie, Wiseman, & French, 2012; Wagenmakers et al., 2016) have spurred intense debate about the nature and importance of replication – especially within the field of psychology (Cesario, 2014; Earp & Trafimow, 2015; Ebersole et al., 2016; Finkel, Eastwick, & Reis, 2017; Maxwell, Lau, & Howard, 2015; Pashler & Wagenmakers, 2012; Stroebe & Strack, 2014; Zwaan et al., 2018). The debate has generally led to increased efforts to solidify the role of replication within psychological research practice (Zwaan et al., 2018). Several journals have begun to encourage submission of replication reports (e.g., Lindsay, 2015; "Replication studies | Royal Society Open Science," n.d.; Simons, 2014; see Martin & Clarke, 2017 for a review). Furthermore, funding bodies are starting to explicitly direct grant resources toward replication efforts (e.g., "NSF Invites Grant Applications Related to Reproducibility in Neuroimaging," n.d.; "Replication Studies," n.d.). Perhaps the clearest signal of sustained changes in research practice is the increase in published replication studies (see https://curatescience.org/app/replications for a comprehensive list of recent replication studies in psychology). Funders, researchers, and journals are increasingly willing to finance, perform, and publish replication studies to improve the reliability of scientific knowledge.

Although the concept of replication is a central value of empirical science, not every replication study is equally valuable. For example, most researchers will intuitively agree that a study proposing 20 direct replications of the Stroop-effect (Stroop, 1935), a phenomenon which is replicated in hundreds of psychology classrooms every year, will not be the most informative scientific project to perform if the goal is to simply verify that the Stroop-effect exists. If replication of empirical findings is considered important, but the value of replication varies from claim to claim, this raises the question of when a

With exeption of the first and last author, all authors are listed in alphabetical order, and their contribution is categorized accoring to the CRediT taxonomy (Brand, Allen, Altman, Hlava, & Scott, 2015). A preprint version of this manuscript has been made available on MetaArxiv: https://doi.org/10.31222/osf.io/2gurz. A presentation of the model presented in this manuscript has been made available on YouTube: https://www.youtube.com/watch?v=0j-nKDCPcRQ.

The authors made the following contributions. Peder Mortvedt Isager: Conceptualization, Investigation, Writing - Original Draft, Writing - Review & Editing, Visualization; Robbie C.M. van Aert: Investigation, Writing - Review & Editing; Štěpán Bahník: Investigation, Writing - Review & Editing; Mark J. Brandt: Investigation, Writing - Review & Editing; K. Andrew DeSoto: Investigation, Writing - Review & Editing; Roger Giner-Sorolla: Investigation, Writing - Review & Editing; Joachim I. Krueger: Investigation, Writing - Review & Editing; Marco Perugini: Investigation, Writing - Review & Editing; Ivan Ropovik: Investigation, Writing - Review & Editing; Anna E. van 't Veer: Investigation, Writing - Review & Editing, Supervision; Marek Vranka: Investigation, Writing - Review & Editing; Daniël Lakens: Conceptualization, Investigation, Writing - Review & Editing, Supervision, Funding acquisition.

Correspondence concerning this article should be addressed to Peder Mortvedt Isager, Den Dolech 1, Atlas 9.417, 5600 MB, Eindhoven, The Netherlands. E-mail: pederisager@gmail.com

replication of an empirical finding is valuable enough to the scientific community to be worth performing.

Scientists operate under resource constraints. Scarcity of time and money means that there will be more claims that could be replicated than we currently have the resources to replicate. A researcher may be interested in the replicability of more claims than they have the time and money to address. A journal editor may want to issue a call for replications on important claims in a special issue, but is unsure which study proposals to prioritize for review and publication. A funding agency may receive more proposals for replication studies than they can support. As one example, in 2016 the Dutch science funder NWO decided to spend 3 million euro exclusively on replication grants ("Replication Studies," n.d.). The call initially ran for 3 years, and each year, only around 10% of submitted proposals could be funded, while many proposals received high evaluations from peers. In these cases we need to evaluate which among several potential replications would be the most valuable to conduct. This may be especially important for fields that have failed to replicate studies from past decades, and now realize their empirical foundations are less stable than assumed. Consequently, we need guidelines for which claims are more and less in need of replication, so that we can direct limited funding and working hours towards the most pressing replication efforts.

In this article, we propose a formalized definition of *replication value* to guide the decision of which claims to select for replication when a choice between several candidates must be made. We begin by reviewing proposed methods for study selection in replication research and justifications for study selection in published replication reports, and we summarize the factors that feature prominently in this literature. We then present a formalized definition of *replication value* based on decision theory, a central tenet of which is optimizing decision making for expected utility gain. With this goal in mind, we discuss how *replication value* can be used to evaluate the utility of replicating a particular claim, relative to a set of candidate claims. Further, we suggest how to construct formulas for estimating *replication value* quantitatively. Finally, we discuss the most important challenges to implementing our approach for study selection in replication research.

Our goal is not to provide a single set of rules for deciding what to replicate in all circumstances. Study selection is a complicated decision problem that will likely require different approaches depending on the specific purpose of replication and the person or group who is replicating. Our goal is to provide a general structure for the decision problem "what is (most) worth replicating?" to help researchers to consider what information is important, and which trade-offs need to be made, when making this decision (Clemen, 1996). By using a principled method, the decision of which study to replicate becomes transparent and can be openly discussed.

2. What factors influence replication study selection?

Researchers have explored to great depths how to conduct replication studies and interpret replication results (e.g., Baribault et al., 2018; Brandt et al., 2014; Frank et al., 2017; LeBel et al., 2018; Maxwell et al., 2015; Morey & Lakens, 2016; Westfall, 2016). The question of what we should be replicating has received comparatively less attention. In responses to a recent article by Zwaan et al. (2018), arguing for the importance of performing replication studies, some authors raised the importance of justifying the choice for which claims to replicate. Study selection, they propose, could be based on a cost-benefit analysis (Coles et al., 2018), a Bayesian decision-making framework (Hardwicke et al., 2018), or on a random selection process (Kuehberger & Schulte-Mecklenbeck, 2018). In response to these commentaries, Zwaan et al. (2018) state:

> "... we do not think that special rules for selecting replication studies are needed, or even desirable. [...] Idiosyncratic interests and methodological expertise guide the original research questions that people pursue. This should be true for replication research, as well."

Although it is important to allow for some degree of idiosyncrasy when selecting claims to replicate, we believe transparently communicating which claims are deemed valuable to replicate is important (cf. Giner-Sorolla, Amodio, & van Kleef, 2018). Publication is a strong extrinsic incentive for researchers to conduct research, and there is currently a great deal of uncertainty about whether journals would even publish replication studies. Given that replication studies are rewarded less than original research (Koole & Lakens, 2012), the additional uncertainty about whether any replication study would be seen as valuable by editors could further reduce the probability that researchers will choose to perform a replication study even if they are intrinsically motivated to do so. Furthermore, some researchers might not have strong idiosyncratic interests. They might be primarily motivated to perform a replication study that makes the biggest possible contribution to the scientific knowledge base. It seems unlikely that leaving the selection of replication studies entirely up to idiosyncratic interests will be the most efficient way to encourage researchers to conduct and publish replication studies. If we want to guide researchers to claims that would be important to replicate, this raises the question of which factors make a claim important to replicate.

In the following sections we review three sources of information about which factors may affect the need for replication. First, we review factors commonly mentioned in theoretical discussions of replication study selection. Second, we review attempts to develop quantitative models of replication importance, and we examine commonalities between factors

mentioned in these proposals. Third, we examine stated justifications for the selection of a claim by authors of replication studies. The main purpose of the following sections is to collate existing viewpoints on the factors that make replications valuable. It is important to note that this is not a systematic review. We have limited ourselves to a discussion of factors that are primarily mentioned in psychological research. A more systematic and comprehensive review would likely uncover additional factors that play a role in replication study selection.

2.1 Theoretical discussions of replication study selection.

Theoretical discussions of replication study selection have considered a number of different criteria for selection. There are discussions that primarily argue for targeting valuable research topics for replication. The underlying intuition is that when a claim impacts scientific theory, clinical practice, or public policy and understanding, the stakes of being right or wrong about the claim are raised. The higher the impact, the more we should want to know whether a claim is supported by evidence. Makel, Plucker, and Hegarty (2012, p. 541) suggest that *"the replication of important studies that impact theory, important policies, and/or large groups of people would provide useful and provocative insights"*. They also suggest that the citation count of the original research article gives an indication of this underlying impact, and tentatively offer a simple heuristic for deciding when a study should be replicated: *"as an arbitrary selection, if a publication is cited 100 times, we think it would be strange if no attempt at replication had been conducted and published"* (Makel et al., 2012, p. 541). Coles et al. (2018) propose to develop a decision theoretical framework for replication study selection, which should encompass evaluations of impact on theory and society (cf. Hardwicke et al., 2018). The desire to concentrate replication efforts on valuable claims is also explicitly stated in the editorial policies of many journals (Block & Kuckertz, 2018; "JESP Registered Reports Guidelines," n.d.; "Journal of Personality and Social Psychology - APA Publishing | APA," n.d.; "Peer-review policy | Nature Research," n.d.; Lindsay, 2017).

Then there are discussions that focus on the uncertainty of the to-be-replicated claim in the current literature. The intuition here is that replication can hardly be considered valuable if the claim has already been convincingly corroborated or falsified in the past. Field et al. (2019) and Pittelkow, Hoekstra, Karsten, and van Ravenzwaaij (2020) propose a procedure based on Bayes factors to quantify the relative ambiguity of different claims in order to target the most ambiguous claims for replication. Hardwicke et al. (2018) propose a similar approach, in which the Bayesian evaluation scheme could also be extended to incorporate how much information about the claim one would be able to gain through replication. "Information" could here capture both statistical uncertainties due to low sample size and imprecise estimates, and lack of credibility due to suspicions of questionable research practices such as p-hacking or publication bias. In other words, imprecise and biased data are less *informative* about a claim than precise and unbiased data.

A more general framework for study selection in experimental research, still focusing on uncertainty given the existing literature, has been proposed by authors within the field of molecular and cellular cognition (Landreth & Silva, 2013; Matiasz et al., 2018; Matiasz, Wood, Wang, Silva, & Hsu, 2017; Silva, Landreth, & Bickle, 2014; Silva & Müller, 2015). The framework combines rules for causal identification with Bayesian evidence (Matiasz et al., 2018, 2017) in an attempt to quantify the replicability (or consistency) and convergence of causal claims across experiments (see Silva et al., 2014, for an extensive introduction to the framework). The aim of this approach is to concentrate replication studies on tests of causal claims that are supported by weak or inconsistent evidence in the present literature.

While discussions often focus on either value or uncertainty, several authors have argued for selection strategies that take both factors into account (Brandt et al., 2014; Heirene, 2020; Mackey, 2012; Royal Netherlands Academy of Arts and Sciences, 2018). Field et al. (2019) and Pittelkow et al. (2020) propose qualitative evaluation of factors related to value, such as the theoretical merits of the research question, in addition to their Bayesian assessment procedure. Hardwicke et al. (2018) suggest that the information gain framework could be incorporated into an *"expected value analysis"* in which the value of the research topic is also taken into account.

Finally, some have argued that the value of replication also depends on the quality of the research design and feasibility of the replication study. Hardwicke et al. (2018) argue that research designs that cannot distinguish between different relevant hypotheses are not worth replicating, because they will not lead to information gain if conducted. Replication studies have low information gain when the quality of the replication study design is poor (Pittelkow et al., 2020), when the study is too costly (Coles et al., 2018), or when it cannot be conducted due to feasibility constraints (Field et al., 2019).

2.2 Factors included in proposals to quantify replication value

We have solicited additional perspectives on factors that contribute to replication study selection by asking researchers interested in replicability to create a quantitative formula for replication value[2]. In January 2016 a public invitation

[2]Note that this project was undertaken prior to the development of the formal model presented in this article. Thus, these researchers did not necessarily assume the definition of replication value that is

was shared in an online blog post (Lakens, 2016) and distributed through mailing lists, which led to eight teams of researchers who each created a quantitative replication value operationalization. For a detailed overview of the different operationalizations that were generated, see supplementary "RV formula" documents on OSF (https://osf.io/asype/).

There was substantial variation in the rationale for each operationalization, as well as in the specific factors that were considered. Yet, at a more general level, all formula proposals contained some index quantifying the value of the research topic (e.g., citation impact, field-weighted citation impact, journal impact factor, Altmetric Attention score), and some index quantifying the uncertainty of existing knowledge (e.g., p-value of existing tests, Bayesian posterior evidence, sample size, preregistration status, presence of inconsistencies in reported statistical results). This demonstrates both a consensus on the relevance of value and uncertainty in the study selection process, and a recognition of the many ways these factors can be operationalized.

2.3 Self-Reported Justifications for Selecting Studies for Replication

In addition to reviewing theoretical discussions of replication study selection and soliciting proposals for replication value formulas, we also surveyed self-reported justifications for study selection described by researchers who published replication studies. The first author conducted a literature review of study selection justifications in 85 replication reports (Isager, 2018). The reports were collected from the Curate Science database (LeBel et al., 2018), and were supplemented by a small number of more recent replication studies not mentioned in the database at the time of review.

Of those studies that specified a justification for their study selection (68 out of 85 reports), the justification was catalogued and categorized. Factors related to the value of the research topic (citation impact, theoretical importance, citation in textbooks, influence on public policy, etc.) was mentioned in 52 out of 68 reports. Factors related to the uncertainty of existing research (lack of replication, imprecise estimates, prevalence of questionable research practices etc.) was mentioned in 51 out of 68 reports. Many reports considered a combination of factors related to both value and uncertainty (see table of quotes in Isager, 2018). Some justifications also explicitly mentioned low costs and feasible study designs as criteria for replication study selection (4 out of 68 reports; see e.g., Errington et al., 2014; Open Science Collaboration, 2015)[3]. In addition to these factors, study selection was often motivated by personal preferences. For example, in 16 out of 68 reports, study selection was motivated at least partly by the research interests of the replication authors (e.g., a replication was conducted as a first step in a broader effort to extend on an existing study design).

Overall, our review suggests that researchers often consider four factors when deciding what would be worth replicating: (1) the value of the research topic, (2) the uncertainty about our current state of knowledge about the claim, (3) the quality of the proposed replication study, or the ability of the replication study to reduce uncertainty about the claim, and (4) the costs and feasibility of running a particular replication study. These factors can also be recognized in statements by journals who explicitly invite replication studies, such as the *Journal of Personality and Social Psychology*:

> "Major criteria for publication of replication papers include (i) theoretical significance of the finding being replicated, (ii) statistical power of the study that is carried out, and (iii) the number and power of previous replications of the same finding" ("Journal of Personality and Social Psychology - APA Publishing | APA," n.d.).

Building on the recommendations from many previous authors, we argue that when considering which finding is most worth replicating, we should ideally take all of these factors into account. Fortunately, there already exist formal theoretical frameworks for taking informed decisions based on the value and uncertainty of different options. Building on ideas by Coles et al. (2018) and Hardwicke et al. (2018), we will in the next section develop a formal model of replication study selection based on principles from utility theory.

3. Formalized definition of replication value

We model replication study selection in the structural causal model framework developed by Pearl (2009, definition 7.1.1). Figure 1 and Table 1 present the causal assumptions, structural equations, and verbal summaries for all terms mentioned in the text. For clarification, all terms from Figure 1 and Table 1 are italicized whenever mentioned in the text.

Our proposed model represents a decision process, and we define *replication value* based on decision theory (see Raiffa & Schlaifer, 1974 for an introduction). We assume that the goal of replication is to maximize the *marginal gain in expected utility* (or usefulness) of scientific claims after replication. In our model, we consider expected utility for science as a whole, but it could possibly be extended to consider costs and benefits for the individual scientist. Based on this, we model the process of deciding "which claim in a given set of claims would we gain the most utility by replicating?" In other words, we assume a decision-maker who has already decided to conduct a replication (as opposed to testing a novel claim, etc.). The *expected utility of a finding before replication*

proposed here.

[3] It may be fair to assume that feasibility constraints played a role in all reports, whether it is mentioned or not, since studies are only conducted if they are considered feasible to conduct.

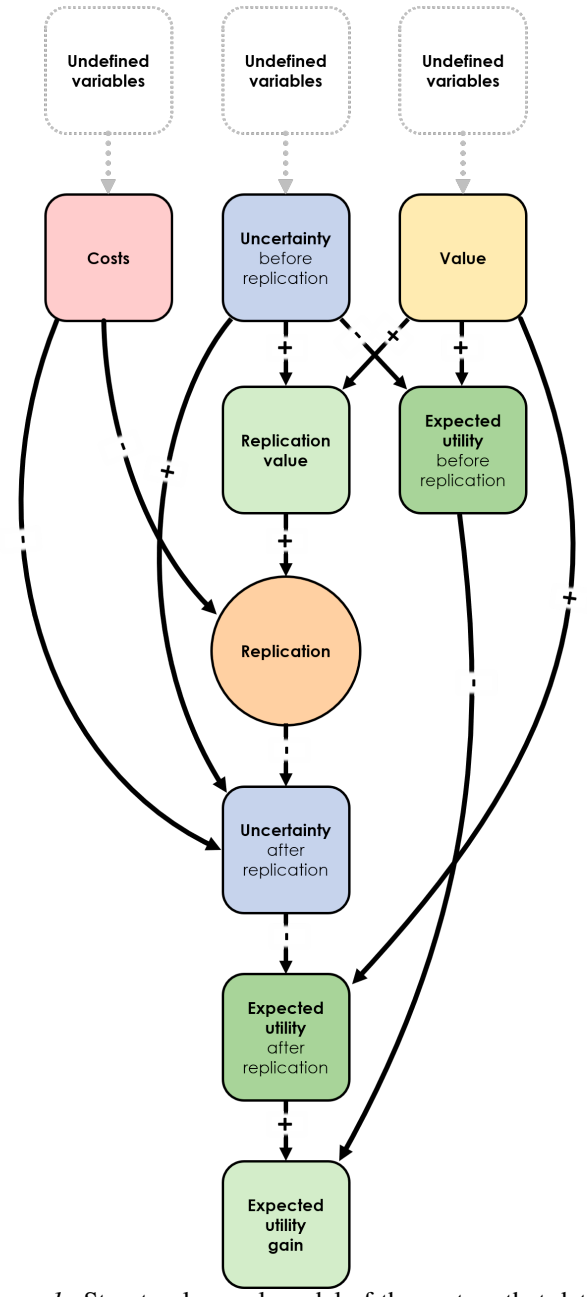

Figure 1: Structural causal model of the system that determines replication value. Arrow direction signals the causal direction of effects. Time flows from the top to the bottom of the figure; variables (nodes) closer to the top are determined earlier in time than nodes closer to the bottom (e.g., the value of "Costs" is determined before the value of "Replication"). The "+" and "-" signs on the arrows indicate whether the effect is positive or negative. Consult Table 1 for variable definitions and the structural equations that determine the value of each variable in the graph.

Name	Definition	Structural equation
Undefined variables (u)	A set of undetermined (exogenous) factors external to the model.	$u=f()$
Costs (C)	The costs of performing the planned replication study.	$C = f(u)$ Scale: $\{0 \leq C \leq \infty\}$
Uncertainty *before replication* (Un_{pre})	Uncertainty about the claim before replication.	$Un_{Pre} = f(u)$ Scale: $\{0 \leq Un_{Pre} \leq 1\}$
Value (V)	Value of the scientific claim	$V = f(u)$ Scale: $\{0 \leq V \leq \infty\}$
Expected utility *before replication* (EU_{pre})	Expected utility of the claim before replication.	$EU_{Pre} = V \times (1 - Un_{Pre})$
Replication value (RV)	Replication value: The potential increase in EU_{Gain} if all remaining uncertainty was removed.	$RV = V \times Un_{Pre}$
Replication (R)	Carrying out a replication study of the claim.	$R = f(RV, C)$ Scale: R = {"true", "false"} If $\uparrow RV$ then $\uparrow P(R =$ "true") If $\uparrow C$ then $\downarrow P(R =$ "true")
Uncertainty *after replication* (Un_{post})	Uncertainty about the claim after replication.	$Un_{post} = f(R, C, Un_{Pre})$ Scale: $\{0 \leq Un_{Post} \leq 1\}$ If $R =$ "true" then $\downarrow Un_{Post}$ If $\uparrow C$ then $\downarrow Un_{Post}$ If $\uparrow Un_{Pre}$ then $\uparrow Un_{Post}$
Expected utility *after replication* (EU_{post})	Expected utility of the claim after replication.	$EU_{post} = V \times (1 - Un_{Post})$
Expected utility gain (EU_{gain})	Marginal gain in expected utility after replication.	$EU_{gain} = EU_{Post} - EU_{Pre}$

Table 1: Structural equations for the structural causal model in Figure 1. The "Name" column corresponds to the node names inside Figure 1 (abbreviations in parentheses). The "Definition" column gives the verbal definition of each variable. The "Structural equation" column describes how the value of each variable in the model is causally determined by other variables in the model. The structural equations use the abbreviated variable names from the "Name" column. For any given structural equation, the variables on the right hand side of the equation correspond to those variables that point towards the variable in question inside Figure 1. The only exception are the undefined variables (u), which denote factors that are not specified by the model, but that nonetheless influence the value of variables in the model. Structural equations defined as a non-specific function $f()$ are not specified in the model. All we can formally say in these cases is that some function of the variables contained inside $f()$ can be used to determine the variable in question.

is a function of two factors: the *value* of the research claim (e.g., how important it would be to know whether smoking causes cancer) and the *uncertainty of our knowledge about the claim before replication* (e.g., how confident we are based on existing research whether smoking causes cancer). The assumed function of a well-designed *replication* is to reduce *uncertainty after replication*, which in turn increases the *expected utility of the scientific claim after replication*. Thus, our goal is to identify and perform replication studies that can substantially reduce uncertainty about claims that would be valuable to know the truth status of. If we incorporate the costs of a replication in the model, there is a point where the benefits of performing an additional replication study no longer outweigh the costs. In the remainder of this section we will explain this model in more detail and provide a formal definition of *replication value*.

In the model, *value*, *uncertainty* (*before* and *after replication*) and *costs* are all a function of *undefined variables* that are specified outside of the model (Pearl, 2009, definition 7.1.1). In other words, the model does not specify how *value*, *uncertainty*, and *costs* should be determined. However, even though a formal causal definition does not follow from our model, we can still say something about which variables are likely to be contained in our set of undefined variables, and the function with which they should be combined to determine *value*, *uncertainty*, and *costs*.

The *value* of a claim is defined as the importance of gaining certain knowledge about whether the claim is true or false[4]. The *value* of a research claim is usually related to the impact of the claim. This can include (but is not limited to) the pure ideal of gaining knowledge, the theoretical implications of the particular claim, or its potential for application. The more valuable the research claim is (to researchers, practitioners, or the general public), the higher the expected utility of the claim will be, and the more valuable a replication of research examining this claim will be. Ignoring some extreme cases where society would feel it is better not to know something, we assume that we can represent the value of having scientific knowledge on a scale from zero (no value) to infinity (infinitely valuable).

The *uncertainty* about a claim (before and after replication) is related to the probability that the claim is true, given some knowledge we have about the claim. Quantitatively, we express uncertainty on a scale from 0 (completely certain) to 1 (completely uncertain). If the probability *P*("smoking causes cancer"|knowledge) = 1, we have no uncertainty about the truth value of this claim (we know that it is true). If the probability *P*("smoking causes cancer"|knowledge) = 0, we also have no uncertainty about the claim (we know that it is false). Conversely, if we think it is equally likely that smoking causes cancer and that smoking does not cause cancer then the probability *P*("smoking causes cancer"|knowledge) = 0.5, and we are completely uncertain about the claim[5]. There are many reasons we might be uncertain about a claim. For example, the current evidence base may be sparse or ambiguous, effects relevant to the claim may have been imprecisely measured, the validity of designs in the existing empirical literature may be low, or existing studies might not reduce uncertainty due to publication bias and other factors that increases the prevalence of false positive findings (e.g., Lodder, Ong, Grasman, & Wicherts, 2019). The more uncertain we are about a claim, the lower the expected utility of the claim will be.

To the extent that we can quantify the *value* of scientific claims and the *uncertainty* of current knowledge, *expected utility* can be defined as the product of *value* and *1 - uncertainty* (see Table 1 for structural equations), where *1 - uncertainty* represents our certainty, or lack of uncertainty, about the truth value of a claim based on existing research. If we are completely certain that smoking causes cancer before replication then $Un_{Pre} = 0$, which implies $1 - Un_{Pre} = 1$ and $EU_{Pre} = V \times 1 = V$ (abbreviations and structural equations are spelled out in Table 1). In words, under complete certainty the *expected utility* of a claim simply equals the *value* of the claim. Conversely, if we are completely uncertain about whether smoking causes cancer before replication then the potential *value* of this knowledge might be very high, but the *expected utility* is still zero ($EUPre = V \times 0 = 0$). This explains why we do empirical research: We reduce the uncertainty about scientific claims we find valuable in order to increase the expected utility of these claims.

As defined in the introduction, *replication* refers to studies for which any outcome would be considered diagnostic evidence about a claim from prior research (for a more comprehensive definition, see Nosek & Errington, 2020). The function of *replication* in our model is to reduce *uncertainty about a claim after replication* (e.g., by reducing sampling error). By reducing uncertainty, replication increases the *expected utility of scientific claims after replication*, which increases the *expected utility gain*. In the model, replication is represented as an action on a binary scale, in which we can either conduct the replication (*replication*="true") or not (*replication*="false"). The quality of a replication study is, in our model, simply defined as the ability of the replication study to reduce uncer-

[4]More comprehensive definitions of value could be construed. For example, we might want to differentiate between the value of becoming certain that the claim is true vs. the value of becoming certain that the claim is false, or we might want to attach a negative value to being wrong about a claim.

[5]A more comprehensive definition could consider the probability of various belief states (e.g., correct rejection of claim vs. correct acceptance of claim vs. type 1 error vs. type 2 error), and should be able to model the fact that we can be misled by biased data such that the probability of drawing the correct conclusion about a claim is less than 50%.

tainty (represented by the effect size on the negative arrow *replication → uncertainty after replication*, in Figure 1). In other words, a high quality replication study leads to a larger reduction in *uncertainty after replication* than a lower quality replication study.

If our goal is to select the replication study that maximizes *expected utility gain*, our main problem is that *expected utility gain* is partially defined by *expected utility after replication*. Because this variable is determined after *replication*, we would need to conduct the replication study to determine *expected utility gain*, which defeats the purpose of using *expected utility gain* to determine which study should be replicated. However, if we are willing to make some assumptions about the effect of *replication* on *uncertainty* (*replication → uncertainty after replication* in Figure 1), it is possible to estimate *expected utility gain* based only on variables determined before *replication*. Given a claim with a set *value* and *uncertainty before replication*, the *replication value* of the claim is defined as the maximum possible gain in expected utility we could achieve through replication. It is essentially identical to the concept of "expected value of perfect information" from utility theory (Clemen, 1996, Chapter 12). *Replication value* indicates how much expected utility would increase after replication by removing all remaining uncertainty about a claim. If we assume that we could perform replication studies until all uncertainty about the claim has been removed ($Un_{Post} = 0$) then *replication value* (*RV*) becomes equivalent to *expected utility gain* (EU_{Gain}) since:

$$
\begin{aligned}
&EU_{Gain} = \\
&EU_{Post} - EU_{Pre} = \\
&V \times (1 - Un_{Post}) - V \times (1 - Un_{Pre}) = \\
&V \times (1 - 0) - V \times (1 - Un_{Pre}) = \\
&V - 0 - V + V \times Un_{Pre} = \\
&V \times Un_{Pre} = \\
&RV
\end{aligned}
$$

(abbreviations and structural equations are spelled out in Table 1).

In reality, a replication study can never completely remove uncertainty. Therefore basing *replication value* on the assumption that uncertainty is completely removed following replication will lead us to consistently overestimate *expected utility gain*. However, as long as the amount of uncertainty reduced is independent of the *replication value* of the claim, rank-order *replication value* will still be an unbiased estimator of rank-order *expected utility gain* across studies[6]. If our goal is to find the claim with the highest *expected utility gain* from a set of replication candidates, accurate rank-order estimates are all we require. However, we must then be willing to accept that we cannot use replication value to evaluate whether one study is twice as important to replicate as another, and other questions that require an interval scale variable. All else equal, *replication value* is highest for valuable claims that we are very uncertain about before replication. Conversely, *replication value* will be low for highly uncertain claims that are not worth knowing, and for valuable claims that we are already quite certain about.

It is possible to further extend our consideration of which replication study will lead to the highest *expected utility gain* by also considering the costs of the replication study. If studies *A*, *B*, and *C* all have the same *replication value*, but replications of each study differ in their costs, and we have the resources to replicate either only study *A* or both studies *B* and *C*, then all else equal we will gain most utility if we replicate studies *B* and *C*, instead of study *A*. In utility theory this idea is known as marginal utility per dollar. We choose to perform the replication study that provides the largest increase in scientific knowledge per dollar spent on the study. All else equal, the lower the cost of a replication study, the higher the gain in utility per dollar. Note that "per dollar" is a simplistic turn of phrase in this setting, since costs can also refer to non-monetary resources such as the amount of expertise we need to gain, or the amount of work-hours we have to spend.

Sometimes the *costs* of a replication study are so high that it is not feasible to replicate the study (e.g., access to the required population would take decades or more money than is available). That the cost of a study can preclude replication is represented by the negative arrow *costs → replication* in Figure 1. When a study is feasible, we can usually spend resources to improve the quality of the replication and increase the reduction in uncertainty. This can be done for instance by recruiting more participants to increase statistical power, or by conducting more extensive pilot work to validate measures and perform manipulation checks. This is represented by the negative arrow from *costs → uncertainty after replication* in Figure 1.

Once we take costs and the ability of the replication to reduce uncertainty into account in our study selection strategy, we can consider not only the maximum increase in expected utility that could be gained (*replication value*) but also the

[6] As long as *uncertainty after replication* is marginally independent of *replication value*, *uncertainty after replication* will simply introduce positive noise at random every time *replication value* is used to predict *expected utility gain*. The average positive shift is cancelled out if we consider only the rank-order of these variables. All we are left with then is noise due to random variation in the effect size *Replication → Uncertainty after replication* across claims. This random noise will tend to distort the rank-order of *expected utility gain* relative to the rank-order of *replication value* across claims, making *replication value* a less reliable estimator of *expected utility gain*. However, since the noise is random it will not bias the rank-order estimates in any particular direction.

predicted increase in expected utility after performing a specific replication study. In utility theory, this idea is called the expected value of sample information (Clemen, 1996): How much will the expected utility of our decisions based on claims increase if we add the results of a replication study to our scientific knowledge? All else equal, we would replicate the claims where expected utility increases the most following replication.

In the following sections we will discuss the possibility of estimating *replication value* quantitatively, and we consider some practical challenges of using *replication value* as a tool for choosing a study to replicate from among several candidates. For simplicity, we will omit considerations of *costs* in this discussion, and we will assume that rank-order *replication value* is an unbiased estimator of rank-order *expected utility gain* (i.e. we assume that *replication value* is independent of the size of the causal effect *replication* → *uncertainty after replication*, in the model in Figure 1).

4. Quantitative formulas for estimating replication value

Starting from the model defined in the previous section, we argue that it is both possible and desirable to develop quantitative formulas for estimating *replication value*. Formula values can be used as a basis for formalized replication study selection procedures (e.g., Pittelkow et al., 2020). A formalized procedure means the steps that together describe how selection between candidate studies will be performed are clearly defined and standardized (e.g., "the *n* studies with the highest *replication value* based on formula *Y* will be chosen for replication"). Such procedures are transparent about how studies will be selected. They can hence be applied consistently to all candidate studies. Different stakeholders might disagree on which selection procedure would be the most valid or efficient. However, a transparent and formalized decision process should at least make it easy to identify sources of disagreement, and make it possible to resolve disagreements by modifying the *replication value* formula or selection procedure. Finally, because quantitative estimates of (rank-order) *replication value* are easier to derive than evaluations based on qualitative review of the literature supporting a claim, study selection procedures based on quantitative estimates of *replication value* can be applied even in cases where the number of replication candidates makes qualitative evaluation unfeasible.

To quantify *replication value* we first need to operationalize the *value* and *uncertainty of original claims before replication*. This will be challenging, as *value* and *uncertainty* are both multi-faceted constructs (much like "intelligence" or "socioeconomic status"), whose state likely depends on a combination of several observable variables. In addition, since *value* is subjective, the *value* of a claim (and, by extension, the *replication value* of the claim) will depend on who is doing the evaluation. Resolving these measurement problems is beyond the scope of this paper. Here we simply suggest a few quantitative variables that are highly likely to be related to *value* and *uncertainty* in many contexts.

The scientific and societal impact of a claim are widely considered to be important indicators of the claim's *value* (Isager, 2018; Mueller-Langer, Fecher, Harhoff, & Wagner, 2019; Royal Netherlands Academy of Arts and Sciences, 2018). Quantitative indicators of *value* might therefore include citation counts (Aksnes, Langfeldt, & Wouters, 2019; Lewandowsky & Oberauer, 2020), Altmetric Attention scores (Bornmann, 2014), journal impact indicators (Garfield, 2006; but see Oh & Lim, 2009), best paper awards, citation by textbooks or clinical guidelines or public policy, reviewer ratings of importance and novelty, etc.[7] An operationalization of *value* could also include a utility function to represent subjective value.

Quantitative indicators of *uncertainty before replication* could include sample size (Fraley & Vazire, 2014), Bayesian posterior belief or Bayes factors (Field et al., 2019; Hardwicke et al., 2018), number of prior replications (Matiasz et al., 2018) prediction market ratings of replicability (Dreber et al., 2015), variance of effect estimates, statistical power of existing studies of the claim, prevalence of reporting errors, statistical bias estimates, etc.

Once it has been decided how to operationalize *value* and *uncertainty before replication*, we will need to decide how to combine these two indicators into an overall estimate of the *replication value* of a claim. Following our model, which is based on decision theory, the two terms should be multiplied (see the structural equation for *replication value* in Table 1).

As a purely hypothetical example, suppose we operationalized the *value* of the claim as a concave utility function of the Altmetric Attention score of the paper the study is published in, and *uncertainty before replication* as a function of the probability given by a prediction market that the claim will replicate. The *replication value* based on these parameters could then be calculated as:

$$RV = f(Altmetric) \times (1 - 2|0.5 - P_{PM}|)$$

where RV is the replication value, $f(Altmetric)$ is a concave function of the Altmetric Attention score, P_{PM} is the prob-

[7] Note that impact metrics are not part of the value construct as such. Increasing the citation count or Altmetric Attention score associated with a claim does not necessarily make the claim more valuable. Such indicators are only valid for measuring value to the extent that we tend to cite valuable claims more often than less valuable claims. Ideally we would quantify indicators that are more directly related to value, such as the importance of the claim for scientific theory, or the amount of human suffering that could be reduced by policy based on the claim.

ability that the claim will replicate given by the prediction market, and the function $(1 - 2|0.5 - P_{PM}|)$ is a transformation of the prediction market probability that the claim will replicate. The transformation is needed to create a measure of *uncertainty before replication* that equals 1 when the prediction market is completely certain either that the study will replicate ($P_{PM} = 1$) or that the study will not replicate ($P_{PM} = 0$), and that equals 0 when the prediction market is maximally uncertain about the replicability of the study ($P_{PM} = 0.5$). Indicators might often need to be transformed to behave in line with the definitions of *value*, *uncertainty before replication* and *replication value* given by the model presented here. For additional examples of how *replication value* could be quantified, consult the supplementary "RV formula" documents on OSF (https://osf.io/asype/).

Several existing quantitative procedures for selecting studies for replication could be viewed as special instances of the model proposed in this paper, given a few additional assumptions. For example, quantitative comparison of replication candidates based on Bayes factors proposed by Field et al. (2019) could be considered an application of our model in which *uncertainty before replication* is operationalized in terms of Bayes factors and *value* is assumed to be constant across claims. In other words, this strategy assumes that all candidate claims are equally valuable, and only uncertainty ought to influence *replication value* estimates.

Conversely, proposed approaches that rely on citation metrics and other indicators of impact to guide replication study selection (e.g., Makel et al., 2012) could be considered an application of our model that operationalizes *value* in terms of impact indicators and holds *uncertainty before replication* constant. In other words, these approaches assume that all candidate claims have an equal degree of *uncertainty before replication*, and only the *value* of the claims should influence *replication value* estimates.

Researchers, journal editors and funding bodies may choose different quantitative operationalizations because their priorities differ. For example, a funding body that wants to support practical applications of claims may opt to quantify *value* as the number of patents or clinical interventions generated based on the knowledge considered. Furthermore, the same funding body might change their definition of *value* based on context. They may adopt one definition for funding instruments that support practical applications, and another for funding instruments that support basic research. Thus, we can acknowledge that the exact determination of *replication value* is subjective and changes based on the context and goals of the research, and still adopt a formalized approach to replication study selection.

Finally, we should note that it is wise to combine quantitative estimation and qualitative evaluation during study selection. First, many factors that determine the uncertainty and value of a claim cannot easily be quantified, such as concerns about questionable research practices used in the original study, or the importance of a certain observational fact for a theory. However, such factors can be qualitatively evaluated by the replicating researcher and inform the decision as to whether a study is worth replicating. Second, replication value does not, by definition, consider if and what kind of replication study would reduce uncertainty about claims from the original study. However, the replicating researcher will of course want to consider factors related to the effect of *replication* on *uncertainty after replication*. For example, it is important to consider whether the original study design is of sufficient quality so that a replication of this design will be informative. Because qualitative assessment tends to be more time-intensive than quantitative estimation, we expect that two-stage selection strategies will be most efficient, in which quantitative replication value formulas are used to create a manageable list of promising candidates that can then be qualitatively evaluated before a candidate is chosen for replication. In fact, selection strategies based on a mix of quantitative and qualitative information have already been proposed (Field et al., 2019; Pittelkow et al., 2020).

5. Challenges and limitations

Throughout this article we have assumed that the goal of replication research is to maximize gain in expected utility of claims through replication. However, utility maximization is not always the goal of replication. Consider the Reproducibility Project: Psychology, the goal of which was to accurately estimate the overall replication rate of empirical findings published in flagship psychology journals (Open Science Collaboration, 2015). This goal is not reconcilable with the decision model we outline here. Accurate estimation of replication success rates depends on random sampling of studies from the target population (Kuehberger & Schulte-Mecklenbeck, 2018). Selecting studies based on *replication value* prevents random sampling of studies and introduces selection bias by design. In other words, the usefulness of the model proposed herein – as well as any specific study selection strategy derived from it – is strictly limited by the goal we have assumed. Researchers aiming to reach different goals will consequently need different decision models and different study selection strategies.

Assuming that the goal of replication is utility maximization, three primary challenges in using *replication value* for study selection are (1) deciding what information is relevant for measuring *value* and *uncertainty before replication*, (2) combining this information into a single judgement about *replication value*, and (3) evaluating the validity of this approach for estimating *expected utility gain*. We know from the literature that multiple sources of information can be used to evaluate

value and *uncertainty before replication*. Some factors feature more commonly than others, such as citation count as an indicator of *value*, and the width of confidence intervals around effect sizes as an indicator of *uncertainty before replication* (Isager, 2018). We need to investigate whether such factors are valid measures of *value* and *uncertainty before replication* in different replication contexts. For example, confidence intervals may not be valid measures of uncertainty when we suspect that data have been selectively or fraudulently reported. Citation impact may be a more valid measure of *value* in some research fields than in others. Furthermore, in most cases, the use of field-weighted citation counts might be preferable to absolute citation counts (Purkayastha, Palmaro, Falk-Krzesinski, & Baas, 2019).

Researchers may legitimately disagree which variables *should* be used to measure *value* and *uncertainty before replication* and what functional form should be used to combine these into an estimator of *replication value*. We should expect that some factors are more relevant in some fields than others. Thus, another important challenge to implementing algorithms for study selection is to identify which factors are most relevant given a particular research field or context, and which kinds of studies ought to be prioritized for replication in particular research fields. As one example, Heirene (2020) proposes factors (e.g., clinical impact) and replication targets (e.g., studies evaluating novel interventions or screening procedures) that are particularly relevant within the field of addiction research. Identifying and explicating such contextual factors will likely be an important precondition to formalized replication study selection in any scientific field.

Once we have decided how we want to operationalize *value* and *uncertainty before replication* and combine these to define *replication value*, we need to verify that *replication value* is a valid and reliable measure of *expected utility gain*. In other words, we need to make sure that replicating the studies with the highest estimated *replication value* consistently causes us to maximize the expected utility of our replication efforts. Partly, this depends on valid operationalizations of *value* and *uncertainty before replication*. However, we also need to know whether *replication value* alone is sufficient to estimate *expected utility gain*, or whether the other causal determinants of *expected utility gain – costs* and effect of *replication* on *uncertainty after replication* – must be measured as well. It is, for example, possible to have a valuable and uncertain claim for which a *replication* will do nothing to reduce uncertainty. Suppose that our uncertainty about a claim stems primarily from the low quality of the original research design used to test that claim, which would presumably be repeated in the replication. In such a case *replication value* becomes a poor predictor of *expected utility gain* since *replication* of a low-quality study design would not reduce our uncertainty about a claim much, regardless of what the *replication value* of the claim is.

Any operationalization of *replication value* will require validation. At the very least, we should make sure that our assessment strategy will often indicate a high *replication value* for claims that we are intuitively confident would be worth replicating, and a low *replication value* for claims we are intuitively confident would not be worth replicating. More severe validation studies would certainly be desirable, though we are not at present sure what such studies would look like.

In practice, we might also want to entertain the idea that quantitative estimates of replication value could be "gamed" to achieve goals not in line with maximizing utility of existing research. Consider a funder who, based on the example formula presented in section 4, sets a threshold replication value that must be achieved before a replication study will receive funding. A team of researchers who have already decided on a study to replicate, and are not interested in exploring alternative candidates, might attempt to artificially inflate the replication value of the original study to meet the funder's criterion. For example, the researchers could add links to the original study in blog- or social media posts to increase the Altmetric score of the article. Or they could try to influence the opinions of the prediction market that assigns the value of P_{PM}. Such practices would almost certainly compromise the validity of replication value estimates for predicting expected utility, in a very similar way to how p-hacking compromises the validity of the p-value as an inferential statistic.

Finally, a decision-theoretical approach to study selection could be extended to include higher level questions such as whether resources are best spent on a replication study or a novel study, or even which research lines should be prioritized given limited resources. A fully developed decision-theoretical model of study selection should allow us to consider the utility of different potential research activities, such as measurement validation, examining computational reproducibility, testing the generalizability of findings, or studying a novel theoretical prediction. The model we propose is a component of such a full model of study selection, focusing on a specific decision, and does not currently assist researchers in other types of decisions that need to be made.

Replication value can only be used to evaluate a number of replication candidates relative to each other. It cannot be used to evaluate whether a replication of an existing study would be more useful than a novel study. Deciding between a replication study and a novel study would require resolving important questions about the goal of data collection, about the factors that determine the importance of a novel research question, and about ways to quantify the uncertainty about a novel theoretical prediction. Although such decision processes occur in practice (e.g., at CERN where only a small set of all possible research questions can be empirically exam-

ined in the Large Hadron Collider), quantifying the value of novel research questions is itself a big (but possibly valuable) challenge for future research.

Similarly, *replication value* can only be used to maximize utility within the set of replication candidates under consideration. It can be used to guide decisions about which candidate in the set to replicate but it does not necessarily help us select a good set of studies to select from, which can limit our ability to achieve the goal of utility maximization. For instance, if a candidate set consists entirely of the least valuable claims in a research field, maximizing expected utility would likely be better achieved by picking a new set than by selecting for high *replication value* claims within the set. Thus, the choice of candidates to compare places an important practical constraint on the usefulness of study selection strategies based on *replication value*.

6. Conclusion

Assuming that many claims are in need of replication, but resources for conducting replication studies are limited, we need to decide which claims to replicate first. For situations when the goal of replication study selection is to maximize the expected utility gain of the replication effort, we propose that several pieces of information are crucial for making this decision - the value of having knowledge about the research claim, the uncertainty of our current knowledge about the claim, the ability of the replication to reduce uncertainty (replication quality), and the costs of conducting the replication. These factors are frequently considered both in theoretical discussions of replication study selection, and during actual study selection in replication projects. Using well-known concepts from the framework of utility theory, we propose a general decision model for study selection in replication research, and a formal definition of *replication value*. We also suggest ways in which quantitative formulas could be derived from this definition and used to generate formalized study selection procedures.

Our decision model should be helpful for anyone who wishes to maximize the *expected utility gain* of replication efforts under resource constraints, including individual replication-oriented researchers and labs (e.g., Feldman, 2021), large-scale collaborations with limited resource capacities (e.g., Paris, IJzerman, & Forscher, 2020), replication funders with limited grant resources (e.g., "Replication Studies," n.d.), and metascientists in the business of developing formal study selection strategies (e.g., Field et al., 2019). In general, we believe that our model will be helpful in structuring the discussion of how replication studies should be selected, because it makes our assumptions about the function and goal of replication research clear and explicit. Clear assumptions, in turn, make it easier to explain and identify sources of disagreement about how a certain quantitative metric is expected to work, which should make future discussion about study selection strategies more productive. Thinking clearly about the value of replication studies should also help individual researchers to more clearly formulate why they are replicating a study, even when their approach to study selection is not as formal as what we propose here. We hope that our model can be used as a foundation for creating concrete study selection procedures that will enhance the transparency, consistency, and efficiency of future replication research.

Disclosures

Conflicts of interest. The authors declare that they have no conflicts of interest with respect to the authorship or publication of this article.

Acknowledgements. This work was funded by VIDI grant452-17-013. We thank all formula proposal authors for their substantial contributions to the early stages of this project, and for their valuable comments throughout the process of writing this manuscript.

References

Agnoli, F., Wicherts, J. M., Veldkamp, C. L. S., Albiero, P., & Cubelli, R. (2017). Questionable research practices among italian research psychologists. *PLOS ONE*, *12*(3), e0172792. https://doi.org/10.1371/journal.pone.0172792

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), 215824401982957. https://doi.org/10.1177/2158244019829575

Banks, G. C., O'Boyle, E. H., Pollack, J. M., White, C. D., Batchelor, J. H., Whelpley, C. E., ... Adkins, C. L. (2016). Questions About Questionable Research Practices in the Field of Management: A Guest Commentary. *Journal of Management*, *42*(1), 5–20. https://doi.org/10.1177/0149206315619011

Baribault, B., Donkin, C., Little, D. R., Trueblood, J. S., Oravecz, Z., van Ravenzwaaij, D., ... Vandekerckhove, J. (2018). Metastudies for robust tests of theory. *Proceedings of the National Academy of Sciences*, *115*(11), 2607–2612. https://doi.org/10.1073/pnas.1708285114

Block, J., & Kuckertz, A. (2018). Seven principles of effective replication studies: Strengthening the evidence base of management research. *Management Review Quarterly*, *68*(4), 355–359. https://doi.org/10.1007/s11301-018-0149-3

Bornmann, L. (2014). Validity of altmetrics data for measuring societal impact: A study using data from Altmetric and F1000Prime. *Journal of Informetrics*, *8*(4), 935–950. https://doi.org/10.1016/j.joi.2014.09.007

Brand, A., Allen, L., Altman, M., Hlava, M., & Scott, J. (2015). Beyond authorship: Attribution, contribution, collaboration, and credit. *Learned Publishing*, *28*(2), 151–155. https://doi.org/10.1087/20150211

Brandt, M. J., IJzerman, H., Dijksterhuis, A., Farach, F. J., Geller, J., Giner-Sorolla, R., … van 't Veer, A. (2014). The Replication Recipe: What makes for a convincing replication? *Journal of Experimental Social Psychology*, *50*, 217–224. https://doi.org/10.1016/j.jesp.2013.10.005

Cesario, J. (2014). Priming, Replication, and the Hardest Science. *Perspectives on Psychological Science*, *9*(1), 40–48. https://doi.org/10.1177/1745691613513470

Clemen, R. T. (1996). *Making hard decisions: An introduction to decision analysis* (2nd ed). Belmont, Calif: Duxbury Press.

Coles, N. A., Tiokhin, L., Scheel, A. M., Isager, P. M., & Lakens, D. (2018). The costs and benefits of replication studies. *Behavioral and Brain Sciences*, *41*, E124. https://doi.org/10.1017/S0140525X18000596

Doyen, S., Klein, O., Pichon, C.-L., & Cleeremans, A. (2012). Behavioral Priming: It's All in the Mind, but Whose Mind? *PLoS ONE*, *7*(1), e29081. https://doi.org/10.1371/journal.pone.0029081

Dreber, A., Pfeiffer, T., Almenberg, J., Isaksson, S., Wilson, B., Chen, Y., … Johannesson, M. (2015). Using prediction markets to estimate the reproducibility of scientific research. *Proceedings of the National Academy of Sciences*, *112*(50), 15343–15347. https://doi.org/10.1073/pnas.1516179112

Earp, B. D., & Trafimow, D. (2015). Replication, falsification, and the crisis of confidence in social psychology. *Frontiers in Psychology*, *6*, 621. https://doi.org/10.3389/fpsyg.2015.00621

Ebersole, C. R., Atherton, O. E., Belanger, A. L., Skulborstad, H. M., Allen, J. M., Banks, J. B., … Nosek, B. A. (2016). Many Labs 3: Evaluating participant pool quality across the academic semester via replication. *Journal of Experimental Social Psychology*, *67*, 68–82. https://doi.org/10.1016/j.jesp.2015.10.012

Errington, T. M., Iorns, E., Gunn, W., Tan, F. E., Lomax, J., & Nosek, B. A. (2014). An open investigation of the reproducibility of cancer biology research. *eLife*, *3*, e04333. https://doi.org/10.7554/eLife.04333

Fanelli, D. (2010). "Positive" Results Increase Down the Hierarchy of the Sciences. *PLOS ONE*, *5*(4), e10068. https://doi.org/10.1371/journal.pone.0010068

Fanelli, D. (2012). Negative results are disappearing from most disciplines and countries. *Scientometrics*, *90*(3), 891–904. https://doi.org/10.1007/s11192-011-0494-7

Feldman, G. (2021). Mass Replications & Extensions (CORE). https://web.archive.org/web/20210301102821/https://mgto.org/registered-replications/.

Fiedler, K., & Schwarz, N. (2016). Questionable Research Practices Revisited. *Social Psychological and Personality Science*, *7*(1), 45–52. https://doi.org/10.1177/1948550615612150

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

Finkel, E. J., Eastwick, P. W., & Reis, H. T. (2017). Replicability and other features of a high-quality science: Toward a balanced and empirical approach. *Journal of Personality and Social Psychology*, *113*(2), 244–253. https://doi.org/10.1037/pspi0000075

Fraley, R. C., & Vazire, S. (2014). The N-pact factor: Evaluating the quality of empirical journals with respect to sample size and statistical power. *PLoS ONE*, *9*(10). https://doi.org/10.1371/journal.pone.0109019

Frank, M. C., Bergelson, E., Bergmann, C., Cristia, A., Floccia, C., Gervain, J., … Yurovsky, D. (2017). A Collaborative Approach to Infant Research: Promoting Reproducibility, Best Practices, and Theory-Building. *Infancy*, *22*(4), 421–435. https://doi.org/10.1111/infa.12182

Garfield, E. (2006). The History and Meaning of the Journal Impact Factor. *JAMA*, *295*(1), 90–93. https://doi.org/10.1001/jama.295.1.90

Giner-Sorolla, R., Amodio, D. M., & van Kleef, G. A. (2018). Three strong moves to improve research and replications alike. *Behavioral and Brain Sciences*, *41*, e130. https://doi.org/10.1017/S0140525X18000651

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

Hardwicke, T. E., Tessler, M. H., Peloquin, B. N., & Frank, M. C. (2018). A Bayesian decision-making framework for replication. *Behavioral and Brain Sciences*, *41*, e132. https://doi.org/10.1017/S0140525X18000675

Heirene, R. (2020). *A call for replications of addiction research: Which studies should we replicate & what constitutes a "successful" replication?* (Preprint). PsyArXiv. https://doi.org/10.31234/osf.io/xzmn4

Isager, P. M. (2018). What To Replicate? Justifications Of Study Choice From 85 Replication Studies. https://doi.org/10.5281/zenodo.1286715

JESP Registered Reports Guidelines. (n.d.). Elsevier.

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the Prevalence of Questionable Research Practices With Incentives for Truth Telling. *Psychological Science*, *23*(5), 524–532. https://doi.org/10.1177/0956797611430953

Journal of Personality and Social Psychology - APA Publishing | APA. (n.d.). *https://www.apa.org*. https://www.apa.org/pubs/journals/psp/index.

Kitcher, P. (1995). *The advancement of science: Science without legend, objectivity without illusions* (1. Pb. publ). New York: Oxford Univ. Press.

Koole, S. L., & Lakens, D. (2012). Rewarding Replications: A Sure and Simple Way to Improve Psychological Science. *Perspectives on Psychological Science*, *7*(6), 608–614. https://doi.org/10.1177/1745691612462586

Kuehberger, A., & Schulte-Mecklenbeck, M. (2018). Selecting target papers for replication. *Behavioral and Brain Sciences*, *41*. https://doi.org/10.1017/S0140525X18000742

Lakens, D. (2016). The 20% Statistician: The Replication Value: What should be replicated? *The 20% Statistician*.

Landreth, A., & Silva, A. J. (2013). The need for research maps to navigate published work and inform experiment planning. *Neuron*, *79*(3), 411–415. https://doi.org/10.1016/j.neuron.2013.07.024

LeBel, E. P., Borsboom, D., Giner-Sorolla, R., Hasselman, F., Peters, K. R., Ratliff, K. A., & Smith, C. T. (2013). PsychDisclosure.Org: Grassroots Support for Reforming Reporting Standards in Psychology. *Perspectives on Psychological Science*, *8*(4), 424–432. https://doi.org/10.1177/1745691613491437

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

LeBel, E. P., & Peters, K. R. (2011). Fearing the Future of Empirical Psychology: Bem's (2011) Evidence of Psi as a Case Study of Deficiencies in Modal Research Practice. *Review of General Psychology*, *15*(4), 371–379. https://doi.org/10.1037/a0025172

Lewandowsky, S., & Oberauer, K. (2020). Low replicability can support robust and efficient science. *Nature Communications*, *11*(1), 358. https://doi.org/10.1038/s41467-019-14203-0

Lindsay, D. S. (2015). Replication in Psychological Science. *Psychological Science*, *26*(12), 1827–1832. https://doi.org/10.1177/0956797615616374

Lindsay, D. S. (2017). Preregistered Direct Replications in Psychological Science. *Psychological Science*, *28*(9), 1191–1192. https://doi.org/10.1177/0956797617718802

Lodder, P., Ong, H. H., Grasman, R. P. P. P., & Wicherts, J. M. (2019). A comprehensive meta-analysis of money priming. *Journal of Experimental Psychology: General*, *148*(4), 688–712. https://doi.org/10.1037/xge0000570

Machery, E. (2020). What Is a Replication? *Philosophy of Science*, *87*(4), 545–567. https://doi.org/10.1086/709701

Mackey, A. (2012). Why (or why not), when and

how to replicate research. In G. Porte (Ed.), *Replication research in applied linguistics* (Vol. 2146, pp. 21–46). New York: Cambridge University Press.

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, G. N., & Clarke, R. M. (2017). Are Psychology Journals Anti-replication? A Snapshot of Editorial Practices. *Frontiers in Psychology*, *8*. https://doi.org/10.3389/fpsyg.2017.00523

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

Matiasz, N. J., Wood, J., Wang, W., Silva, A. J., & Hsu, W. (2017). Computer-Aided Experiment Planning toward Causal Discovery in Neuroscience. *Frontiers in Neuroinformatics*, *11*, 12. https://doi.org/10.3389/fninf.2017.00012

Maxwell, S. E., Lau, M. Y., & Howard, G. S. (2015). Is psychology suffering from a replication crisis? What does "failure to replicate" really mean? *American Psychologist*, *70*(6), 487–498. https://doi.org/10.1037/a0039400

Moonesinghe, R., Khoury, M. J., & Janssens, A. C. J. W. (2007). Most Published Research Findings Are FalseBut a Little Replication Goes a Long Way. *PLoS Medicine*, *4*(2), e28. https://doi.org/10.1371/journal.pmed.0040028

Morey, R., & Lakens, D. (2016). Why Most Of Psychology Is Statistically Unfalsifiable. https://doi.org/10.5281/ZENODO.838685

Mueller-Langer, F., Fecher, B., Harhoff, D., & Wagner, G. G. (2019). Replication studies in economicsHow many and which papers are chosen for replication, and why? *Research Policy*, *48*(1), 62–83. https://doi.org/10.1016/j.respol.2018.07.019

Nosek, B. A., & Errington, T. M. (2020). What is replication? *PLOS Biology*, *18*(3), e3000691. https://doi.org/10.1371/journal.pbio.3000691

Nosek, B. A., Spies, J. R., & Motyl, M. (2012). Scientific Utopia: II. Restructuring Incentives and Practices to Promote Truth Over Publishability. *Perspectives on Psychological Science*, *7*(6), 615–631. https://doi.org/10.1177/1745691612459058

NSF Invites Grant Applications Related to Reproducibility in Neuroimaging. (n.d.). *Association for Psychological Science*. https://www.psychologicalscience.org/policy/nsf-invites-grant-applications-related-to-reproducibility-in-neuroimaging.html.

Oh, H. C., & Lim, J. F. (2009). Is the journal impact factor a valid indicator of scientific value? *Singapore Medical Journal*, *50*(8), 749–751.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716–aac4716. https://doi.org/10.1126/science.aac4716

Paris, B., IJzerman, H., & Forscher, P. S. (2020). *PSA 2020-2021 study capacity report* (Preprint). PsyArXiv. https://doi.org/10.31234/osf.io/v9zma

Pashler, H., & Wagenmakers, E. (2012). Editors' Introduction to the Special Section on Replicability in Psychological Science: A Crisis of Confidence? *Perspectives on Psychological Science*, *7*(6), 528–530. https://doi.org/10.1177/1745691612465253

Pearl, J. (2009). *Causality: Models, reasoning, and inference*. Cambridge, U.K. ; New York: Cambridge University Press.

Peer-review policy | Nature Research. (n.d.). https://www.nature.com/nature-research/editorial-policies/peer-review#criteria-for-publication.

Pittelkow, M.-M., Hoekstra, R., Karsten, J., & van Ravenzwaaij, D. (2020). *Replication Crisis in Clinical Psychology: A Bayesian and Qualitative Re-evaluation* (Preprint). PsyArXiv. https://doi.org/10.31234/osf.io/unezq

Purkayastha, A., Palmaro, E., Falk-Krzesinski, H. J., & Baas, J. (2019). Comparison of two article-level, field-independent citation metrics: Field-Weighted Citation Impact (FWCI) and Relative Citation Ratio (RCR). *Journal of Informetrics*, *13*(2), 635–642. https://doi.org/10.1016/j.joi.2019.03.012

Raiffa, H., & Schlaifer, R. (1974). *Applied statistical decision theory* (6. print). Boston: Div. of

Research, Graduate School of Business Administration, Harvard Univ.

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

Replication Studies. (n.d.). https://www.nwo.nl/en/researchprogrammes/replication-studies.

Replication studies | Royal Society Open Science. (n.d.). https://royalsocietypublishing.org/rsos/replication-studies.

Ritchie, S. J., Wiseman, R., & French, C. C. (2012). Failing the Future: Three Unsuccessful Attempts to Replicate Bem's "Retroactive Facilitation of Recall" Effect. *PLoS ONE*, *7*(3), e33423. https://doi.org/10.1371/journal.pone.0033423

Royal Netherlands Academy of Arts and Sciences. (2018). *Replication studies Improving reproducibility in the empirical sciences,*

Scheel, A. M., Schijen, M., & Lakens, D. (2019). Positive result rates in psychology: Registered Reports compared to the conventional literature.

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

Silva, A. J., Landreth, A., & Bickle, J. (2014). *Engineering the next revolution in neuroscience: The new science of experiment planning*. Oxford: Oxford University Press.

Silva, A. J., & Müller, K.-R. (2015). The need for novel informatics tools for integrating and planning research in molecular and cellular cognition: Figure 1. *Learning & Memory*, *22*(9), 494–498. https://doi.org/10.1101/lm.029355.112

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. *Psychological Science*, *22*(11), 1359–1366. https://doi.org/10.1177/0956797611417632

Simons, D. J. (2014). The Value of Direct Replication. *Perspectives on Psychological Science*, *9*(1), 76–80. https://doi.org/10.1177/1745691613514755

Stroebe, W., & Strack, F. (2014). The Alleged Crisis and the Illusion of Exact Replication. *Perspectives on Psychological Science*, *9*(1), 59–71. https://doi.org/10.1177/1745691613514450

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, *18*(6), 643–662. https://doi.org/10.1037/h0054651

Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin*, *76*(2), 105–110. https://doi.org/10.1037/h0031322

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

Westfall, J. (2016). Designing multi-lab replication projects: Number of labs matters more than number of participants. *Cookie Scientist*.

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