Deciding What to Replicate

Deciding What to Replicate

Peder Mortvedt Isager

Department of Industrial Engineering and Innovation Sciences
Eindhoven University of Technology
Netherlands

This project was funded by Netherlands Organization for Scientific Research (NWO) VIDI Grant 452-17-013. The work described in this thesis has been carried out at the Human-Technology Interaction group, Eindhoven University of Technology.

A catalogue is available from the Eindhoven University of Technology Library

ISBN: 978-90-386-5496-6

NUR: 741

Keywords: replication, replication value, study selection, expected utility, validity, causal inference,

Cover: The front cover illustrates the central problem dealt with throughout this thesis. The replicating researcher must choose their replication target from a vast and disorganized research literature. Cover design by Ole Gunnar M. Isager.

Deciding What to Replicate

PROEFSCHRIFT

ter verkrijging van de graad van doctor aan de Technische Universiteit Eindhoven, op gezag van de rector magnificus prof.dr.ir. F.P.T. Baaijens, voor een commissie aangewezen door het College voor Promoties, in het openbaar te verdedigen op dinsdag 24 mei 2022 om 11:00 uur

door

Peder Mortvedt Isager

geboren te Lørenskog, Noorwegen

Dit proefschrift is goedgekeurd door de promotoren en de samenstelling van de promotiecommissie is als volgt:

voorzitter:	prof.dr. I.E.J. Heynderickx
1e promotor:	prof.dr. C.C.P. Snijders
copromotor(en):	dr. D. Lakens
	dr. A.E. van 't Veer
leden:	prof.dr. Denny Borsboom
	(Universiteit Amsterdam)
	prof.dr. Rolf Zwaan
	(Erasmus Universiteit Rotterdam)
	dr.ir. Krist Vaesen

Het onderzoek of ontwerp dat in dit proefschrift wordt beschreven is uitgevoerd in overeenstemming met de TU/e Gedragscode Wetenschapsbeoefening.

Summary

Replication – the act of repeating existing research designs to examine whether results can be independently reproduced – is a core element of the scientific method. The importance of replication for ensuring research quality has however been woefully neglected in many areas of science, particularly in psychology. The research community's interest in replication has been revived in the last decade, following several high-profile and large-scale replication efforts that suggest replication success rates in the field are much lower than previously assumed. The overconfidence in past results up until this point has been due in large part to publication practices that simultaneously incentivize the publication of false positive results and stifle replication attempts designed to root out false positives from the literature. In response to this less-than-ideal situation, researchers have taken great strides to facilitate replication as a research practice. Today, there are journals actively encouraging submission of replication reports, funders are slowly beginning to earmark resources for replication study proposals, and methodologically minded scholars have come far in analyzing the merits of different kinds of replication study designs.

This thesis is devoted to a question that follows in the wake of this increased focus on replication research; which studies are the most important to replicate? The importance of this question follows from three basic premises. (1) Researchers have limited resources available for conducting replication studies. (2) The number of studies that have never been replicated greatly exceeds the resources currently available for replication. (3) Non-replicated studies vary in how much they need to be replicated. These premises together imply that most replicating researchers will have more studies to choose from than they can feasibly replicate, and the choice of study matters for whether resources allocated for replication research are spent efficiently.

To answer the question of *which* studies are most important to replicate, we must have a clear understanding of *what* makes a study important to replicate. **Chapter 2** presents a conceptual analysis to answer this question, building on insights from past research and discussions of replication study selection. A theoretical model is presented that frames the question of replication study selection as an economical assessment of the value of information. *Replication*

value is defined as the maximum expected utility we can expect to gain by replicating an original study. The replication value of a study is determined by how valuable we perceive the studied claims to be, and our uncertainty about the claims after seeing the original study results. Highly valuable claims which we remain highly uncertain about after an original study is conducted have a high replication value and should be prioritized for replication if our goal is to maximize the expected utility of research output.

Chapter 4 outlines a possible quantitative operationalization of replication value, dubbed RV_{Cn}. In this operational definition, the value of a claim is operationalized in terms of citation impact. The uncertainty about a claim is operationalized in terms of the sample size of the original study (this operationalization is only valid under a specific definition of test validity, which is presented in **chapter 3**). An equation is proposed that combines citation and sample size information into an overall estimate of replication value. A quantitative definition of replication value allows for clear and precisely defined study selection rules and makes it possible to sift large bodies of literature for potentially high-value replication targets. However, to grant these benefits RV_{Cn} must be valid and reliable. To assess this, the strength and direction of the causal relationships proposed in chapter 4 must be corroborated empirically.

Chapter 5 explores the feasibility of applying RV_{Cn} in a set of empirical social psychology fMRI studies. It first reports efforts to generate a representative candidate set of replication targets in social fMRI research. It then explores the feasibility and reliability of estimating RV_{Cn} for the targets in our set, resulting in a dataset of 1358 studies ranked on their value of prioritizing them for replication. Finally, the chapter assess the face validity of this strategy for identifying the studies that would be the most important to replicate. Chapter 5 demonstrates how RV_{Cn} could be implemented in practice and provides a general framework for exploring the feasibility of formal study selection strategies. However, the chapter stops short of providing rigorous validation of RV_{Cn} as a measure of replication value, which is left for future research.

Chapter 6 takes stock of the work presented in the previous four chapters, solidifies the linkage between replication value and established theory on the *value of information*, and considers possible future avenues for research into replication value and replication study selection. In conclusion, this thesis (1) establishes a formal framework for discussing what makes a study more or less worth replicating, (2) demonstrates how quantitative operationalizations of replication value may be derived from this framework, and (3) demonstrates how the feasibility and validity of such operationalizations may be tested.

Contents

Chapter 1

Introduction

1.1 Problem introduction

Professor X receives a big grant to conduct empirical research in cancer biology. Professor X has recently become aware of research that suggests the reproducibility of published findings in this field may be compromised (Begley and Ellis, 2012; Errington et al., 2014; Wen et al., 2018; eLife, 2017). In an effort to improve on the current state of affairs, professor X decides to direct their grant resources towards replication of published preclinical studies in cancer research. Research staff is hired, lab space is acquired, and orders are made for the necessary instruments and materials. All that is left to do before work can commence is to identify and choose which studies to focus the replication effort on.

However, a quick search in bibliometric databases suggests several thousand empirical articles have been published in cancer biology in the last decade, the vast majority of which will likely consist of non-replicated original studies. This leaves professor X and their team with a problem. On the one hand, all non-replicated empirical studies would in principle be candidates for their replication effort. On the other hand, the grant does not provide sufficient resources to replicate thousands of studies. In fact only one or a handful of replication efforts can be carried out. The team must consider a very important practical question. Which studies should be prioritized for the replication effort? In other words, which studies are the most important to replicate?

The very structure or the question "which studies are most important to replicate?" suggests that the value of a replication effort is something that can be thought of in quantitative terms. *Most important* implies that there should be some way to quantify, rate or at least rank-order the relative difference in importance between different potential replication efforts. The existence

of a *most important* study implies there is also a *least important* study, and so on. Researchers may not consciously rank-order studies in this way normally. However, their ability to do so is still evident from the fact that, when pressed, they do choose among candidates for replication using quantitative and qualitative criteria to justify why the chosen study should be considered particularly worthwhile to replicate (compared with an explicit or implicit list of alternative studies, Isager, 2018).

That we can quantify relative replication importance and use quantitative estimates to devise strategies for study selection is a fundamental assumption of this thesis, as well as a number of other attempts to devise study selection guidelines in replication research (e.g., Field et al., 2019; Pittelkow et al., 2020; Matiasz et al., 2018). But what exactly are we trying to quantify? What information determines whether research is important to replicate, and how could this information be estimated and used for replication study selection in practice? Finally, for any given study selection strategy, how would we know that it works as intended? These are the primary research questions of this thesis.

By developing and validating principled strategies for replication study selection, we unlock a powerful tool for research coordination that all stakeholders in the replication research space could make use of. First and foremost, researchers like professor X could use such strategies to efficiently decide which studies to concentrate time and grant resources on. Similarly, funding bodies that issue replication grants could use the same strategies to decide which proposed replication efforts to fund in the first place. Furthermore, a formal theory of replication study selection is a very effective communication device. When the rationale behind replication study selection is clearly specified it becomes much easier for all stakeholders to discuss, argue, and come to agreement about what research would be most important to direct grant resources, research time, and journal space towards. Such coordination will eventually benefit the most important stakeholder in science; the general public. Whenever science is publicly funded, it is the duty of the scientific community to ensure that the public funds invested in research are spent responsibly and efficiently. Principled study selection strategies represent an important step towards ensuring such efficiency in replication research.

1.2 Expected value of information

The idea of quantifying the potential value of information in order to make research priorities more efficient was already considered over one hundred years ago by Charles Saunders Peirce (Peirce, 1967; Wible, 1994):

> The doctrine of economy, in general, treats of the relations between utility and cost. That branch of it which relates to research consid-

> ers the relations between the utility and the cost of diminishing the probable error of our knowledge. Its main problem is, how, with a given expenditure of money, time, and energy, to obtain the most valuable addition to our knowledge.

Following this assertion, Peirce proceeds to lay out a quantitative framework for calculating the utility associated with various potential research efforts, and to show how such a framework can be used to identify the research effort that would be the most efficient to conduct, assuming the goal is to maximize increase in utility.

Peirce is far from alone in thinking about knowledge gain in economic terms. Indeed, an entire branch of economics, housed under the broader umbrella of utility theory, is devoted to analysis of the *value of information* (*VoI*. Clemen, 1996; Raiffa et al., 1961; Wilson, 2015; Eckermann et al., 2010). The basic premise for all *VoI* analysis is a decision scenario with multiple possible decisions and multiple decision outcomes that can be valued more or less by the decision-maker. The decision outcomes will usually depend both on what the decision-maker decides to do, and on the state of the decision-makers world. As an example, we may or may not bring an umbrella to work (two potential decisions), and it may or may not rain on our way to work (two possible states of the world). There are thus four potential decision outcomes (bring umbrella and it rains, bring umbrella and it does not rain, do not bring umbrella and it rains, do not bring umbrella and it does not rain) which we attach different utilities to (we prefer to bring an umbrella when it rains and to leave it at home when it does not rain).

The more certain we are about the state of the world, the easier it becomes to optimize decision making; if we have perfect knowledge about when it will rain, we always know when to bring an umbrella to work. The state of the world is often something we are not perfectly certain about. However, we can often gather more information to reduce uncertainty about the state of the world (e.g., we can invest in meteorological research to obtain more accurate information about which days it will rain). The goal of *VoI* analysis is to assess the value of reducing uncertainty about the world. Value of information is quantified as the increase in expected utility following decreases in uncertainty. *VoI* analysis is usually used to guide decisions in applied settings (e.g., Heath et al., 2020; Clemen, 1996, see example on p. 448), where value can be quantified as monetary gain, avoidance of monetary loss, lives saved (or other quantifications of quality of life; e.g., Whitehead and Ali, 2010), adverse effects avoided, etc., and where the decision that is going to be made based on the research can be clearly defined (e.g., subject a patient to this or that intervention, buy stocks in this or that company, seed a hurricane or not, etc.).

Since replication is an act of information gathering, it makes sense to consider the value of replication as a specific instance of value of information, and to

use principles from utility theory to help us make decisions about what to replicate. However, the exact decisions research will inform, and the concrete utility of different decision outcomes, is usually not obvious in basic social and behavioral research, which is perhaps partly why *VoI* analysis is not more widely adopted for replication study selection in these areas. This thesis will offer an alternative implementation of *VoI* logic to determine which studies to prioritize for replication that can potentially overcome the problem of fuzzy decision contexts. The exact relationship between traditional *VoI* analysis and the framework put forward in this thesis is formally worked out in the final chapter of the thesis.

1.3 Replication in the last decade of psychological science

In the past decade, there has been a renewed surge of interest in replication research – particularly in close/direct replication (LeBel et al., 2018; Schmidt, 2009) – in fields such as psychology, behavioral economy, and areas in social, behavioral, and medical science (Zwaan et al., 2018; Plucker and Makel, 2021; Button et al., 2013; Blaszczynski and Gainsbury, 2019; Heirene, 2021; Sale and Mellor, 2018; Murphy et al., 2021; Poldrack et al., 2017). The push for increased focus on replication is downstream consequence of the "replication crisis" – an ongoing period in the social sciences where the robustness of published empirical research in the field is being called into questions following high-profile failures to replicate highly impactful studies in psychology (e.g., Ritchie et al., 2012; Wagenmakers et al., 2016; Ranehill et al., 2015; Hagger et al., 2016), estimates of low replication success in several fields (e.g., Open Science Collaboration, 2015; Klein et al., 2014; Camerer et al., 2018; Errington et al., 2014) evidence of widespread publication bias (Scheel et al., 2019; Franco et al., 2014) and other questionable research practices (Gopalakrishna et al., 2021), etc.

Replication is seen as one tool in a larger toolkit to improve the quality and reliability of social and behavioral research (Munafò et al., 2017; Nosek et al., 2012). This makes sense. Replication can improve the reliability of research in multiple ways. It provides additional data on a phenomenon, improving the precision of original estimates. If performed by an independent team of researchers, it can aid in verifying that observed results are genuinely produced by the study methods, and are not influenced by personal biases in the original research team. If performed under variations in conditions thought not to matter for the phenomenon of interest it can also be used to study the robustness and boundary conditions of an effect (Baribault et al., 2018). For the past decade, metascientific research has primarily focused on working out best practices for performing high quality replication research, including preregis-

tration of methods, open data and materials, clarifying the closeness of the replication effort, clarifying the interpretation of replication results, etc. (e.g., Brandt et al., 2014; Morey and Lakens, 2016; LeBel et al., 2018; Nosek and Errington, 2020). This line of research has been quite successful, and in just a few years we have gained substantial knowledge about what can and cannot be learned from particular replication study designs. However, as scientists increasingly value replication studies but have limited resources to replicate the exponentially growing literature of original non-replicated research, they must also address the important problem of deciding which out of several potential replication efforts to spend their limited resources on.

In principle, we may wish to replicate every original study in the literature, ideally even multiple times. After all, there is a risk for any study that the observed results are a statistical fluke that are not representative of the population in the long run. Similarly, any study result could be influenced by particulars of the study context - the time, the place, the exact sample of subjects chosen, etc. Beyond the challenge random and contextual variation presentes, without replication any study results must be accepted based on trust in the original study authors' ability to summarize their research accurately, honestly, and transparently. Replication is a cornerstone of science because it provides a method for replacing this fundamental reliance on trust. Taking nobody's word for it, any study could thus potentially benefit from close replication by independent authors.

In practice, however, we are currently generating original research findings at a rate that far exceeds resources available for replication. The scientific literature is expanding at an exponential rate (Bornmann and Mutz, 2015). At the same time, replication of original research remains a rarity in several research fields, such as psychology (Makel et al., 2012), economics, (Mueller-Langer et al., 2019) and cognitive neuroscience (Poldrack et al., 2017). In these fields, this state of research has led to an ever-increasing pool of non-replicated original research findings. The size of this pool far exceeds the resources we currently have available for conducting replications. Therefore, while replication may in principle be something we wish to subject all scientific research to, in practice researchers, funders, and other stakeholders in science will need to decide which studies, findings, and claims are the most important to replicate.

In psychology, early efforts to address the problem of study prioritization include the Psychfiledrawer top-20 list of studies nominated for replication (Psychfiledrawer, 2014). Any researcher in the community could nominate any study they felt required replication to be included on the Psychfiledrawer list. Visitors to the site could then "upvote" each nomination into the top-20 list and discuss the merits of each nomination openly. The Psychfiledrawer was a laudable early effort to establish bottom-up coordination of replication efforts in psychology by providing a platform for the community to nominate,

rank-order and discuss the importance of potential replication candidates. Unfortunately, the platform was eventually discontinued.

Early efforts to coordinate replication priorities also include a collaborative effort by multiple authors of the Open Science Collaboration to define and test quantitative formulas for estimating *replication value* that could be used to identify particularly important replication targets. These early efforts provided the foundations for the work presented in this thesis, and the formulas developed by the original Open Science Collaboration team are included in the supplementary materials of chapter 2: https://osf.io/asype/). More recent efforts to formalize strategies for replication study selection have also focused on the idea that replication value can be quantified (Field et al., 2019; Pittelkow et al., 2020; Makel et al., 2012).

Authors of replication manuscripts in psychology sometimes implicitly express a similar belief when they justify the selection of a study based on quantitative heuristics such as the number of citations, the number of participants, the width of confidence intervals, etc. (see Isager, 2018, for a review). However, early efforts to operationalize and estimate replication value have been severely limited by the lack of a formal, unifying explanation of the goal replication study selection is supposed to achieve. In lieu of an underlying theoretical rationale it is impossible to assess whether any criteria used for replication study selection are appropriate.

1.4 Thesis outline

1.4.1 Thesis scope and aims

In this thesis I develop a formal model of the decision process underlying replication study selection. This model makes it possible to derive a transparent and formalized definition of *replication value*. The decision model is outlined in its entirety in chapter 2, and is founded on principles adopted from VoI analysis. In this model the goal of replication study selection is clearly stated, all key factors involved in the study selection process are identified, and the relationships between factors are formally defined. A further aim of this thesis is to demonstrate how the model can be used to justify different strategies for replication study selection that are feasible to implement in practice. Chapter 4 presents an estimator of replication value based on observable quantitative information. Importantly, this estimator can be directly related to the decision model developed in chapter 2, yielding a straight-forward rationale for how the estimator is intended to work. Chapter 4 also outlines and discusses the measurement assumptions that must be made explicit to connect the latent constructs in chapter 2 with observable variables. This discussion demanded a novel definition of measurement validity, which is presented in chapter 3.

The definition of validity utilized in this thesis is a minor improvement on the realist definition of validity proposed by Borsboom et al. (2004). A final goal of this thesis is to assess whether to study selection strategies based on the estimator proposed in chapter 4 in practice. An attempt to implement the estimator in the context of social fMRI research is presented in chapter 5. In summary, the overarching aim of this thesis is to provide a strategy for replication study selection that targets a specific and clearly stated research goal, that is built on formal decision theory, that is practically feasible to implement, and that can (in principle) be falsified by data.

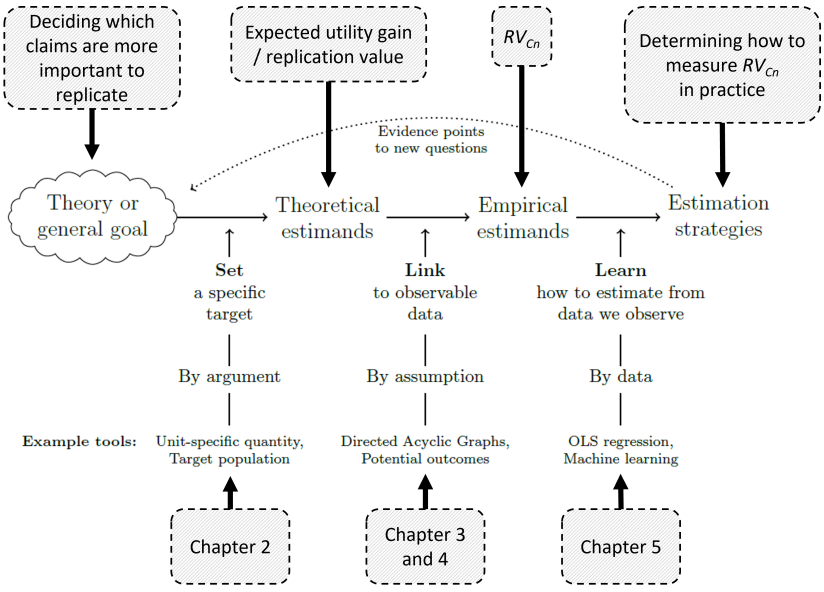

Figure 1.1: Diagram of critical measurement choices in quantitative social science, adapted from Lundberg et al. (2020). Dashed grey boxes illustrate how each element of the model is related to the chapters and central problems and concepts presented in this thesis. The only aspect of the model left for future research is the collection of validating evidence that could be used to corroborate the arguments and assumptions laid out in this thesis ("evidence points to new questions" in figure).

In general terms, the first aim of the thesis is then to define 'replication value' as an estimand, and the second aim of the thesis is to operationalize an estimator of this estimand. The overall line of research presented in this thesis closely follows the general framework for estimand definition proposed by Lundberg et al. (2020, see figure 1.1). In this thesis, deciding which claim to prioritize for replication forms the general goal. Chapter 2 provides the ultimate theoretical

estimand (expected utility gain) and proximal theoretical estimand (replication value) of interest. Chapter 4 connects these estimands to an empirical estimand which we name RV_{Cn} (see the section on chapter 4 below for further details). Finally chapter 5 explores how best to calculate RV_{Cn} in real-world data.

1.4.2 Chapter 2

Summary: The goal of chapter 2 is to define replication study selection as a decision problem and to identify the meaning of replication value in terms of this problem. Importantly, the chapter formally defines the goal of replication research (to increase the utility of scientific claims), and it lays out how replication study selection can help reach that goal in more or less efficient ways. Having formalized the goal of replication study selection, the chapter goes on to formally define replication value as a proxy of expected utility gain - the theoretical estimand of interest in replication study selection. Chapter 2 also includes a comprehensive review of empirical estimators that have been proposed and utilized to justify replication study selection in the published literature.

Central research questions:

1. What makes something important to replicate?
2. What does *replication value* mean?
3. What are the key factors determining replication value?

Contributions: The model developed in chapter 2 provides the theoretical bedrock for subsequent chapters in this thesis. Moreover, the model provides a theoretical framework for developing and justifying study selection strategies that other researchers could utilize and build upon. By clearly stating the assumed goal of replication study selection and by formalizing the underlying decision problem, chapter 2 provides the necessary foundations for a constructive and transparent discussion of which studies are most important to replicate.

1.4.3 Chapter 3

Summary: Chapter 3 proposes a novel definition of test validity that expands on the definition proposed by Borsboom et al. (Borsboom et al., 2004). The chapter argues that a test is valid for measuring an attribute if (a) the attribute exists, and (b) variation in the attribute is d-connected to variation in the measurement outcomes. A clear definition of test validity is a fundamental precondition for being able to evaluate whether any proposed measure of replication value is valid or not. Chapter 3 thus introduces the core measurement

assumptions that in chapter 4 is used to connect the theoretical estimand of replication value with empirically observable variables. The original plan for chapter 4 was to adopt the definition of validity from Borsboom et al. (2004) wholesale. However, a refinement of the definition was needed in order to explain how the casual relationship between sample size and uncertainty could be considered valid when sample size (the measured attribute) is the causal parent of uncertainty (the target attribute). Thus, while chapter 3 may seem somewhat dissociated from the overall thesis topic, it is fundamental to the measurement rationale adopted in subsequent chapters.

Central research questions:

1. What does test validity mean?
2. Can a causal ancestor act as a valid estimator of its causal descendants?

Contributions: The d-connection definition of validity serves as the logical backbone of the measurement model presented in chapter 4. Beyond the direct relevance for this thesis, chapter 3 represents a general improvement on the earlier definition by Borsboom et al. (2004) by resolving the logical disconnect between the stated goal of measurement (derive procedures that are sensitive to variation in target attributes) and the unnecessarily strict requirement that the target attribute must cause the measured attribute. By relaxing the assumption of causal direction and by couching the definition of validity within the language of structural causal modeling (Pearl, 2009), our general understanding of what constitutes valid measurement in any given context should be greatly increased.

1.4.4 Chapter 4

Summary: Chapter 4 links the theory of replication study selection developed in chapter 2 to quantitative observable data. Building on the model developed in chapter 2 and the definition of validity outlined in chapter 3, chapter 4 proposes to operationalize replication value as a function of the citation impact of the article in which a claim is reported and the sample size of the study used to test the claim (RV_{Cn}). The chapter lays out the rationale and measurement assumptions used to justify the validity of RV_{Cn} as an indicator of expected utility gain, it suggests how RV_{Cn} could be implemented for replication study selection, and it provides some preliminary evidence for the validity of RV_{Cn}. The chapter ends by discussing more rigorous ways in which the validity of RV_{Cn} could be tested.

Central research questions:

1. How can replication value be estimated empirically?

2. What auxiliary measurement assumptions must be added to justify the empirical indicator, and how likely are these to hold?
3. What potential estimation issues can already be anticipated?

Contributions: Because the assumptions under which RV_{Cn} functions as intended are made explicit, it can also be made explicit when RV_{Cn} will *not* function as intended. Thus, chapter 4 provides the first demonstration of a falsifiable strategy for replication study selection. The chapter serves several purposes. First, it provides a first example of how the theoretical rationale in chapter 2 can be used to develop concrete study selection strategies, and what additional measurement assumptions must be added for the strategy to be falsifiable. Future researchers could use chapter 4 as scaffolding to develop new and improved indicators of replication value, or subject existing alternative indicators to the same form of analysis to make these falsifiable as well. Second, the chapter provides the necessary foundations for validation of RV_{Cn}. If RV_{Cn} turns out to be an accurate estimator of replication value, it could have a substantial impact on the efficiency of resource allocation in replication research, which should in turn serve to increase the impact of replication studies.

1.4.5 Chapter 5

Summary: Chapter 5 explores how RV_{Cn} can be implemented in practice, by applying the general study selection procedure proposed in chapter 4 to published research in the field of social fMRI research. The chapter touches on many practical questions related to implementation of RV_{Cn}, such as how the initial set of candidates to select from can be obtained, whether the information needed to calculate RV_{Cn} can be reliably obtained, how time-consuming the data collection effort is likely to be, whether undergraduate research assistants can conduct the data collection, and whether additional quantitative indicators of value and uncertainty can be obtained.

Central research questions:

1. Is the information required to calculate RV_{Cn} available in practice?
2. Are citation impact estimates reliable across sources?
3. Can study sample size be reliably coded by undergraduate student assistants?

Contributions: Feasibility is a necessary precondition for implementing a given study selection strategy, but dedicated feasibility reports are rarely offered for existing strategies (but see Field et al., 2019). Chapter 5 provides an example of how such feasibility reports could be carried out. Chapter 5 also clears a practical hurdle on the path to validation of RV_{Cn} by confirming

that the data needed to calculate RV_{Cn} can be collected in a feasible and reliable manner for even large (>1000) sets of studies. The procedures for testing feasibility developed in this chapter could easily be repeated to test the feasibility of implementing RV_{Cn} in other research fields. Moreover, chapter 5 could serve as a template which could be adapted to examine the feasibility of other replication study selection strategies.

1.4.6 Chapter 6

Chapter 6 takes stock of the work presented in the previous four chapters and considers possible future avenues for research into replication value and replication study selection. In this final chapter, the direct relationship between the model outlined in chapter 2 and *VoI* analysis is made explicit, demonstrating that replication value is equivalent to the concept of value of perfect information from decision theory. Using decision tree modeling, the definitions of value and uncertainty given in chapter 2 can be made much more concrete. From these definitions it can be shown that replication value is identical to the expected value of perfect information, while expected utility gain is equivalent to the expected value of sample information. Chapter 6 also highlights potential sources of bias in measurements of replication value that may be introduced when the model is applied in practice, discusses the clash between the formal and vernacular meaning of "validity", and considers how RV_{Cn} may be subject to Goodhart's law. Finally, chapter 6 considers the broader issue of research efficiency as an emerging metascientific line of research, and the contributions of this thesis to that research line.

Chapter 2

Deciding what to replicate: A decision model for replication study selection under resource and knowledge constraints[1]

2.1 Introduction

The goal of science is the advancement of knowledge (Kitcher, 1995). To achieve this goal, scientists need to generate novel claims[2] 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

[1]This chapter has been published in Psychological Methods as Isager, P. M., van Aert, R. C. M., Bahník, Š., Brandt, M. J., DeSoto, K. A., Giner-Sorolla, R., Krueger, J. I., Perugini, M., Ropovik, I., van 't Veer, A. E., Vranka, M., & Lakens, D. (2021). Deciding what to replicate: A decision model for replication study selection under resource and knowledge constraints. *Psychological Methods*. https://doi.org/10.1037/met0000438

[2]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 et al., 2018; Field et al., 2019; Hardwicke et al., 2018; Heirene, 2020; Kuehberger and Schulte-Mecklenbeck, 2018; LeBel et al., 2018; Mackey and Porte, 2012; Schmidt, 2009; Zwaan et al., 2018).

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 replicability of significant claims (Tversky and Kahneman, 1971), (b) published articles report an implausibly high rate of positive claims (Fanelli, 2010, 2012; Scheel et al., 2019), (c) there are many scientific practices that can increase the false-positive rate in published reports (e.g., Simmons et al., 2011), and (d) such practices may be relatively common (Agnoli et al., 2017; Banks et al., 2016; Fiedler and Schwarz, 2016; John et al., 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 and Peters, 2011; Nosek and 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 et al., 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 et al., 2012; Hagger et al., 2016; Nosek et al., 2012; Open Science Collaboration, 2015; Ranehill et al., 2015; Ritchie et al., 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 and Trafimow, 2015; Ebersole et al., 2016;

Finkel et al., 2017; Maxwell et al., 2015; Pashler and Wagenmakers, 2012; Stroebe and 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; Royal Society Open Science, 2020; Simons, 2014; see Martin and Clarke, 2017, for a review). Furthermore, funding bodies are starting to explicitly direct grant resources toward replication efforts (e.g., Association for Psychological Science, 2018; NWO, 2019). 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 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 (NWO, 2019). 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.

2.2 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 and 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 and 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 et al., 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 and 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.

2.2.1 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 and Kuckertz, 2018; JESP, 2018; American Psychological Association, 2021a; 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 et al. (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 and Silva, 2013; Matiasz et al., 2017, 2018; Silva et al., 2014; Silva and Müller, 2015). The framework combines rules for causal identification with Bayesian evidence (Matiasz et al., 2017, 2018) 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; of Arts and Sciences, 2018; Mackey and Porte, 2012). 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).

2.2.2 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[3]. In January 2016 a public invitation was shared in an online blog post (Lakens, 2016b) 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.

2.2.3 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),

[3]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 proposed here.

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)[4]. 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" (American Psychological Association, 2021a).

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.

[4]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.

2.3 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 2.1 and 2.2 present the causal assumptions, structural equations, and verbal summaries for all terms mentioned in the text. For clarification, all terms from figure 2.1 and 2.2 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 and 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* 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[5]. The *value* of a research claim is

[5]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

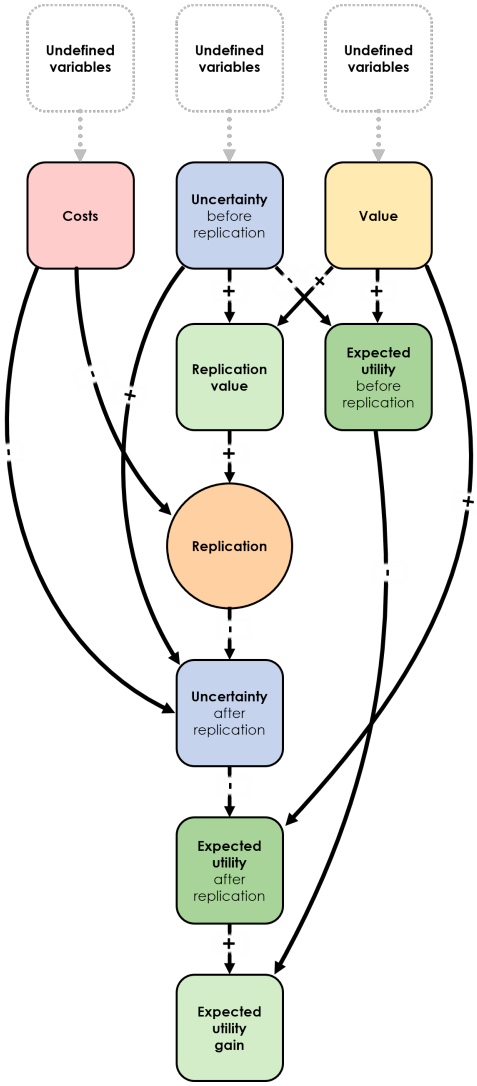

Figure 2.1: Structural causal model of the system that determines replication value. Arrow direction signals the causal direction of effects. Time flows from the top to the bottom of the figure; variables (nodes) closer to the top are determined earlier in time than nodes closer to the bottom (e.g., 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 figure 2.2 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.	$EU_{gain} = EU_{Post} - EU_{Pre}$

Figure 2.2: Structural equations for the structural causal model in figure 2.1. The "Name" column corresponds to the node names inside figure 2.1 (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 2.1. 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.

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[6]. 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 et al., 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 figure 2.2 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 figure 2.2). 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

value to being wrong about a claim.

[6] 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%.

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 and 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 uncertainty (represented by the effect size on the negative arrow *replication* → *uncertainty after replication*, in figure 2.1). 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 2.1), 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 figure 2.2).

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[7]. 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 2.1. 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

[7] 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.

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 2.1.

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 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 2.1).

2.4 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; of Arts and Sciences, 2018; Mueller-Langer et al., 2019). Quantitative indicators of *value* might therefore include citation counts (Aksnes et al., 2019; Lewandowsky and Oberauer, 2020), Altmetric Attention scores (Bornmann, 2014), journal impact indicators (Garfield, 2006; but see Oh and Lim, 2009), best paper awards, citation by textbooks or clinical guidelines or public policy, reviewer ratings of importance and novelty, etc.[8] 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 and 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 figure 2.2).

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:

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

[8]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.

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

2.5 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 and 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 et al., 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 examined 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*.

2.6 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 et al., 2020), replication funders with limited grant resources (e.g., NWO, 2019), 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.

Chapter 3

Test validity defined as d-connection between target and measured attribute: Expanding the causal definition of Borsboom, Mellenbergh & van van Heerden[1]

3.1 Introduction

Borsboom et al. (2009) give the following general definition of test/measurement validity: "*Validity is a property of measurement instruments that [...] codes whether these instruments are sensitive to variation in a targeted attribute*" (Borsboom et al., 2009, p. 135). A formalized definition is offered by Borsboom et al. (2004; henceforth BMH), who state that "*A test is valid for measuring an attribute if (a) the attribute exists and (b) variations in the attribute causally produce variation in the measurement outcomes*" (BMH, p. 1061). The latter definition implies that a measure can only be valid when

[1]This chapter has been made available on PsyArXiv as Isager, P. M. (2020, September 28). Test validity defined as d-connection between target and measured attribute: Expanding the causal definition of Borsboom et al. (2004). https://doi.org/10.31234/osf.io/btgsr

the attribute of interest is a cause of measurement outcomes. However, it is not clear from BMH why valid measurement should be restricted to this specific causal scenario.

This article argues that variations in an attribute need not *causally produce* variation in the measurement outcomes in order for the measurement instrument to be "*sensitive to differences in the attribute*". It is enough that the measure and the attribute are d-connected, which we can acknowledge by modifying the definition offered by BMH. What follows is an outline of such a modified definition, as well as a discussion of which simple causal scenarios can be considered valid measurement. For simplicity, the term *measured attribute* will be used as a general term for referring to a "test", a "scale", a "measurement tool", a "measurement instrument", or any other term referring to an observed variable that is used to estimate the values of a latent (or unmeasured) *target attribute*.

3.2 D-connection definition of test validity

Definition: *A measured attribute is valid for measuring a target (unmeasured) attribute if (a) the target attribute exists and (b) if the target attribute is d-connected to (i.e. not d-separated from) the measured attribute, such that variation in the measured attribute is statistically associated with variation in the target attribute (for a definition of d-separation, see Pearl, 2009, definition 1.2.3; Hernán and Robins, 2020, fine point 6.1). In other words, a measurement instrument is valid if, in the true causal graph, there is an "open path" between the instrument and the attribute that we want the instrument to measure.*

D-connection between target attribute and measured attribute ensures that the instrument either gives us information about what has happened to the target attribute in the past, or about what is going to happen to the target attribute in the future. This proposed modification of BMH's definition has substantial implications for when a test would be considered valid for measuring an attribute. Importantly, validity is no longer restricted to situations where the measured attribute is caused by the target attribute. Valid measurement would also include cases where the target attribute is caused by the measured attribute, cases where measured and target attribute are not in a direct causal relationship but share a causal ancestor, and cases where measured and target attribute share a causal descendant that is conditioned on.

The rationale for preferring this d-connection definition over BMH's definition goes as follows: Criterion (b) in BMH's definition clearly implies that measurement involves gaining information about the values of one variable (the target attribute) by observing the values another variable (the measured attribute). Such information gain is possible when the target attribute causes

the measured attribute. However, it is also possible in all cases where the target attribute is d-connected to the measured attribute. If there is no other justification for criterion (b) in BMH's definition, then we should relax this criterion to allow as valid all causal scenarios where observation of the measured attribute yields information about the target attribute.

The following sections discuss six basic examples of valid and invalid measurement according to the d-connection definition of validity. After this follows a consideration of why one might prefer measured attributes that adheres to BHM's definition, even if d-connection is considered the fundamental criterion for measurement validity. Finally, example statements from BMH are discussed where the d-connection definition leads to radically different conclusions.

3.3 Examples

3.3.1 Target attribute *A* causes measured attribute *M*

Target attribute A and measured attribute M are d-connected when A is a causal parent of M (see figure 1A). In this case, variation in M gives us information of what happened to A in the past. This situation is what is covered in the definition given by BMH and is perhaps the most typical causal scenario for measurement instruments in many sciences. As an example, an MRI scanner can act as a measure of the spatial distribution of neural activity because it is sensitive to blood flow/oxygenation in various regions of the brain. It is a valid measure of neural activity in regions of the brain because neural activity A has a causal effect on blood flow M in the same region.

3.3.2 Measured attribute *M* causes target attribute *A*

A and M are also d-connected when M is a causal parent of A (see figure 3.1B). In this scenario, variation in M gives us information about what will happen to A in the future. As an example, consider a company that wants to hire applicants that will perform well in their job. The company does not know what the applicants' future job performance A will be, but they do have access to the applicants' past work experience M. If past work experience has a causal effect on future job performance, then M is a valid measure of A in this case.

Note that the example above would not be considered a case of valid measurement according to BMH's definition of validity, since variation in the target attribute "future job performance" does not *causally produce* variation in the measured attribute "past work experience". We return to this discrepancy between the d-connection definition and BMH's definition in the discussion.

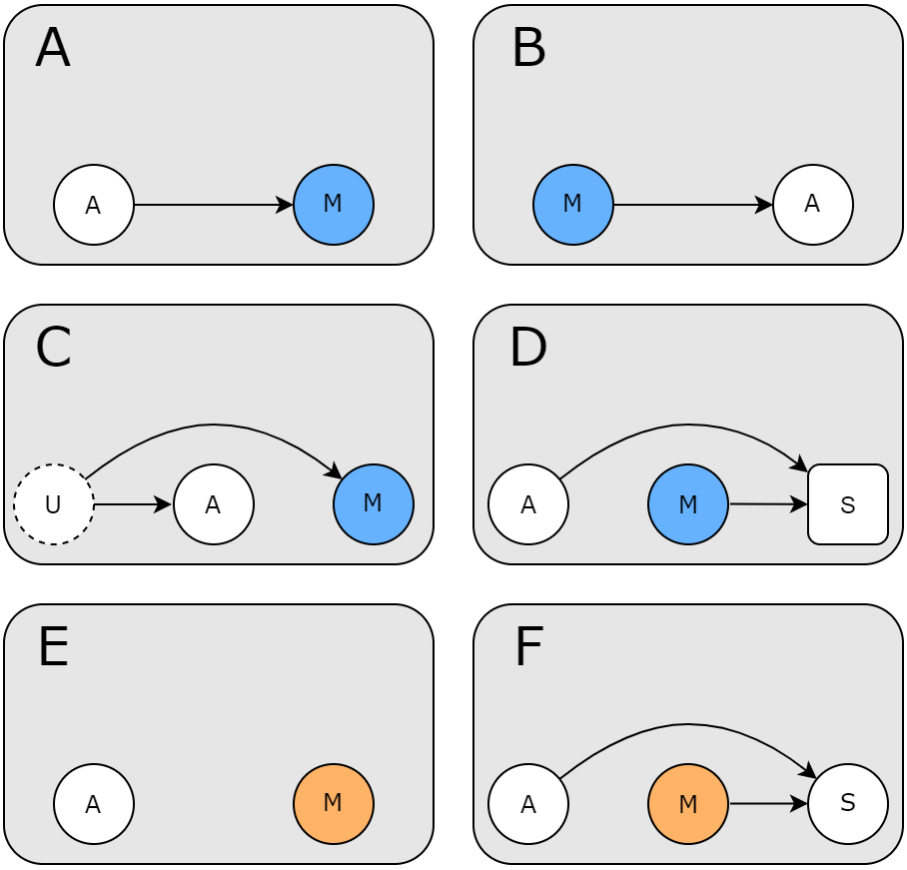

Figure 3.1: Causal models for various valid (1A-1D) and invalid (1E-1F) measurement scenarios. A = target attribute, M = measured attribute, U = unmeasured non-target attribute, S = shared causal descendant. The square in panel D indicates that S is being conditioned on. Blue coloring of M in panel 1A to 1D indicates that M is a valid measure of A. Orange coloring of M in panel 1E and 1F indicates that M is an invalid measure of A.

3.3.3 Target attribute *A* and measured attribute *M* are both caused by some third attribute *U*

M can be a valid measure of A even when there is no direct causal relationship between M and A. One example of this is when M and A share a causal parent U (see figure 1C). In this case, M gives us information about what happened to U in the past. Information about what happened to U in turn gives us information about what will happen to A in the future, since U causes A. As an example, consider again the company that wants to use past work experience M to predict future job performance A. However, in this scenario there is no direct relationship between past work experience and future job performance. Both are simply caused by the applicants' education U. However, past work experience M is still valid for measuring future job performance A since past work experience M is an indicator of the education U an applicant has received, which in turn is an indicator of their future job performance A.

3.3.4 Target attribute *A* and measured attribute *M* both cause some third attribute *S* that is conditioned on

Another example of when M is valid for measuring A - even when there is no direct causal relationship between them - is when M and A both have a causal influence on some shared descendant variable S, and we condition on S before using M to measure A (see figure 3.1D). This is also known as *conditioning in a collider* (Rohrer, 2018). In causal modeling terms, the variable S in figure 1 is called a collider variable (Pearl et al., 2016). Variables on the path on either side of a collider are d-separated from each other (Pearl, 2009, definition 1.2.3) unless they are conditioned on. When S is conditioned on, the marginally blocked path becomes unblocked (Pearl et al., 2016). In general terms, we use the combined knowledge of the values of M and S to reason back to what A must likely have been.

As an example, suppose we intend to use a person's height M as a measure of their hand-eye coordination skills A. In the general population these attributes are unrelated. However, suppose we restrict our measure to professional basketball players in the Women's National Basketball Association (WNBA). In this case, we know that both good hand-eye coordination and being tall are important for performing well enough in basketball to play in the WNBA. Thus, WNBA players either have good coordination skills, are tall, or both. Women who are neither tall nor particularly sleight of hand will very likely not make it into the WNBA. Thus, if we observe a shorter-than-average WNBA player we can infer that they must likely have better-than-average hand-eye coordination to make up for their height disadvantage. Otherwise, they would not have made it into the WNBA. Consequently, height M is a valid measure of hand-eye coordination A conditional on being a WNBA player S (assuming

the causal relationships described above are true).

Valid measurement is not restricted to the examples above. Any causal scenario in which A and M are d-connected (assuming A exists) represents a scenario where M is a valid measure of A.

3.3.5 Invalid measurement

Invalid measurement occurs whenever the target attribute A does not exist, since a measured attribute M cannot be d-connected to a non-existing target attribute. Invalid measurement also occurs in situations where A exists, but A and M are d-separated. This happens, for example, when M and A have no direct causal relationship and do not share any causal ancestors or descendants (figure 3.1E), and when A and M share a descendant which is not conditioned on (figure 1F).

Suppose again that we are using height to measure hand-eye coordination, but we are not conditioning on any common outcome S. In that case, M would contain no information about A and would not be a valid measure of A. In other words, whether a measurement instrument is valid depends not only on properties of A and M, but also on properties of all variables that play a part in the d-connection between A and M. This holds even if we assume BMH's more restrictive definition of test validity, since we can obviously alter the context of measurement to manipulate whether variation in target attribute A will causally produce variation in measured attribute M (e.g. by conditioning on mediating variables).

3.3.6 Measuring target attribute scores vs. measuring the effect of treatment on the target attribute

Although the d-connection definition of validity allows several causal scenarios to be considered valid measurement, there may be good reasons to prefer measurement tools where A causes M.

One reason is that the validity of a test depends on whether the goal of the test is to predict target attribute values, or to detect change in the target attribute as a function of some treatment. In all scenarios in figure 3.1, we treat measurement as synonymous with prediction. The goal is to observationally estimate (or "measure", or "predict", or "gain accurate information about") the true values of A. Given that the causal model is true in each case, M will be valid for estimating A in figure 3.1A-D.

However, suppose we are conducting a randomized experiment where we want to measure the causal effect of some treatment T on attribute A. Our target attribute of interest is no longer A per se. Rather, we want to measure *change*

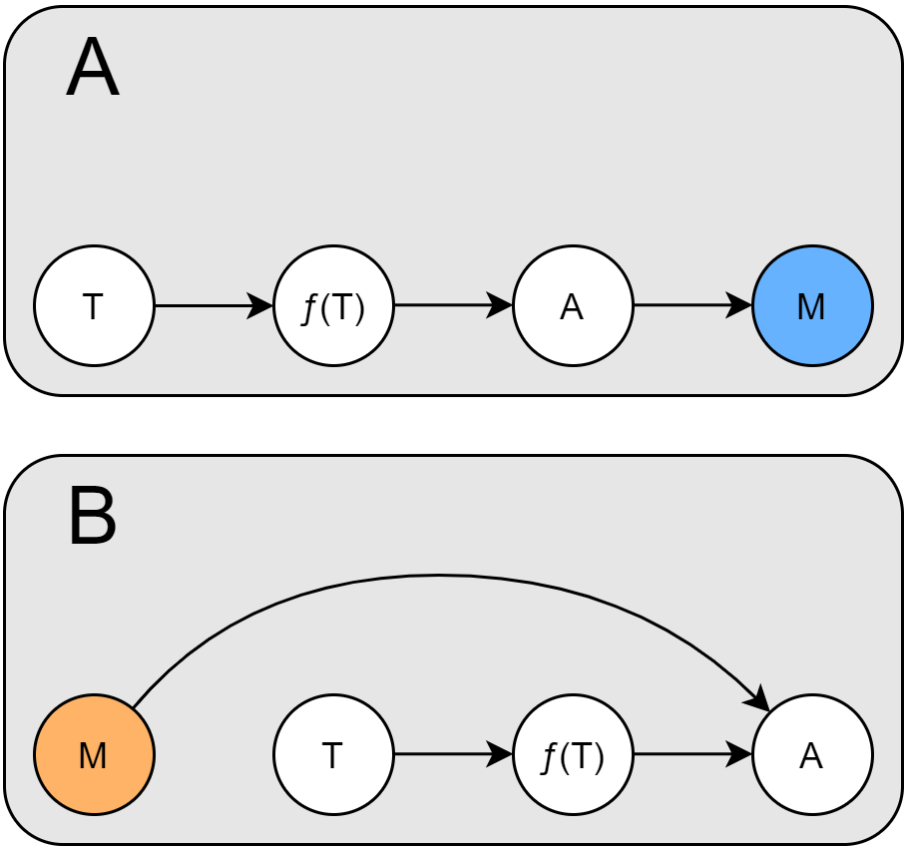

Figure 3.2: Measurement validity when the goal is to measure change in an attribute as a function of a given change in treatment. A = attribute, M = measured attribute, T = treatment, $f(T)$ = change in attribute as a function of treatment. Blue coloring of M in panel 2A indicates that M is a valid measure of A. Orange coloring of M in panel 2B indicates that M is an invalid measure of A.

in A with respect to changes in T. In technical terms, we want the form of the function $f(T)$ in the structural equation that determines A, $A = f(T, U)$, where U is all unmeasured causes of A that are either experimentally controlled or randomized between groups in our experiment. For example, in a linear model we would want to know the value of the direct effect of T, βT, in the linear structural equation for A, $A = \beta T + \epsilon$.

Figure 3.2 displays the same causal models for the relationship between A and M as figure 3.1A and 3.1B (see figure 3.2A and 3.2B respectively), but now shows what happens if our target attribute is changed to $f(T)$. In figure 2A our measured attribute M is still valid since it remains d-connected with $f(T)$. However, in figure 3.2B, even though M is d-connected to A, it is no longer d-connected to the target attribute $f(T)$ and ceases to be a valid measure. As an example, imagine an employer who wants to know the effect of a training program T on future job performance A. The employer has access to information about employee's past work experience M, and she knows that work experience is a cause of future work experience. In that case, past work experience M is a valid measure of future job performance A. However, M is not valid for measuring the effect of the training program T on performance $f(T)$. Regardless of how efficient the training program T is, it cannot change the applicant's past work experience M, so M is completely insensitive to the effect of the training program $f(T)$, and consequently it is insensitive to changes in A following change in T.

The advantage of a measurement instrument caused by the target attribute ($A \rightarrow M$) is that M will be d-connected to both A and any treatment effect $f(T)$ on A (figure 2A). Thus, M will remain valid regardless of whether our goal is simple prediction of A or testing the effect of some treatment T on A. We can always use the principle of d-connection to determine if M is a valid measure. However, when A is not a cause of M we need to be very conscious about what target attribute we are interested in measuring.

3.3.7 Using a measured attribute for conditioning on a target attribute that is a confounder.

Causal scenarios of the form $A \rightarrow M$ are also superior to other types of valid measurement when we want to estimate a causal effect ($X \rightarrow Y$) and A is a confounder which we need to condition on to accurately estimate the causal effect of X on Y. In this case, a measured attribute M can be used to (partially) block open paths through A, but *only* if M is a causal descendant of A (Greenland and Pearl, 2014).

Figure 3.3 demonstrates the issue of using a measured attribute to condition on a target attribute in two different causal scenarios. Suppose we want to know if people who attend a job training program X increase their job performance Y, and that this effect is confounded by job performance before

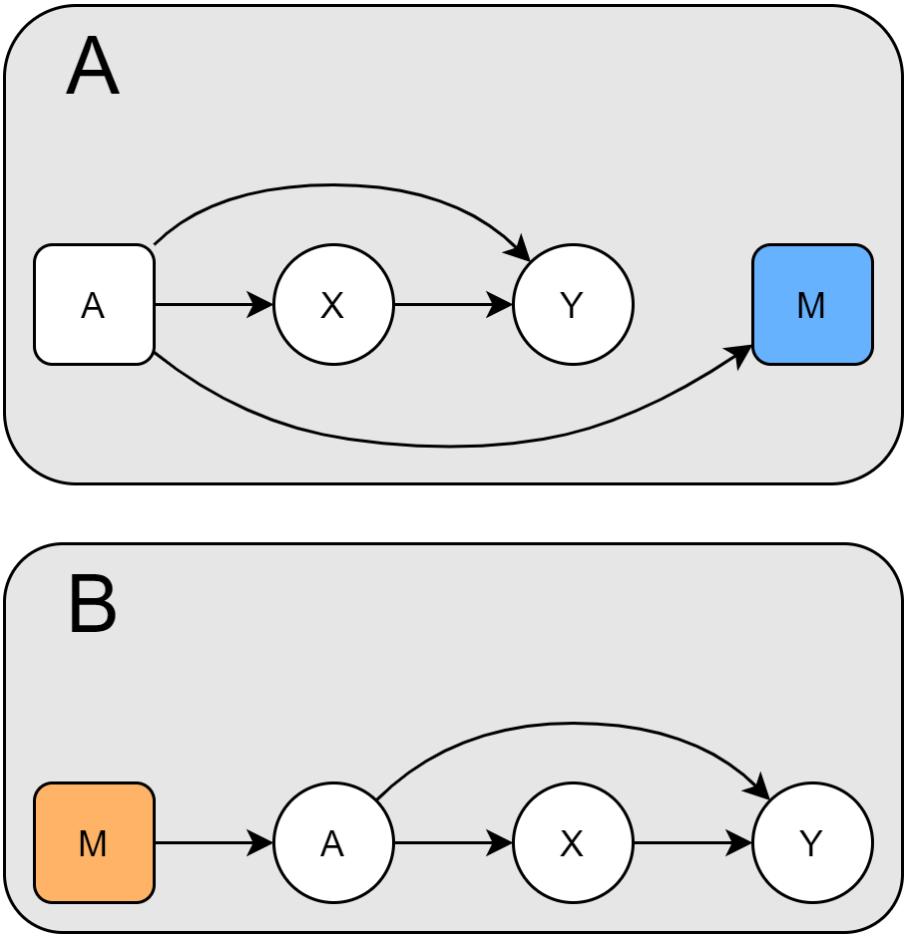

Figure 3.3: Measurement validity when the goal is to utilize the measurement for conditioning on a confounder. A = confounding variable for the relationship $X \rightarrow Y$, M = measured attribute, X = cause, Y = outcome. Square nodes represent those nodes that are (partially) conditioned on by conditioning on M. Blue coloring of M in panel 3A indicates that M is can (partially) block the confounding path between X and Y through A. Orange coloring of M in panel 3B indicates that M is not valid for blocking the confounding path through A.

training A. For example, people who already perform well at their job may be more likely to both attend the job training program, and to perform well at their job in the future. Suppose we cannot observe employee pre-training job performance A directly. However, we have access to supervisor ratings of pre-training performance M (figure 3.3A). If performance A is a cause of ratings M and we condition on M, we will at least partially condition on A and block (some of) the back-door path between X and Y. Contrast this with using past work experience M as a measure of pre-training job performance A (figure 3.3B). In this case we are no longer using a causal descendant of A as the measured attribute. Thus, even if M is valid for measuring A according to the d-connection definition, M cannot be used to condition on A and block the back-door path between X on Y. Whether this ability to condition on the target attribute should be a required property of a valid measurement instrument is left as an open question here. If one were to argue for this, one would then have to admit that the goal of measurement is no longer simply to construct instruments that "are sensitive to variation in a targeted attribute" (Borsboom et al., 2009).

3.4 Points of similarity and divergence between BMH and the d-connection definition of validity

In terms of philosophical foundations, the d-connection definition of test validity is identical to the definition of test validity offered by BMH. Validity is still considered to be a concept within the domain of ontology (as opposed to epistemology). Validity is still considered entity separate from reliability and measurement bias (however, zero reliability does imply no validity because a completely unreliable measure implies d-separation. I.e., variance in a completely unreliable measure cannot, by definition, be sensitive to variance in the attribute). Existence of the target attribute, and the causal relationship between target and measured attribute, are still the fundamental criteria used to determine validity. Finally, several issues raised by BMH are equally applicable to the d-connection definition, such as the distinction between intraindividual and interindividual measurement structures, and whether validity is best thought of as binary (true/false) or continuous entity.

The only difference between BMH's definition and the d-connection definition is whether validity holds only when the target attribute is a cause of the measured attribute, ($A \rightarrow M$), or whether validity holds under any d-connected relationship between target- and measured attribute. Still, this one modification of the definition leads to very different conclusions about validity in several scenarios discussed by BMH.

As one example, consider the dissatisfaction of BMH with the statement by Guilford (1946) that "a test is valid for anything with which it correlates". BMH argue that "*... the likelihood of encountering zero correlation in real life is exceedingly small, and especially in the social sciences, everything tends to correlate with everything (Meehl, 1978). Therefore, the upshot of any line of thinking that sees correlation as a defining feature of validity is that everything is, to some degree, valid for everything else. This absurdity does not arise in a causal theory because it is not the case that everything causes everything else.*" (BMH, p. 1066). In contrast, the d-connection definition of test validity is in practice closely aligned with the statement by Guilford (1946). However, Guilford's statement does not suffice as a definition of test validity. A test "*is valid for anything with which it correlates*" only if the correlation is caused by d-connection between the test and the target attribute, and not by statistical noise or violation of causal identifiability conditions (Hernán and Robins, 2020). In addition, a test can be valid for measuring attributes with which the test does not correlate. One example of such a case is BMH's example of meter stick measurements of rods of equal length. Another is when the measure and attribute are strongly but non-linearly related (e.g., a measured attribute M whose structural equation is the sine of A). In general, since validity is a statement about ontology (i.e., about causal processes in the real world), the language of causality must always be invoked to explain *why* correlation is (sometimes) an indication of valid measurement.

As another example of when the d-connection definition deviates from BMH, consider BMH's's statement that "*Height and weight correlate about .80 in the general population, but this does not mean that the process of letting people stand on a scale and reading off their weight gives one valid measurements of their height*". According to the d-connection definition, weight is a valid measure of height in this scenario, so long as height exists and is d-connected with weight. That is, so long as we can establish a d-connection relationship as the source of the correlation between height and weight, either will be a valid measure of the other. Some may object that we could easily imagine scenarios where a change in weight implies no change in height. If we run Mike Teavee through the taffy puller, he will become taller without gaining any weight. However, this is equivalent to imagining an intervention on height and wanting to know the effect of that intervention, which we can account for using the d-connection definition. If height A and weight M are both caused by common genes (U; see figure 3.1C), then weight is not valid for measuring a treatment effect $f(T)$ on height. However, weight is still valid for measuring raw scores of height in the population. If you know how much a person weighs, you have (imperfect) information about how tall they are, even if some members of the population have been run through the taffy puller.

Finally, consider the statement made by BMH that in formative models "*the observed indicators are not considered to be causally affected by the latent variable but, rather, to cause such a latent variable. In this case, it is diffi-*

cult to see how these observed indicators could be conceptualized as measures of the attribute in question because the arrows between the attribute and the observations run in the opposite direction" (BMH, p. 1069). According to the d-connection definition of test validity, observed indicators are valid measures of target attributes in both latent variable models (figure 3.1A) and formative models (figure 3.1B). To appreciate why this makes sense, consider an extreme case where we know the exact structural equation that assigns values to a formative target attribute, and we have accurately measured all causal parents that enter into this equation. For example, many happily measure out 500 grams of sugar using a measuring cup, even though volume is a causal determinant of mass. We can do this because we know that mass equals volume times density, and we trust that the density of sugar is roughly accounted for by the scale printed on our measuring cup. In this case we can calculate values on our target attribute (weight of sugar in the cup) with near perfect certainty using only one measured cause of the attribute (volume of sugar in the cup) and an additional measurement assumption (sugar density is a constant that is corrected for). Surely this estimation procedure can still be considered a valid measure of the target attribute.

Note that in all the examples just mentioned, the d-connection definition is still in line with the more general definition of validity offered in Borsboom et al. (2009). That is, a measurement is "*sensitive to differences in the attribute*" as long as the measured and target attribute are d-connected.

3.5 Implications of accepting the d-connection definition of validity

Before accepting a d-connection definition of test validity, consider some of the stranger implications of the d-connection definition of test validity. The d-connection definition suggests that there is no substantive difference between terms such as "measure", "estimate", "predict", "determine", and "compute". It implies that we can measure what has not yet happened (unless one argues that what has not happened does not exist and, by extension, does not meet the requirement that the target attribute must exist). It also implies that a measurement instrument can be valid even if there is no direct causal path between the instrument and the attribute it purports to measure.

The d-connection definition requires no particular causal proximity between measured and target attribute. The causal path from the target A to the measured attribute M may include any number of mediators (e.g. $A \rightarrow B \rightarrow C \rightarrow D \rightarrow E \rightarrow M$). Causal proximity is still important for measurement because a long mediating path implies many sources of error, which is important for measurement reliability. But causal proximity does not factor into whether a measurement is considered valid or not. In addition, the d-

connection definition does not contain an explicit requirement of quantitative structure, and it is therefore somewhat detached from the classical conception of measurement utilized in the physical sciences (Michell, 2003).

The reader must critically evaluate whether these implications are acceptable. Based on my personal experience discussing these ideas with colleagues, I suspect many will be skeptical. This is good. The goal of this article is not to argue that the d-connection definition of validity represents the final word in the debate about what constitutes test validity. Rather, the goal is to add to that debate by highlighting what seems to be a conceptual issue in BMH's framework, and by proposing a modification (d-connection) that would seem to resolve this issue. Whether this modification is sensible or presents conceptual issues of its own remains a topic for discussion – the outcome of which will surely deepen our understanding of test validity even further.

Chapter 4

Replication value as a function of citation impact and sample size[1]

4.1 Introduction

After close to a century of repeated reminders that replication studies play an important role in establishing robust scientific knowledge, and following a number of high-profile replications published in the last decade (e.g., Open Science Collaboration, 2015; Ritchie et al., 2012; Wagenmakers et al., 2016; Ranehill et al., 2015; Hagger et al., 2016), researchers from many scientific disciplines are increasingly calling for a renewed focus on replication research (Zwaan et al., 2018; Plucker and Makel, 2021; Button et al., 2013; Blaszczynski and Gainsbury, 2019; Heirene, 2021; Sale and Mellor, 2018; Murphy et al., 2021). Given the exponential growth of the empirical literature (Bornmann and Mutz, 2015; Parolo et al., 2015) and the low rates of replication up until this point (Makel et al., 2012; Mueller-Langer et al., 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 often be a large number of replication targets to choose from, and assuming we can

[1]This chapter has been made available on MetaArXiv as Isager, P. M., van 't Veer, A. E., & Lakens, D. (2021, August 24). Replication value as a function of citation impact and sample size. https://doi.org/10.31222/osf.io/knjea

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.

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 et al. (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 et al., 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 using 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 the goal can be achieved by selecting studies based on the target attribute(s), and specify the measurement model that connects the target attribute(s) to our measured indicator(s).

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 (which has been worked out in a previous article, Isager et al., 2020). 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.

4.2 The target attribute of replication study selection.

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

$$RV = f(value, uncertainty) \tag{1}$$

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 equation 2).

4.3 Defining relevant measurement properties.

To create an indicator that is useful for replication study selection, we must ensure that the indicator is valid for measuring the target attribute, and that it measures the target attribute in a reliable and unbiased way. To accurately estimate replication value stakeholders will thus need to find valid, reliable and unbiased operationalizations of value and uncertainty. Following Isager (2020),

we consider an indicator to be *valid* for measuring replication value if (a) the attribute replication value exists and (b) replication value is d-connected to the indicator (i.e., the two variables are joined by an unblocked path in the causal model that describes their true relationship; see fine point 6.1 in Hernán and Robins, 2020).

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 (American Psychological Association, 2021b). 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 (American Psychological Association, 2020; Isager et al., 2020). 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 that term refers to something more specific here, Borsboom et al., 2004). For an indicator of replication value to be useful for replication study selection, it must have high measurement quality. That is, it must be valid *and* reliable *and* unbiased[2].

4.4 Operationalizing an indicator of replication value.

Many replication value indicators have already been proposed. Of these, most focus exclusively on either the value- or the uncertainty-side of equation 2. Some suggested indicators take both value and uncertainty into account (Isager et al., 2020, supplementary materials), but suffer from at least one of two problems. In some cases, indicators are operationalized in obviously problematic ways. For example, if an indicator uses p-values in their operationalization of uncertainty (e.g., https://osf.io/x73rk/) it is not possible to differentiate between ambiguous statistical evidence (high uncertainty) and strong evidence

[2]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.

for the null (low uncertainty. See e.g., Field et al., 2019). In other cases, indicators depend on information that is not easily available in practice (such as whether the claim studied is considered "surprising"; https://osf.io/v8nkd/). Indicators that are difficult to calculate will be difficult to use and to validate, since validation depends on our ability to collect 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 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:

$$\begin{aligned} RV &= value \times uncertainty \\ &\approx f(citation\ count) \times f(sample\ size) \end{aligned} \tag{3}$$

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

4.4.1 Citation count as an indicator of value

We define the value of a claim as the stakes involved in decision outcomes based on the claim. Suppose a mining company is considering whether to establish a new mine. Research suggests "there is gold in them there hills". If the claim is *correct* and the company chooses to believe it, the new mine will turn a huge profit. If the claim is *false* and the company chooses to believe it, the mine will be a complete waste of resources. Thus, the value of the claim "there is gold in them there hills" depends on the interaction 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 expected utility of being correct about the truth status of the claim (relative to the value of being wrong). Value is usually related to the impact of the claim in science and society. A claim may be impactful for many reasons. Scientific claims can be used to build theories and research lines, transform clinical treatment, guide education, inform public policy, etc. The potential benefits of such applications of scientific 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 when applying this claim (e.g., the health and safety of drug

dependent patients) the value of having accurate knowledge about the claim is high. Formally, we could treat value as a formative attribute constructed from multiple forms of impact. Whenever we state that a claim is valuable we usually mean that the claim has been impactful in one or several ways. Finding a valid measure of value thus entails operationalizing one or more of these impact attributes. Stakeholders can differ in how they operationalize these attributes, depending on what goal(s) they value.

Here we propose to focus on scientific impact, and operationalize impact in terms of citation count, which is generally considered a valid indicator of scientific impact (Aksnes et al., 2019). Scientific impact can be roughly defined as the impact of a claim on future research decisions (such as whether to run a follow-up study, build on a previous finding, etc.). If we can assume that researchers tend to cite articles as support for important research decisions, then citation count will be sensitive to variation in value that is caused by variation in scientific impact (figure 4.1). Based on comprehensive reviews of the literature on citation behavior (Aksnes et al., 2019; Bornmann and Daniel, 2008) and empirical studies of the relationship between citation count and perceived scientific impact by researchers (Ioannidis et al., 2014; Radicchi et al., 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 perceived value of the claims in that article (Isager et al., 2020). Citation count also has several other desirable measurement properties. Its interpretation is relatively straight-forward. A meticulous record of citation count is kept by many bibliometric sources, which means that reliability of the count across sources can be studied (Martín-Martín et al., 2018). Bibliometric recordkeeping also means that citation counts can be obtained with little effort in any scientific discipline, which is crucial for being able to utilize citation count for study selection in practice.

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 stakeholders value, such as societal or clinical impact. Impact on these dimensions is less likely to generate citations from articles published in scientific outlets listed in major bibliometric databases (Aksnes et al., 2019; Eck et al., 2013). Consequently, a traditional citation count metric will likely not be sensitive to variation in value that is caused by variation in these dimensions of impact (see dashed circles in figure 4.1). When these other sources of impact are the most important to our overall definition of value, citation count is going to be a poor measure of value (which, by extension, will lead to poor estimates of replication value).

Second, even in cases where scientific impact is of primary interest, articles are cited for a myriad different reasons that need not have anything to do with the scientific impact of claims put forward in them. Non-relevant influences on citation count include (but are not limited to) the time of publication, the

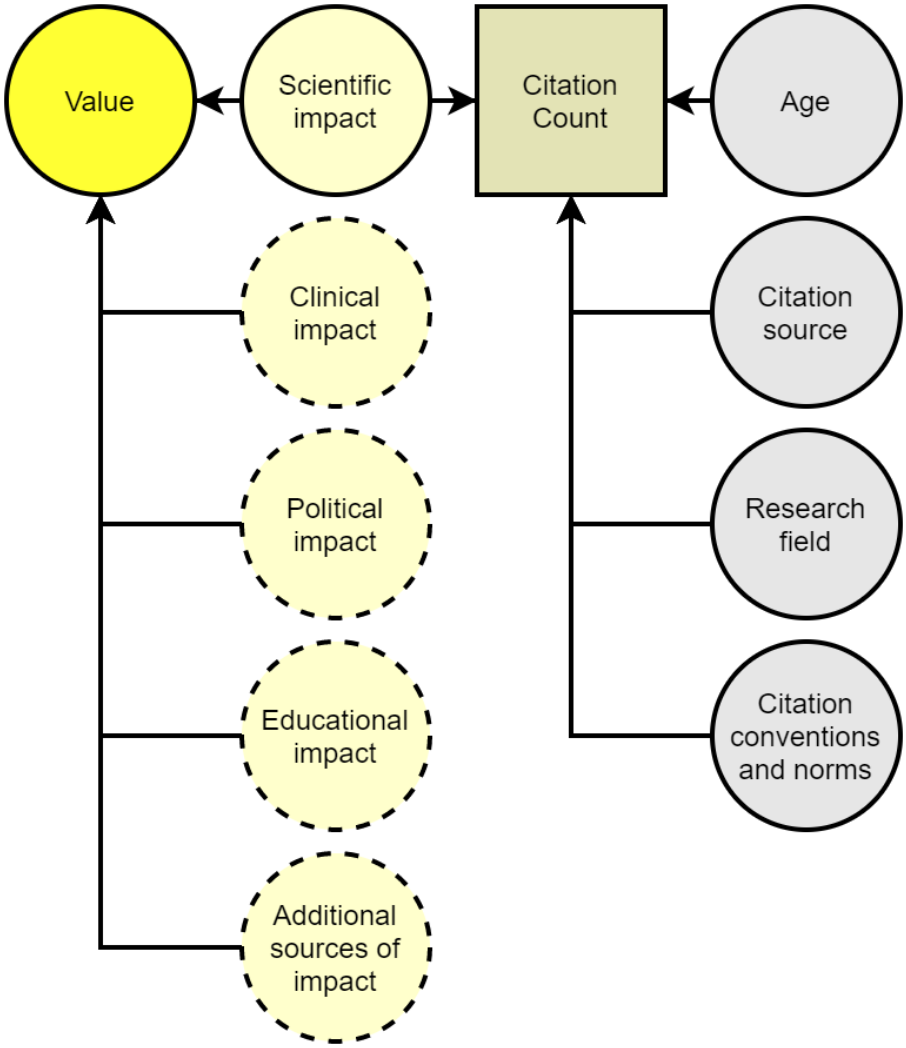

Figure 4.1: Proposed causal relationship between *value* and *citation count* (*citation count* is outlined by a square to signal it is observable). The figure can be interpreted as a directed acyclic graph model, where 'value' and 'citation count' are d-connected through 'scientific impact', and all 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 (the unmeasured attributes U traditionally used in causal graph models). 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.

bibliometric source that is doing the counting, arbitrary citation conventions within scientific fields, the language in which an article is written, reputation of article authors, preference for citing personal acquaintances, bandwagon effects, self-citations to increase academic standing, etc. (Bornmann and Daniel, 2008; Aksnes et al., 2019). To the extent that such factors are independent of article 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 4.1).

It is worthwhile to consider whether noise factors could be controlled to increase reliability of the measurement. In some cases this will be challenging. For example, while it might be possible to classify whether citations occured due to arbitrary citation 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). 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 the article (Bornmann and Daniel, 2008; Wang et al., 2013), the source of the citation count (Martín-Martín et al., 2018), and the research field the target claims are part of (Waltman and van Eck, 2013).

Replication can only increase the expected utility of research decisions that have not yet been made. Therefore, we specifically want to know the *future* scientific impact of a replication target. If citation impact is taken as an indicator of simultaneous scientific impact, it follows that what we are interested in is the *future* citation impact of a replication target article. Current total citation count of an article is a measure of past citation impact, but is likely predictive of future citation impact (Wang et al., 2019). However, *total* citation count is not a useful predictor of future citations as soon as we start comparing target articles across different publication years, as older articles have had more time to be cited. As an example, imagine two articles, A and B, that are both cited exactly 10 times every year. However, at the time of comparison study A is 3 years old and has a total citation count of 30, while study B is 9 years old and thus has a total citation count of 90. If we would use A and B's total citation count as a linear predictor of their relative future citation count we will erroneously conclude that B will receive three times as many citations as A. It is therefore sensible to adjust the two totals for their respective ages and estimate the *yearly citation rate* (i.e. the slope, or the 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 the number of years since the article was published. That is, we operationalize the value of a claim as the *average citations per year* of the paper that claim was reported in. In practice this assumes that past average yearly citation rate is a good predictor of the future 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 4.2A 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 (yellow bars) and dividing by the number of years. As expected, the prediction of the actual citation trajectory (grey line) is very accurate.

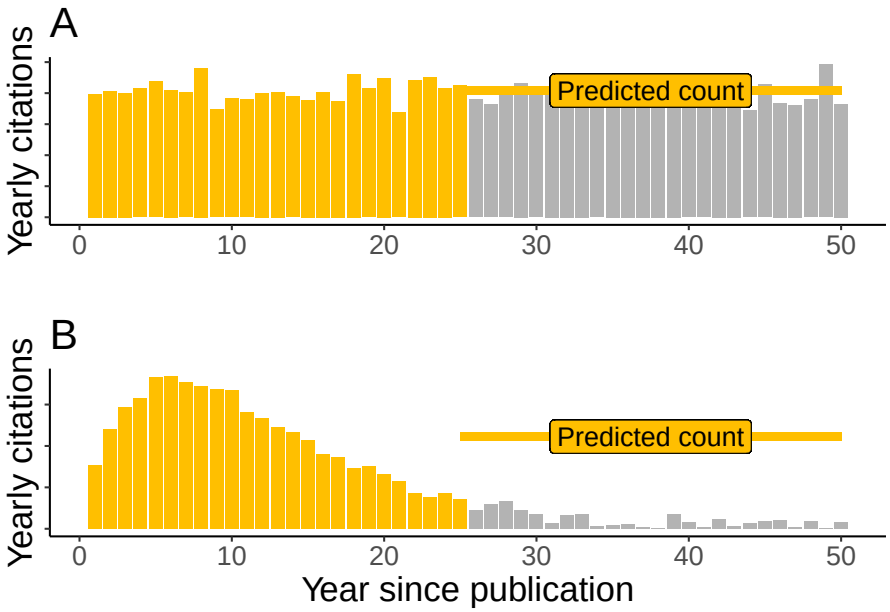

Figure 4.2: Simulated 50 year citation rate for two imagined articles. The yellow bars display the citation rate for the first 25 years in the article's history. The grey bars represent the future rate (year 26-50) that we want to predict. The yellow labeled line displays the predicted trajectory of 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 article is not constant over time. Figure 4.2B displays a more realistic 50-year citation trajectory (Parolo et al., 2015). In this context, predicting the future trajectory based on past average yearly citation count leads to a considerable overestimate of future citations. In fact, average yearly citation count will systematically over- or underestimate the future citation trajectory when-

ever the trajectory is not converging towards the past average yearly citation count. We therefore recommend that, whenever possible, prediction of future citation impact should be based on more sophisticated prediction techniques that utilize the entire distribution of past citation rate and other bibliometric features of the target article (e.g., Chakraborty et al., 2014; Yuan et al., 2018)[3]. However, these methods require advanced expertise to 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, 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 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 reference coverage and in their exact citation counts for the same reference (Martín-Martín et al., 2018). However, rank-order correlations between citation counts from different sources appear to be very high (Martín-Martín et al., 2018; Burgers, 2019). Thus, as long as the same source is consistently used for all replication candidates under consideration, the relative rank-order difference in citation count between two candidates should be highly similar regardless of which source we use.

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

One prominent difficulty with FWCI is determining the reference class to normalize citation counts against. That is, when taking the average citation count of a research field, which articles in the literature belong to that field? While it is common to use Web of Science field categories for this delineation, these are considered too heterogeneous to accurately control for variations in citation practices (Waltman and van Eck, 2013). Van Eck, Waltman, van Raan, Klautz, and Peul (2013) have demonstrated that FWCI fails to account for within-field variations in citations practices in medical science, and this can lead to a severe underestimation of the impact of clinical research. Another challenge when using FWCI to measure value is that some systematic differ-

[3]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 et al., 2015) and other hard-to-predict fluctuations in citation rate.

ences in citation counts between fields may genuinely be due to differences in the value society places on knowledge produced in different fields. For example, according to Web of Science, the field of oncology (cancer research) is nearly four times the size of dermatology (skin condition research) in 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 factor journal in the world (292 as of 2019). If the research field of oncology contains more researchers than dermatology - leading to more papers getting published and more citations generated, on average, for each published paper - this could reflect the fact that research in oncology is, on average, considered more valuable than 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 and 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. 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), S denotes the source the citation count is retrieved from, and Y denotes the age of the article in years (1 is added 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]. The Value of an article published less than a year ago simply equals the weighted

[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 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.

citation count of the article ($w(C_S) \div (0+1) = w(C_S)$), the Value of an article published one year ago equals half the weighted citation count of the article ($w(C_S) \div (1+1)$), and so on. Thus, Value is not identical, but closely related to the average yearly (weighted) citation count of the article.

We could substitute equation (4) with a more intricate estimator and likely get a more accurate impact estimate (Martin, 2011). However, there is an effort/accuracy tradeoff involved in replication study selection. The benefits gained by more accurate estimates of value need to be worth the additional efforts expended in collecting those estimates (e.g., in reviewing documentation, interviewing key stakeholders, running case studies, etc. Klautzer et al., 2011), since all effort could be avoided by simply choosing a replication candidate quickly based on random chance or personal interest, which would still lead to 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.

4.4.2 Sample size as an indicator of uncertainty

Uncertainty, like value, is a multi-determined attribute. We may be uncertain about a claim for a variety of reasons. The study design(s) used to test the claim may lack internal or external validity, we may have a high prior that counteracts existing research, statistical power might be too low to detect effect sizes of interest, original findings may not have been independently replicated (or independent replication has failed to reproduce the original study results), we may suspect that the evidence base is influenced by publication bias, selective reporting, p-hacking, or fraud, and so on. Many factors related to uncertainty could be quantified. For example, we can express our uncertainty about parameter estimates using confidence intervals or other variability indices, and we can express uncertainty about the relative likelihood of data given different hypotheses using 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 beliefs about claims. Alternatively, we could quantify uncertainty in terms of entropy using principles from information theory (DeDeo, 2018). If our uncertainty depends on results from multiple studies, we could combine their estimates through meta-analysis. The list goes on. However, the information required to compute informative Bayesian posteriors, Shannon entropy, etc., is often difficult to curate from published research reports, which makes the task of estimating uncertainty based on these indicators difficult in practice. To offer an indicator of uncertainty that is more easily derivable we here propose to estimate uncertainty about a claim via the sample size of the existing study (or studies) that investigate the claim.

Establishing the validity of sample size for measuring uncertainty requires a small chain of measurement assumptions (figure 4.3). First, given a set of findings comprising the original and the replication literature, we propose to approximate uncertainty as the precision with which a parameter relevant to the claim has been estimated (e.g., for the claim "stretching reduces the risk of athletic injury" we might be interested in the precision of risk ratio estimates from relevant randomized controlled trials). Precision of parameter estimates is obviously only one factor that makes up our overall uncertainty about a claim (Isager et al., 2020). Even so, it seems reasonable to assume that (all else being equal) increasing precision of a parameter estimate also decreases the uncertainty about claims based on that parameter. Statistically we can define precision of the estimate as the standard error of the parameter estimate[5].

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

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

where σ is the standard deviation of the estimate, and n is the corresponding sample size (i.e. the number of participants). When σ is constant, the equation is proportional to $1 \div \sqrt{n}$. In other words, we assume that the standard deviation is the same for all replication targets compared, and we assume that no other variance components but participant variance are relevant to our standard error estimate (Westfall et al., 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 estimated in multiple studies, we should set n to the total sample size over all studies (see appendix A 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 important factors that also contribute to uncertainty (dashed circles in figure

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

[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 $n = 1$.

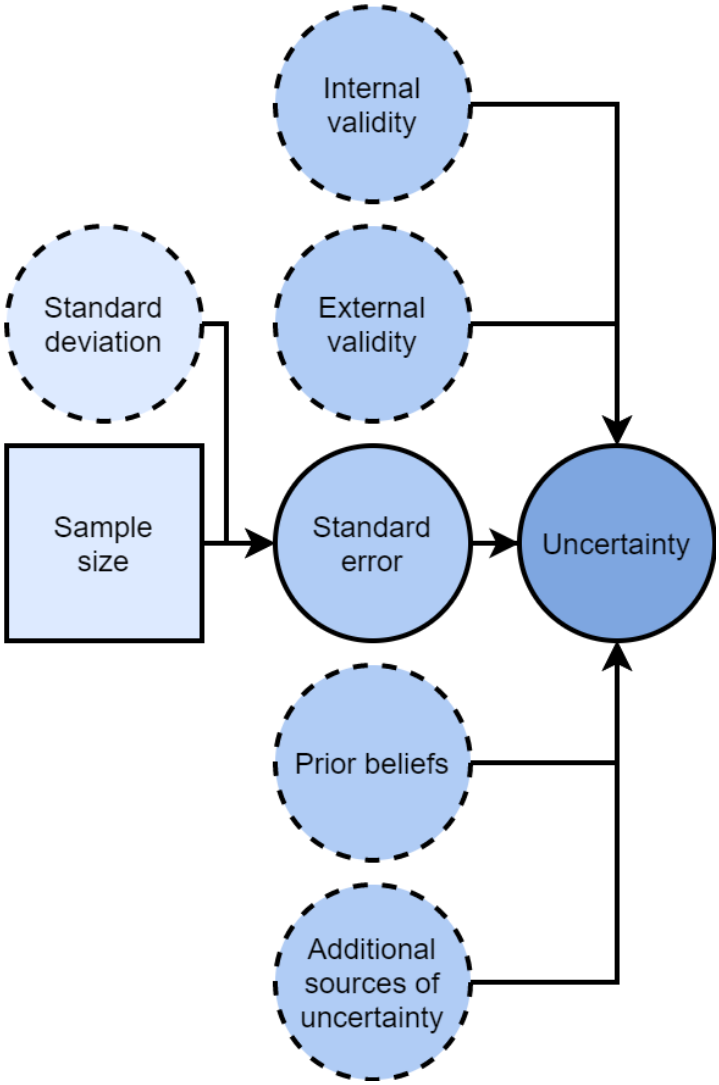

Figure 4.3: Proposed causal relationship between 'uncertainty' and 'sample size' ('sample size' is outlined by a square to signal it is observable). The figure can be interpreted as a directed acyclic graph model, where 'sample size' has a causal effect on 'uncertainty' by reducing the 'standard error' 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 (the unmeasured attributes U traditionally used in causal graph models). 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.

4.3). For example, we may be highly uncertain about a claim if the experiment used to support it is poorly designed, poorly conducted, or if the claim suggests that experimental effects can be generalized to a widely different context. This uncertainty may be completely independent of the sample size of the study. In addition, there are instances where participant sample size is not a strong determinant of the standard error. For example, participant sample size is a poor estimator of the standard error in mixed model designs where the random effect of participants is low compared with other random factors, such as stimulus (Westfall et al., 2014; Rouder and Haaf, 2018; DeBruine and Barr, 2021). The same happens in repeated measures designs where the number of repeated samples per participant can matter more than the number of participants per se, depending on the within-subject correlation (see appendix B for a method of correcting the sample size estimate in such cases). In general, because sample size is a cause of uncertainty, all determinants of uncertainty that are independent of sample size will tend to reduce the reliability of sample size as a measure of uncertainty. This problem is mitigated in more complex uncertainty estimators such as Bayesian posteriors, which can incorporate several independent causes of uncertainty into the estimate.

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

4.4.3 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:

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

where $\mathrm{RV_{Cn}}$ 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 4.4 represents the measurement model for $\mathrm{RV_{Cn}}$. It summarizes the causal assumptions that justify the use of $\mathrm{RV_{Cn}}$ as a measure of expected utility gain. Assuming that the casual relationships in the model holds, $\mathrm{RV_{Cn}}$ is d-connected with, and hence a valid measure of, expected utility gain. However, due to known unmeasured relationships (the U nodes in figure 4.4) $\mathrm{RV_{Cn}}$ can not be expected to be perfectly reliable. The exact reliability of $\mathrm{RV_{Cn}}$ is an important empirical question for future studies. Clearly, we must require a certain level of reliability in order to consider $\mathrm{RV_{Cn}}$ a *useful* measure of expected utility gain.

Figure 4.5 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 4.5 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.
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 either of these assumptions is violated, RV_{Cn} will cease to be a useful indicator of replication value.

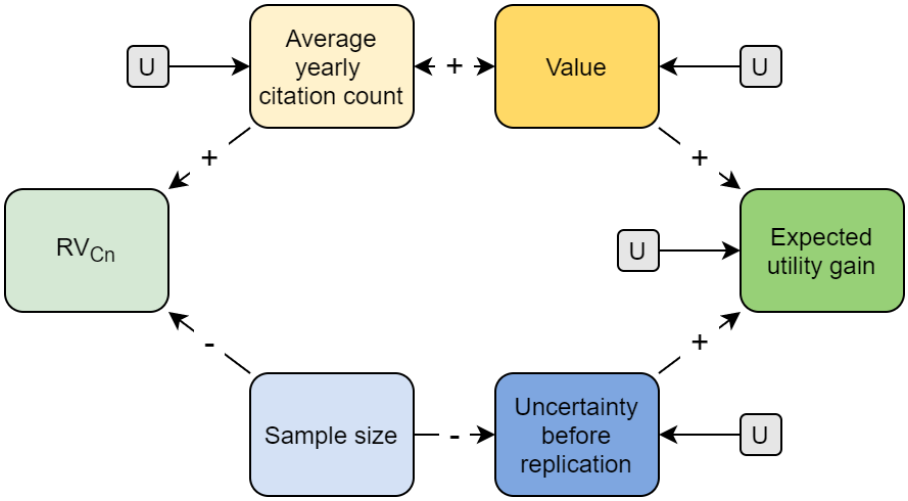

Figure 4.4: Measurement model justifying the validity of RV_{Cn} as a measure of *expected utility gain*. RV_{Cn} is d-connected with *expected utility gain*, as required for valid measurement by the d-connection definition of validity (Isager, 2020). The causal relations *value* $\rightarrow$ *expected 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 4.1 and figure 4.3. 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.

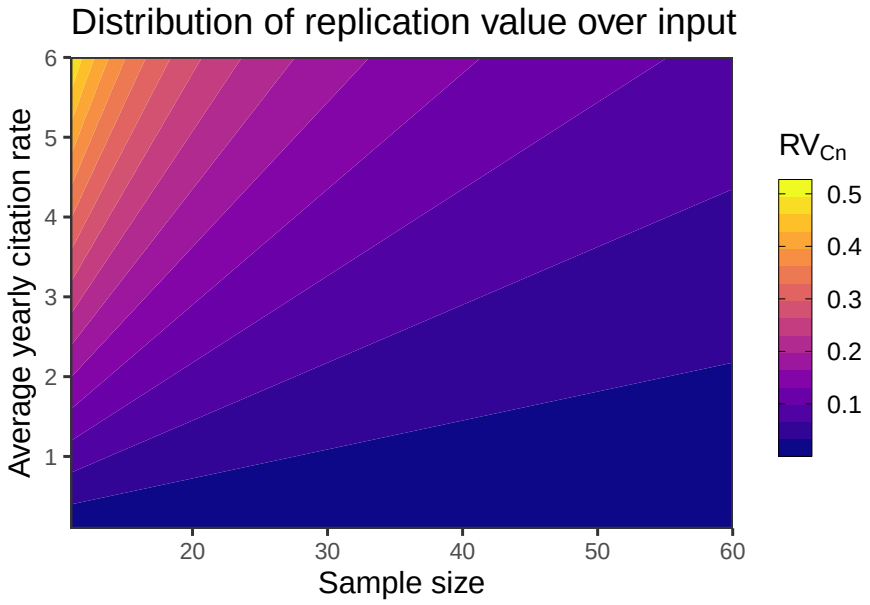

Figure 4.5: Distribution of $\mathrm{RV}_{\mathrm{Cn}}$ for a range of input parameters.

4.5 General study selection strategy

Having provided the measurement rationale for RV_{Cn}, we must now specify how RV_{Cn} can be implemented in a study selection strategy. In line with recommendations by Field et al. (2019), we propose a procedure for study selection that combines RV_{Cn} with a more comprehensive evaluation of expected utility gain. In-depth evaluation allows for 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 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 worthwhile to replicate. A more detailed evaluation process can then be applied efficiently to the highest RV_{Cn} candidates before a replication target is finally selected. Replication study selection thus follows a general four-step procedure, outlined in figure 4.6.

First, a set of candidate replication targets is curated. This set should contain all claims that are relevant to our interests and expertise. Targets that cannot be replicated due to feasibility-constraints can be excluded from the initial candidate set for efficiency. For example, suppose we want to form an initial candidate set of all experimental studies on the psychoactive effects of cannabis in healthy human participants. Curation of this set 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 set of candidates matches our research interests and feasibility constraints. Second, RV_{Cn} is calculated for all replication targets in the candidate set, so that replication targets can be rank-ordered relative to each other based on their RV_{Cn} estimates. Third, a subset consisting of the highest RV_{Cn} targets in our candidate set is selected for closer inspection. Assuming RV_{Cn} is a valid estimator of expected utility gain, this should increase the probability that the studies evaluated further are those most likely to lead to high expected utility gain if replicated. We can then invest a more substantial amount of time collecting both qualitative and quantitative information about which studies in this subset would be most worth replicating. How this evaluation should proceed is a topic of discussion (KNAW, 2018; Field et al., 2019; Pittelkow et al., 2020), and should probably be adapted depending on the goals of the stakeholder. However, the model put forth in Isager et al. (2020) suggests that evaluation should always include some consideration of (1) the perceived value of the research claim, (2) uncertainty about the truth status of the claim prior to replication, (3) the ability of the planned replication study design to reduce uncertainty about the claim, and (4) the estimated

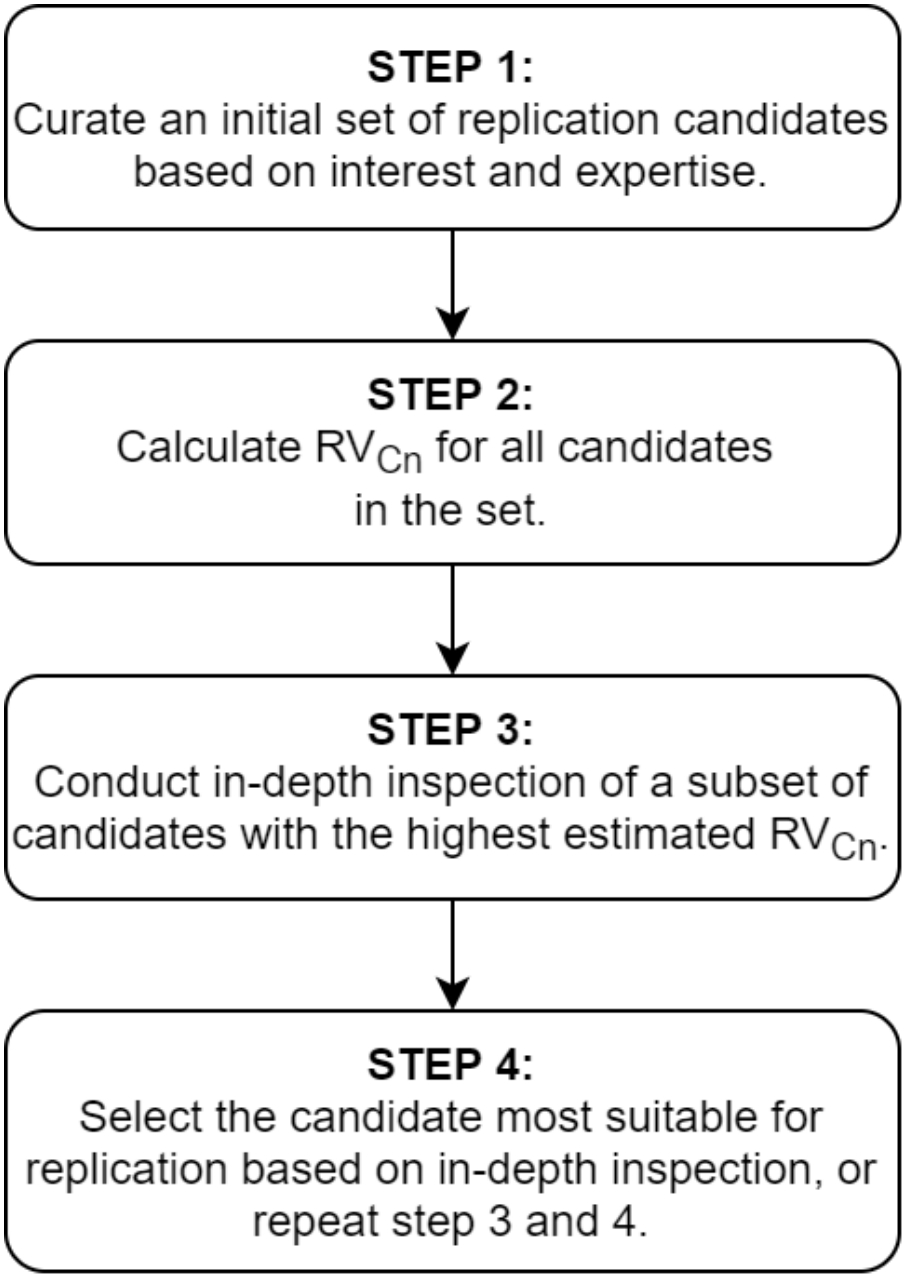

Figure 4.6: A general procedure for replication study selection. Quantitative estimation of replication value through RV_{Cn} (or through any other operationalization of replication value) can be implemented for study selection as part of this four-step procedure.

costs of conducting the replication effort. Empirical research to explore this issue in further detail is currently underway (Pittelkow et al., 2021). Finally, the replication target that is considered the most promising after comprehensive evaluation is prioritized for replication. Alternatively, if detailed evaluation suggests no study in the subset would be worthwhile to replicate, step 3 can be repeated until a suitable target is found.

4.6 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.

4.6.1 Methods

4.6.1.1 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 Curate Science replication dataset (https://curatescience.org/app/replications). As of 2020-10-20, the Curate Science replication dataset contained information about 1127 replications of 202 original studies, primarily from the field of psychology. For these original studies we estimated RV_{Cn} using sample size information available in the Curate Science dataset, and using publication year and citation count information from Crossref (https://www.crossref.org/. Citation counts were extracted from crossref 2020-10-20 using the rcrossref package in R. Chamberlain et al., 2020). Due

to missing DOI information, RV_{Cn} could not be calculated for 35 studies. The final sample of replicated studies contained 167 studies from 145 articles.

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 estimated RV_{Cn} for each article using the sample size available in the meta-analysis tables, and using publication year and citation count information from Crossref (citation counts were extracted from crossref 2020-10-20). To retrieve citation count information, the DOI of each study had to be retrieved, which was accomplished by applying a text-mining procedure to the bibliometric information available in the meta-analysis tables and reference lists. However, limited bibliometric information in the meta-analysis tables led to 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 4071 studies from as many articles.

4.6.1.2 Statistical analyses

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

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

4.6.2 Results

Statistical results are presented in table 4.1. Cube-root transformed distributions of the variables of interest are presented in figure 4.7. RV_{Cn} was, on

Table 4.1: 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	

average, substantially greater in the sample of replicated studies than in the comparison sample (figure 4.7D). 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 4.7A), even after adjusting for article age (figure 4.7B). Conversely, replicated studies had substantially lower sample size, on average (figure 4.7C).

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 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}.

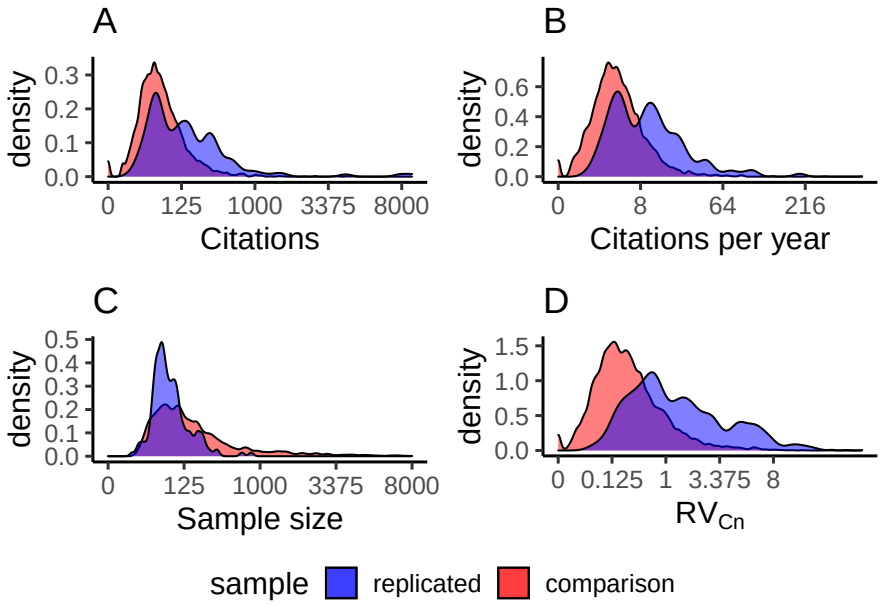

Figure 4.7: Distributions of various parameters in the comparison sample of psychological findings (red) and the sample of replicated findings in psychology (blue). The scale in all plots has been transformed by taking the cube root of the true values, which preserves the overall shape of the distribution 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.

Our design is limited in a number of ways. The citation data used for these analyses was collected in 2020 - years after the decisions to replicate were made. It is unclear what effects this delay may have. In addition, our comparison sample consisted entirely of empirical studies that are included in meta-analyses, which might differ from empirical studies that are not in several aspects, including citation count and sample size. Furthermore, RV_{Cn} could not be computed for a majority of the comparison sample due to missing DOI information, and cases may not be missing at random. Thus, the 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 study was selected for replication and select matched non-replicated controls from the same year and research field. Finally, the external validity of the sample of replicated studies 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, 2015). All studies replicated as part of this effort used the same general study selection procedure (see https://osf.io/ezrsc/ for details about the sampling strategy).

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

4.7 General discussion

To develop any strategy for efficient replication study selection, we need to (1) specify the goal we are trying to achieve, (2) specify how this goal is realized by the strategy we are proposing, and (3) empirically validate that the proposed strategy works as intended. These three steps must be addressed in sequential order. We cannot specify how a given strategy achieves a goal

without first specifying the goal, and we cannot validate whether a strategy works as intended before we can specify how the strategy is intended to work. Isager et al. (2020) address the first of these steps. In this article, we have focused on the second. Given the goal of maximizing expected utility gain (and the assumption that this attribute can be approximated by replication value) we have formalized an observable indicator - RV_{Cn}. We have formally specified the assumptions under which study selection based on RV_{Cn} will maximize expected utility gain, and we have discussed a number of obfuscating factors that limit the usefulness of RV_{Cn}.

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 incorporate information about both value and uncertainty, which is important whenever the goal of selection is to maximize expected utility gain. Should RV_{Cn} be a valid and reliable measure of expected utility gain, it holds enormous potential for improving the transparency and efficiency of resource allocation in replication research. The indicator could be utilized by any researcher, funder, journal, or other stakeholder who wishes to direct limited resources towards important replication targets. Because calculation only requires information about citation impact and sample size RV_{Cn} should be applicable for study selection across a variety of scientific disciplines. The strong measurement rationale on which RV_{Cn} is built also makes it possible to improve and adapt the indicator to different contexts. For example, in disciplines where sample size is a poor measure of 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).

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 rcrossref

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 validated. That is, we must empirically corroborate that RV_{Cn} is sensitive to variation in the actual expected utility gained from replication efforts. The lack of high-quality indicators of expected utility gain prevents straight-forward benchmarking. As a substitute we could attempt to identify plausible indicators of expected utility gain and compare RV_{Cn} to these, like is done in the study reported above. However, interpretation of results will be less straight-forward in such studies because we must add the auxiliary assumption that whatever attribute we compare RV_{Cn} with is itself a good measure of expected utility gain (Meehl, 1997). Alternatively, RV_{Cn} could be validated by corroborating the individual assumptions it is built on. The constituent causal hypotheses in the chain that leads RV_{Cn} estimates to be a valid measure of expected utility gain (figure 4.4) could be separated and tested in isolation. In principle, falsifying any single causal assumption would falsify the validity of RV_{Cn} overall, since valid measurement requires that all the causal relationships hold. Finally, instead of examining the validity of RV_{Cn} in isolation we could train machine-learning algorithms to predict indicators of expected utility gain based on a broad range of bibliometric and statistical information, as has recently been done to understand predictors of research *replicability* (Altmejd et al., 2019; Yang et al., 2020). This should yield a better understanding of the predictive power of citation impact and sample size relative to other types of information we could be considering.

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 accept the causal hypotheses outlined in figure 4.4, then by definition replication RV_{Cn} must be somewhat unreliable due to the influence of unmeasured noise factors U in the measurement process. The influence of the various noise factors will likely depend on the diversity of the candidate set we are considering. Consider a set of studies, all from the same subfield, all published within the span of a few years, and all utilizing similar study designs. Now compare this to a set consisting of replication candidates from across a wide time-span and many scientific fields, studying a wide range of topics with widely different study designs. Variation in the noise factors included in figure 4.1 and figure 4.3 will likely be much higher

in the latter set than in the former. Thus, the reliability of RV_{Cn} will likely decrease as the diversity of replication targets increases, and there might even be a limit to 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, consider the possibility that researchers prefer to cite articles that provide strong support for the claims tested within (i.e., citation count is reduced by uncertainty about the 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 4.8). Similarly, overestimation of expected utility gain can happen if research which we are more certain will *not* replicate tends to get cited more often. Recent research indicates the latter situation might often be the case (Serra-Garcia and Gneezy, 2021). Note that we could in principle repeat this thought experiment for any two variables in figure 4.4 not already joined by a causal arrow. Whenever the causal relationship implied by such an arrow seems plausible and causes systematic differences between the values of RV_{Cn} and expected utility gain, we should be worried about bias in the measurement.

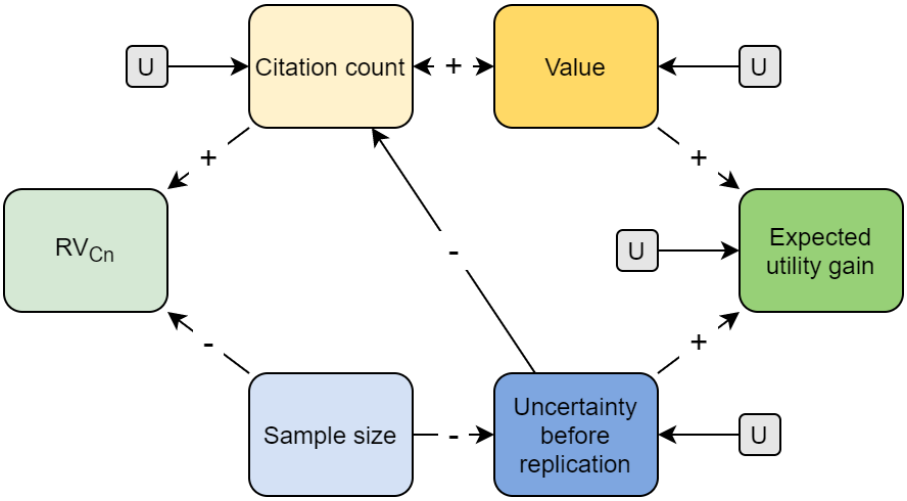

Figure 4.8: Causal scenario which will lead RV_{Cn} to become a biased measure of expected utility gain due to the influence of *uncertainty before replication* on *citation count*.

The 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 the feasibility and quality of the study selection criteria used. This assessment can only happen if we first specify the goal

of replication and formalize how the goal is achieved through the selection strategy we are proposing. Only then can we evaluate whether the strategy works as intended. For existing indicator-based strategies where the rationale behind the indicator used remains unclear (e.g., Makel et al., 2012; Field et al., 2019; Isager, 2018; Isager et al., 2020, supplementary materials), the measurement model should be formalized so that validation and comparison of different strategies can take place. The 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 us to examine whether a given strategy is likely to work as intended. In doing so we enable a more transparent discussion of what is important to replicate, which will be crucial for coordinating future replication efforts and direct resources to where they are needed the most. Through such discussion and coordination we increase the overall impact of replication studies on scientific discourse and strengthen the argument for why replications should be conducted and rewarded.

Chapter 5

Selecting Studies for Replication in Social Neuroscience: Exploring a Formal Approach[1]

5.1 Introduction

Close replication of original research results is essential for ensuring that the results can be reliably produced by different researchers (Poldrack et al., 2017). The practice is especially important in subfields of social and cognitive neuroscience where correlations between brain activity and behavioral outcomes are often of primary interest. These research designs tend to be highly vulnerable to error-rate-inflation and overestimation of effect size due to a combination of (1) low statistical power (Szucs and Ioannidis, 2017), (2) substantial researcher degrees of freedom (Carp, 2012; Botvinik-Nezer et al., 2020), and (3) incentives to publish statistically significant results (Button et al., 2013). In addition, neuroimaging research is generally vulnerable to generalizability issues that arise as a result of the complicated measurement pipeline. Unsurprisingly, rates of successful replications in the field are low (Boekel et al., 2015). And yet, close replications – which protect against persistence of false-positives – are not common practice (Poldrack et al., 2017; Huber et al., 2019; Ashar et al., 2021). As a consequence, many original studies are currently in need of replication. At the same time, the costs of data collection – and, by extension,

[1]This chapter is in preparation as Isager, P. M., Lakens, D., & van 't Veer, A. E., Selecting Studies for Replication in Social Neuroscience: Exploring a Formal Approach.

replication – is high (Poldrack et al., 2017). With limited resources and many non-replicated studies to choose from, the field should consider which studies in the published literature would be the most important to replicate, so that resources directed towards replication can be spent optimally.

Various formal strategies for replication study selection have been developed in recent years (Field et al., 2019; Matiasz et al., 2018; Isager et al., 2021; supplementary formula documents in Isager et al., 2020). If effective, such strategies have a great potential for increasing the transparency and efficiency of replication study selection in neuroimaging research. When criteria for study selection are made transparent, it becomes easier to discuss which replication studies are worthwhile to fund, conduct, and publish. Additionally, when important-to-replicate targets can be identified more easily the overall contribution of replication as a research activity increases. By increasing the efficiency of coordination and resource spending in replication research, formal study selection strategies present a major step forward towards the important goal of making replication part of mainstream research practice (Zwaan et al., 2018).

However, no formal study selection strategy has been tested for application in social neuroscience (or any other field for that matter). To be applicable, a strategy must meet two basic conditions. First, it must be feasible to apply the strategy in practice. That is, the information needed to execute the strategy must be possible to obtain given reasonable time and resource constraints. Most formal study selection strategies are based on a combination of statistical, bibliometric, and substantive information about the candidate replication targets, which is often not easy to access (e.g., Tay et al., 2020; Sullivan and Feinn, 2012; Furukawa et al., 2006; Glasziou et al., 2008; Federer et al., 2018). The feasibility of existing strategies for application in any particular area of research is therefore uncertain. Second, provided that the strategy is feasible to apply we must validate that the strategy is actually helping us reach our prespecified research goals. All feasible selection strategies lead to *some* prioritization of replication targets, but whether prioritized targets truly are those more in need of replication is an empirical question.

In this article we explore how to apply a particular study selection strategy (Isager et al., 2021) to fMRI research in social neuroscience. Because a strategy must be feasible to apply before it can be validated, we focus mainly on the issue of establishing the feasibility of selection strategies. By *establishing feasibility* we simply mean that we will explore whether a certain strategy can be carried out in practice (can we identify the set of replication targets, can we collect reliable estimates of the necessary data, and so on) given reasonable time- and resource constraints. We also provide a brief qualitative assessment of the targets recommended to us by the strategy, noting potential issues that future validation studies may want to examine more carefully.

The immediate goal in this article is to understand how best to implement this

strategy in social fMRI research so that the strategy can be validated for use in this field and potentially used for efficient study selection by future researchers. However, our implementation efforts could also be used as a framework to test the feasibility of alternative study selection strategies (e.g., Field et al., 2019). In addition, we aim to curate and openly share a large dataset of metadata about empirical studies from the social fMRI literature that could form the basis for any replication study selection in this field, regardless of the selection strategy utilized. Finally, by exploring the range of information relevant to replication study selection in social fMRI, we hope to leave researchers in this field better equipped to make well-informed decisions about which original research to prioritize for replication.

5.2 A four step approach to selecting studies for replication

To decide on a method for replication study selection, we must first settle on a goal, and a rationale for why selecting certain studies helps us reach this goal more efficiently. We here adopt the formal decision model for replication study selection proposed by Isager et al. (2020). According to this model, the goal of a replication effort is to maximize the *expected utility* of knowledge gained. *Expected utility gain* can be approximated by the *replication value* of the target claim we want to replicate. Replication value is a function of the *value* of having accurate knowledge about the replication target, and our *uncertainty* about the truth status of the target based on available evidence prior to replicating. Research claims that are highly valuable, and about which we are highly uncertain, will have a high replication value, and should be prioritized for replication in order to maximize expected utility gain.

We have previously proposed an operational definition of replication value (Isager et al., 2021), in which value is operationalized as the average yearly citation impact of the article in which a claim is reported, and uncertainty is operationalized as the sample size used to investigate the claim. Replication value is then operationalized as the indicator RV_{Cn}:

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

where $RV\ Cn$ denotes a particular operationalization of replication value, C stands for citation impact, n stands for the total number of participants included in the study, $w()$ stands for the weighting function that should be applied to the citation impact, s denotes the source the citation data is retrieved from, and Y stands for the age of the article in years. The equation assumes that average yearly citation impact is causally influenced by scientific impact, which itself is a determinant of the value of a replication target.

Sample size (partially) determines the standard error of relevant parameter estimates, which in turn is a determinant of the uncertainty about a replication target.

A four-step procedure for replication study selection based on RV_{Cn} is then proposed (see figure 5.1). First, an initial set of replication candidates is identified based on the research interests and resource constraints of the replicating researcher. As with every systematic review of the literature, the scope needs to be broad enough to encompass all claims of interest to the researchers, but narrow enough so that the review process becomes feasible. Second, RV_{Cn} is calculated for each replication target included in the set to create an initial estimate of rank-order expected utility gain. Third, some subset of the targets with the highest RV_{Cn} is inspected in-depth. This step functions to quality-control RV_{Cn} estimates, evaluate a broad range of factors relevant to value and uncertainty (e.g., Field et al., 2019; KNAW, 2018; Heirene, 2021), and consider feasibility given available resources. Similarly, the ability of the planned replication study design to reduce uncertainty should be considered for each candidate during this step (Isager et al., 2020). Fourth, once the researcher feels they have sufficient knowledge to make an informed choice, the candidate deemed most worthwhile to replicate is selected. Alternatively, if the researcher thinks no candidate would be worth replicating upon closer inspection, step 3 and 4 can be repeated for the remaining candidates in the larger set.

5.3 Exploring the feasibility of using RV_{Cn} for study selection in Social Neuroscience

RV_{Cn} represents a promising step towards more efficient coordination of replication efforts in social fMRI research. However, it is not clear whether RV_{Cn} is feasible to apply for study selection in social fMRI. The current report aims to address the many practical questions related to application of RV_{Cn} that are currently unresolved. How can we determine an initial set of replication candidates? Does it matter from which source we collect citation impact information? Can we code sample size for all candidates in a feasible and reliable way? What additional insights could be gained about expected utility gain from the in-depth evaluation process that follows quantitative ranking?

Our exploration focuses on the first two steps of the four-step procedure listed in figure 5.1. We report the results of our attempt to implement these steps in practice, including our method for collecting a sample set of replication candidates (step 1), our method for collecting the citation impact and sample size data necessary to calculate RV_{Cn}, the reliability of our methods for generating accurate citation count and sample size estimates, and the distribution of RV_{Cn} for our set of candidates (step 2). In appendix D and E we also

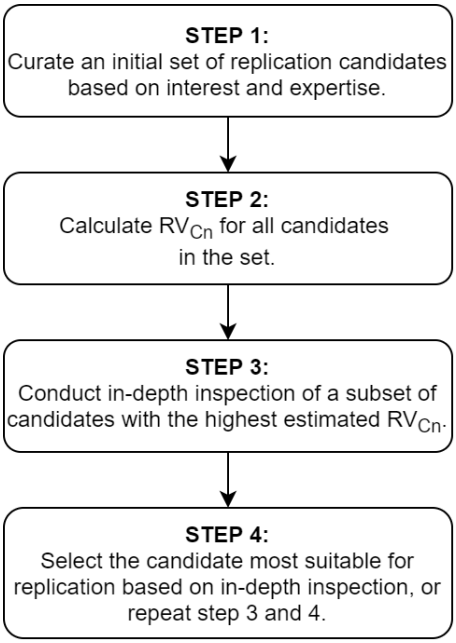

Figure 5.1: General study selection procedure in which the RV_{Cn} indicator is implemented.

briefly summarize unsuccessful pilot efforts to identify and collect additional quantitative information for the candidates in our set. Finally, we also provide a brief qualitative evaluation of the recommendations produced by RV_{Cn} to better understand what sort of studies are being recommended, and to get a sense of what factors one might want to consider in a comprehensive implementation of step 3. We end the article by generating hypotheses and offering suggestions for studies that could be undertaken to test the validity of RV_{Cn}, having established that it can feasibly be calculated in social fMRI research.

5.3.1 Step 1 - Determining an initial set of candidates

5.3.1.1 Eligibility criteria

To test the feasibility of calculating RV_{Cn} we first set out to determine a suitable set of candidate articles given our interest to perform a replication in social neuroscience. We would prefer if this replication effort would be focused on studies that would be highly worthwhile to replicate, yet reviewing the entire literature is not feasible. We restricted our search for replication targets to fMRI research within social neuroscience, placing no further restrictions on topics and subfields of interest. We further restricted our candidate set based on article age. We reasoned that recently published empirical research will have exerted less of its full potential impact on the field compared to older research, such that a replication of recent research will more likely be conducted "in time" to still prevent unproductive follow-up research (when the original research is non-replicable) and stimulate productive follow-up research (when the original research is replicable). We therefore restricted ourselves to articles published in the last eleven years (2009-2019 at the time this decision was made).

Social fMRI articles were collected using two separate search strategies which we further detail below. From this initial set of articles we then excluded articles we did not believe would be feasible for us to replicate given our expertise and available resources, which meant excluding animal model research, highly invasive study designs, imaging methods outside our area of expertise, research on patient groups, etc. Since we did not know all the reasons why a study might not be possible for us to replicate, exclusion criteria were not predetermined, but were exploratorily derived through inspecting keyword information in our initial candidate set. To ensure transparency a written record of the decision rationale for each excluded keyword has been made openly available on OSF (https://osf.io/mtx72/).

5.3.1.2 Search strategy

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

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

The search was conducted on 2019-02-21. 5636 articles were identified via this search strategy.

Searching only field-specific journals is bound to miss many important articles published in general topic journals such as PLOS ONE, PNAS or Neuroimage. To identify social neuroscience articles in other journals we also searched the entire WoS database for articles containing the keywords "social" and "fMRI" in either title or abstract. Empirical articles containing the relevant keyword information were identified by submitting the following search term to WoS:

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

The search was conducted on 2019-02-21. 2706 records were identified via this search strategy.

5.3.1.3 Selection process

Unsurprisingly, the two search strategies yielded overlapping results, as articles published in the four social neuroscience journals we identified also frequently contained the keywords "social" and "fMRI". After removing duplicate records, the two search strategies yielded 7413 unique empirical articles in total (see figure 5.2). Basic bibliometric information about each article, including author-provided keywords, were downloaded for all articles.

Authors PMI and AvtV reviewed the initial set of articles and determined for each article whether author-provided keywords indicated that replication would require access to samples or equipment that would not be feasible for us to obtain (the complete list of excluded keywords, categories, and exclusion rationales can be found at https://osf.io/mtx72/). All articles whose keywords, titles or abstracts matched keywords in our list were excluded. Our final set of candidates contained 2268 empirical articles. These articles were considered our initial candidate set.

5.3.1.4 Exploration of sample representativeness

Once the final set of candidate records was determined, we explored the available bibliographic information to ensure that the sample indeed seemed representative of the field of social fMRI research. To verify that the initial candidate set curated was representative of the population we wanted to sample from, two complementary validating questions were explored. First, does the sample adequately sample the strata of our population of interest? That is, are research topics and themes common in the population represented in the sample, and does the sample seem sufficiently diverse in terms of topics, themes, and subfields, given our expectation of diversity in the population sampled from? Second, does the sample contain strata from populations we did not intend to sample from? That is, are topics and themes prevalent in the data that obviously do not belong in the research field we are trying to sample from?

To address these questions we explored a number of bibliometric indicators of the research topics contained within a set of articles, including (1) the journals in which articles in the initial set were published, (2) the WoS field categories (Clarivate Analytics, 2020) assigned to articles in the initial set, and (3) topic-related keyword information gathered from the articles themselves. The following sections summarize the analyses of each indicator in turn.

The analyses reported in the following sections only summarize a subset of all bibliometric information available for our initial candidate set of articles. Additional information includes all WoS Core Collection fields, the open access status of each article, and a range of citation metrics. The full dataset, including all bibliometric variables and a variable codebook, are available on OSF (https://osf.io/f7zdq/).

5.3.1.4.1 Distribution of articles over journals We explored the distribution of journals in our data, including the topic specializations implied by the journal titles and the frequency with which articles were published in each unique journal. We addressed our two validating questions by examining (1) whether there was a substantial spread of articles over general and specialty journals, as we would expect in a sample of social fMRI articles, (2) whether

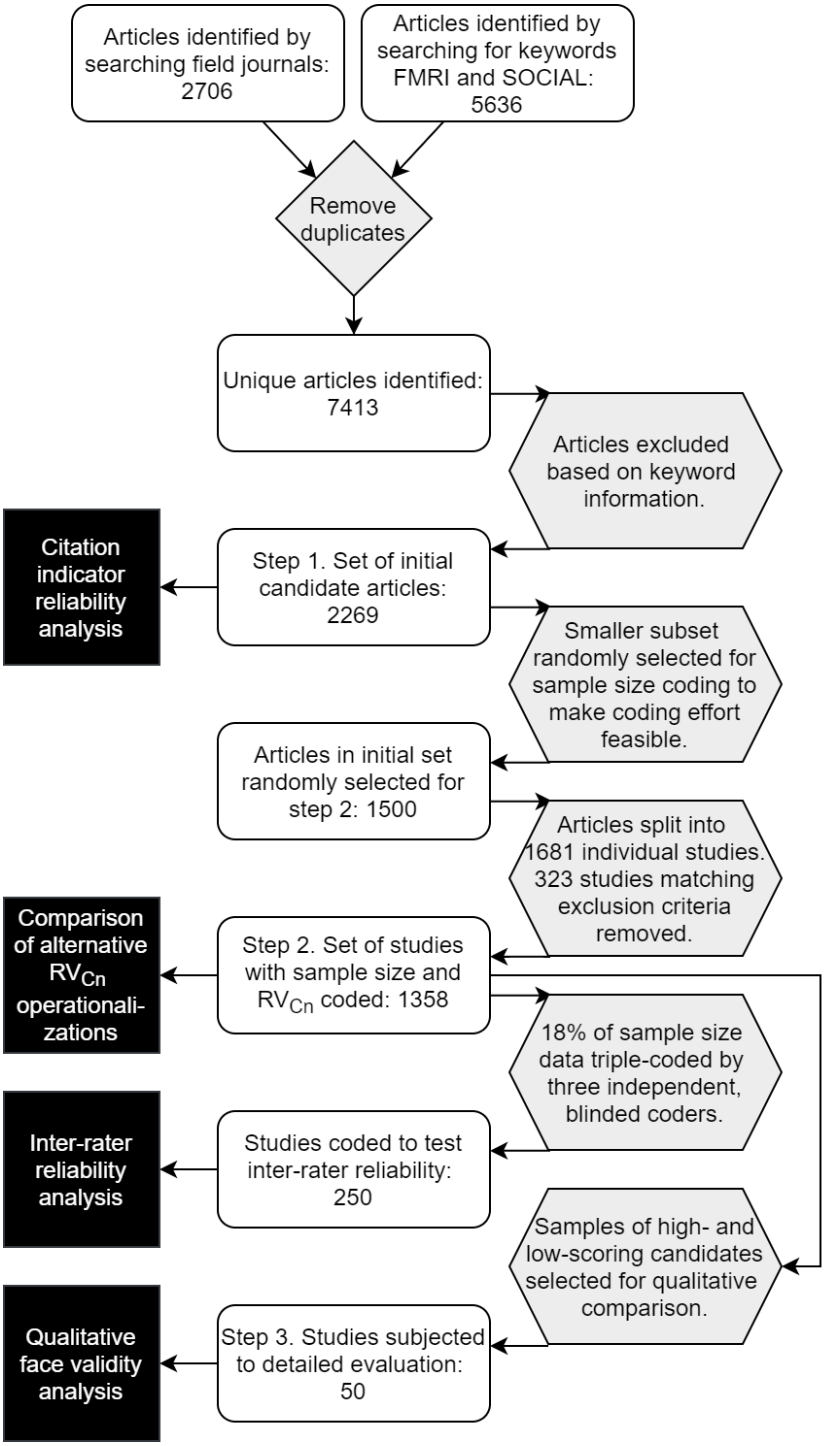

Figure 5.2: Overview of candidate selection process and data points available for each respective analysis reported below.

journals known to be prevalent in social fMRI research were frequently published in, and (3) whether journals obviously not related to social neuroscience research were infrequently published in.

The records included in our dataset were published in 329 unique journals, consistent with our expectation that social neuroscience is a broad and loosely connected discipline of researchers from many subfields, who publish in a variety of specialty- and general-topic journals. table 5.1 displays the name and frequency of the 20 journals most frequently published in (0.71 of all articles in the set were published in these 20 journals). Unsurprisingly, two of the four social neuroscience journals targeted by our search strategy were also among the most prominent journals in the candidate set (Social Cognitive and Affective Neuroscience, and Social Neuroscience). As to why the other two preselected journals were not featured, the journal "Socioaffective Neuroscience & Psychology" only contributed 19 articles to the initial candidate set, of which 17 were excluded based on keyword information. The journal "Behavioral Neuroscience" contributed 810 articles to our initial candidate set, of which 805 were excluded based on keyword information. Besides the preselected journals, the sample appeared to be dominated by journals that were either general-topic (PLOS One and PNAS) or general neuroscience/psychology (e.g. Neuroimage, Frontiers in Psychology, Cortex), which is broadly in line with our expectations about which journals ought to be prevalent in social fMRI research.

Inspecting the full distribution of unique journal names yielded similar observations. Most journals appeared to be either specialty journals within various subfields of psychology and neuroscience or general topic journals whose field of interest might plausibly overlap with social neuroscience research. Only a small minority of journals seemed clearly outside the normal scope of social fMRI research (ACM Computing Surveys, Physical Review E, The Accounting Review, as well as a small number of journals dedicated to statistics).

5.3.1.4.2 Distribution of articles over Web of Science field categories We explored the distribution of WoS field categories linked to our data, including the range of categories mentioned and the frequency of each category. We addressed our two validating questions by examining (1) whether a substantial spread of categories from psychology and neuroscience were covered, as we would expect in an interdisciplinary research field, (2) whether categories expected to be prevalent in social fMRI research were frequent in our dataset, and (3) whether categories obviously not related to social neuroscience research (e.g. "engineering, petroleum") were infrequent in the data.

The records in our dataset were classified as being members of 178 unique WoS categories. table 5.2 displays the name and frequency of the 20 WoS categories most frequently tagged (0.865 of all articles in the set were sorted under these 20 Web of Science categories). This distribution is largely consistent with what we would expect to see in studies sampled from social neuroscience research,

Table 5.1: Journals which the articles in our initial candidate set were most frequently published in.

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

with many categories covered and with categories such as "Neurosciences; Psychology; Psychology, Experimental" and "Multidisciplinary Sciences" being among the most common. However, it is somewhat surprising that categories such as "Psychology, Social" and "Neuroimaging" are not more prevalent in a dataset that is supposed to contain fMRI studies of social psychological phenomena.

Inspecting the full range of WoS category labels yielded similar observations. Most labels were clearly related to areas of behavioral science, psychiatry, neuroscience and biology, consistent with the interdisciplinary branches of social fMRI research. Only a few category labels seemed obviously unrelated to social neuroscience ("Engineering, Mechanical, Information Science & Library Science", "Ophthalmology", "Physics, Fluids & Plasmas; Physics, Mathematical", and "Urology & Nephrology"). Of the articles sorted under these categories in our data, only one article (Colman and Vukadinović Greetham, 2015) seemed truly unrelated to social neuroscience on closer inspection.

5.3.1.4.3 Frequently co-occurring article keywords Journal- and WoS category information encodes the research topics covered in our candidate set on a course level, since any journal or field category could include a wide variety of research topics (e.g., research topics as diverse as empathy, face perception, peer influence, and working memory all feature in the latest journal issue of Social Neuroscience; Vol 16, issue 3). To supplement the preceding analyses of journal and WoS field categories with more fine-grained information, we utilized the statistical visualization software VOSviewer (van Eck and Waltman, 2010) to extract commonly mentioned terms from the titles and abstracts of all studies, and we studied whether terms co-occurred in line with our prior knowledge of terminology in different subfields of social neuroscience. Additional analyses of keywords retrieved from the Centre for Science and Technology Studies (CWTS, https://www.cwts.nl/) are reported in appendix C.

All data included in the initial candidate set were subjected to analysis in VOSviewer (co-occurrence map with parameters set to binary counting, minimum number of occurrences set to 15, maximum number of keywords set to 200. Age-related and generic terms were excluded. The list of excluded keywords and map files to recreate the reported co-occurrence map can be found on OSF: https://osf.io/f7zdq/). figure 5.3 displays the co-occurrence map between commonly mentioned keywords in our dataset (online interactive version of the figure: https://bit.ly/3yDPMup).

The VOSviewer co-occurrence map corroborates the findings of the previous analyses. Themes commonly studied in social neuroscience frequently co-occur in the titles and abstracts of articles in our data. Further, it seems that individual topics could be organized into larger categories based on keyword co-occurrence clusters (represented as keyword colors in figure 5.3, van Eck

Table 5.2: Web of Science field categories most frequently tagged in our initial candidate set.

Field	Frequency
Neurosciences; Neuroimaging; Radiology, Nuclear Medicine & Medical Imaging	345
Neurosciences; Psychology; Psychology, Experimental	333
Neurosciences	291
Neurosciences; Psychology	225
Multidisciplinary Sciences	222
Behavioral Sciences; Neurosciences	112
Neurosciences; Psychology, Experimental	97
Psychology, Multidisciplinary	81
Behavioral Sciences; Neurosciences; Psychology, Experimental	77
Psychology, Experimental	38
Psychology, Social	27
Audiology & Speech-Language Pathology; Linguistics; Neurosciences; Psychology, Experimental	19
Psychology, Developmental; Neurosciences	18
Endocrinology & Metabolism; Neurosciences; Psychiatry	13
Psychiatry	13
Psychology, Biological; Neurosciences; Physiology; Psychology; Psychology, Experimental	12
Anatomy & Morphology; Neurosciences	10
Neuroimaging	10
Neurosciences; Pharmacology & Pharmacy; Psychiatry	10
Neurosciences; Physiology	9

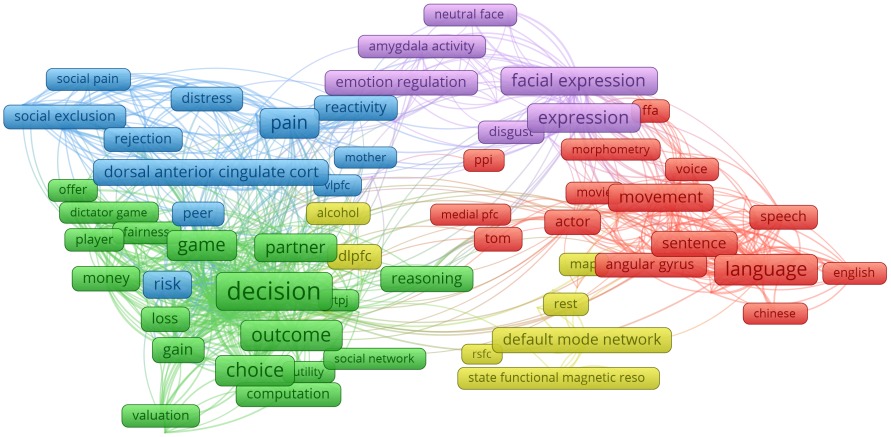

Figure 5.3: VOSviewer co-occurrence map of substantive keywords retrieved from the title and abstract of articles in our dataset. Colors represent VOSviewer-defined clusters of closely related keywords. See van Eck and Waltman (2014) for further details on clustering in VOSviewer.

and Waltman, 2014). As expected from a set of articles sampled from social neuroscience, these categories center around themes such as face perception (purple cluster), judgement and decision-making (green cluster), language (red cluster), and social pain/ostracism/exclusion (blue cluster). The default mode network (yellow cluster) also has clear ties to social neuroscience research (Li et al., 2014).

5.3.1.5 Overall evaluation of step 1 implementation

Converging lines of evidence suggest that our search strategy and selection process was successful in curating a dataset both representative of, and exclusive to our target population of healthy human social fMRI research. Note that our sampling and selection process was largely constructed to overcome the problem that social fMRI is not a well-defined bibliometric category. Determining an initial set of candidates will likely be more straightforward when the field of interest aligns more closely with a well-defined bibliometric category (e.g., a WoS field category).

5.3.2 Step 2 - Deriving quantitative replication value estimates

Having determined a set of candidate articles to consider for replication, we next set out to quantitatively estimate the replication value for each replication target in this set (see figure 5.1, step 2). Following Isager et al. (2021) we chose RV_{Cn} as our operationalization of replication value (equation (1)). This means we operationalize value as a function w of target article citation impact C derived from a particular source S, divided by publication year (Y). We operationalize uncertainty as the sample size (n) of the target study.

5.3.2.1 Operationalizing value as citation impact

In practice, we need to specify what function w, type of citation impact C, source S, and sample size n we intend to use for calculating RV_{Cn}. To examine the impact of choosing one specification over another, we studied the reliability of citation impact estimates across a range of impact types C, sources S, and functions w. Two qualitatively different types of citation impact C were collected; traditional academic citation indexes and Altmetric attention scores. Altmetric attention scores were collected using the *rAltmetric* package in R (Ram, 2017, download date: 2020-10-30). Altmetric attention scores are a weighted count of news- and social-media attention an article has received (Altmetric, 2021). For traditional citation impact, we collected data from multiple sources, including WoS (collected 2020-11-07 using the WoS web interface), Crossref (collected 2020-10-30 using the rCrossref package in R. Chamberlain et al., 2020), Scopus (collected 2020-10-30 using the rScopus package in R. Muschelli, 2019), and CWTS (collected 2020-10-28 from the CWTS database by author TvL). WoS, Crossref and Scopus citation counts are all unweighted raw counts of incoming citations of an article. CWTS citation counts consist only of incoming citations that are not self-citations. We also collected field- and age-normalized citation counts from the CWTS database (for details about the normalization procedure, see Waltman et al., 2011). Thus, our data contained three different functions w of citation impact (raw count, self-citations subtracted, and field/age-normalized). Publication year data Y was collected from the WoS database.

5.3.2.2 Operationalizing uncertainty as sample size

Following the rationale of Isager et al. (2021), we operationalized the uncertainty about a claim before replication in terms of the standard error of effects supporting the claim, which we then approximated as the sample size of the study in which the claim is reported. Sample size was here defined as the total number of participants in a study for which fMRI data was reported (i.e., the

number of participants that were not excluded from all fMRI analyses). Sample size was manually coded by undergraduate students at Leiden University and Eindhoven University of Technology.

We originally aimed to collect additional information about uncertainty (e.g. information about statistical analyses, experimental design, whether the study was exploratory or preregistered, number of existing replications, etc.) that could be used to derive alternative operationalizations of replication value. We subsequently planned to compare estimates from the RV_{Cn} indicator with other proposed indicators of replication value (e.g., Field et al., 2019, which requires information about bayes factors). However, following two pilot studies aimed at identifying additional uncertainty information (see appendix D and E), we concluded that additional information relevant to uncertainty-assessment in social fMRI would not be feasible for us to collect. The primary reason for this was the difficulties we faced in identifying the main finding of interest in each study, which would be necessary in order to know what test statistics, design features, replication studies etc. would be relevant for the replication target. The secondary reason was that, even if a main finding could be identified, both finding the corresponding result and finding a consistent method of reporting results across the field was difficult. In the end, sample size was the only operationalization of uncertainty we were able to move forward with in this study.

5.3.2.3 Collecting and inspecting the reliability of RV_{Cn} input

In exploring how RV_{Cn} should be calculated based on the quantitative information listed above we first studied the reliability of citation impact estimates across sources and weighting schemes to gauge how much these factors matter for the final citation impact estimate. Second, we considered the influence of age on citation impact, and we estimated how well this influence is mitigated by dividing citation impact by article age. Third, we explored the feasibility and reliability of coding sample size for target studies in the candidate set. Finally, based on the preceding analyses we designed two alternative replication value indicators and examined the divergence in their respective estimates.

5.3.2.3.1 Reliability of citation impact across sources To better understand the reliability of citation impact C across sources S, we explored the strength of association between a variety of citation metrics (table 5.3).

Table 5.3: Frequency of various citation metrics available for our data. Web of Science citation counts were originally available for all articles, but some could not be retrieved when the citation count data was updated in 2020.

Metric	Description	N
WoS	Web of Science Core Collection Times Cited Count, updated 2020-11-07	2105
Crossref	Crossref citation counts, downloaded 2020-10-30	2253
Scopus	Scopus citation counts, downloaded 2020-10-30	2238
CWTS	Total Citation Score. CWTS citation counts - excluding self-citations, downloaded 2020-10-28	2220
CWTS normalized	Total Normalized Citation Score. CWTS citation impact of article relative to the primary cluster to which the article belongs. The score represents how many more times the article is cited relative to the average citation count of an article in its cluster from the same year. I.e. An article that is cited 10 times, and that belongs to a cluster in which articles of the same age are cited 4 times on average, will receive a tncs score of 10/4=2.5	2220

Metric	Description	N
Altmetric	Altmetric attention score, downloaded 2020-10-30. The number of missing Almetric attention scores were high compared with other citation metrics. This is because attention scores are not tracked for an article until Altmetrics have detected at least one mention of the article. Traditional citation metrics separate between articles that are tracked but cited 0 times and articles that are not tracked, but Altmetrics makes no such distinction. Missing Altmetric attention scores appeared randomly distributed over the distribution of traditional citation count from all sources.	1874
Total	Number of articles for which all citation metrics were available	1706

All metrics were retrieved within a timespan of two weeks to ensure that there would be no differences in citation impact from different sources due to time-lag. Due to the skewed distribution of all citation metrics, and because we are primarily concerned with the rank-ordering of the records (Isager et al., 2020) Spearman's rho correlation was used to assess the strength of association between metrics.

In addition, we expected WoS, Crossref, Scopus, and CWTS to be highly correlated measures of the same underlying construct - the raw academic citation impact of an article. To test this expectation, we subjected the citation data from these sources to an intraclass correlation analysis (model = two-way fixed effects, type = single rater, definition = consistency, Koo and Li, 2016) using the ICC function in the R package *psych* (Revelle, 2021, ICC3 output reported).

Figure 5.4 displays the distributions of all citation metrics. All metrics are heavily right skewed. The distributions of raw citation counts are highly overlapping across sources (Figure 5.4A). CWTS citation counts are more heavily skewed towards zero than raw counts from other metrics, likely due to the fact that CWTS subtracts self-citations from the total citation count.

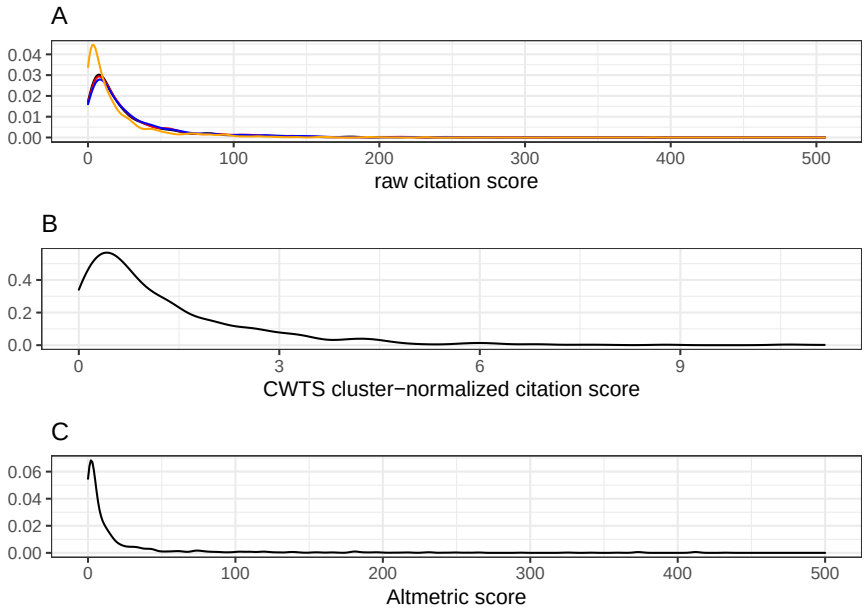

Figure 5.4: Density distribution of citation metrics. **A)** The distribution of raw citation counts from Web of Science (black), Crossref (red), Scopus (blue) and CWTS (orange). **B)** The distribution of CWTS citation impact, normalized by research field/cluster. **C)** The distribution of Altmetric attention scores.

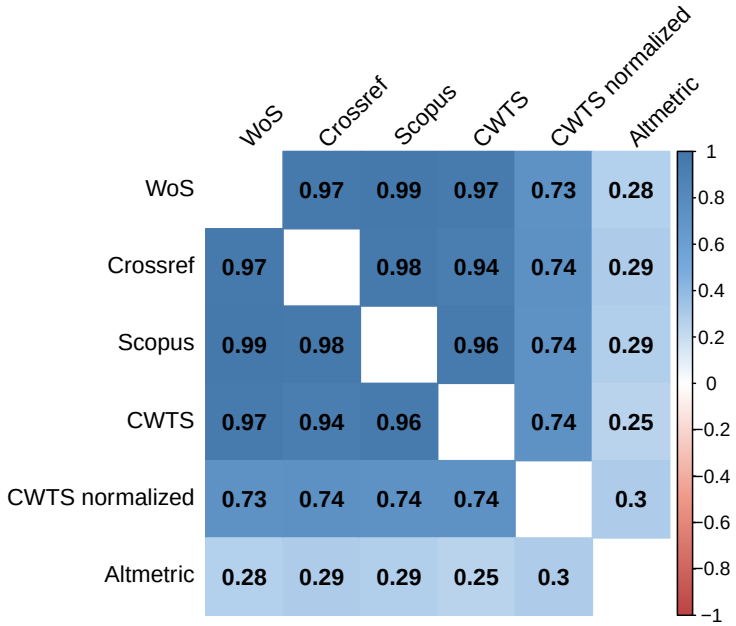

Figure 5.5: Matrix of bi-variate correlations between the citation metrics available for the articles in our dataset.

Figure 5.5 displays the rank-order correlations between various citation metrics. The correlation between raw citation counts from any two sources was very high (always >0.936). The inter-rater reliability between these metrics was similarly high, ICC = 0.97, CI95%[0.968, 0.971]. Even though self-citations are subtracted from CWTS citation counts, these scores were only marginally less correlated with scores from the three other sources, compared to intercorrelations between the other sources.

As expected based on the prior literature (Costas et al., 2015) the correlations between Altmetric scores and all other metrics were consistently quite low. The correlation between normalized and non-normalized citation counts was consistently high across sources, though substantially lower than the inter-correlation between different raw citation counts. This suggests that it matters little for RV_{Cn} estimates which source S is used, but it does matter whether one chooses raw or field-normalized citation count as the operationalization of $w(C)$, and it would matter substantially whether one chooses to use traditional citation count or news/social-media impact as the operationalization of C. The reliability of Altmetric attention scores as estimates of news/social-media impact remains unclear, as we had no other metrics for this kind of impact to compare against.

5.3.2.3.2 Age and citation count. Because total citation count is a metric that accumulates over time, it is strongly influenced by publication age. The upshot is that value estimates based on raw citation count will tend to be overestimated for older claims (Isager et al., 2021). In other words, total citation count will give the impression that older claims are more valuable than younger claims, even if there is no change in value of claims studied over time. To prevent such systematic measurement error RV_{Cn} attempts to adjust for publication age by using average yearly citation count as a measure of value.

To explore the effectiveness of this method for age adjustment, we examined how the correlation between age and citation count changed as raw citation count was transformed into average yearly citation count. We focus on WoS citation count data in this analysis, but the reported pattern of results is highly similar regardless of which citation source is used (see appendix F). We similarly examined the effect of age-averaging on Altmetric attention scores. In addition, we explored the relationship between age-averaged citation count and age/field-normalized CWTS citation count, which could be considered a superior method of age adjustment. If age-averaging is an effective method for age adjustment, age-averaged citation count should correlate more strongly with CWTS normalized scores than raw citation count.

We computed pairwise spearman correlations between publication age, WoS citation count, Altmetric scores, WoS citation count divided by years since publication, Altmetric scores divided by years since publication, and CWTS normalized citation count.

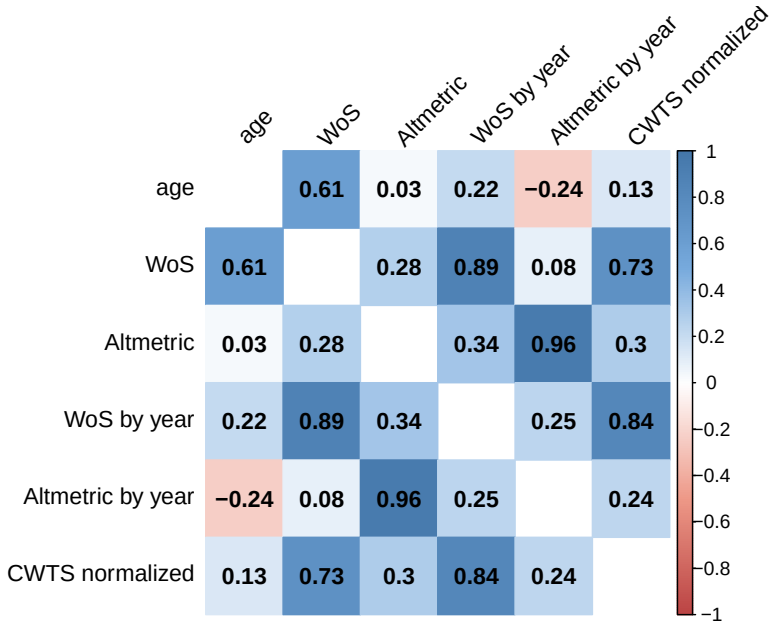

Figure 5.6: Matrix of bi-variate correlations between age and citation indices.

Figure 5.6 displays the correlation coefficients between all variables of interest. As expected, there was a strong correlation between age and raw WoS citation count (rho = 0.608). The correlation between citations and age dropped substantially when citation count was divided by years since publication. However, a meaningful residual correlation between average yearly citation rate and publication age remains (rho = 0.218). This suggests that dividing total citation count by the number of years since publication is an imperfect age adjustment method. Averaging over age works best if citation time accumulates at a constant rate, but this rate is very likely *not* constant for most articles (Isager et al., 2021). Encouragingly, however, averaging citation count by age does increase the correlation between citation count and CWTS normalized scores, whose method of age correction is superior to averaging because it does not depend on a constant accumulation rate. Interestingly, even CWTS scores are weakly positively correlated with age, suggesting that perfect adjustment for article age is challenging. In summary, taking the average yearly citation count seems to be an efficient method for age adjustment in traditional citation metrics.

Unexpectedly, there was only a negligible correlation between age and Altmetric attention score in our data (rho = 0.034). It is not entirely clear why Altmetrics, which also accumulates over time, is not dependent on publication age. It may be due to differences in the processes that generate accumulation of traditional citations and altmetric scores, respectively. For WoS citations to accumulate, citing articles must themselves be published, which means the accumulation of citation impact is a slow process stretched out over many years. In contrast, Altmetric data such as blog citations, news report citations and retweets can accumulate quickly, and may also subside more quickly as the article fades from immediate public attention. Thus, Altmetric scores may not be dependent on age beyond the first few weeks or months following publication (Chi et al., 2021). This hypothesis is further supported by the fact that the correlation between Altmetric scores and traditional citation count seems largely independent of age-correction. When there is no effect of age on Altmetric scores, there is no need to control for age, and dividing the scores by age creates an artificial negative effect of age on the Altmetric score (rho = -0.239). The upshot is that we will tend to overestimate the citation impact of recently published articles. Unless we purposefully want to bias our study selection towards more recent articles it would be more appropriate to use raw Altmetric scores than age-corrected Altmetric scores when RV_{Cn}.

5.3.2.3.3 Sample size inter-rater reliability There does not, as far as we know, exist any tool for extracting sample size from research articles automatically. Applications such as Statcheck (Nuijten et al., 2017) may allow for this in the future through detection of relevant degrees of freedom in articles. However, we were not able to utilize Statcheck for our data collection efforts because articles in our dataset rarely report statistics in the format required for

the algorithm to work. Sample size for each study in our dataset was therefore coded manually. In this section we first report our procedure for extracting a subset of the full set of candidate articles for which sample size could feasibly be coded. We then report our procedure for coding the sample size of each study included in this subset. Finally, we report the results of an inter-rater reliability analysis designed to investigate the ability of coders to code sample sizes reliably and without error.

From the outset it was unclear to us how challenging it would be to code sample size information for the studies in our candidate set. However, manually coding sample size for all studies in the full set of 2268 candidate articles was assumed to be costly and time consuming from the outset. In lieu of prior information about the speed and reliability with which sample size can be manually coded, we tentatively aimed for a subset of 1000 coded studies. We sampled 1000 articles at random from the full set of 2268 articles and began the process of splitting these into individual studies for sample size coding. However, in the early stages of coding it became clear that many studies would have to be excluded due to not meeting our initial selection criteria upon closer inspection. To ensure that the final set of candidates would include at least 1000 studies, an additional sample of 500 articles were sampled at random from the full set. The exact code used to draw the sample is available on OSF (https://osf.io/rxukq/). After removing articles that matched our initial exclusion criteria (e.g., single non-fMRI studies from multi-study articles, such as De Vries et al., 2018, study 4) the sample size was coded for each fMRI study in the article.

Coding was primarily performed by a team of three undergraduate research assistants. For each article we identified the number of studies reported in the article. For each study we recorded the number of participants who contributed any fMRI data to analyses reported in the study (even if their data were excluded from some analyses). We did not code more detailed sample size information such as the number of stimuli and trials used in each study. Although such information is obviously important for accurate estimation of overall statistical uncertainty (Westfall et al., 2014) piloting efforts suggested it would be too difficult to extract this information for over 1000 studies. For further details about how coders were instructed to proceed with sample size coding, see the supplementary coding instructions (https://osf.io/j3pxf/).

The 1500 articles contained 1681 individual studies, of which 323 matched our exclusion criteria. The final dataset contained 1358 individual studies from 1283 unique articles. On average, coders reported that coding the sample size of a single article would take a few minutes, but time taken was not normally distributed. Most studies could be coded in only a minute or so when sample size and exclusion criteria were clearly summarized in either the study abstract or the "participants" subsection of the methods section. A smaller subset of studies would take several minutes to code, usually because study authors

would not report sample size clearly, or would not report clearly if data were excluded, meaning the entire methods and results sections would have to be read before sample size could confidently be coded.

Manual coding introduces human error. In addition, discrepancies between coders can emerge when it is not clear whether participants should be excluded from a study, whether a study should be excluded from the candidate set, etc. In order to ensure that sample size estimates were reliably coded, a subset of 250 studies, randomly selected from the larger set of 1358, were double-coded by independent coders and subjected to an inter-rater reliability analysis. Two additional coders (one additional undergraduate student - the undergraduate coder - and the first author - the PhD coder) re-coded the sample size for each study in this subset. While coding, all coders were blind to the sample size provided by other coders. To study inter-rater reliability, we subsequently calculated the percentage agreement between each of the coders, and we calculated the intraclass correlation coefficient between coders (model = one-way fixed effects, type = single rater, definition = absolute agreement) using ICC function in the R package *psych* (ICC3 output reported).

Overall, there was a high but imperfect agreement between the three coders (percentage exact agreement = 0.772). The undergraduate double-coder and the PhD coder had a slightly higher agreement rate (percentage exact agreement = 0.884) than either one had with the original undergraduate coders (percentage exact agreement between original coders and undergraduate coder = 0.816, percentage exact agreement between original coders and PhD coder = 0.828). The intraclass correlation coefficient between raters was high, ICC = 0.825, CI95%[0.795, 0.851]. Figure 5.7 displays the variation in sample size between the coders, plotted on log scale.

Coders disagreed in 57 cases. All disagreements between coders were resolved by the PhD coder after inspecting comments by the other coders. In most cases, one coder was clearly correct and the other clearly incorrect. In cases where the correct sample size was genuinely ambiguous (e.g., when study exclusion procedures were not clearly explained) the PhD coder had final say in which sample size would be considered correct. In addition to the cases of disagreements identified in the data used for inter-rater reliability analysis, one additional sample size coding error in the full set of 1358 studies was detected and corrected at a later time during the analyses. Figure 5.8 displays the distribution of sample size in our data after resolving coder disagreements (mode=20, median=24, frequency of *n<=10*=37, *11-20*=479, *21-30*=365, *31-40*=184, *41-50*=97, *51-60*=60, *61-70*=27, *71-80*=25, *81-90*=10, *91-100*=10, *n>100*=64).

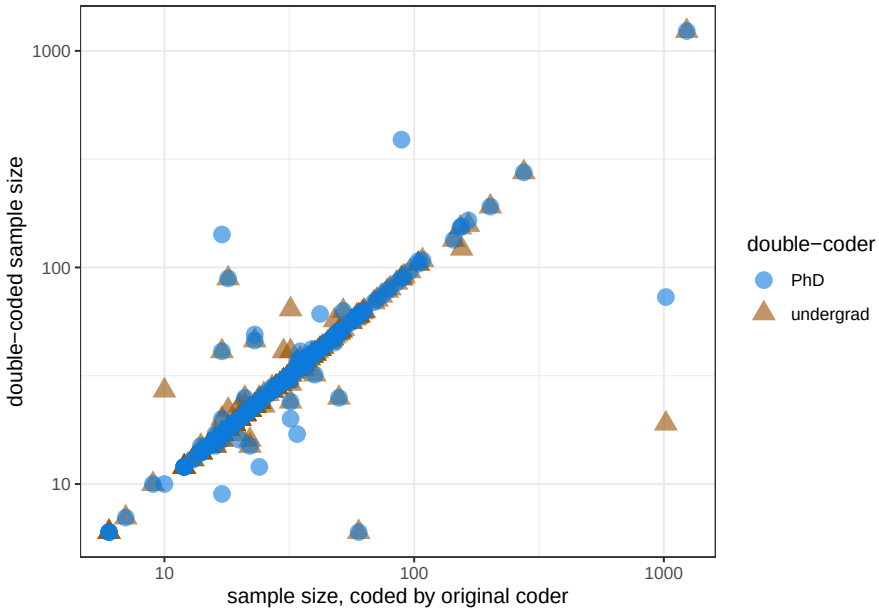

Figure 5.7: Variation in sample size between coders. Sample size is plotted on log scale. The original sample size coded is represented on the x-axis. Double-coded sample size values are represented on the y-axis. Blue circles represent values from the PhD-student coder. Brown triangles represent values from the undergraduate student coder.

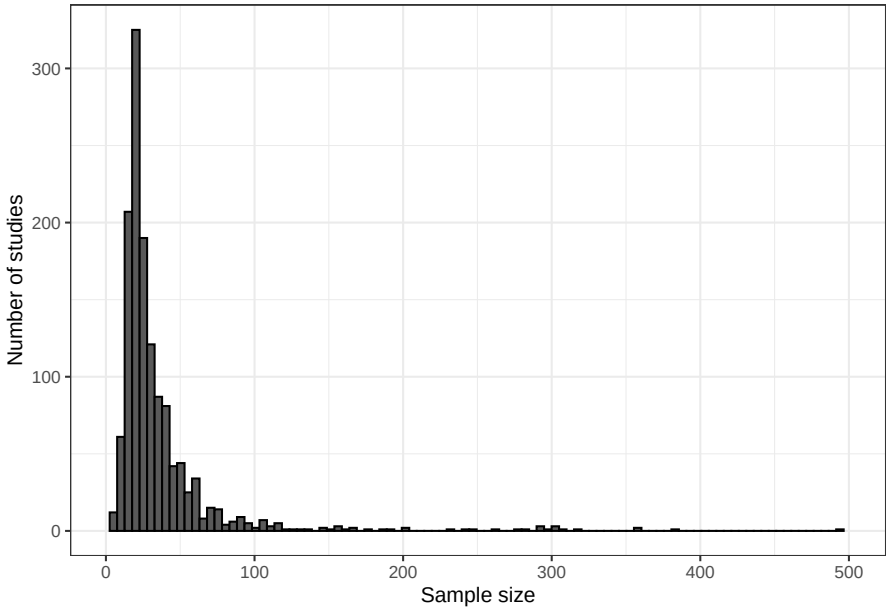

Figure 5.8: Distribution of sample sizes in the dataset. For visualization purposes, the x-axis limit is set to n=500, excluding 10 cases where n>500.

5.3.2.4 Calculating and comparing alternative operationalizations of RV_{Cn}

Having established that accurate citation count and sample size estimates can be reliably collected, we proceeded with the actual calculation of RV_{Cn} Reliability checks revealed that traditional citation counts were not strongly associated with Altmetric attention scores, regardless of citation source. Replicating researchers might justifiably want to use either or both of these metrics to estimate value. Therefore, we decided to compare the results of two alternative operationalizations of replication value; one indicator measured value via the Web of Science citation count of the articles (RV_{WoS}), while the other indicator measured value via Altmetric score of the articles (RV_{Alt}). Both indicators used sample size as a measure of uncertainty. Note that we could also construct and compare additional operationations based on these data, for example based on field-normalized citation impact or a combination of citation impact and Altmetric score, but for illustration purposes we here focus on the two least correlated value estimates.

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

$$RV_{WoS} = \frac{C_{WoS}}{\sqrt{Y+1}} \times \frac{1}{\sqrt{n}} \tag{2}$$

where C_{WoS} denotes the Web of Science citation count of the article a study is reported in, Y denotes the article age in years, and n denotes the sample size of the study after exclusion.

RV_{Alt} was calculated in the following way:

$$RV_{Alt} = C_{Alt} \times \frac{1}{\sqrt{n}} \tag{3}$$

where C_{Alt} denotes the Altmetric attention score of the article, and n denotes the sample size of the study after exclusion. Because exploratory analyses revealed that Altmetric attention scores are not correlated with article age in our data, we did not average C_{Alt} over publication year in this replication value indicator. Because Altmetric attention scores were not available for all reports in our dataset, C_{Alt} could only be calculated for 1156 of 1358 studies.

Importantly, we calculated both RV_{WoS} and RV_{Alt} under the assumption that no study in our candidate set is a replication of another study in the set, implying that no studies should be combined in the estimate of n. This assumption is very likely false for some candidates, in which case it would have been more appropriate to combine the sample size from the original study and it's replications (see appendix A)S. However, detecting such cases is extremely

challenging because original and replicated studies are not linked in bibliometric records. Because lack of replication research in fMRI research (Poldrack et al., 2017) implies that only very few articles in our dataset would be replications of one another, we found it acceptable to proceed with calculation under the assumption that there were no replications in the data.

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

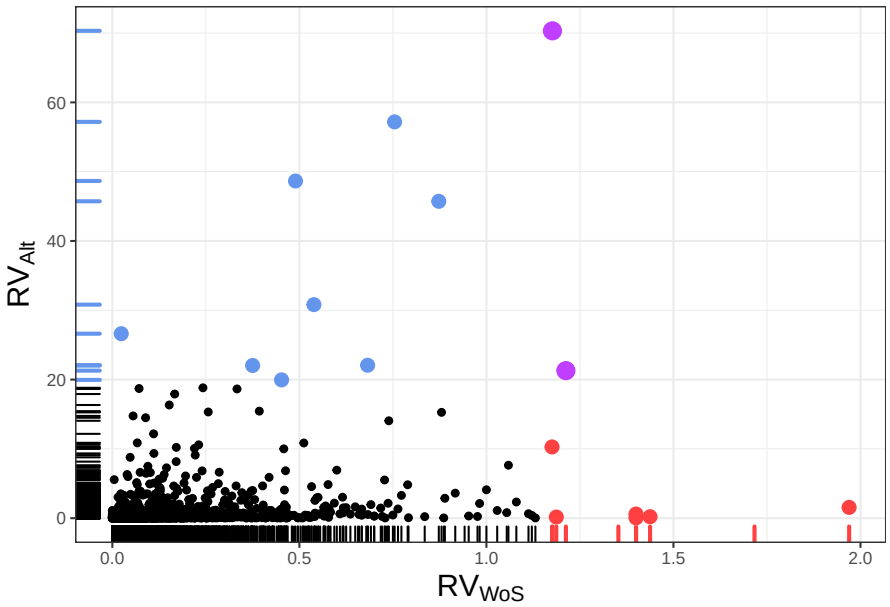

Figure 5.9: Scatter plot visualizing the relationship between RV_{WoS} and RV_{Alt}. Distribution of RV_{WoS} estimates are visualized as bars on the x-axis. Distribution of RV_{Alt} estimates are visualized as bars on the y-axis. Blue bars (and dots) represent the 10 highest RV_{Alt} scores. Red bars (and dots) represent the 10 highest RV_{WoS} scores. Purple dots represent scores that are among the 10 highest scores on both estimators. Two of the ten studies with the highest RV_{WoS} scores are not included in the scatter plot because their RV_{Alt} scores could not be computed due to missing Altmetric attention scores.

Figure 5.9 displays the distribution of RV_{WoS} and RV_{Alt} and their associa-

tion. Overall, both distributions are highly skewed (see bars on axes in figure 5.9), which is expected given that sample size, Web of Science citation count and Altmetric attention scores are all highly skewed as well (see figure 5.4 and figure 5.8). Overall rank-order correlation between estimators was moderate, $\rho = 0.401$, CI95%[0.348, 0.45]. Since our replication selection strategy involves selecting a study from the subset of the highest indicator-ranked studies, we also considered rank-order overlap between the two indicators for the highest-ranked studies from each indicator. Of the ten highest-ranked studies from each indicator, only two studies (Tamir and Mitchell, 2012; Kassam et al., 2013) were ranked among the top ten in both indicator rank-orderings (purple-colored points in figure 5.9). Besides these, these top ten studies based on RV_{WoS} (red bars and points in figure 5.9) had substantially lower scores on RV_{Alt} than the top ten studies based on RV_{Alt} (blue bars and points in 9), and vice versa. This weak association between RV_{WoS} and RV_{Alt} is as expected. Traditional citation impact and altmetric attention scores are generally thought to measure different aspects of impact and are known to be weakly associated.

In other words, quantitative recommendations for which studies to replicate will vary substantially based on whether traditional or altmetric citation impact is used to estimate replication value, because these impact metrics measure different attributes. To the extent that they measure scientific impact, they measure non-overlapping aspects of scientific impact. Different stakeholders may prefer either operationalization, depending on what aspects of impact they find most relevant. If a stakeholder has no preference, they should either combine both sources of impact in their estimate of C, or focus on studies with a high value on both indicators.

5.3.2.5 Overall evaluation of step 2 implementation

In summary we return to the research questions posed in the introduction. First, does it matter from which source we collect citation impact estimates? It depends. For academic citation impact, exact count may vary between sources, but relative citation rank-order appears very reliable across sources. So long as the same source is used consistently for all replication targets evaluated, choice of source has little consequence. Field-normalizing citation counts seems to have a more substantial impact on citation rank-order, though correlation between raw and field-normalized counts was also quite high for our data. Altmetric attention scores are only weakly correlated with traditional citation counts,which has a substantial impact on RV_{Cn} estimates. This is to be expected since Altmetric attention scores and traditional citation count metrics measure different value-related attributes, and it means that, in practice, researchers need to think carefully about which of these metrics more closely represents their personal definition of value. Finally, it seems that the recommendation to average citation impact over article age in the RV_{Cn} esti-

mate (Isager et al., 2021) is an effective but imperfect method of adjustment for article age. However, depending on the citation metric chosen, averaging over age may itself introduce systematic measurement error in the scores. We therefore urge researchers to consider alternative age adjustment methods when these are available.

Second, can we code sample size for all candidates in a feasible and reliable way? Indeed it seems so. However, (1) this process was more challenging than we had first anticipated, (2) inter-rater reliability of manual coders is high, but not perfect, and (3) we were only able to code sample size coarsely as 'number of participants in study', omitting information often used in the calculation of the standard error such as number of trials and stimuli (Westfall et al., 2014). Our main obstacle in the coding process was the non-standardized reporting of participant exclusions, which we needed to derive the final sample size of the study design. Most often, exclusions and final sample size would be reported in the 'participants' subsection of the methods section. However, information about participant exclusions would frequently not appear until the results section. In several cases, exclusions would be reported in several places. For example, participants excluded after psychiatric assessment might be mentioned in the methods section, while participants excluded due to movement artifacts might appear as part of the reported results. Frequent inconsistencies like these meant that double-checking usually had to take place, which slowed down the coding effort and led to several coding errors. In the end, we were not able to code sample size for all studies in the entire set of 2268 articles given our time- and resource constraints, but a substantial subset could be coded by a small roster of undergraduate research assistants in less than 100 hours, all told. We estimate that the entire set could likely be coded in less than 200 hours. In cases where it is not feasible for researchers to carry out step 2 for the entire set of candidates derived in step 1 (see figure 5.1), randomly selecting a feasible subset of all candidates in step 2 seems to us a reasonable way to proceed, since it at least gives every candidate an equal chance of being evaluated further. Of course, alternative methods for narrowing down the candidate set could also be considered. A simple procedure would be to define the field of interest more narrowly than we did here, leaving fewer replication targets in the initial set of candidates to begin with.

5.3.3 Step 3 - In depth review of recommended candidates

To assess the face validity of RV_{WoS} and RV_{Alt} recommendations, we followed up the quantitative analyses with a brief qualitative review. The goal of this review was twofold. First, we wanted to see whether quantitative replication value estimates would conform to our own intuitions about replication value. Would it usually seem obvious to us that studies which received a

high RV_{WoS}/RV_{Alt} estimate would be more worthwhile to replicate than studies that received a low estimate? Second, we wanted to diagnose potential sources of noise and systematic error in the replication value estimation procedure, leading to a high replication value estimate without actually warranting a replication (e.g., an article may be highly cited for reasons other than the empirical studies it reportes, by utilizing a novel method, espousing an ideological claim, etc.). Importantly, the goal of this section is not to identify targets for replication per se, because we are not yet confident that RV_{WoS} and RV_{Alt} are estimating replication value as intended during step 2. Rather, the goal is to identify potential issues with validity, reliability and measurement error that future validation studies of RV_{Cn} may want to follow up on. Therefore, we do not commit to the same depth of qualitative analysis here that we would normally recommend during full-scale execution of step 3.

5.3.3.1 Review procedure

Because RV_{WoS} and RV_{Alt} rank order was only moderately correlated with each other, we included the 10 highest and 10 lowest estimates from both indicators in our review. In addition, we included the 10 lowest non-zero estimates from the RV_{WoS} distribution, because we suspected that RV_{WoS} scores of 0 often simply reflect a paper too young to have picked up citations yet. In total, 44 unique studies were included in our face validity review (6 studies were among the highest or lowest scores for both indicators).

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

5.3.3.2 Central outcomes of the review process

The review process was completed in roughly one week. Due to issues with blinding, which are detailed below, we elected not to increase the sample of reviewed studies beyond the 44 initially reviewed. In future feasibility studies, and in a full-scale implementation of step 3, researchers may want to review a larger subset of candidates, which is of course possible so long as one has the

time and resources available. It should be noted that a full-scale implementation of step 3 will likely involve a more in-depth review of each candidate than was conducted here, which would entail a much longer review process for the same number of studies.

The face validity review led to three primary observations. First, detailed inspection of quantitative replication value estimates turned out to be highly important for quality control. We discovered two cases of erroneously coded sample sizes. In the one case, the correct sample size of 16 was coded as 1, likely due to a typographical error. In the other case, missing data exclusion was not taken into account and the coded sample size of 561 was corrected to 114. In addition to manual errors in sample size encoding, we noticed eight studies that were not obviously connected to social neuroscience, and one study that should clearly have been filtered out by our initial step 1 exclusion criteria because it did not utilize fMRI for imaging. Finally, in one case we had incorrectly labeled a single two-session repeated measures study as two separate studies.

Second, correspondence between RV_{WoS}/RV_{Alt} rank order and our intuitions about the replication value of the claims (based on title and abstract information) was not immediately obvious. In some cases, it seemed natural to us that the studies estimated as having the lowest/highest replication value should belong to that group; in other cases it did not. In hindsight, lack of blinding on part of the reviewers made interpretation difficult because it prevented unbiased interpretation of sample size and citation impact information during review. Normally, this information would have been natural to consider when forming an intuitive judgement about replication worthiness, but in the present analysis we could not consider it without confounding the comparison between formula estimates and intuitive judgements. Therefore, we cannot at present conclude whether lack of correspondence between intuition and formula estimates signals a problem with RV_{WoS}/RV_{Alt} or simply indicates that the assessment procedure is compromised.

Third, we discovered a number of potential sources of noise and systematic error in the RV_{WoS}/RV_{Alt} estimates during review (which may partly be responsible for the low perceived correspondence between indicator estimates and reviewer intuitions). The most obvious problem was the number of studies utilizing within-subject designs. Variation in study design leads to noisy estimation of true replication value, since both RV_{WoS} and RV_{Alt} exclusively use number of participants to estimate uncertainty and thus ignore reduction in uncertainty due to repeated measures of each participant. More problematic still, the use of within-subject designs seemed to be much more common among the highest ranked studies. This makes intuitive sense. A within-subject design is likely to require a lower number of participants on average than a between-subject design, since lack of participants can be compensated for by repeated measures, which means study design and sample size is likely

to be correlated. If so, RV_{WoS} and RV_{Alt} will tend to overestimate the true replication value of within-subject designs compared to between-subject designs. Similarly, there seemed to be consistent differences between studies in the higher and lower groups with respect to study design more generally. Studies with larger sample size often concerned large-scale collaborations studying genetic effects and resting state fMRI, while smaller sample size studies were more often single-lab studies measuring BOLD contrasts during a task manipulation. If certain research areas utilize within-subject research designs more often than others, then research from these areas will tend to get favored by RV_{WoS} and RV_{Alt} simply due to systematic failure to account for study design in the uncertainty estimate. In future applications of RV_{Cn}-based study selection we therefore recommend that the sample size estimate is corrected for the study design used, ideally for all studies during step 2 or, if this is unfeasible, at least as part of evaluation during step 3 (see appendix B for technical details on how such correction can be achieved).

Finally, we discovered a few cases of disconnect between the replication targets reported in an article and the apparent reason why the article received attention. In one case, an article containing both a literature review and an empirical study seemed to be cited primarily due to the literature review (Dimoka et al., 2011). In another case, a study on human navigation appeared to receive a large Altmetric score primarily due to speculative news reports claiming that GPS use can "turn the brain off", which does not clearly follow from the study's conclusions (Javadi et al., 2017). While one could argue that correctness of reporting is unrelated to news media impact per se, it is unclear whether the solution to dubious interpretation of study claims is replication of the original study. Formally, it is unclear how much we would be able to reduce uncertainty by replicating. Researchers should always consider the likely effect of replication on uncertainty during step 3 since replication value indicators, by definition, are not sensitive to this determinant of expected utility gain (Isager et al., 2020).

Our face validity review thus identifies several issues that are important to consider when conducting in-depth review of replication targets, including coding errors in RV_{Cn} estimates, potential systematic mismeasurement of certain study designs and topics, and the appropriateness of replication for reducing uncertainty about highly cited claims. Future research should give us a better understanding of which factors to consider during in-depth review of replication candidates (e.g., Pittelkow et al., 2021).

5.4 General discussion

The overall aim of this exploratory study was to test the feasibility of implementing the four-step replication study selection procedure based on RV_{Cn}

proposed by Isager et al. (2021) in social fMRI research - a field of interest that spans several thousand replication targets. Based on the exploratory research reported in the previous sections, we believe the basic feasibility of the procedure has been established. It was possible to construct an initial set of relevant empirical claims (step 1 in figure 5.1). By excluding articles based on keyword information, non-relevant research topics and study populations could be filtered out of the candidate set. Further, reliable estimates of both sample size and citation impact could feasibly be derived for each study in order to calculate RV_{Cn} (step 2 in figure 5.1). Undergraduate student coders were capable of providing reliable estimates of sample size, though their accuracy was not perfect. Traditional citation count metrics were highly rank-order correlated, meaning there is little difference in which source S is used in the calculation of RV_{Cn}. Altmetric attention score was only weakly correlated with traditional citation impact in this field, and might represent a qualitatively different way of measuring value. Whichever measure is preferred, both Altmetric scores and traditional citation counts could easily be extracted using free and open source applications (e.g., Ram, 2017; Chamberlain et al., 2020). Finally, in-depth review of the highest ranking indicator estimates from step 2 appears to be an important method of quality control before a candidate is selected for replication. Whether these conclusions generalize to application of RV_{Cn} in other disciplines is an open question which will need to be empirically examined.

The validity of RV_{Cn} for estimating expected utility remains an open question. The current exploration suggests several factors that could compromise the validity of RV_{Cn}. To help facilitate future validation of this indicator, we briefly suggest a few different study designs that validating researchers might consider.

5.4.1 Suggestions for future research

Ideally, we would validate any quantification of replication value by benchmarking the quantitative estimates against some gold standard for measuring expected utility gain. However, such benchmark validation will be highly difficult to implement in practice. First, no gold standards exist. Second, even if a standard could be formalized in terms of observable variables, one would then need to actually conduct the replication in each case to determine the post-replication utility that was gained. Only when this is done for a large number of claims could the association between the gold standard (whatever it may be) and RV_{Cn} be studied reliably. Validation through benchmarking is therefore at present unfeasible. However, they might become possible once the number of replications in the literature increases.

In lieu of gold standards, we could instead investigate whether RV_{Cn} is associated with other operational measures that are hypothesized to predict expected

utility gain. That is, we can attempt to provide criterion validation of RV_{Cn}. This works under the assumption that the measures RV_{Cn} is compared with have a causal structure that makes them both valid measures of expected utility gain and correlated with RV_{Cn} (one does not necessarily follow from the other per the definition of validity proposed in Isager, 2020). For example, we would expect RV_{Cn} to predict which studies are chosen for replication in practice under the assumption that both RV_{Cn} and replication authors' selection criteria are caused by the expected utility of the replication effort (Isager et al., 2021).

If replication authors are capable of predicting the expected utility gain of different replication targets, RV_{Cn} could also be validated by asking researchers to intuitively estimate the relative replication value of a set of replication candidates from their discipline and then examine the ability of RV_{Cn} to predict these estimates. A validation effort of this kind is currently being developed by the authors of this article. In a similar vein, one could perform more extensive versions of the survey research reported here (see appendix D) to increase understanding of which factors researchers usually consider when selecting a study for replication. Empirical research on this front is currently underway (Pittelkow et al., 2021). Subsequently, one could examine whether RV_{Cn} are associated with these factors in the published literature. Such data would also be able to provide content validation of RV_{Cn} by informing us of which aspects of value and uncertainty are adequately captured by RV_{Cn}, and which are not.

It might also be worthwhile to test the discriminant validity of RV_{Cn} with respect to certain key variables. In particular, it would be useful to ensure that RV_{Cn} is not associated with replicability. That is, RV_{Cn} estimates should not be predictive of whether a study will replicate successfully. RV_{Cn} estimates should be the highest for studies which we are very uncertain whether or not will replicate successfully. Conversely, studies that are very likely to replicate *and* studies that are very unlikely to replicate should *both* receive low RV_{Cn} estimates, all else being equal. One way to examine this predicted non-relationship is to collect the RV_{Cn} estimates of original studies from existing replication efforts where the replication outcome is fairly certain (e.g., Klein et al., 2018), and then test whether RV_{Cn} is unassociated with replicability (Lakens et al., 2018a). Another potential way to test the hypothesis would be to collect estimates of replicability through prediction markets, which are known to successfully predict replication outcomes (e.g., Forsell et al., 2019) and test whether RV_{Cn} estimates are associated with the strength of market beliefs, but unassociated with the direction of beliefs.

5.4.2 Conclusion

The replication study selection procedure proposed by Isager et al. (2021) seems to be a feasible option for replication study selection in social fMRI

research. An initial candidate set can be curated and filtered based on available bibliometric information, and RV_{Cn} can be calculated in a reliable manner using bibliometric information, and sample size information can be reliably coded by (under)graduate coders. However, whether the procedure assessed here is valid for use in replication study selection is currently unclear. Rigorous empirical validation of RV_{Cn} isrequired before the indicator can confidently be utilized for replication study selection. If RV_{Cn} proves to be a valid and reliable measure of expected utility gain, social fMRI researchers (focusing on healthy human subjects) could utilize the data provided in this article to select studies for replication more efficiently. Additionally, the candidate set of studies we have curated could serve as an initial candidate set for any study selection strategy adopted by researchers in our field, regardless of the strategy utilized for selection. Our study procedure could be extended to test the feasibility of implementing other proposed study selection strategies (e.g., Field et al., 2019; Matiasz et al., 2018; supplementary formula documents in Isager et al., 2020), and could be reiterated to examine the feasibility of applying RV_{Cn} in other research contexts. Finally, by exploring and documenting the wealth of information relevant to replication study selection, we increase the ability of researchers to make well-informed decisions about which original research would be the most important to replicate.

Chapter 6

General Discussion

This thesis investigates the potential for using quantitative indicators of replication value to construct efficient study selection strategies. The thesis began by solidifying theoretical foundations on which indicators could be justified (chapter 2 and 3), which were previously left more or less implicit in discussions about which replication studies were valuable to perform. From these foundations a particular quantitative indicator (RV_{Cn}) was then constructed, and additional measurement assumptions needed to complete the rationale for RV_{Cn} was identified and explicated (chapter 4). The feasibility of using RV_{Cn} for study selection in practice was explored by calculating the indicator for a large set of empirical studies from the field of social neuroscience (chapter 5).

The ultimate goal throughout the thesis has been to construct a formal, well-justified, and feasible strategy for selecting replication studies efficiently. All chapters can be viewed as incremental steps in the process of reaching this ultimate goal. In this final chapter I will revisit the research questions posed in the introduction to this thesis and elaborate on the immediate outstanding challenge of providing empirical validation of RV_{Cn}. I will also revisit certain important problems that arose in discussions of chapter 2-4 based on the preprints that were posted online. Finally, I will briefly consider some of the broader metascientific themes touched on in this thesis, including the need for formalization of meta-scientific theory, and the emergence of the meta-scientific subfield of research efficiency analysis.

6.1 Central research questions revisited

Before moving on to discussion of outstanding challenges and future research, I will briefly summarize what the work presented in the preceding chapters can tell us about the research questions posed at the outset of this thesis.

6.1.1 Chapter 2

Central research questions:

1. What makes something important to replicate?
2. What does *replication value* mean?
3. What are the key factors determining replication value?

Chapter 2 provides a clear answer to the question of what makes something important to replicate (though alternative goals of replication research could be formulated). Replications can be considered important when they have a *high expected utility gain*. That is, a replication is most worthwhile when it substantially reduces our uncertainty about a claim we find valuable. *Replication value* is closely related to the expected utility gain of the replication effort. However, it is not exactly the same because the quality of a replication effort is not known before the study is planned. Instead, replication value refers to the *potential expected utility gain* that a high-quality replication would be able to achieve. This makes replication value possible to estimate without specifying in advance how the replication study will be designed, because the only factors required to determine replication value of a claim is the expected value of the claim (to science and/or society) and our pre-replication uncertainty about the truth status of the claim. Later in this chapter I show how the meaning of expected utility gain and replication value can be made even more explicit by formally relating these terms to concepts from value of information analysis through decision tree modeling.

6.1.2 Chapter 3

Central research questions:

1. What does test validity mean?
2. Can a causal ancestor act as a valid estimator of its causal descendants?

Chapter 3 provides a clear and general definition of test validity. The purpose of a measurement instrument is to gain information about a target (unmeasured) attribute. As such, a test is valid for measuring an attribute if (a) the attribute exists, and (b) variation in the attribute is d-connected to variation in the measurement outcomes. In other words, there are many causal relationships between target and measured attribute that can lead to valid measurement, including scenarios where the measured attribute is the causal ancestor of the target. This definition of validity improves on the existing realist definition offered by Borsboom et al. (2004) by resolving the logical inconsistency of requiring the target to be the cause of the measured attribute.

It is not necessary for a target attribute to cause the measured attribute in order for the measurement instrument to give accurate information about the target attribute. The d-connection definition of validity is used in chapter 4 to explain how sample size can be considered valid for estimating uncertainty when sample size is clearly a causal ancestor of uncertainty. Later in this chapter I revisit the terminology of chapter 3 in light of objections raised by readers and consider some changes to avoid future misunderstanding about the necessary conditions for valid measurement.

6.1.3 Chapter 4

Central research questions:

1. How can replication value be estimated empirically?
2. What auxiliary measurement assumptions must be added to justify the empirical indicator, and how likely are these to hold?
3. What potential estimation issues can already be anticipated?

Chapter 4 provides a simple and straightforward method for estimating replication value empirically through information about article citation impact and total study sample size. The resulting indicator is dubbed RV_{Cn}. To formalize the measurement rationale and make RV_{Cn}, measurement assumptions about the causal relationship between citation impact and value, and between sample size and uncertainty, had to be explicated. Now that these assumptions are made explicit, it is possible to explain how RV_{Cn} is supposed to function, and whether it is likely to function as intended in practice. The operationalized indicator proposed in chapter 4 is obviously only one of many possible ways replication value could be estimated. Future research will undoubtedly be able to improve on what should be considered a rather crude measurement instrument for the target attribute of interest. A concrete step to improving RV_{Cn} is to identify and find potential solutions to sources of error in the measurement of expected utility gain. Chapter 4 anticipates several potential sources of measurement error that can arise when using RV_{Cn} to measure replication value (see figure 4.1, figure 4.3, and figure 4.8 in chapter 4 for a summary) and potential solutions to several of these are discussed (e.g., averaging citations over article age, using field-normalized citation impact, adjusting the sample size estimate for study design). Later in this chapter I will outline plausible sources of measurement error that cut across operationalizations of replication value.

6.1.4 Chapter 5

Central research questions:

1. Is the information required to calculate RV_{Cn} available in practice?
2. Are citation impact estimates reliable across sources?
3. Can study sample size be reliably coded by undergraduate student assistants?

Chapter 5 provides initial confirmation that RV_{Cn} is feasible to calculate in practice, even for large replication candidate sets. With a small team of undergraduate research assistants, sample size could be reliably coded for over one thousand studies in less than 100 hours. Citation impact estimates could be automatically extracted for all studies from bibliometric databases and the relative impact estimates were highly similar across sources (Altmetric citation impact was substantially different from traditional citation impact, however). The feasibility of calculating RV_{Cn} substantially relies on the ease with which sample size can be coded in any given research area. Feasibility studies may have to be repeated in other fields, though we suspect that the results of chapter 5 are likely to generalize across subfields in cognitive neuroscience. By exploring how RV_{Cn} can be calculated in practice, chapter 5 sets the stage for empirical validation to take place - the final outstanding task of this research program.

6.1.5 The outstanding task of validating RV_{Cn}

The challenge of providing a formal and unifying definition of replication value turned out to be formidable. As did the challenge of outlining the assumed measurement properties of RV_{Cn}, allowing us to say something about how and with what accuracy RV_{Cn} can be expected to estimate replication value. Providing a theoretical framework and working out measurement assumptions was, however, necessary in order to begin the work of empirical validation. Without these foundations in place, it would not have been possible to form clear hypotheses about how a quantitative indicator is supposed to behave, making validation impossible. We cannot validate whether a given strategy works as intended until we can clearly explain how the strategy is intended to work

The task of validating RV_{Cn}, or any other operationalization derived from chapter 2, will likely be as challenging as the task of making validation possible. The central challenge is that there does not yet exist any gold standard measurement of replication value that quantitative indicators could be compared against. Contrast this with efforts to quantify predictors of replication success. Prediction market scores, machine learning algorithms, and surveys used to forecast if an original study is *replicable* can all be compared to actual replication outcomes by running replications of forecasted studies (e.g., Yang et al., 2020). Conversely, for the question of whether the original is valuable enough to replicate, it is not clear how *actual* replication value should be measured, except perhaps through a complex historical analysis of the scientific

record, in which the downstream consequences of replication efforts for subsequent decision-making (in science and society) are compared for high- vs low replication value studies (in the spirit of the actuarial approach to cliometric metatheory proposed by Meehl, 1992). Such an analysis could in theory be undertaken, but would require a great deal of time, resources, and access to detailed historical data on part of the investigator. Validation of RV_{Cn} and other indicators will therefore likely depend on alternative methods of corroboration.

One method for validating RV_{Cn}, would be to examine if it correlates with alternative indicators that are predicted to capture replication value in roughly the same way (i.e., criterion validation). An alternative method would be to determine observable events that should follow from a study having a high replication value, and then examine whether indicators of replication value are able to predict the occurrence of such events. An example of this kind of validation is presented in the empirical section of chapter 4. In chapter 4 it is assumed that claims tend to get replicated because they have a high replication value, implying that an indicator of replication value should be able to predict which studies are chosen for replication. From the assumption that claims with high replication value tend to get selected for replication, it necessarily follows that researchers should have some awareness of the relative replication value of claims. In an ongoing validation effort we are utilizing this assumption to test whether RV_{Cn} is able to predict researchers' beliefs about which studies in a set are more worthwhile to replicate (Isager et al., 2019). We will recruit research participants familiar with the study population sampled in chapter 5. Each participant will be presented with pairs of study summaries from the studies coded in chapter 5. For each pair of studies we will ask the participant to indicate which of the studies they believe is most valuable to replicate. Our hypothesis is that RV_{Cn} should predict which study in each pair is preferred for replication (assuming that researchers' preferences are guided by differences in replication value, and assuming that researchers agree on how value and uncertainty should be defined).

Work is also being carried out to better understand the range of factors considered important by researchers when choosing which target to replicate. In an ongoing research project, led by Merle Pittelkow, the Delphi method (McKenna, 1994) is being employed to generate expert consensus on which factors are relevant to consider during replication study selection (Pittelkow et al., 2021). Studies of this kind will likely be important for future validation of quantitative indicators. They inform us about the kinds of information that a quantitative indicator should ideally be sensitive to and, consequently, leave us better able to foresee potential measurement problems with proposed indicators. As an example, RV_{Cn} is not sensitive to the influence of questionable research practices on study results, even though this factor clearly motivates study selection in certain cases (Isager, 2018).

The ongoing research listed above will need to be supplemented by additional

validation studies to ensure rigorous validation of the usefulness of quantitative metrics for replication study selection. Consider for instance that both studies outlined above assume researchers are capable of making well-informed judgements about the replication value of claims. However, there is no guarantee that this is the case. Researchers may not be looking to maximize expected utility gain, or they may simply decide to replicate the first candidate that comes to mind without considering potential alternatives. In this case it is no longer reasonable to assume that RV_{Cn} predicts researcher preferences, or that expert consensus is useful for devising study selection strategies. It is also likely that indicators like RV_{Cn} are valid for measuring replication value in some contexts and invalid in others. For example, citation practices may be predictive of value in one field but not another, and different researchers may legitimately hold different notions of how value is to be defined, which will have downstream consequences for what each researcher considers valid estimation of value. By understanding these nuances of the replication study selection process, we can better understand when specific study selection strategies are more and less efficient, and which strategy might be the most efficient for a particular researcher in a particular context.

6.2 Reflections on earlier chapters

6.2.1 Chapter 2 and 4: On the formal relation between replication value and value of information, and consequences for the interpretation of RV_{Cn}

In discussions following the preprint publication of Isager et al. (2020), several colleagues have raised the question of whether the model proposed in chapter 2 is related to *value of information* (*VoI*) analysis from decision theory. This relationship was specified in chapter 2, but was not formally demonstrated there due to the still imprecise definitions of value and uncertainty. In this section I will demonstrate the direct relationship between *VoI* and replication value. I will show how the concepts of value, uncertainty, expected utility gain and replication value introduced in chapter 2 can be directly translated into terms from *VoI* analysis. The model in chapter 2 is by implication simply a special case of ordinary *VoI* analysis, in which replication functions as the specific method for changing the outcome probabilities involved in the decision model.

The clarification of the relationship between replication value and *VoI* presented below allows us to clarify and correct the concept definitions introduced in chapter 2, which have certain inherent limitations. Specifically, it is not clear from the definition of uncertainty proposed in chapter 2 how we could model *false certainty*, which could easily arise through publication bias and

questionable research practices, and which replication could help to reduce. For example, a replication may cause us to reject the results of an original study and return us from a state of false certainty to a state of uncertainty. Similarly, the definition of value offered in chapter 2 offers no way of differentially modeling the value of being certain that the claim is true vs. being certain that the claim is false. Consider the claim "drug X cures HIV". If the claim is true, we can use the information to save lives that would otherwise be lost. If the claim is false, however, we are left stranded with no way to treat patients. It is not obvious that the value of these two knowledge states is the same. Here, I will show how we can resolve the conceptual issues by redefining value and uncertainty in terms of well-defined concepts from *VoI* analysis.

To demonstrate the equivalence between replication value and *VoI*, I will utilize decision tree modeling, which is a common way to represent problems in decision analysis (e.g., Clemen, 1996; Raiffa et al., 1961). Decision tree modeling requires explicit specification of the potential outcomes involved in the decision scenario. Considering the outcomes associated with making a scientific claim turns out to be the key to unifying the two frameworks and improving the model put forward in chapter 2.

6.2.1.1 Calculating replication value using decision trees – example case

To demonstrate the relationship between replication value and *VoI*, consider the following fictional example. Suppose we are considering whether to implement a novel cancer treatment policy at our local hospital. If the treatment is implemented and is effective, it could potentially save 100 lives per year that would otherwise have succumbed to cancer (given the proposed treatment efficacy and the number of cancer patients admitted to the hospital). However, if the treatment is implemented but is not effective, it would replace old effective treatment options with a new ineffective one, and 300 people would succumb to cancer that would otherwise have survived. If we decide not to implement the treatment, everyone is simply given the old treatment options, and no additional lives are saved or lost. Thus, our claim is "the cancer treatment is effective"[1], which will lead to different patient outcomes depending both on whether it is true or not, and whether we believe it is true or not.

Given the track record of past cancer treatments and the general state of uncertainty about the mechanisms of cancer development, the prior probability that any given cancer treatment will be effective is quite low (30%). To figure out if this particular cancer treatment is effective, we consult the one existing

[1] In reality, we would probably claim that the treatment works and has a certain level of efficacy, thus invoking a claim about effect size, and we would probably place a different value on the treatment depending on treatment efficacy. However, for demonstration purposes I here simplify treatment efficacy to a binary parameter; the treatment either works perfectly or it does not work at all.

study that has investigated it. The study suggests that the treatment is effective. The study was designed in such a way that the type 2 error rate was controlled at 20%, but at the expense of a very high type 1 error rate of 60% due to a combination of several researcher degrees of freedom. What is the value of the claim "the cancer treatment is effective" given current evidence from this study? What is the replication value of this claim? And what is the expected utility gain of a replication that could maintain type 2 error at 20% while reducing type 1 error to 5% through elimination of researcher degrees of freedom?

In order to accurately calculate replication value and expected utility for this claim, we model the decision process as a decision tree. This process begins by identifying the potential outcomes involved in the decision process. In the case of inferring the truth status of a binary claim like "the cancer treatment is effective", there are four potential outcomes to consider. First, the claim may be either true or false; the cancer treatment either works or it does not. When the claim is false, we either correctly conclude that it is false, or we incorrectly conclude that it is true (type 1 error). When the claim is true, we either correctly conclude that it is true, or we incorrectly conclude that it is false (type 2 error). This "true state" times "decision" interaction is what forms the tree structure in our decision tree (see figure 6.1A).

Each outcome in the tree has a certain value attached, which in this example is represented as human lives saved and lost as a consequence of the treatment policies that will be adopted (figure 6.1A, right-hand nodes). Thus, *value* from chapter 2 would be better defined as the discrete distribution of whatever quantity of interest that is gained or lost over the four relevant outcomes in the replication decision scenario. In other words, value is really a set of values distributed over all relevant decision outcomes. Under this conception of value it is possible to model decision scenarios where a type 1 error is more costly than a type 2 error. The example scenario discussed here is such a case.

Each outcome also has a certain probability attached, which is a function of both the prior probability that the claim is true, and the likelihood of making a correct inference about the truth status of the claim given available empirical evidence (figure 6.1A). It is this total distribution of probability over the outcomes that together forms the definition of *uncertainty*, as opposed to the single probability-like parameter defined in chapter 2. In other words, uncertainty is a set of probabilities distributed over all relevant decision outcomes, where the total probability over all outcomes in the set must sum to 1 (some outcome will happen). By defining uncertainty in this way, we can now model cases where we are less than 50% likely to be correct about the truth status of the claim. The example scenario discussed here is such a case.

In our example, we know that current evidence concludes the claim is "true". If our strategy is to enforce policy based on whatever the evidence concludes, we only need to consider the outcome branches associated with a "true" claim

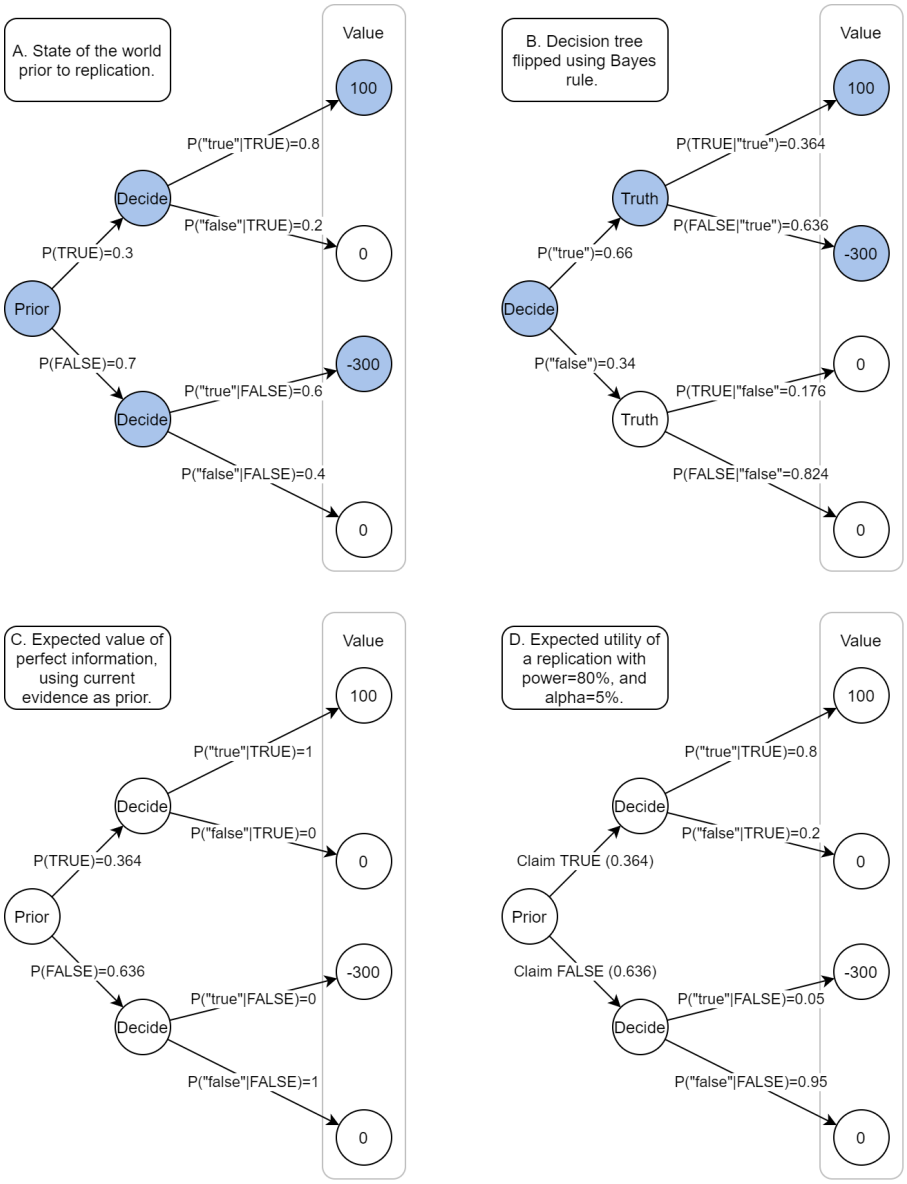

Figure 6.1: The decision tree representing a fictional cancer treatment example. Value is expressed in lives saved (compared with current treatment strategy). Capital TRUE/FALSE represents the actual truth status of the claim. Lower case "true"/"false" represents what truth status is suggested by the empirical evidence.

decision when calculating replication value and expected utility, since the evidence does in fact suggest that the claim is "true"[2]. This is better represented by flipping the decision tree using Bayes rule (figure 6.1B; see Clemen, 1996, chapter 12 for details on how to "flip" decision trees). In this tree, we start with the probability of deciding whether a claim is "true" or "false" on the evidence. Then we consider how likely the effect is to really be true or false for each decision. In our example, we can ignore the 0.66 prior probability that evidence will suggest the claim is "true" since we have looked at the evidence and confirmed that it does suggest the claim is true. In other words, $P("true")$ is 1. Thus, only the probabilities $P(TRUE|"true")$ and $P(FALSE|"true")$ matter for calculating expected utility in this case, since $P("false") = 1 - P("true") = 0$.

The expected utility of the claim given the existing evidence base is then calculated as the sum of the value of each outcome o times the probability of that outcome happening (Briggs, 2019). This means that the equations for EUPre and EUPost defined in chapter 2 should be revised slightly:

$$EU_{pre/post} = \sum_{o \in O} V(o)P(o) \tag{6.1}$$

where $EU_{pre/post}$ stands for the expected utility (before or after replication), O stands for the set of all outcomes o. $V(o)$ stands for value of the outcome, and $P(o)$ stands for the probability of the outcome happening.

In our example, $EU_{pre} = 100 \times 0.364 - 300 \times 0.636 = -154.4$. In other words, in the long run we expect that a decision to treat based on our prior beliefs and existing evidence will cause the avoidable death of about 154 people on average[3].

To calculate the replication value of the claim, we utilize a concept from *VoI* known as the *expected value of perfect information* ($EVPI$). After redefining value, uncertainty, and EU_{pre} as specified above, replication value becomes exactly equivalent to $EVPI$. That is, $EVPI$ is the amount of expected utility we would gain by becoming completely certain about the truth status of the claim (for further elaboration, see e.g., Raiffa et al., 1961; Clemen, 1996; Wilson, 2015; Eckermann et al., 2010), which is identical to the conceptual definition of replication value proposed in chapter 2.

In order to calculate $EVPI$, we first need to set up a new decision tree, using $P(TRUE|"true")$ and $P(FALSE|"true")$ from the decision tree in figure 6.1B

[2]In some scenarios, we might want to consider other branches instead. For example, we could decide not to utilize the new treatment unless we are 90% confident it works, in which case we would only consider the "false" branches unless $P(TRUE|"true") > 0.9$.

[3]This assumes that our strategy is simply to treat if the evidence suggests that the treatment works. We could also calculate the expected utility of the optimal strategy based on available evidence. E.g., if the risks of treating outweigh the benefits, we could choose not to treat even if evidence suggests the treatment is effective. In our example, EU_{pre} would then equal 0, since we could always decide not to treat until EU_{pre} is positive.

as the prior likelihood of whether the claim is true or false. We then imagine that we collect enough evidence that both type 1 and 2 errors become 0 (figure 6.1C), and calculate the expected utility of this imagined "perfect world" tree. The expected utility of our example claim given perfect evidence is 36.4. $EVPI$ is the difference between the expected utility of the claim given perfect information and the expected utility of the claim given current information:

$$RV = EVPI = EU_{perfect} - EU_{pre} = 36.4 - (-154.4) = 190.8 \tag{6.2}$$

Compared to a strategy of following currently available evidence, in the long run we expect to save about 191 additional lives on average by first becoming completely certain about whether the treatment works or not[4]. In other words, the replication value in this case is 191 lives. It is interesting to note here that the use of replication value for replication study selection proposed in chapter 2 is very reminiscent to the suggestion by Clemen (1996) for how to use $EVPI$ in decision making:

> A third step in the structuring of a probabilistic model would be to calculate the $EVPI$ for each uncertain event. This analysis would indicate where the analyst or decision maker should focus subsequent efforts in the decision-modeling process. That is, if $EVPI$ is very low for an event, then there is little sense in spending a lot of effort in reducing the uncertainty by collecting information. But if $EVPI$ for an event is relatively high, it may indeed be worthwhile to put considerable effort into the collection of information that relates to the event. Such information can have a relatively large payoff by reducing uncertainty and improving the decision maker's EMV. In this way, $EVPI$ analysis can provide guidance to the decision analyst as to what issues should be tackled next in the development of a requisite decision model. – Clemen (1996)

Now suppose we would like to calculate the actual expected utility gain of a replication attempt that cannot give us perfect information about treatment efficacy, but that could maintain type 2 error at 20% while reducing type 1 error to 5%. Given the new definitions of value and uncertainty, it can be shown that expected utility gain as defined in chapter 2 is exactly equivalent to the *expected value of sample information* ($EVSI$) from *VoI*. $EVSI$ is the amount of expected utility we would gain by changing the distribution of probabilities

[4]Of course, in the single example we will eventually either lose 300 lives, save 100 lives, or do nothing. The 190.8 lives referred to in the $EVPI$ must be interpreted in relation to an imagined long run of decisions with identical values and probabilities attached. Thus, $EVPI$ = 190.8 lives saved means something like "if we eliminated all uncertainty before making a decision in an infinitely large number of decision scenarios with these exact decision outcomes, values, and probabilities attached, the number of lives on average over all scenarios would approach 190.8".

for various outcomes (uncertainty) in a certain way (for further elaboration, see e.g., Raiffa et al., 1961; Clemen, 1996; Wilson, 2015; Eckermann et al., 2010). Reduction in uncertainty following a replication could be represented as a reduction in type 1 error, type 2 error, or both. Moreover, there is nothing stopping us from modeling failures to reduce uncertainty, or even increases in uncertainty, such as when a fraudulent or flawed replication study shifts our beliefs away from true knowledge about the claim. The current definitions of value and uncertainty thus gives us a much more powerful way to model expected utility gain than the definitions proposed in chapter 2.

In order to calculate $EVSI$, we first set up a new decision tree, using $P(TRUE|"true")$ and $P(FALSE|"true")$ from the decision tree in figure 6.1B as the prior likelihood of whether the claim is true or false. We then plug in the expected type 1 and 2 error rates that will result from the replication in combination with existing evidence (figure 6.1D). $EVSI$ is the difference between the expected utility after replication and the expected utility of the claim given current information:

$$EU_{gain} = EVSI = EU_{post} - EU_{pre} = 19.5 - (-154.4) = 173.9 \tag{6.3}$$

In words, we expect to save about 174 additional lives on average via the information gained by our replication. $EVPI$ creates an upper bound on $EVSI$. In our example, it is not possible to design a replication study for which EU_{gain} is larger than 190.8.

We can also consider the costs involved in running the replication above, assuming costs and value are measured on the same scale, or that some multiattribute strategy for comparing costs and benefits can be constructed. Suppose for example that in order to run the replication study imagined above, we would need to spend money that would otherwise go to treating patients, such that the monetary costs of running the study implies that 10 lives are lost due to lack of resources for treatment. The *net utility gained* is then the utility gained minus the costs of running the study, or $173.9 - 10 = 163.9$.

Figure 6.2 summarizes the definitions proposed above. It is intended as a supplement to figure 2.2 provided in chapter 2 (Isager et al., 2020).

6.2.1.2 Consequences for RV_{Cn} as defined in chapter 4.

Redefining replication value in the manner proposed here resolves the issues caused by the previous definitions. False certainty and differential value of becoming certain that the claim is true vs false can now be modelled explicitly, and replication study selection can be reasoned about using familiar concepts from *VoI* analysis. Having worked out a more satisfying definition of terms, we must now consider what this implies for the interpretation of RV_{Cn}, which was built on the definitions of value and uncertainty from chapter 2.

Parameter	***Definition***	***Equation***

Value (V)	The distribution of an outcome variable of interest over discrete decision outcomes.	V={
		V(TRUE|"true"),
		V(TRUE|"false"),
		V(FALSE|"true"),
		V(FALSE|"false")
		}

Uncertainty (U)	The distribution of probabilities over discrete decision outcomes. P must sum to 1.	U={
		P(TRUE|"true")*P(TRUE),
		P(TRUE|"false")*P(TRUE),
		P(FALSE|"true")*P(FALSE),
		P(FALSE|"false")*P(FALSE)
		}

Costs (C)	The costs of running a replication study which will change the distribution of P in a specified way, specified in the same units as V.	
Expected utility (EU)	Sum over all outcomes *o* in *O* of the value of an outcome *o* times the probability of outcome *o* happening.	V*P
Replication value (RV)	Equal to expected value of perfect information (EVPI) for the decision tree that describes the set of outcomes *O*.	$EU_{perfect}$ - Eu_{pre}
Expected utility gain (EU_{gain})	Equal to expected value of sample information (EVSI) gained by replication for the decision tree that describes the set of outcomes *O*.	EU_{post} - Eu_{pre}
Net expected utility gain	Expected utility gained after subtracting the costs of running the replication.	EU_{gain} - C

Figure 6.2: Model parameters from figure 2.2 in chapter 2, redefined so as to be consistent with *VoI* concepts.

The definition of RV_{Cn} in chapter 4 assumes that the distribution of value and uncertainty over decision outcomes can each be described using a single parameter(or, in the case of value, as a single set of attribute parameters (lives saved, money earned, CO_2-emissions reduced, etc.). On this assumption, it makes sense to use a single variable like citation impact to measure value, and sample size to measure uncertainty. Since the defining value and uncertainty as distributions over all possible outcomes violates this assumption, we must consider what this implies for the interpretation of RV_{Cn}. It is possible to collapse the distribution into a single parameter and still meaningfully describe the decision problem at hand, but only under certain reductionistic assumptions.

If *value* is collapsed into a single parameter, it could be defined simply as 'the difference in value between being right and being wrong about the truth status of the claim'. This definition preserves the idea that value is related to the outcome of deciding whether to believe in a claim - the stakes involved in being right or wrong. However, note that we are now forced to assume that the stakes are the same regardless of whether the claim is true or false; it does not matter whether the treatment is efficient or not – the value gained by becoming certain that it works is the same as the value of gain by becoming certain that it does not work. This assumption may not always be reasonable.

It is not entirely obvious that a metric like citation impact is well-suited to measure this collapsed definition of value. If the drug works, the study corroborating this fact will likely receive a high citation count. However, if the drug does not work, it is not at all obvious that the study corroborating this fact will be cited in the same way. In other words, citation counts may not actually respect the assumption that certainty is equally valuable regardless of whether the claim is true or false.

To collapse *uncertainty* into a single value, it could be defined simply as 'the probability of being correct about the truth status of the claim'. This definition preserves the idea that uncertainty has to do with type 1 and 2 error probabilities and the prior probability that the claim is true. In fact, to calculate the probability of being correct about the truth status of the claim, we utilize all this information since:

$$P(correct) = P("true"|TRUE) \times P(TRUE) + P("false"|FALSE) \times P(FALSE) \tag{6.4}$$

However, in collapsing the definition of uncertainty in this way we are again forced to assume that the utility of the claim being true or false is the same, since there is no longer a way to separately calculate the utility of being correct about a true claim and being correct about a false claim. Thus, we also can no longer compare cases where type 2 error is high to cases where type 1 error is high and reason about which case might be more worth replicating. Collapsing

uncertainty into a single value forces us to treat error reduction of any kind as equally valuable, and leads to incorrect reasoning whenever this assumption does not hold. As an example, suppose we would compare the cancer treatment scenario above (treatment A) with an almost identical treatment (treatment B). The only difference between treatment A and B that past research on treatment B has lower power (0.2) and a lower type 1 error rate (0.343) than treatment A. Substituting the type 1 and 2 error rates for treatment B in figure 6.1A decision tree yields a lower expected utility and replication value than for treatment A ($EU_{pre} = -128.6$, $RV = 171.4$). However, because $P(correct) = 0.52$ for both treatment A and B, and their value distributions are identical, we would be forced to assume that the expected utility and replication value is the same in both cases if we used only $P(correct)$ to model uncertainty.

It is interesting to consider how a metric like sample size is related to $P(correct)$. On the one hand, it does not take into account any prior information about $P(TRUE)$ and $P(FALSE)$. On the other hand sample size will be related to either $P("true"|TRUE)$ or $P("false"|FALSE)$ or both (depending on the method used to set the type 1 and type 2 error rates. Maier and Lakens, 2021), since collecting a larger sample size allows one to design a more informative study with lower error rates. Future studies could investigate the measurement properties of this new target attribute that sample size is proposed as a measure of. However, it is also worthwhile to consider whether Bayesian decision analysis (e.g., as proposed by Hardwicke et al., 2018) might be better suited to quantify uncertainty as defined in this chapter.

6.2.2 Chapter 2: On potential violation of assumptions in the structural causal model

The assumptions supporting the structural causal model in chapter 2 may not necessarily hold, and are subject to scrutiny. When these assumptions do not hold, it is important to understand the consequences of their violation for our overall ability to maximize expected utility gain through study selection. One benefit of representing replication study selection as a causal graph model is that we can introduce additional causal relations into the model and gain a clear visual understanding of their consequences for model assumptions. This is useful for modeling potential measurement problems that can arise when replication study selection is carried out in a practical setting and additional factors besides those mentioned in chapter 2 come into play. One example of such a measurement problem is the problem discussed in chapter 4 (see chapter 4, figure 4.8); once value is operationalized as article citation count, it is possible that citation count is also influenced by the citing authors' uncertainty about the original research, which would lead replication value to underesti-

mate the true expected utility gain. Similar problems may arise through other pathways as well. Such problems ought to be explored by future critics of the model proposed in chapter 2. In this section I will consider a few concrete measurement problems related to the quality of the replication research design. My goal is to show that the assumptions in chapter 2 can be questioned, and to demonstrate how the causal graph model provides a useful tool for analyzing the downstream consequences of particular assumption violations.

Replication design quality is an umbrella term meant to capture all the elements of a replication study design that together determine how much uncertainty is reduced following a conducted replication attempt. This could include the sample size of the replication, the validity of the experimental manipulation, the validity of the measurement instruments, the validity of control conditions, the appropriateness of the statistical analysis, the level of experimental control, potential for selection bias, blinding procedures, manipulation checks, etc. Through some complicated and unknown function, these factors are combined to determine our uncertainty about the tested claim after replication (see figure 6.3A). In other words, in order to reduce uncertainty after replication we must both decide to do a replication *and* run a replication study design that is of high enough quality to reduce uncertainty. The latter is not a trivial matter, and the recent history of replication in psychology includes several debates over whether beliefs about some substantive claim (facial feedback, ego depletion, etc.) ought to be updated based on replication evidence given certain details about the replication study design (e.g., Noah et al., 2018; Drummond and Philipp, 2017).

On its own, replication design quality is a noise factor. Replication design quality has a net positive effect on expected utility gain but has no effect on replication value (figure 6.3A) which implies that the replication value construct is inherently insensitive to variation in expected utility that is caused by variation in replication design quality. However, introducing design quality into our model also opens the door to measurement bias once we acknowledge that the quality of the replication design in large part depends on the quality of the original study design.

Since a replication (at least a close replication) essentially involves duplicating most aspects of the original research design, the strengths and flaws of the original design will tend to determine the quality of the replication study design (except those explicitly altered in the replication, such as the sample size). Simultaneously, the original study design quality greatly influences uncertainty prior to replication, by the exact same logic that the replication study design influences uncertainty after replication. This opens a biasing path between replication value and expected utility gain (figure 6.3B). On the one hand we would like to replicate claims with high prior uncertainty, because these are claims where we could potentially reduce uncertainty a lot (and gain a lot of utility). However, high uncertainty claims may be uncertain because the study

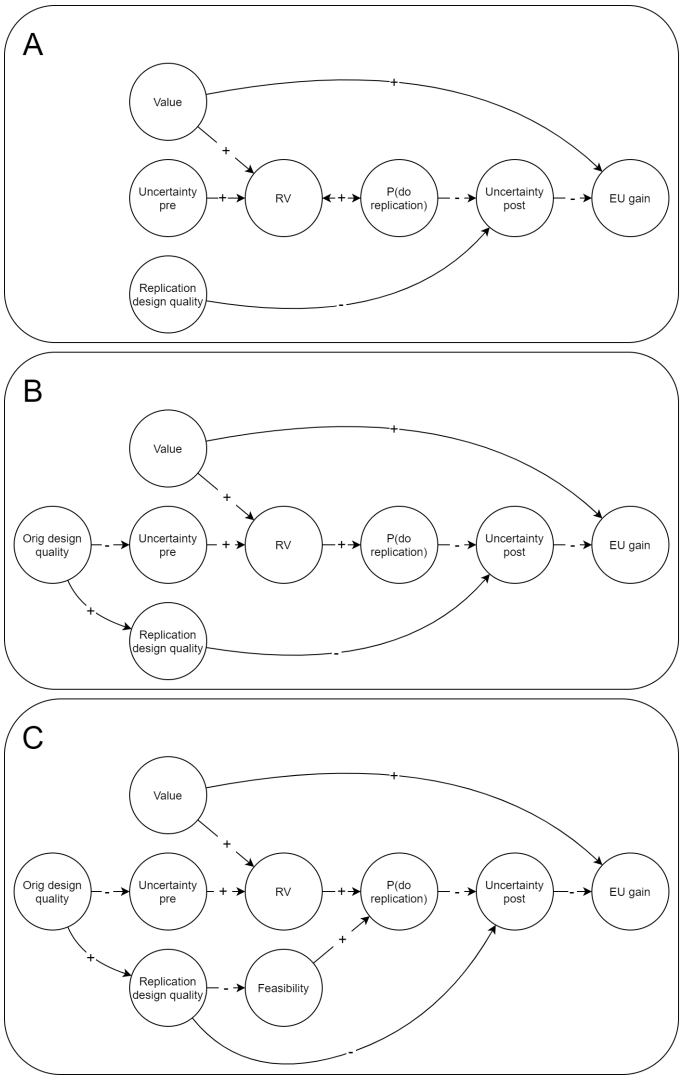

Figure 6.3: Structural causal models visualizing the potential relationship between *replication design quality* and the decision model proposed in chapter 2. **A)** *Replication design quality* represents a noise factor in the estimation of *EU gain* through its influence on *Uncertainty post*, and is otherwise independent of all variables in the model. **B)** *Replication design quality* mediates measurement bias caused by *Orig design quality*. Low *Orig design quality* increases *RV* but decreases *EU gain*. **C)** *Replication design quality* causes measurement bias. Low *Replication design quality* increases *RV* but decreases *EU gain*.

designs used to test them are of low quality. This is a problem because the amount of utility we are expected to gain depends on the replication study design, which is going to inherit the low quality of the original design. This puts us in a bind. Prioritizing highly uncertain claims for replication commits us to replicating low-quality study designs. This is obviously not desirable. This kind of bias is an example of a violation of the assumption stated in chapter 2 (Isager et al., 2020) that the effect of replication on expected utility is independent of the decision to replicate.

The same kind of bias could also originate from replication design quality itself. Consider the possibility that as the quality of the replication study design increases, the study becomes progressively less feasible to conduct (figure 6.3C). Increasing the number of participants, sampling diversely from the population, running more ecologically valid experimental manipulations, and other efforts to increase study design quality will likely increase the time, resources, and expertise required to run the study. Decreased feasibility, in turn, will likely decrease the probability that the replication study is run. This puts us in the same bind as before. If higher feasibility increases the probability of selecting a study for replication, and if more feasible studies tend to imply lower-quality study designs (as compared to less feasible studies), the same studies selected for replication will tend to be those that reduce our uncertainty about claims the least.

The obvious solution to the problems outlined above is to find a method for estimating the quality of the replication study design. In practice this will likely be difficult, which is why chapter 2 attempted to side-step the issue via assumptions to begin with. However, the examples above show that the assumption of independence between effect of replication and probability of doing replication may be violated in practice. In other words, side-stepping the estimation of a replication's effect on uncertainty may not always be an acceptable solution. The magnitude of the problem will of course depend on the magnitude of the causal effects that make up the bias. If biasing effects are small in practice, the assumptions put forward in chapter 2 might still, for all intents and purposes, hold. However, if we suspect substantial bias, we should carefully consider whether replication value is a useful tool for study selection, or if additional information is necessary to devise a selection strategy that maximizes expected utility gain. Future theoretical work should examine other scenarios in which the assumptions of chapter 2 could be violated.

6.2.3 Chapter 3: On the conflict between the formal and vernacular definitions of test validity.

Following the preprint publication of chapter 3, readers have raised various objections to the d-connection definition of test validity. Some of these were included in the subsequent revision of chapter 3, such as the detachment of the

d-connection definition from the classical definition of measurement utilized in the physical sciences (Michell, 2003). However, the most commonly raised objection, which has not yet been adequately considered, has to do with the separation of validity from the other important measurement concepts; reliability and bias. Following the rationale offered by Borsboom et al. (2004), it is perfectly reasonable to consider validity as separate from reliability and bias. However, this narrow conceptualization of validity is budding hard against the conventional, vernacular meaning of validity, which seems to cause a great deal of confusion among critics. It simply does not "sit right" with many that a biased and unreliable measurement instrument can be considered valid for measuring a target attribute, even though this is a perfectly reasonable and logical thing to do *given that we assume the d-connection definition of validity.*

As an example, consider the use of a weight scale to measure a person's height. On the d-connection definition of validity, this measurement instrument is perfectly valid so long as we can assume that variation in height tends to cause variation in weight (if a person grows taller, they don't simply stretch out their current body mass, but adds additional mass which increases their body weight). However, it does not *feel* right to call weight a valid measure of height. Why not? Perhaps it stems from the fact that we could easily imagine how this measure of height will be unreliable, since a person could gain or lose substantial weight without changing their height at all. Perhaps we feel dissatisfied because we could easily imagine much better measurement instruments for the same target attribute. Perhaps there is something about the causal distance between weight and height that troubles us; we may feel that a measure of height should work more like a meter stick, tapping the target attribute "directly" instead of roughly estimating it through a proxy variable.

Whatever the case may be, it is evident that the conventional interpretation of validity is more similar to what Borsboom et al. (2004) refers to as "quality". That is, in common research vernacular, validity means something like the overall "goodness" or "appropriateness" of the test, which is a function of the causal relationship between measurement and target, but also clearly depends on the test being reliable and unbiased. Divorcing the validity term from this vernacular use does not come naturally to many, which leads to confusion. Specifically, there is a tendency to voice disagreement with chapter 3 on the grounds that d-connection does not by itself ensure that a test is "good". The disagreement is an illusion, however, because chapter 3 never argues that d-connection can ensure overall test "goodness". Thus, in chapter 3, a "valid" test does not automatically mean a "good" test. It simply means that two necessary conditions for a test to be "good" are satisfied - the target attribute exists and is d-connected with the measured attribute. The fulfillment of these conditions is all that is meant by "validity" in chapter 3.

Conceptual clarity is perhaps the most important goal in concept formation.

However, it is not the only goal. In pursuit of clarity, the d-connection definition of validity has been overly neglectful of validity´s conventional meaning, and has perhaps strayed too far from Gerring's criterion of familiarity:

> The degree to which a new definition "makes sense," or is intuitively "clear," depends critically upon the degree to which it conforms, or clashes, with established usage-within everyday language and within a specialized language community. If a term is defined in a highly idiosyncratic way it is unlikely to be understood, or retained. - Gerring (1999)

Fortunately, the conceptual confusion can be remedied easily. The vernacular meaning of test validity very closely resembles what Borsboom et al. (2004) refer to as test *quality*. To avoid confusion, we could simply substitute *validity* for *quality* to let the former concept retain much of its conventional meaning. We will then need a new name for the concept referred to as *validity* in chapter 3. A simple solution would be to split this concept into its constituent conditions of existence of the target attribute, and d-connection between target and measured attribute. These conditions, along with reliability and bias, together form the necessary and sufficient conditions for a test to be valid. In other words, a test is valid if (1) the target attribute exists, (2) variation in the target attribute is d-connected to variation in the measured attribute, (3) estimates of the target attribute are (sufficiently) reliable, and (4) estimate of the target attribute are sufficiently unbiased. Chapter 3 is then simply claiming that d-connection is a necessary (but not sufficient) condition for valid measurement to take place. Reorganizing the measurement framework of Borsboom et al. (2004) in this way should help increase conceptual familiarity and reduce confusion about what *valid measurement* entails.

6.2.4 Chapter 4: On Goodhart's Law in relation to RV_{Cn}

> Any observed statistical regularity will tend to collapse once pressure is placed upon it for control purposes. - Goodhart (1984)

> When a measure becomes a target, it ceases to be a good measure. - Strathern (1997)

On the assumption that RV_{Cn} is a valid indicator of expected utility gain in principle, it is worthwhile to consider what might happen as the indicator becomes adopted by the research community. Throughout this thesis, RV_{Cn} has been treated as a passive measurement procedure, simply designed to access information about a particular feature of the research process (expected utility gain). However, the research process is a complex system and, like any other

complex system, has a tendency to react and change when subjected to measurement (Meadows, 2008, chapter 5). A pernicious such reaction often occurs when indicators are used for performance measurement and control. In such cases, the system becomes incentivised to maximize measured performance at whatever cost, and cheating or "gaming" the performance metric is unfortunately often an effective strategy. This tendency is known as Goodheart's law. A well-known example in research practice is the degradation of citation metrics through citation rings and other forms of post-production misconduct, which is motivated by the use of citation metrics to hire, promote, and control academic survival (e.g., Biagioli, 2016). Since RV_{Cn} is also proposed as a form of control mechanism - it essentially helps to control what will and will not be replicated - it is interesting to consider whether there is a danger that it too may be subject to Goodheart's law.

The most immediate way in which RV_{Cn} might suffer from degradation due to indicator gaming is through gaming of the indicators that are used as input to the indicator. When citation count is widely gamed by researchers for selfish purposes it becomes a less useful measure of scientific impact, which will at best hurt the reliability of RV_{Cn} and at worst will invalidate the indicator altogether. Similar problems may also apply to sample size as a measure of uncertainty if sample size becomes valued for its own sake (e.g., as a measure of journal quality. Fraley and Vazire, 2014) at the expense of the overall quality of the research design.

Whether RV_{Cn} itself will become subject to gaming is less obvious, because it is not obvious that a high or low RV_{Cn} is unequivocally good or bad. A high RV_{Cn} may indicate that a study has had substantial scientific impact, but it also implies that the study results are not very conclusive. A low RV_{Cn} may indicate a high-quality research design, but may also indicate that a study has not been very impactful. Thus, in principle RV_{Cn} should be more resistant to gaming than more straight-forward performance metrics. However, RV_{Cn} may in practice be used in ways that leaves it vulnerable to Goodhart's law. Suppose a funder makes a rule that only proposed studies with an RV_{Cn} over a certain amount will receive funding. The goal of the funder is to maximize the expected utility gain of the money they invest in research. However, researchers applying for funding may become concerned with pursuing RV_{Cn} for its own sake, or for the sake of acquiring funding, which leads to promotion, which leads to academic survival. RV_{Cn} is not a perfect measure of expected utility gain, and it is perfectly possible to artificially increase the RV_{Cn} of a replication effort. One method would be to (1) publish an original study with as low sample size as possible, (2) inflate the citation impact of this study through self-citation etc., and (3) apply for funding once RV_{Cn} reaches the desired level. Alternatively, RV_{Cn}-based thresholds for funding and replication may lead researchers to exploit measurement error in an effort to maximize RV_{Cn}. For example, within-subject designs may be preferred for replication over between-subject designs simply because RV_{Cn} is overestimated

for within-subject designs. One could adjust the RV_{Cn} estimate for study design of course, but researchers may prefer not to do so because measurement error is actually helpful for the goal of acquiring funding. Other examples of incentivized gaming could easily be construed. Because RV_{Cn} may be subject to gaming in practice, monitoring the use and misuse of RV_{Cn} will be an important aspect of validating the metric in the future.

6.3 The value of formalizing concepts in metascience

Taking a step back from the specific problem of selecting a study for replication, a more general aim of this thesis has been to explore how meta-scientific concepts and theory can be formalized. I began my PhD studies in metascience in 2017, about the same time as the problem of weak and verbal theory in the "soft" sciences was beginning to attract serious attention by metascientists. In the years preceding, the field of metascience was chiefly concerned with problems in methodology and research practice related to the replication crisis, such as analytic flexibility, publication bias, lack of open data and materials, and lack of focus on direct replication. However, in recent years there has been an increased focus on weak theory as a potential driver of low replicability (Muthukrishna and Henrich, 2019). A number of calls to formalize theory in areas of social science have been published (e.g., Smaldino, 2016; Muthukrishna and Henrich, 2019; Smaldino, 2019; Oberauer and Lewandowsky, 2019; Devezer et al., 2021; Fried, 2020), as have suggestions for how formalization could systematically be carried out (e.g., Borsboom et al., 2021; Navarro, 2021; Robinaugh et al., 2021).

In my own lab, we became intimately familiar with the downstream consequences of weak theory through our work on introducing equivalence tests in psychological research practice (Lakens et al., 2018b). The frequentist version of an equivalence test is a variation on traditional null-hypothesis testing which offers researchers the possibility to test whether an effect can be considered trivially small for practical purposes. The procedure is designed to combat the common problem in psychology that absence of a significant null-hypothesis test result is interpreted as evidence for the absence of any effect; absence of evidence is interpreted as evidence of absence. Informative application of equivalence tests depends on the researchers' ability to specify an informative smallest effect size of interest. This is typically a scenario in which theory should aid the researcher by placing constraints on what effect sizes are expected. However, most psychological theories place no boundaries on what effect sizes are expected. Consequently, specification of the smallest effect size of interest is currently the biggest roadblock to application of equivalence tests in psychology.

Lack of formal theory was also an important early obstacle in developing formal strategies for replication study selection. When my PhD project began, several quantitative indicators of replication value had already been proposed, and a manuscript outlining the underlying rationale behind various indicators had been drafted (see supplementary documents in Isager et al., 2020). However, initial work on validating the various indicators quickly revealed deep conceptual questions about replication value that had not yet been answered, or even addressed. The intuition driving indicator development at the time was the idea that replication value should be some function of the impact of the empirical findings (usually quantified in terms of number of citations), and the quantity and quality of empirical evidence used to test the claim (quantified in various ways). Why should impact and empirical evidence be the only factors involved in estimating replication value? How should these factors be combined into an estimate of replication value? Moreover, what was the measurement rationale justifying the use of various observed variables (citation counts, p-values, observed power, sample size, etc.) as indicators of impact and quantity of evidence? What was the similarity between the different quantitative indicators proposed that could justify assuming that they were tapping the same latent construct? At the time, no theory of replication study selection existed that would allow us to answer these questions. Formalizing such a theory quickly became one of the main goals of this thesis, resulting in the model presented in chapter 2. Once such a model was in place and sufficiently formalized, it allowed for a much clearer analysis of how various selection strategies ought to function and - given their intended function - how they should be designed.

As an example, consider the proposal by Field et al. (2019) and Pittelkow et al. (2020) to use the Jeffreys-Zellner-Siow (JZS) Bayes factor to select t-test results in need of replication. Stated very briefly, the strategy is to assign a higher replication priority to statistical results based on how close to 1 the Bayes factor of the test statistic is. Based on the theoretical model presented in chapter 2, we can reason clearly about how this selection strategy is intended to help maximize expected utility gain. Selecting studies for replication based on the JZS Bayes factor makes sense given that this metric is a reasonable quantitative measure of our overall uncertainty about a claim before replication. The crucial assumption is that the claims with the JZS Bayes factor closest to 1 represent those studies for which we can reduce uncertainty the most through replication. Further, we also need to assume that the result of a single t-test is a reasonable unit of selection, given that the true replication target we are after is the overall claim about which we are uncertain. For studies in which the overall claim is related to a single test result this might be reasonable, but less so in cases where a single overall claim is supported by a combination of multiple test results. We can also immediately identify limitations in selecting studies purely through Bayesian evidence. The strategy does not take the value of the potential replication targets into account,

and does not consider the ability of a replication to reduce uncertainty in each case, even though we know on the theory that these factors are important for determining expected utility gain. Hence, we can also explain why assessing these factors qualitatively as a separate step in the strategy makes sense, as Field et al. (2019) propose.

Replication study selection is not the only subfield of metascience who could benefit from further theory- and concept formalization. Metascientific research often aims to (1) identify interventions (replication, preregistration and registered report publication, preprinting, open data sharing, abandoning statistical significance, etc.) that either promote desirable research practices or prevent undesirable research practices, and (2) anticipate the effects of such interventions (reduction in false positive findings in published literature, increased diversity in hiring and promotion, increased transparency, reduced prevalence of fraud etc.). In other words, metascientific research often assumes specific causal relationships between researcher incentives, research practices engaged in, the effect of such practices on the scientific literature, and the effects of certain interventions on research practice. In addition, metascientific research very often relies on assumptions about the purpose of scientific research. Essentially, any argument that a certain research practice is questionable, neglected, or flawed must rest on some assumption of what research practice should *ideally* be like. In many cases, however, assumptions about mechanisms and ideal practice are left partially or completely implicit, which can impede constructive discussion and progress on metascientific topics (Devezer et al., 2021).

Consider as an example the registered report format (Chambers and Tzavella, 2020), in which submitted empirical articles are accepted in principle, based on a preregistered analysis plan, before data is collected. Is the purpose of this publication process (1) to prevent editors and reviewers from engaging in results-based publication bias, (2) to reduce the researcher degrees of freedom of the submitting authors, (3) to reassure readers that preregistration was carried out faithfully by having the editor oversee the process as a neutral third party, (4) to make sure that reviewer comments can benefit the research design before it is executed, (5) to nudge submitting authors to formalize their substantive hypotheses and more clearly link them to data, all of the above, or something else? Further, why are any of the outcomes above, if achieved, desirable? What scientific goals are they supposed to help us achieve, and how - in formal terms - is this achieved by the preregistration process? Are there any conditions under which the registered report format can *hinder* scientific goals from being achieved (Chambers, 2019)? Lack of formal treatment of these questions makes life difficult for both advocates and critics of the registered report format. For the critic it is difficult to understand what exactly is being advocated for, and for the advocate it is difficult to understand exactly what aspects of the format are being criticised, and whether those aspects were assumed by anyone to begin with (Scott, 2013; Chambers, 2019). By working

out the entire range of costs and benefits associated with the registered reports format, and by considering how costs and benefits interact in moving towards or away from some assumed research goal, we are in a better position to consider the overall utility of the format (Soderberg et al., 2020).

Fortunately, there are clear signs that metascience is beginning to heed its own call for increased focus on formal theory development. Examples include recent efforts to formalize the function of research practices like replication (Tunç and Tunç, 2020; Earp and Trafimow, 2015) and preregistration (Lakens, 2019), as well as efforts to formally define the functional relationships between various metascientific concepts and phenomena (McElreath and Smaldino, 2015). I expect that focus on formally defining the central concepts and relations in metascience research will only intensify in the years to come. In doing so we will greatly improve our ability to talk and reason about all topics with which metascience is concerned.

6.4 Research efficiency: An emerging metascientific area of research.

The question of which targets to prioritize for replication is one example of an emerging class of metascientific problems that concerns the *efficiency* of research practice. These problems start off from an assumption about a desired outcome (or set of outcomes) of research we wish to maximize and some assumptions about resource limitations researchers operate under, and then ask how the researcher can maximize some desired outcome given their resource constraints. It is important to recognize that efficiency only has meaning in relation to some desirable outcome. What we consider to be the desired outcome defines what we mean by efficiency, and efficiency can thus mean different things to different stakeholders. It is also important to recognize that in practice we will usually have many outcomes we wish to achieve at once, and that efforts to maximize one desired outcome can come at the expense of other desirable outcomes (Peterson and Panofsky, 2021). For example, replication increases the reproducibility of research but will demand resources that could be spent on exploring novel research topics and synthesizing the existing literature. Thus, efficiency-related problems include not only how to maximize one specific outcome given available resources, but also how to appropriately balance resources over a number of different activities that can have positive *and* negative influences on a number of desired outcomes.

Problems that deal with making research efficient differ from the class of problems that deal squarely with research *quality.* Quality-related problems tend to start from an assumption about *ideal* research practice, followed by a comparison of the ideal with current practice, and ask how real and imagined research practices could move actual research practice closer to or further away from

the ideal. In this context, research practices are usually considered in isolation as categorically good or bad, where *good* means closer to ideal practice. Consider for instance the longstanding literature on the misinterpretation of the p-value in empirical research (Bakan, 1966; Goodman, 2008). In an ideal world, researchers would simply interpret p-values correctly, and in line with their statistical philosophy. In practice, researchers often misinterpret the meaning of the p-value. The question is then how these misinterpretations lead us away from ideal research practice (e.g., by increasing type 1 error rates) and what can be done to remedy the situation (e.g., increased statistical education). The same problem of discrepancy between real and ideal research practice forms the basis for a wide range of metascientific topics, such as replication success rates, publication bias and the file-drawer effect, failures to align policy with research evidence, the quality of study designs, the quality of statistical analysis, open science initiatives, and questionable research practices. In each case we are in some way considering the alignment between current research practice and some ideal state of research practice.

Research on quality-related problems has been very successful in the sense that we have gained a good understanding of how the quality of research could *in principle* be improved, and the metascience community has been quite successful in pushing for the adoption of certain practices intended to increase the quality of research, such as preregistration, open science practices, and direct replication. Perhaps due to this success, metascience is increasingly facing the problem that solutions to quality-related problems often have practical costs. This tends to transform problems about how to increase quality in principle to problems about how to increase quality in practice, given limited resources and trade-offs between multiple quality-increasing efforts we could engage in. For example, lowering the type-1 and type-2 error error rates as much as possible would be ideal in principle. In practice, lowering error rates either requires spending resources which are limited (participants, time, equipment, etc.), or one type of error could be reduced by increasing the other. For example, we may set a lower alpha level to reduce the type 1 error of a test, but the likelihood that a true effect fails to reach significance increases as a consequence. These practical constraints have motivated research into optimal thresholding and balancing of error rates (Maier and Lakens, 2021).

Replication study selection is a good example of an efficiency-related problem that has emerged from solutions to a quality-related problem. In the past decade, metascientific research on replication was by and large focused on how replications ought to be designed and incorporated into research practice in order to increase the reliability of published research (Brandt et al., 2014; Open Science Collaboration, 2015). Today, we know much about how replication studies ought to be designed and interpreted (e.g., Baribault et al., 2018; Tunç and Tunç, 2020; Muradchanian et al., 2021; LeBel et al., 2018) and we are slowly seeing them become a more regular part of mainstream research practice. As a consequence, practical problems emerge such as deciding when to

spend time and resources replicating vs. conducting a novel study, whether to replicate a single resource-intensive study or several less resource-demanding ones, and which out of several replication targets to prioritize.

Issues related to the efficiency of research are not limited to replication research however. In fact, the question of how to maximize research quality given limited resources is central to many problems throughout metascience and applied statistics, including (but not limited to) the following:

- Given a set amount of resources, how big a sample size can we collect, how big a sample size should we collect, and is the study worth doing given the sample size that could be collected? (e.g., Lenth, 2001; Lakens, 2021)
- Given relevant decision outcomes, what thresholds for type 1 errors, type 2 errors, relative Bayesian evidence, etc. are appropriate? Relatedly, given a certain sample size, what is the optimal balance between type 1 and type 2 errors? (e.g., Maier and Lakens, 2021; Kim and Choi, 2021).
- Given a certain set of equivalent causal models, which of several possible studies we could run would let us reject the most alternative explanations from the equivalence class? (Matiasz et al., 2017)
- Given a finite amount of grant resources and a set of research proposals, which research proposals should be funded? Relatedly, how should funding be distributed across researchers and labs in order to achieve desired research outcomes? (Smaldino et al., 2019)
- How do we determine which effect sizes are "practically relevant" and worth following up on in future research? (e.g., Torgerson et al., 1995; Lakens et al., 2018a; Anvari et al., 2021)
- Is the registered report format an inherently slower and more labor-intensive format than traditional research reports, or is the labor simply shifted from the end of the research process to the beginning? (e.g., Chambers, 2019)

Central to each question above is the assumption that resources are finite. Sample sizes cannot be made infinitely large. Type 1 and 2 errors cannot be made infinitely small. We cannot conduct or fund all studies we can imagine. Forced to consider the practical constraints of limited resources and multiple desirable options, it makes sense to consider not only which research practices increase the quality of research, but also which practices increase quality the most given a finite set of resources and a set of available, mutually exclusive options.

There are good reasons to think that research efficiency as a metascientific topic will only become more important in the years to come (see Peterson and Panofsky, 2021, for a critique of this view, though the critique could also be read as an argument in favor of the importance of efficiency as a topic for debate and study). First, consider the emergence of large-scale data collection

collaborations in several areas of social science in the past decade (e.g., Alipourfard et al., 2021; Moshontz et al., 2018; Coles et al., 2019; Klein et al., 2014; Frank and Ben, 2021; Botvinik-Nezer et al., 2020; Wagge et al., 2019). Such collaborations usually target substantial data collection efforts - sometimes involving hundreds of researchers - towards a small selection of study designs (e.g., Forscher et al., 2020) to overcome various methodological issues in single-lab research. When resources get concentrated in this way, it becomes all the more important to ensure that the knowledge we generate will be worth the resources invested. Second, consider the growing number of calls to increase quality in all areas of research practice, including theory (Muthukrishna and Henrich, 2019) measurement (Borsboom, 2006; Flake and Fried, 2020) generalizability (Yarkoni, 2020), transparency (Nosek et al., 2016), reproducibility (Munafò et al., 2017), etc. Since raising the quality of research often comes at a cost to the researcher, it is important that metascience consider possible unintended consequences of proposed interventions to increase quality (Peterson and Panofsky, 2021). E.g., researchers might be tempted to switch to more feasible data collection methods and study designs such as online surveys and cross-sectional designs (e.g., Kuehberger and Schulte-Mecklenbeck, 2018), which could reduce the practical efficiency of policies that ought, in theory, to increase research quality. Third, consider that recent calls to increase research quality also generates a need for more efficient coordination in years to come. That is, if we want improved theoretical rigor, improved measurement instrument quality, etc. we need researchers within subfields to come together, identify the central quality issues relevant to their field, and coordinate their resources to tackle these issues as a community. Metascience ought to be at the forefront in studying how coordination could be organized to increase research efficiency.

Beyond generating a unifying framework for the study of replication study selection, this thesis also contributes to the broader field of research efficiency in important ways. First, it identifies a research goal (increasing the expected utility of scientific claims) that could be extended and serve as an end point for other research processes as well. Second, the thesis highlights important interdisciplinary bridges that can be of value to any future research on research efficiency. For example, application of causal graphical models for representing research decision scenarios should be useful for modeling a broad range of metascientific interventions. Another example is the application of *VoI* concepts (Clemen, 1996) for modeling and comparing the expected utility of different possible research decisions. In principle, if we were able to accurately estimate the *EVSI* of various research activities, such as preregistering a study design, increasing sample size, validating a measurement instrument, or conducting a novel study, then we should be able to extend the model in chapter 2 to formalize and compare the expected utility of these research activities. A topic to which *VoI* concepts should be more immediately applicable is the problem of alpha justification (Lakens et al., 2018b), since *EVSI* max-

imization is simply the formal combination of "weighing the relative cost of errors" and "incorporating prior probabilities" proposed by Maier and Lakens (2021). Model generalization will be more challenging for research practices whose function is complicated and may serve multiple goals (e.g., open access publishing).

6.5 Conclusion

As replication slowly becomes adopted into mainstream research practice, deciding which studies to focus replication efforts on is a problem faced by a growing number of researchers. In this thesis I explore solutions to this problem by establishing a unifying framework to justify study selection strategies, by demonstrating how a formal study selection strategy can be derived from this framework given certain additional measurement assumptions, and by exploring how the selection strategy can be carried out in practice. Left for future research is the crucial task of validating RV_{Cn} and comparing its performance with other potential indicators of replication value. It is my sincere hope that this thesis will have provided sufficient foundations for such work to commence.

It is also my hope that this thesis can help facilitate a more nuanced discussion of efficiency in metascience more generally. Scientists face resource constraints not only when deciding what to replicate, but also when deciding whether to replicate or pursue a novel line of research, which line of research to pursue, whether to run a small-scale study on a self-determined topic or join forces with a larger consortium, which research design to opt for, how much resources to invest in data collection, whether to invest all resources into a single project or distribute funds over several projects, etc. Whenever resource constraints force us to choose between several courses of action that all have desirable properties, it is important that we carefully consider what goals we are hoping to achieve. Furthermore, it is important that we actively debate the relative importance of different goals in different contexts, and consider how different research activities may be more or less helpful (or even harmful) for reaching the various goals of scientific research. Only through careful consideration of what our goals should be and how a given research practice relates to these goals can we say anything meaningful about whether engaging in that research practice is *efficient*. Every working scientist funded by the public has a responsibility to make sure that their research is carried out efficiently, in a manner that ensures high quality research for the resources the public has invested. Subsequently, every metascientist advocating for a certain (change to) research practice has a responsibility to make sure this research practice will not only be helpful for reaching a particular goal in principle, but that it will be efficient in practice given all the goals deemed important in science. Thus, it should be of great interest to the metascientific community to discuss problems of efficiency. This

thesis provides one example of how such discussion could be carried out.

Appendix A

Updating replication value once a replication is conducted

A.1 Calculating replication value for a meta-analytic estimate

When replications of a replication target have already been performed, we will usually want to combine the information from these replications in our replication value estimate. Similarly, once we have replicated a chosen replication target, we may want to combine the information from the original study and our replication to consider if further replication is warranted, or if it would be better to focus new resources on a different replication target. A straight-forward way to calculate RV_{Cn} based on combined evidence from several studies would be to calculate the meta-analytic variance estimate for the studies.

For a fixed effects meta-analysis, RV_{Cn} can be estimated in the following way:

$$RV_{fixed} = \frac{w(C_S)}{Y+1} SE_M = \frac{w(C_S)}{Y+1} \sqrt{\frac{1}{\sum_{i=1}^{k} W_i}} = \frac{w(C_S)}{(Y+1)\sqrt{\frac{1}{\sum_{i=1}^{k} W_i}}} \quad \text{(A.1)}$$

where RV_{fixed} is the estimate of replication value, C is the citation count of the original article reporting on the target claim, Y is the number of years since the original article was published, SE_M is the standard error of the summary

effect for the fixed effect meta-analysis (see Borenstein et al., 2009, equations 11.4 and 11.5), W is the inverse variance weight of each included study (see Borenstein et al., 2009, equation 11.2), and i denotes a particular study in the set k included in the RV_{fixed} estimate.

Equation A.1 can still be used for calculating RV_{fixed} whether or not we want to assume that the standard deviation is equal across all candidates and use only sample size to estimate the standard error for each study. When we make the assumption of equal standard deviations, the equation stays identical, but we must change the variance estimate provided to the inverse variance weight W (see Borenstein et al., 2009, equation 11.2) from $\frac{\sigma^2}{n}$ to $\frac{1}{n}$. The inverse variance weight then simply becomes the sample size, since $\frac{1}{Var} = \frac{1}{\frac{1}{n}} = n$.

In many situations, however, it would be more appropriate to calculate the variance for a random effects meta-analysis, because there is often true effect size heterogeneity which will influence the variance estimate (Borenstein et al., 2009, chapter 13)[1]. For a random effects meta-analysis, RV_{fixed} can be estimated in the following way:

$$RV_{fixed} = \frac{w(C_S)}{Y+1} SE_{M*} = \frac{w(C_S)}{Y+1} \sqrt{\frac{1}{\sum_{i=1}^{k} W_i*}} = \frac{w(C_S)}{(Y+1)\sqrt{\frac{1}{\sum_{i=1}^{k} W_i*}}} \quad \text{(A.2)}$$

where RV_{rand} is the estimate of replication value, C is the citation count of the original article reporting on the target claim, Y is the number of years since the original article was published, ${SE_M}^*$ is the standard error of the summary effect for the mixed effect meta-analysis (see Borenstein et al., 2009, equation 12.8 and 12.9), W is the inverse variance weight of each study including τ^2 (see Borenstein et al., 2009, 2013, equation 12.6, 12.7), and i is a given study in the set k included in the RV_{rand} estimate.

While the random effects model is theoretically straightforward to calculate for a set of studies, there are two practical obstacles to using random effects variance in the estimate of RV_{Cn}:

1. In addition to variance estimates, which can be derived using only the sample size, we need to determine the effect sizes of interest in order to calculate the between-study heterogeneity estimate τ^2 (see Borenstein et al., 2009, equation 12.2 and 12.3).
2. We need a sufficient sample of studies in order to reliably estimate τ^2 (Borenstein et al., 2009, page 84).

[1]However, when we only allow close replications into the meta-analytic estimate, we only expect theoretically close effects to be included (LeBel et al., 2018), which should imply low effect size heterogeneity. This means that, in practice, the difference between RV estimates based on fixed-effects and random-effects models should be low whenever close replication results are combined.

In addition to the practical difficulties of deriving random effects precision estimates, it can also be difficult to determine which among a set of findings should be combined in a meta-analysis (Sharpe, 1997; Esteves et al., 2017). Because closely related findings are rarely linked to each other in meta-data, identifying such findings will currently require manual inspection by the replicating researcher. However, platforms like CurateScience could perhaps make automatic identification of replications possible in the future (LeBel et al., 2018).

One might reasonably ask whether the citation count of all replications should also be combined in the meta-analytic replication value estimate. On the one hand, more studies entail a larger literature, which in theory could increase the overall impact and visibility of the claims studies, and perhaps citation count would reflect such increases. However we regard it as likely that replication and original studies are usually cited together, or at least for similar reasons, which means that each replication's citation count provides highly overlapping information about the underlying value of the replication target. We therefore only include the citation count of the original study in the definitions of RV_{fixed}/RV_{rand}, though we recognize the appropriateness of this choice is a largely unresolved empirical question.

A.2 Example: Applying RV_{fixed} to studies on the Stroop effect

The R script containing the data material and exact calculations used to produce the numbers reported in this section can be found on OSF (https://osf.io/e35pu/).

Suppose we would like to calculate RV_{fixed} for the classic Stroop effect (Stroop, 1935). The Stroop effect is an extremely impactful finding, and one of the most cited publications in psychology. On the other hand, the original results have been consistently replicated in many research efforts [e.g., MacLeod (1991); Ebersole et al. (2016); Verhaeghen and De Meersman (1998), not to mention psychology classrooms around the world. Considering whether to, at this point, commit further resources to replicating the Stroop effect, we need to consider our uncertainty about the Stroop effect given the total weight of evidence from both the original study as well as from replications.

As of 2021-08-14, the citation count of the original Stroop effect (Stroop, 1935) was 9423 according to Crossref, and the age of the publication at that time was 83 years. There are three separate studies reported in Stroop (1935). Study 2 directly tests the well-known interference effect of word meaning on color naming that most later replications have been based on (MacLeod, 1991; Ebersole et al., 2016).

Study 2 includes data from 100 participants, but we should adjust this sample size for the fact that Stroop (Stroop, 1935, Study 2) is a within-subject design (see supplementary material 2). Unfortunately, like many repeated measures experiments, Stroop does not report the correlation between dependent measures, which is necessary to accurately calculate the standard error and effect size of a repeated measures experiment (Dunlap et al., 1996). However, we can estimate the within-subject correlation from data generated by a similar Stroop paradigm. For example, a close replication of the original Stroop paradigm was performed by Burns et al. (Burns et al., 2019). For the conditions relevant for the replication of Stroop (1935), Study 2, the within-subject correlation in this study is 0.932, 95%CI[0.901, 0.954]. With this correlation estimate we can convert the within-subject effect size to a corresponding between-subject effect size that would have the same amount of precision. The adjusted sample size is (100*2)/(1-0.932) = 2958.737 (see appendix B, equation B.1). The replication value for Stroop (Stroop, 1935, Study 2) thus becomes:

$$\frac{w(C_S)}{Y+1} \times \frac{1}{\sqrt{n}} = \frac{9423}{86+1} \times \frac{1}{\sqrt{2958.737}} = 1.991 \tag{A.3}$$

Suppose we would like to update this replication value estimate after replications of the Stroop effect are performed. A collection of close replications of the original Stroop paradigm can be found in Verhaeghen and De Meersman (1998). Study designs and sample characteristics (within the young group) were similar to Stroop (1935), Study 2 in all but two of the studies reported in this meta-analysis (in two cases, subjects were told to read the words, not name the colors. Dulaney and Rogers, 1994; Park et al., 1996).

We can track change in replication value as replications accumulate by recalculating equation A.1 after every successive replication attempt, including in each calculation all replication studies published up until that point. Assuming equal standard deviations, the only parameter changing between successive replications is the sample size. Therefore, replication value will always decrease monotonically under these assumptions[2]. Figure A.1 displays the reduction in replication value with every replication reported in Verhaeghen and De Meersman (1998), in the order by which these replications were published.

[2]Monotonic decrease may not hold under different assumptions. For example, if we instead use equation A.2 to update replication value, replication value could in theory increase after a replication if the effect size heterogeneity τ^2 increases substantially.

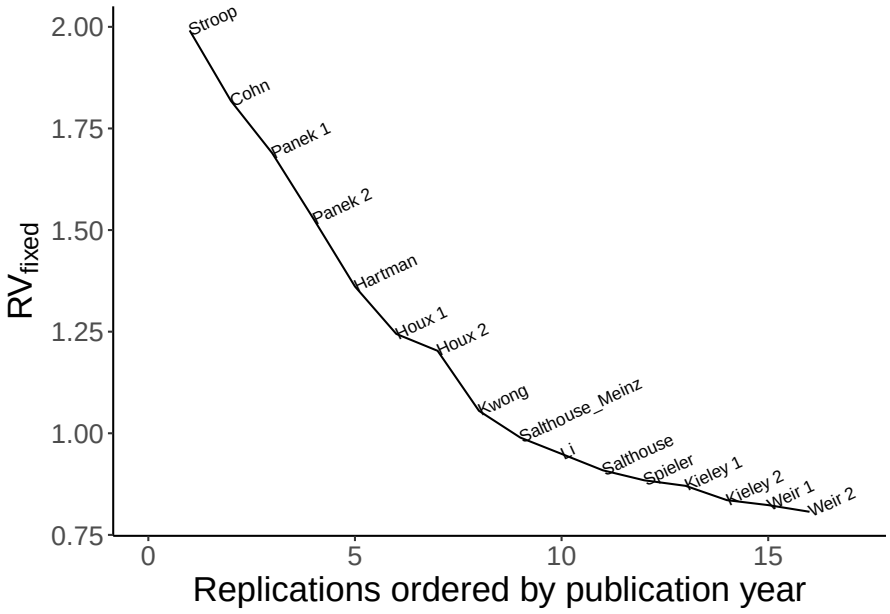

Figure A.1: Cumulative replication value of the Stroop effect over time, derived by recalculating equation A.1 after every successive replication attempt. Studies included are reported in Verhaeghen and De Meersman (1998), table 1, with the exception of Dulaney and Rogers (1994), and Park et al. (1996).

Appendix B

Converting within-subjects sample size to between-subjects sample size

A statistical limitation that arises when we approximate standard error from sample size alone is that we ignore the statistical design of the studies we intend to compare, which is usually important for accurately estimating the standard error (Borenstein et al., 2009, chapter 8). As one example, a paired samples *t*-test has better precision for the difference score that is calculated between the two measurements than an independent samples *t*-test, partly because the participants contribute twice as many data points in the paired design, and partly because of within-subject correlation (Lakens, 2016a).

We can attempt to mitigate this violation, however, by converting the sample sizes from within-subject designs into the sample size that would achieve the same precision in a between-subjects design. In a replication candidate set composed of both within-subject and between-subject designs, converting within-subject sample size in this way should increase RV_{Cn} measurement quality by reducing measurement bias caused by variation in study design.

$$n_B = \frac{n_W a}{1 - \rho} \tag{B.1}$$

where n_W is the sample size of the within-subject design[1], ρ is the within-subject correlation, a is the number of groups that each subject contributes

[1]Note that we have replaced the capital N in Maxwell and Delaney (2004) with the

data points to, and n_B is the estimated sample size that a between-subject study would need to reach the same level of precision.

The population parameter ρ is usually estimated from the within-subject correlation r in the sample. A practical issue is that this value is very rarely reported in published manuscripts. In these cases, it is possible to calculate r from summary statistics. For example, if we have access to the t-value, Cohens $d_{average}$ (see Borenstein et al., 2009, equation 4.18), and the sample size n_W, we can calculate r by solving for r in Dunlap et al. (1996), equation 3:

$$r = \frac{2t^2 - d^2_{average} n_W}{2t^2} \tag{B.2}$$

Or, if we have access to the standard error of the difference and the standard error of both groups, we could calculate r by solving for r in Lakens (see Lakens, 2013, equation 8):

$$r = \frac{SD_1^2 + SD_2^2 - S^2_{diff}}{2SD_1SD_2} \tag{B.3}$$

If we do not have access to these summary statistics or the raw data, we could estimate ρ based on r in conceptually similar studies. If there are no realistic reference points for ρ whatsoever, we could potentially consider setting ρ to 0. n_W will still receive a correction in this case from being multiplied by a. Note however, that this is a very conservative assumption, and unlikely to be realistic in most cases. More importantly, the choice of 0 over any other arbitrary value of ρ is motivated purely by a desire to be conservative in our assumptions about the strength of ρ, though it is not clear why one would want to be conservative.

lowercase n used to refer to sample size throughout this article. The two symbols refer to the same entity.

Appendix C

Distribution of CWTS citation cluster keywords

We acquired additional bibliometric information from the Centre for Science and Technology Studies (CWTS, https://www.cwts.nl/) about citation clusters in our data (a proxy for scientific subfields based upon clustering of publications based on citation relationships, Waltman and van Eck, 2019). Each citation cluster can be thought of as a data-defined research subfield, where articles that cite similar articles tend to end up in the same cluster/subfield. The clusters were generated independently of our study, based on all bibliometric information in the CWTS database. For each article in our dataset, we retrieved information about its corresponding CWTS cluster, including how many articles in the entire CWTS database were included in this cluster. The CWTS cluster algorithm also generates key-terms that describe the most prevalent research topics dealt with in each cluster. This allows for a higher-resolution analysis of the research topics covered in our dataset. We expected to see a wide range of CWTS clusters included in our data, and we expected the cluster labels would primarily denote terms related to social phenomena and their influence on cognition and neural activity.

The records in our dataset were sampled from 162 unique CWTS citation clusters. The size of each cluster varied substantially (min = 829 records, median = 12354 records, max = 39770 records). The full distribution of cluster size visualized in figure SM1.

To better understand the scientific topics covered in these citation clusters, we inspected the category labels assigned to each cluster by CWTS. In total, the citation clusters were associated with 774 unique labels. Table 2 displays the frequency of the 50 most frequently mentioned category labels in our data.

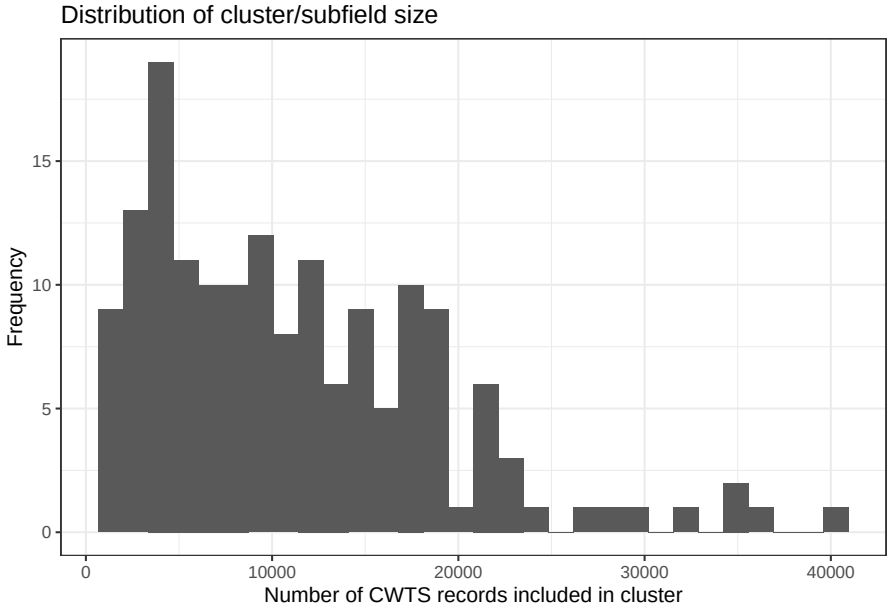

Figure C.1: Distribution of the size of CWTS clusters to which the articles in our dataset belong. The size of each cluster refers to the total number of records associated with that cluster over the entire CWTS database, which includes millions of articles from all over science. Thus, the size of the cluster refers to the population size of the cluster, not how many articles are included in the cluster in our data.

Table C.1: Frequency table of the 50 most prevalent cluster labels in our dataset. Label frequencies are identical for many labels because multiple labels are used to describe a single cluster. For example, the terms "face recognition", "face processing", "prosopagnosia", "unfamiliar face", and "facial identity" all appear 118 times because they are all used to describe one, and only one, cluster to which 118 articles in our data belong.

	Label	Frequency
50	intertemporal choice	364
46	decision making	358
47	delay discounting	358
48	impulsivity	358
49	iowa gambling task	358
45	imitation	318
41	action observation	258
42	empathy	258
43	mirror neuron	258
44	motor imagery	258
40	attentional bias	210
39	fear	207
36	emotional face	203
37	facial expression	203
38	social anxiety	203
31	default mode network	180
32	fmri	180
33	fmri data	180
34	functional connectivity	180
35	resting state	180
30	alzheimer	125
25	face processing	118
26	face recognition	118
27	facial identity	118
28	prosopagnosia	118
29	unfamiliar face	118
21	n400	109
22	primary progressive aphasia	109
23	semantic dementia	109
24	visual word recognition	109
16	contamination	67
17	disgust	67
18	disgust sensitivity	67
19	moral dilemma	67
20	moral judgment	67

	Label	Frequency
10	death anxiety	65
11	mind	65
12	mortality salience	65
13	ostracism	65
14	social exclusion	65
15	terror management	65
5	autobiographical memory	63
6	expressive writing	63
7	generativity	63
8	mental time travel	63
9	rumination	63
2	false belief	60
3	infant	60
4	month old infant	60
1	effect	52

While most cluster labels describe substantive topics, the inclusion of "effect" in the top 50 most prevalent labels should highlight that not all labels assigned by the cluster algorithm are equally informative for understanding cluster content. The distribution of substantive topic labels is largely consistent with our expectations. Terms like "fmri", "fmri data", "imitation", "empathy", "facial expression", "social anxiety", and "social exclusion" are frequently used to describe subfields to which the articles in our dataset belong. Conversely, we did not observe any labels that were obviously unrelated to our a-priori expectations about which topics should be covered in a sample of social fMRI articles.

Close inspection of the cluster label data revealed an important confuscating factor, however. While cluster labels are descriptive of the whole cluster as it appears in the CWTS database, they are not necessarily descriptive of the subset of that cluster included in our candidate set. As a prominent example, the citation cluster described by the labels "intertemporal choice", "decision making", "delay discounting", "impulsivity" and "iowa gambling task" is the most prevalent cluster in our data. However, the labels used to summarize the 13168 CWTS records in this cluster are not necessarily representative of the minority of articles from the cluster that are included in our dataset. Although the term "iowa gambling task" is descriptive of the cluster as a whole, only one out of the 358 articles from the cluster in our dataset mentions the Iowa gambling task. Inferring about the contents of our dataset based on these labels alone could therefore be misleading. However, correspondence in topics between frequently occurring CWTS cluster labels and frequently co-occurring keywords in VOSviewer (see figure 3 in main manuscript) is encouraging.

Appendix D

Consulting field experts to identify potential quantitative indicators of uncertainty.

To better understand what information is important for assessing overall uncertainty about findings from fMRI research, we constructed a survey to probe experts in fMRI research about which information they use to assess the quality and quantity of evidence for fMRI findings in their field. The purpose of the survey was twofold. First, we wanted an opportunity to discover quantitative indicators of uncertainty we had not previously considered, and that might be feasible to code in our data. Second, we wanted to compare the reported importance of sample size for evaluating uncertainty in comparison with other information researchers might also be using.

D.1 Methods

Pilot data collection was carried out on a convenience sample of colleagues of the first (Peder) and second (Anna) author. 13 researchers responded to the survey. All participants were researchers with, or in the process of completing, a PhD, who had experience with collecting and analyzing fMRI data. The purpose of this sampling restriction was to ensure that all participants had sufficient prior knowledge of fMRI methodology to give informative answers to the survey items they were presented with.

The survey was created in Qualtrics (https://www.qualtrics.com/). The survey and all data collected are available on OSF (https://osf.io/f7zdq/). The survey contained open-ended items encouraging researchers to list whatever information they considered important for assessing evidence. The survey also contained a number of closed-ended questions asking researchers to rate (on a visual analogue scale from 1:100 with 1 being the least important and 100 being the most important) and rank-order the importance of the following factors for judging the quality and quantity of evidence in support of a finding (this list of factors were generated by the authors after internal discussion about which factors could plausibly be used to evaluate uncertainty about social fMRI research):

1. The total sample size collected for the study.
2. The percentage of participants that were excluded (after they met the inclusion criteria for participating in MRI research).
3. The statistical power of the study to detect effect sizes of interest.
4. The size of the effect (e.g. Cohen's d for condition differences, Pearson's r for brain-behavior correlations, or percentage signal change for raw BOLD signal differences).
5. Cluster extent of relevant cluster(s).
6. The p-value for relevant cluster(s).
7. Whether the finding is a main effect or an interaction.
8. How participants were assigned to conditions, if relevant (e.g., randomly, single/double blind, etc.).
9. In cases where you know of a replication study, the result(s) of a replication study,
10. In cases where you know of a replication study, whether the replication is a close (direct) or conceptual replication.
11. In cases where you know of a replication study, whether the replication is conducted by an independent team or not.
12. Whether the finding is based on within-subjects measurements or between-subjects measurements.
13. Peak Z-value for relevant clusters.
14. Open access to the underlying empirical data that were analyzed.
15. Whether the study has been preregistered.
16. Whether there are statistical errors in the results reported (e.g., the degrees of freedom do not correspond to the other reported statistics, the total sample size does not equal the sum of the group sample sizes, etc.).
17. Whether the finding has a strong connection with theory.
18. Whether the finding is predicted a priori or discovered during data exploration.
19. How participants were sampled from the population (e.g., stratified random sampling, snowball sampling, convenience sampling, etc.).
20. Whether the finding is unexpected (e.g., counterintuitive), or in line with

what we already know.

For each factor, we also asked for open-ended comments to better understand how the information was being used by researchers to assess evidence. For example, after asking researchers to rate the importance of "the percentage of participants that were excluded", we also asked participants to "indicate in what way you believe this information is related to the quality and quantity of evidence in support of a finding". We used the participants' responses on the items related to "the total sample size collected for the study" as a preliminary validation of whether sample size relates to uncertainty in the way assumed by Isager et al. (2020).

D.2 Results

The open responses by participants did not reveal novel quantitative indicators of uncertainty that we had not already considered, and that would be feasible to collect for all studies in our data.

There seemed to be broad agreement among experts that sample size is important for evaluating the quality and quantity of evidence for a typical fMRI finding. Several experts freely offered sample size as a piece of information they would be evaluating when assessing the credibility of a finding (before seeing our list of potentially important factors). Sample size also received the second highest median rating out of all factors (see table SM2-1 and figure SM2-1, and the highest average rank-order out of all factors (only "the results of a replication study" received an equally high rank. See table SM2-1 and figure SM2-2). In addition, statistical power, partially a function of sample size, was consistently highly rated and ranked by experts, and one expert explicitly pointed to the relationship between sample size and power in their comments ("Sample size is the easiest way to increase statistical power"). Finally, when asked specifically about the importance of sample size, there seemed to be broad agreement that a higher sample size generally entails higher credibility, in line with the assumptions of Isager et al. (2021). However, two experts described feeling less confident about findings supported by a very high sample size, due to the elevated risks of overinterpreting trivially small and meaningless effects (a problem often referred to as "the crud factor", Meehl, 1990; Orben and Lakens, 2020). Besides sample size (and statistical power) participants seemed to consistently agree on the importance of a few other factors (see table SM2-1, figure SM2-1 and SM2-2).

There seemed to be broad agreement among experts that sample size is important for evaluating the quality and quantity of evidence for a typical fMRI finding. Several experts freely offered sample size as a piece of information they would be evaluating when assessing the credibility of a finding (before

seeing our list of potentially important factors). Sample size also received the second highest median rating out of all factors (see table SM2 and figure SM2-1, and the highest average rank-order out of all factors (only "the results of a replication study" received an equally high rank. See table SM2 and figure SM2-2). In addition, statistical power, partially a function of sample size, was consistently highly rated and ranked by experts, and one expert explicitly pointed to the relationship between sample size and power in their comments ("Sample size is the easiest way to increase statistical power"). Finally, when asked specifically about the importance of sample size, there seemed to be broad agreement that a higher sample size generally entails higher credibility, in line with the assumptions of Isager et al. (2021). However, two experts described feeling less confident about findings supported by a very high sample size, due to the elevated risks of overinterpreting trivially small and meaningless effects (a problem often referred to as "the crud factor", Meehl, 1990; Orben and Lakens, 2020). Besides sample size (and statistical power) participants seemed to consistently agree on the importance of a few other factors (see table SM2, figure SM2-1 and SM2-2).

Overall, we cautiously interpret these results as preliminary validation of correspondence between the rationale of Isager et al. (2020) and how experts actually use sample size when evaluating uncertainty. However, we stress that the low sample size and exploratory nature of this pilot calls for replication before any firm conclusions can be drawn.

Table D.1: Median rating and rank for all factors asked about in the pilot survey.

Factor	N ratings	Median rating	N rankings	Median rank
effect predicted or exploratory	9	88	10	7
statistical errors	9	86	10	9.5
sample size	11	84	10	4
replication result	10	84	10	4
open data available	9	83	10	8
strongly connected to theory	10	81	10	4.5
statistical power	10	78.5	10	5.5
replication close or not	10	78	10	NA
preregistered	11	77	10	9.5
condition assignment	11	75	10	13
replication independent or not	10	74.5	10	NA
within or between design	9	71	10	13
effect size	9	67	10	5

Factor	N ratings	Median rating	N rankings	Median rank
cluster p-value	9	60	10	13
effect unexpected	9	59	10	10
cluster extent	10	51	10	12
participants excluded	10	45	10	15
cluster peak Z-value	9	37	10	11.5
participant sampling	10	24	10	15.5
main effect or interaction	9	17	10	12.5

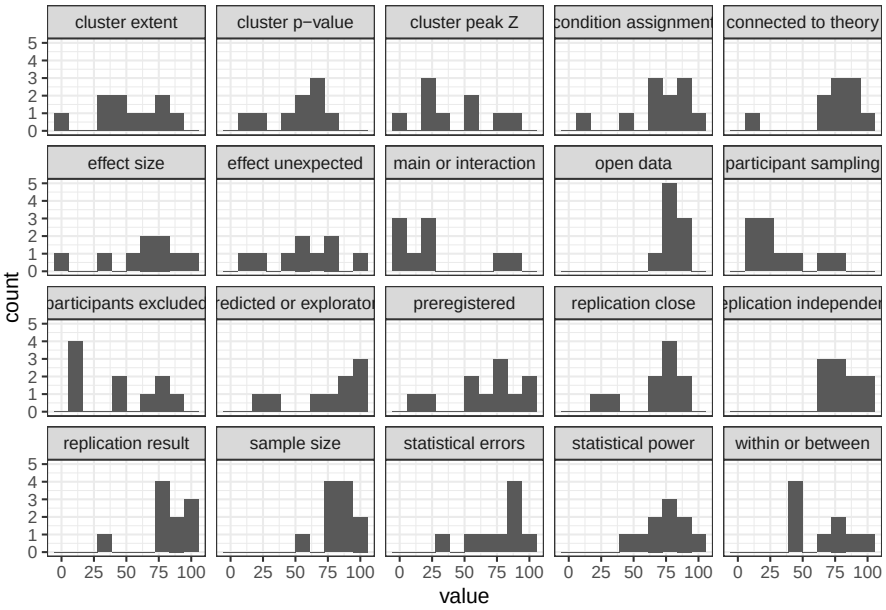

Figure D.1: Histograms of ratings for each of the 20 factors presented to participants. All factors were rated on a visual analogue scale from 1:100 with 1 being the least important and 100 being the most important.

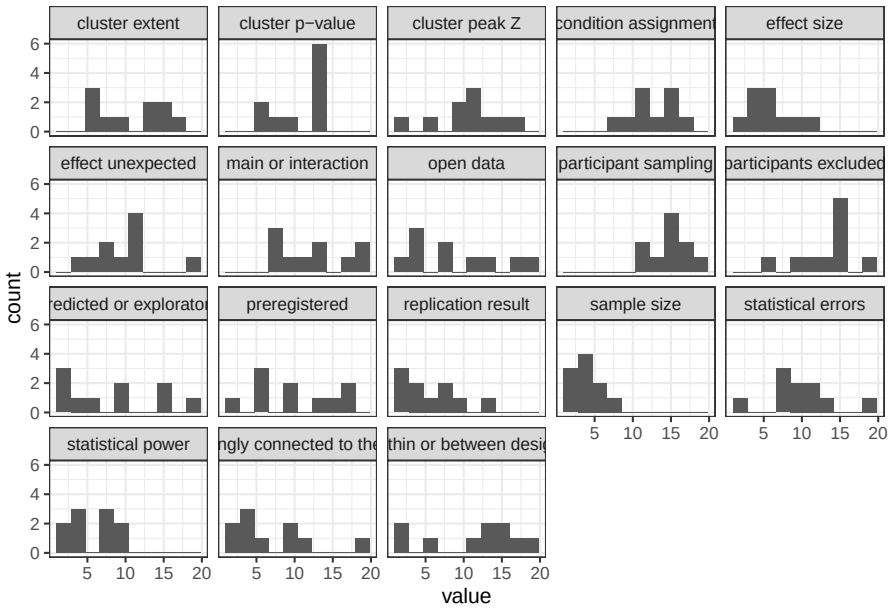

Figure D.2: Histograms of relative rank-order ratings for the 18 factors that participants were asked to rank ("replication close or not" and "replication independent or not" were not included for rank-ordering). Participants assigns one value of rank from 1-18 to all factors, where rank 1 indicates the most important factor, and rank 18 indicates the least important factor.

Appendix E

Identifying the "main claim/finding" for each study

Some information that is relevant for assessing replication value is related to individual empirical findings within studies (e.g., the precision of the estimate for a particular effect size of interest). If we want to use such information to compare the replication value of two studies, we first need to decide which findings from each study to use for our comparison. For example, consider the implication of using the standard error of the mean to calculate RV_{Cn}. If we do not approximate the standard error via the total sample size of the study, standard error is related to a particular mean estimate within the study. Since a study may report many mean estimates, it may be related to any number of standard errors. Thus, it is no longer enough to decide which studies to include as replication candidates. We now also have to decide which specific findings from these studies to consider, because RV_{Cn} estimates depend on statistical information from these findings.

We conducted a study to try and identify the *main finding* of each study in our set of replication candidates. The main finding was defined as the reported finding which is centrally highlighted in either the abstract or conclusion section of the article in which the study is reported, and which seemed to be the focus point of the study design. For example, the finding that the fusiform face area is reliably and selectively activated by images of faces is the main finding used to support the more general claim that faces are processed in a specific spatial location within the human brain. The ultimate goal of this study, in conjunction with the pilot study reported in supplementary material SM2, was to identify indicators of statistical uncertainty for each main finding (such

as standard error of the mean, and Bayesian posterior evidence) from which different estimators of replication value could be constructed, calculated, and compared with RV_{Cn}.

Main findings for each paper had to be coded manually. We developed a general coding procedure, instructing coders on where in the paper to look for mentions of the main finding, and what would indicate that something is a main finding. Three co-authors (PMI, AvtV, LG) then applied this procedure to a small set of studies within our candidate set to test the feasibility of the coding effort. All data and materials from this small coding effort is available on OSF (https://osf.io/953du/). Below follows a brief summary of our own conclusions.

Our pilot suggested that main findings from each study could indeed be identified. Identification was relatively time-intensive. A main finding could be identified within a few minutes on average, but overall time taken varied considerably around this average estimate. Some studies included the main claim in the title, in which case coding could take seconds. Other studies required coders to consult several sections of the article to verify that a claim in question was indeed the main claim of the paper. In these cases coding could take several minutes. In every case, the main finding of the study was mentioned in the abstract of the article in which the study appeared.

With respect to identifying statistical information for each finding, however, we quickly realized that this would become challenging. By and large, main findings were associated with a number of different statistical results. Consider the following, example:

> In two experiments, we used a functional magnetic resonance (fMR)-repetition suppression paradigm to demonstrate that distinct frontal–parietal–temporal regions are sensitive to processing the scenarios or what participants imagined was happening in an event (e.g. medial prefrontal, posterior cingulate, temporal–parietal and middle temporal cortices are sensitive to the scenarios associated with future social events), people (medial prefrontal cortex), objects (inferior frontal and premotor cortices) and locations (posterior cingulate/retrosplenial, parahippocampal and posterior parietal cortices) that typically constitute simulations of personal future events. This pattern of results demonstrates that the neural substrates of these component features of event simulations can be reliably identified in the context of a task that requires participants to simulate complex, everyday future experiences. - Szpunar et al. (2014)

It is clear that many statistical results are being utilized in this statement, and it is not clear which, if any, would be more appropriate to serve as the

results on which a replication value estimate is based. Many of the findings identified in our pilot had a similar structure to the example above. We suspect this finding structure will be common in the field of social fMRI, where hypotheses are often of the form "what does neural activity look like for task/manipulation/stimulus/group X?" and therefore relate to multiple aspects of the fMRI data collected. For the purposes of collecting statistical data for replication value estimation, it appears it would not be enough to simply identify the main finding of each study in our dataset. We would also have to determine, for each finding, which empirical results to extract statistical information from and how to resolve the common case where a finding is related to multiple statistical results. Due to the labor intensity this implies, we determined not to proceed with the coding of main findings in this project.

Appendix F

Age-citation correlation matrices for all citation sources

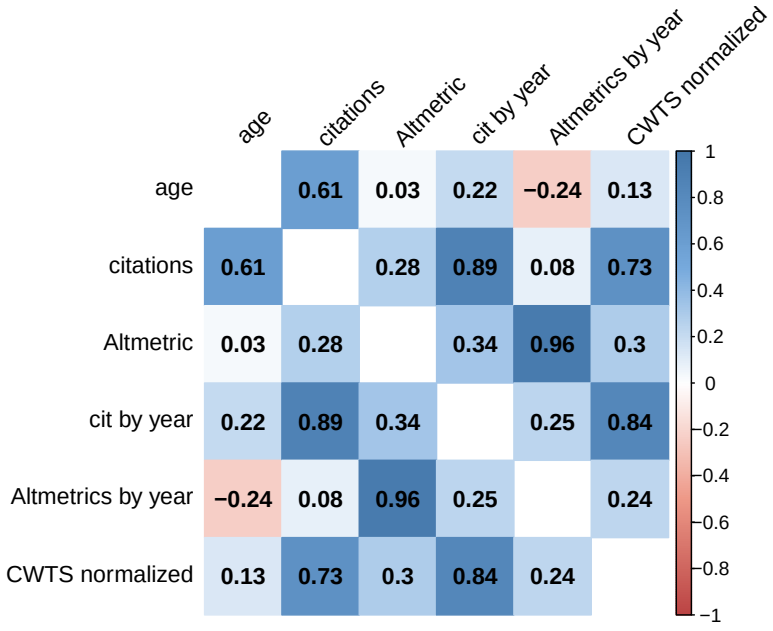

Figure F.1: Matrix of bi-variate correlations between age and Web of Science citation count.

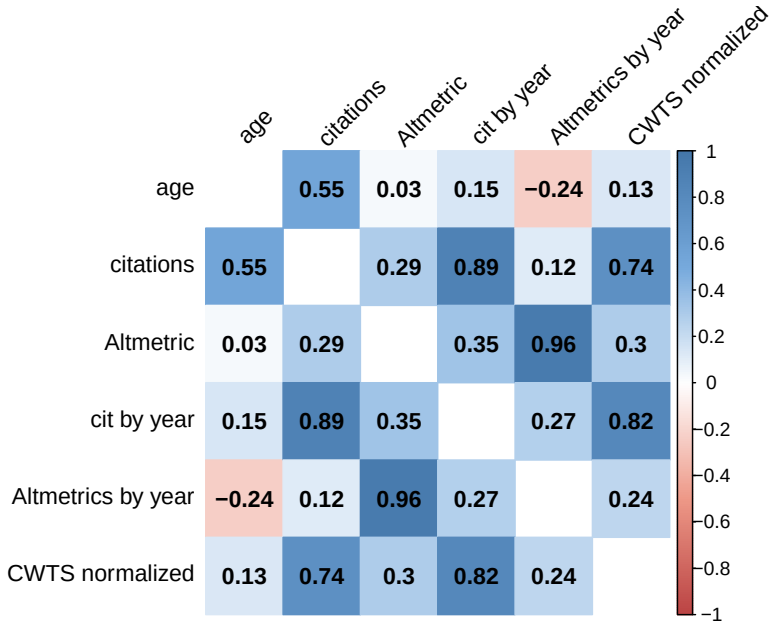

Figure F.2: Matrix of bi-variate correlations between age and Crossref citation count.

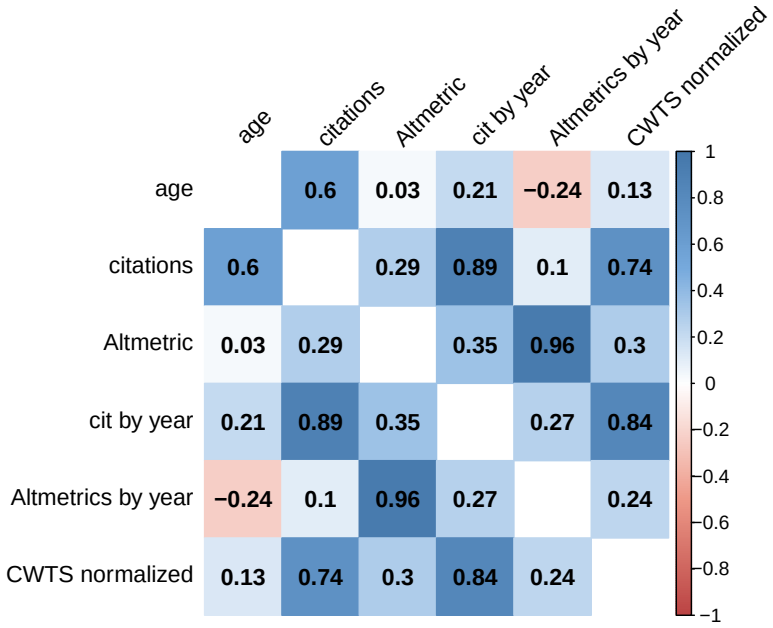

Figure F.3: Matrix of bi-variate correlations between age and Scopus citation count.

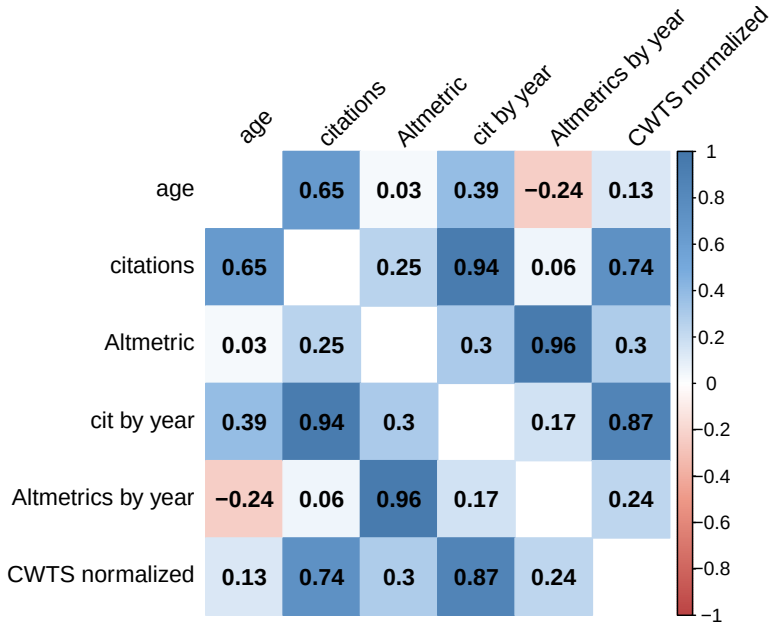

Figure F.4: Matrix of bi-variate correlations between age and CWTS citation count.

Acknowledgements

In considering all the people who have touched my life over the past four years, anything I could say would fall short of capturing he gratitude I feel. I will settle for thanking a small group of people for a brief list of things. There are more people. There are more things. All are dearly remembered.

To my bachelor and master's supervisors, Thomas, Siri, and Marie, and to Jostein. For fostering my early love of science and for teaching me how to do it properly.

To Chris. For providing feedback that always deepened my understanding of issue at hand, for captaining this PhD project safely to shore, and for providing the invisible hand of governance that has let us PhD students survive and thrive at TU/e and at the HTI group.

To my friends and colleagues at TU/e and in the HTI group, for the brown bags and the daily lunches in the cantina. For the cookie Wednesdays, birthday cakes, outings, Christmas lunches, and PhD defense concerts. To Peter for always knocking on the door to remind us to join drinks at Intermate. To Armin for the pleasant conversations. To Laura and Karin for guiding me through my first teaching experience. To Alain, Minha, Elcin, Hanne, Alejandro, Emir, Stephan, and everyone else who have made me feel at home in Eindhoven.

To Amy, Chris, Farid, Jarda, Nick, Noah, Patrick, Sarah, Sophia, and all the amazing visitors to the Red Square lab. For a million memories I will carry with me always. For friendship, knowledge, and adventure. It has been an honor and a privilege to get to know you all.

To my family (bipeds and quadrupeds), my girlfriend, and my friends back in Norway. For all the support and encouragement. For all the visits, Zoom calls and Skype sessions. For being there in the bad times as well as the good. For always reminding me that I am loved.

To Anna. For our endlessly enjoyable collaboration, for the all the advice and support, for giving me a real sense of mastery, and for always making me feel like a scientific peer. The serendipity of how this collaboration

came to be will never seize to baffle me. https://twitter.com/annaveer/status/1045658532245385221 - now we're here. I still have the fMRI machine by the way.

To Anne and Leo. My academic older siblings. For all the days in the office, drinking Trader Joe's chai and Club Mate, stepping over Leo sleeping on the floor, drawing models on the whiteboard, and discussing every meta-scientific topic known to man. For the office Whiskey and the beers at Intermate and Hubble. For the conferences, lab retreats, trips, and meet-ups around the world. For teaching me about a thousand different topics and for introducing me to a hundred different scholars. For being my role models in that most critical period of academic development that is the PhD. I would not be half the scientist I am today without you.

To Daniel. My hero. My mentor. My friend. For everything.

Thank you.

Curriculum Vitae

Peder Mortvedt Isager was born on 02-10-1992 in Lørenskog, Norway. He received his high-school education at Eidsvoll VGS in Viken county, Norway. In 2015 he obtained a bachelor's degree in psychology from the University of Oslo, Norway. His bachelor thesis, focusing on the relation between physical position and social status perception, was published as a journal article, and was presented at an international conference. In 2017 he obtained a master's degree in cognitive neuroscience from the University of Oslo. His master thesis, focusing on reward perception in healthy humans under the influence of an opioid substance, was presented at several international conferences.

After briefly working as a research assistant at the Center for Social and Affective Neuroscience in Linköping, Sweden, he relocated in September 2017 to Eindhoven University of Technology to pursue his PhD. His project was embedded in the VIDI-funded project "Increasing the Reliability and Efficiency of Psychological Science". His research focused on replication and replication study selection, and the application of causal modeling to formally define concepts and theoretical arguments metascience and psychometrics. In addition, he has worked on additional topics within metascience with local and international collaborators, and his research outputs have been published in international journals. Since 2018 he has been a board member of the Data and Methods committee of the Psychological Science Accelerator, and in 2022 he was appointed Assistant Director for Data. Since 2020 he has been employed as assistant professor at Oslo New University College.

Publications

Publications included in the dissertation

Isager, P. M., van 't Veer, A. E., & Lakens, D. (preprint 2021). Replication value as a function of citation impact and sample size. https://doi.org/10.31222/osf.io/knjea

Isager, P. M. (preprint 2021). Test validity defined as d-connection between target and measured attribute: Expanding the causal definition of Borsboom et al. (2004). https://doi.org/10.31234/osf.io/btgsr

Isager, P. M., van Aert, R. C. M., Bahník, Š., Brandt, M. J., DeSoto, K. A., Giner-Sorolla, R., . . . Lakens, D. (forthcoming publication in Psychological Methods). Deciding what to replicate: A formal definition of "replication value" and a decision model for replication study selection. https://doi.org/10.31222/osf.io/2gurz

In preparation

Isager, P. M., Lakens, D., van Leeuwen, T. N., Grandpierre, L., & van 't Veer, A. E. (in preparation). Selecting Studies for Replication in Social Neuroscience: Exploring a Formal Approach.

Publications not included in the dissertation

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

Buchanan E. M., Crain S. E., Cunningham A. L., . . . , Isager, P. M., . . . et al. (2021). Getting Started Creating Data Dictionaries: How to Create a Shareable Data Set. Advances in Methods and Practices in Psychological Science, https://doi.org/10.1177/2515245920928007

Coles, N., Tiokhin, L., Scheel, A., 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

Lakens, D., Adolfi, F. G., Albers, C. J., Anvari, F., Apps, M. A., Argamon, S. E., . . . , Isager, P. M., . . . & Zwaan, R. A. (2018). Justify your alpha. Nature Human Behaviour, 2(3), 168-171. https://doi.org/10.1038/s41562-018-0311-x

Lakens, D., McLatchie, N., Isager, P. M., Scheel, A. M., & Dienes, Z. (2020). Improving inferences about null effects with Bayes factors and equivalence

tests. The Journals of Gerontology: Series B, 75(1), 45-57, https://doi.org/10.1093/geronb/gby065

Lakens, D., Scheel, A. M., & Isager, P. M. (2018). Equivalence testing for psychological research: A tutorial. Advances in Methods and Practices in Psychological Science, 1(2), 259-269. https://doi.org/10.1177/2515245918770963

McIntyre, S., Moungou, A., Boehme, R., Isager, P. M., Lau, F., Israr, A., … & Olausson, H. (2019). Affective touch communication in close adult relationships. In 2019 IEEE World Haptics Conference (WHC) (pp. 175-180). IEEE, https://doi.org/10.1109/WHC.2019.8816093

Moshontz, H., Campbell, L., Ebersole, C. R., IJzerman, H., Urry, H. L., Forscher, P. S., . . . , Isager, P. M., . . . & Chartier, C. R. (2018). The Psychological Science Accelerator: Advancing psychology through a distributed collaborative network. Advances in Methods and Practices in Psychological Science, 1(4), 501-515, https://doi.org/10.1177/2515245918797607

Scheel A. M., Tiokhin L., Isager P. M., Lakens D. (2021). Why Hypothesis Testers Should Spend Less Time Testing Hypotheses. Perspectives on Psychological Science. 16(4), 744-755. https://doi.org/10.1177/1745691620966795

Wang, K., Goldenberg, A., Dorison, C.A. et al. (2021). A multi-country test of brief reappraisal interventions on emotions during the COVID-19 pandemic. Nature Human Behaviour, 5, 1089–1110, https://doi.org/10.1038/s41562-021-01173-x

Bibliography

Agnoli, F., Wicherts, J. M., Veldkamp, C. L. S., Albiero, P., and Cubelli, R. (2017). Questionable research practices among italian research psychologists. *PLOS ONE*, 12(3):e0172792.

Aksnes, D. W., Langfeldt, L., and Wouters, P. (2019). Citations, Citation Indicators, and Research Quality: An Overview of Basic Concepts and Theories. *SAGE Open*, 9(1):215824401982957.

Alipourfard, N., Arendt, B., Benjamin, D. M., Benkler, N., Bishop, M., Burstein, M., Bush, M., Caverlee, J., Chen, Y., Clark, C., Almenberg, A. D., Errington, T., Fidler, F., Fox [SCORE, N., Frank, A., Fraser, H., Friedman, S., Gelman, B., Gentile, J., Giles, C. L., Gordon, M. B., Gordon-Sarney, R., Griffin, C., Gulden, T., Hahn, K., Hartman, R., Holzmeister, F., Hu, X. B., Johannesson, M., Kezar, L., Struhl, M. K., Kuter, U., Kwasnica, A. M., Lee, D.-H., Lerman, K., Liu, Y., Loomas, Z., Luis [SCORE, B., Magnusson, I., Miske, O., Mody, F., Morstatter, F., Nosek, B. A., Parsons, E. S., Pennock, D., Pfeiffer, T., Pujara, J., Rajtmajer, S., Ren, X., Salinas, A., Selvam, R. K., Shipman, F., Silverstein, P., Sprenger, A., Squicciarini, A. M., Stratman, S., Sun, K., Tikoo, S., Twardy, C. R., Tyner, A., Viganola, D., Wang, J., Wilkinson, D. P., Wintle, B., and Wu, J. (2021). Systematizing Confidence in Open Research and Evidence (SCORE). Technical report, SocArXiv. type: article.

Altmejd, A., Dreber, A., Forsell, E., Huber, J., Imai, T., Johannesson, M., Kirchler, M., Nave, G., and Camerer, C. (2019). Predicting the replicability of social science lab experiments. *PLOS ONE*, 14(12):e0225826. Publisher: Public Library of Science.

Altmetric (2021). Altmetric Attention Score. Retrieved April 18, 2021, from https://bit.ly/3dK7nrI.

American Psychological Association (2020). test bias – APA Dictionary of Psychology. Retrieved May 27, 2021, from https://dictionary.apa.org/test-bias.

American Psychological Association (2021a). Journal of Personality and Social Psychology. Retrieved November 1, 2021, from https://www.apa.org/pubs/journals/psp/index.

American Psychological Association (2021b). reliability – APA Dictionary of Psychology. Retrieved November 1, 2021, from https://dictionary.apa.org/reliability.

Anvari, F., Kievit, R., Lakens, D., Przybylski, A. K., Tiokhin, L., Wiernik, B. M., and Orben, A. (2021). Evaluating the practical relevance of observed effect sizes in psychological research. Technical report, PsyArXiv. type: article.

Ashar, Y. K., Clark, J., Gunning, F. M., Goldin, P., Gross, J. J., and Wager, T. D. (2021). Brain markers predicting response to cognitive-behavioral therapy for social anxiety disorder: an independent replication of Whitfield-Gabrieli et al. 2015. *Translational Psychiatry*, 11(1):260.

Association for Psychological Science (2018). NSF Invites Grant Applications Related to Reproducibility in Neuroimaging.

Bakan, D. (1966). The test of significance in psychological research. *Psychological bulletin*, 66(6):423–437.

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

Baribault, B., Donkin, C., Little, D. R., Trueblood, J. S., Oravecz, Z., van Ravenzwaaij, D., White, C. N., De Boeck, P., and Vandekerckhove, J. (2018). Metastudies for robust tests of theory. *Proceedings of the National Academy of Sciences*, 115(11):2607–2612.

Begley, C. G. and Ellis, L. M. (2012). Raise standards for preclinical cancer research. *Nature*, 483(7391):531–533. Bandiera_abtest: a Cg_type: Nature Research Journals Number: 7391 Primary_atype: Comments & Opinion Publisher: Nature Publishing Group Subject_term: Cancer;Drug development Subject_term_id: cancer;drug-development.

Biagioli, M. (2016). Watch out for cheats in citation game. *Nature*, 535(7611):201.

Blaszczynski, A. and Gainsbury, S. M. (2019). Editor's note: replication crisis in the social sciences. *International Gambling Studies*, 19(3):359–361. Publisher: Routledge _eprint: https://doi.org/10.1080/14459795.2019.1673786.

Block, J. and Kuckertz, A. (2018). Seven principles of effective replication studies: strengthening the evidence base of management research. *Management Review Quarterly*, 68(4):355–359.

Boekel, W., Wagenmakers, E.-J., Belay, L., Verhagen, J., Brown, S., and Forstmann, B. U. (2015). A purely confirmatory replication study of structural brain-behavior correlations. *Cortex*, 66:115–133.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., and Rothstein, H. R. (2009). *Introduction to Meta-Analysis*. John Wiley & Sons, Ltd, Chichester, UK.

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.

Bornmann, L. and Daniel, H. (2008). What do citation counts measure? A review of studies on citing behavior. *Journal of Documentation*, 64(1):45–80. Publisher: Emerald Group Publishing Limited.

Bornmann, L. and 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. _eprint: https://asistdl.onlinelibrary.wiley.com/doi/pdf/10.1002/asi.23329.

Borsboom, D. (2006). The attack of the psychometricians. *Psychometrika*, 71(3):425–440. Place: Germany Publisher: Springer.

Borsboom, D., Cramer, A. O. J., Kievit, R. A., Scholten, A. Z., and Franić, S. (2009). The end of construct validity. In *The concept of validity: Revisions, new directions, and applications*, pages 135–170. IAP Information Age Publishing, Charlotte, NC, US.

Borsboom, D., Mellenbergh, G. J., and van Heerden, J. (2004). The Concept of Validity. *Psychological Review*, 111(4):1061–1071.

Borsboom, D., van der Maas, H. L. J., Dalege, J., Kievit, R. A., and Haig, B. D. (2021). Theory Construction Methodology: A Practical Framework for Building Theories in Psychology. *Perspectives on Psychological Science*, page 1745691620969647. Publisher: SAGE Publications Inc.

Botvinik-Nezer, R., Holzmeister, F., Camerer, C. F., Dreber, A., Huber, J., Johannesson, M., Kirchler, M., Iwanir, R., Mumford, J. A., Adcock, R. A., Avesani, P., Baczkowski, B. M., Bajracharya, A., Bakst, L., Ball, S., Barilari, M., Bault, N., Beaton, D., Beitner, J., Benoit, R. G., Berkers, R. M. W. J., Bhanji, J. P., Biswal, B. B., Bobadilla-Suarez, S., Bortolini, T., Bottenhorn, K. L., Bowring, A., Braem, S., Brooks, H. R., Brudner, E. G., Calderon, C. B., Camilleri, J. A., Castrellon, J. J., Cecchetti, L., Cieslik, E. C., Cole, Z. J., Collignon, O., Cox, R. W., Cunningham, W. A., Czoschke, S., Dadi, K., Davis, C. P., Luca, A. D., Delgado, M. R., Demetriou, L., Dennison, J. B., Di, X., Dickie, E. W., Dobryakova, E., Donnat, C. L., Dukart, J., Duncan, N. W., Durnez, J., Eed, A., Eickhoff,

S. B., Erhart, A., Fontanesi, L., Fricke, G. M., Fu, S., Galván, A., Gau, R., Genon, S., Glatard, T., Glerean, E., Goeman, J. J., Golowin, S. A. E., González-García, C., Gorgolewski, K. J., Grady, C. L., Green, M. A., Guassi Moreira, J. F., Guest, O., Hakimi, S., Hamilton, J. P., Hancock, R., Handjaras, G., Harry, B. B., Hawco, C., Herholz, P., Herman, G., Heunis, S., Hoffstaedter, F., Hogeveen, J., Holmes, S., Hu, C.-P., Huettel, S. A., Hughes, M. E., Iacovella, V., Iordan, A. D., Isager, P. M., Isik, A. I., Jahn, A., Johnson, M. R., Johnstone, T., Joseph, M. J. E., Juliano, A. C., Kable, J. W., Kassinopoulos, M., Koba, C., Kong, X.-Z., Koscik, T. R., Kucukboyaci, N. E., Kuhl, B. A., Kupek, S., Laird, A. R., Lamm, C., Langner, R., Lauharatanahirun, N., Lee, H., Lee, S., Leemans, A., Leo, A., Lesage, E., Li, F., Li, M. Y. C., Lim, P. C., Lintz, E. N., Liphardt, S. W., Losecaat Vermeer, A. B., Love, B. C., Mack, M. L., Malpica, N., Marins, T., Maumet, C., McDonald, K., McGuire, J. T., Melero, H., Méndez Leal, A. S., Meyer, B., Meyer, K. N., Mihai, G., Mitsis, G. D., Moll, J., Nielson, D. M., Nilsonne, G., Notter, M. P., Olivetti, E., Onicas, A. I., Papale, P., Patil, K. R., Peelle, J. E., Pérez, A., Pischedda, D., Poline, J.-B., Prystauka, Y., Ray, S., Reuter-Lorenz, P. A., Reynolds, R. C., Ricciardi, E., Rieck, J. R., Rodriguez-Thompson, A. M., Romyn, A., Salo, T., Samanez-Larkin, G. R., Sanz-Morales, E., Schlichting, M. L., Schultz, D. H., Shen, Q., Sheridan, M. A., Silvers, J. A., Skagerlund, K., Smith, A., Smith, D. V., Sokol-Hessner, P., Steinkamp, S. R., Tashjian, S. M., Thirion, B., Thorp, J. N., Tinghög, G., Tisdall, L., Tompson, S. H., Toro-Serey, C., Torre Tresols, J. J., Tozzi, L., Truong, V., Turella, L., van 't Veer, A. E., Verguts, T., Vettel, J. M., Vijayarajah, S., Vo, K., Wall, M. B., Weeda, W. D., Weis, S., White, D. J., Wisniewski, D., Xifra-Porxas, A., Yearling, E. A., Yoon, S., Yuan, R., Yuen, K. S. L., Zhang, L., Zhang, X., Zosky, J. E., Nichols, T. E., Poldrack, R. A., and Schonberg, T. (2020). Variability in the analysis of a single neuroimaging dataset by many teams. *Nature*, 582(7810):84–88. Bandiera_abtest: a Cg_type: Nature Research Journals Number: 7810 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Decision;Decision making;Human behaviour;Scientific community Subject_term_id: decision;decision-making;human-behaviour;scientific-community.

Brandt, M. J., IJzerman, H., Dijksterhuis, A., Farach, F. J., Geller, J., Giner-Sorolla, R., Grange, J. A., Perugini, M., Spies, J. R., and van 't Veer, A. (2014). The Replication Recipe: What makes for a convincing replication? *Journal of Experimental Social Psychology*, 50:217–224.

Briggs, R. A. (2019). Normative Theories of Rational Choice: Expected Utility. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, fall 2019 edition.

Burgers, J. (2019). Citation Counts as a Measure for Scientific Impact.

Burns, D. M., Fox, E. B., Greenstein, M., and Montgomery, D. A. (2019). An old task in new clothes: A preregistered direct replication attempt of enclothed cognition effects on Stroop performance.

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., and Munafò, M. R. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, 14(5):365–376. Number: 5 Publisher: Nature Publishing Group.

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Nave, G., Nosek, B. A., Pfeiffer, T., Altmejd, A., Buttrick, N., Chan, T., Chen, Y., Forsell, E., Gampa, A., Heikensten, E., Hummer, L., Imai, T., Isaksson, S., Manfredi, D., Rose, J., Wagenmakers, E.-J., and Wu, H. (2018). Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behaviour*, page 1.

Carp, J. (2012). On the Plurality of (Methodological) Worlds: Estimating the Analytic Flexibility of fMRI Experiments. *Frontiers in Neuroscience*, 6. Publisher: Frontiers.

Cesario, J. (2014). Priming, Replication, and the Hardest Science. *Perspectives on Psychological Science*, 9(1):40–48.

Chakraborty, T., Kumar, S., Goyal, P., Ganguly, N., and Mukherjee, A. (2014). Towards a stratified learning approach to predict future citation counts. In *IEEE/ACM Joint Conference on Digital Libraries*, pages 351–360.

Chamberlain, S., Zhu, H., Jahn, N., Boettiger, C., and Ram, K. (2020). rcrossref.

Chambers, C. (2019). What's next for Registered Reports? *Nature*, 573(7773):187–189. Bandiera_abtest: a Cg_type: Comment Number: 7773 Publisher: Nature Publishing Group Subject_term: Publishing, Peer review.

Chambers, C. D. and Tzavella, L. (2020). The past, present, and future of Registered Reports. preprint, MetaArXiv.

Chi, A. J., Lopes, A. J., Rong, L. Q., Charlson, M. E., Alvarez, R. D., and Boerner, T. (2021). Examining the correlation between Altmetric Attention Score and citation count in the gynecologic oncology literature: Does it have an impact? *Gynecologic Oncology Reports*, 37:100778.

Clarivate Analytics (2020). Web of Science Core Collection Help. Retrieved July 11, 2021, from https://bit.ly/3dFNgei.

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

Coles, N. A., March, D. S., Marmolejo-Ramos, F., Banaruee, H., Butcher, N., Cavallet, M., Dagaev, N., Eaves, D., Foroni, F., Gorbunova, E., Gygax, P., IJzerman, H., Hinojosa, J. A., Ikeda, A., Khatin-Zadeh, O., Larsen, J. T., Özdoğru, A. A., Parzuchowski, M., Rodriguez-Medina, D. A., Ruiz-Fernandez, S., Som, B., Suarez, I., Trujillo, N., Trujillo, S., van der Zee, T., Villalba-García, C., Willis, M., Yamada, Y., Aczel, B., Hajdu, N., Basnight-Brown, D., Ellsworth, P. C., Gaertner, L., Strack, F., Liuzza, M. T., and Marozzi, M. (2019). The Many Smiles Collaboration: A Multi-Lab Test of the Facial Feedback Hypothesis. preprint, PsyArXiv.

Coles, N. A., Tiokhin, L., Scheel, A. M., Isager, P. M., and Lakens, D. (2018). The costs and benefits of replication studies. *Behavioral and Brain Sciences*, 41.

Colman, E. R. and Vukadinović Greetham, D. (2015). Memory and burstiness in dynamic networks. *Physical Review E*, 92(1):012817. Publisher: American Physical Society.

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

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

DeBruine, L. M. and Barr, D. J. (2021). Understanding Mixed-Effects Models Through Data Simulation. *Advances in Methods and Practices in Psychological Science*, 4(1):2515245920965119. Publisher: SAGE Publications Inc.

DeDeo, S. (2018). Information Theory for Intelligent People. page 15.

Devezer, B., Navarro, D. J., Vandekerckhove, J., and Ozge Buzbas, E. (2021). The case for formal methodology in scientific reform. *Royal Society Open Science*, 8(3):200805. Publisher: Royal Society.

Dimoka, A., Pavlou, P. A., and Davis, F. D. (2011). NeuroIS: The Potential of Cognitive Neuroscience for Information Systems Research. *Information Systems Research*, 22(4):687–702.

Doyen, S., Klein, O., Pichon, C.-L., and Cleeremans, A. (2012). Behavioral Priming: It's All in the Mind, but Whose Mind? *PLoS ONE*, 7(1):e29081.

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

Drummond, A. and Philipp, M. C. (2017). Commentary: "Misguided Effort with Elusive Implications" and "A Multi-Lab Pre-Registered Replication of the Ego Depletion Effect". *Frontiers in Psychology*, 8. Publisher: Frontiers.

Dulaney, C. L. and Rogers, W. A. (1994). Mechanisms underlying reduction in Stroop interference with practice for young and old adults. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, 20(2):470–484.

Dunlap, W. P., Cortina, J. M., Vaslow, J. B., and Burke, M. J. (1996). Meta-analysis of experiments with matched groups or repeated measures designs. *Psychological Methods*, 1(2):170–177.

Earp, B. D. and Trafimow, D. (2015). Replication, falsification, and the crisis of confidence in social psychology. *Frontiers in Psychology*, 6.

Ebersole, C. R., Atherton, O. E., Belanger, A. L., Skulborstad, H. M., Allen, J. M., Banks, J. B., Baranski, E., Bernstein, M. J., Bonfiglio, D. B., Boucher, L., Brown, E. R., Budiman, N. I., Cairo, A. H., Capaldi, C. A., Chartier, C. R., Chung, J. M., Cicero, D. C., Coleman, J. A., Conway, J. G., Davis, W. E., Devos, T., Fletcher, M. M., German, K., Grahe, J. E., Hermann, A. D., Hicks, J. A., Honeycutt, N., Humphrey, B., Janus, M., Johnson, D. J., Joy-Gaba, J. A., Juzeler, H., Keres, A., Kinney, D., Kirshenbaum, J., Klein, R. A., Lucas, R. E., Lustgraaf, C. J., Martin, D., Menon, M., Metzger, M., Moloney, J. M., Morse, P. J., Prislin, R., Razza, T., Re, D. E., Rule, N. O., Sacco, D. F., Sauerberger, K., Shrider, E., Shultz, M., Siemsen, C., Sobocko, K., Weylin Sternglanz, R., Summerville, A., Tskhay, K. O., van Allen, Z., Vaughn, L. A., Walker, R. J., Weinberg, A., Wilson, J. P., Wirth, J. H., Wortman, J., and 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.

Eck, N. J. v., Waltman, L., Raan, A. F. J. v., Klautz, R. J. M., and Peul, W. C. (2013). Citation Analysis May Severely Underestimate the Impact of Clinical Research as Compared to Basic Research. *PLOS ONE*, 8(4):e62395. Publisher: Public Library of Science.

Eckermann, S., Karnon, J., and Willan, A. R. (2010). The Value of Value of Information. *PharmacoEconomics*, 28(9):699–709.

eLife (2017). The challenges of replication. *eLife*, 6:e23693.

Errington, T. M., Iorns, E., Gunn, W., Tan, F. E., Lomax, J., and Nosek, B. A. (2014). An open investigation of the reproducibility of cancer biology research. *eLife*, 3:e04333. Publisher: eLife Sciences Publications, Ltd.

Esteves, S. C., Majzoub, A., and Agarwal, A. (2017). The problem of mixing 'apples and oranges' in meta-analytic studies. *Translational Andrology and Urology*, 6(S4):S412–S413.

Fanelli, D. (2010). "Positive" Results Increase Down the Hierarchy of the Sciences. *PLOS ONE*, 5(4):e10068.

Fanelli, D. (2012). Negative results are disappearing from most disciplines and countries. *Scientometrics*, 90(3):891–904.

Federer, L. M., Belter, C. W., Joubert, D. J., Livinski, A., Lu, Y.-L., Snyders, L. N., and Thompson, H. (2018). Data sharing in PLOS ONE: An analysis of Data Availability Statements. *PLOS ONE*, 13(5):e0194768. Publisher: Public Library of Science.

Feldman, G. (2021). Mass Replications & Extensions (CORE).

Fiedler, K. and Schwarz, N. (2016). Questionable Research Practices Revisited. *Social Psychological and Personality Science*, 7(1):45–52.

Field, S. M., Hoekstra, R., Bringmann, L., and Van Ravenzwaaij, D. (2019). When and Why to Replicate: As Easy as 1, 2, 3? *Collabra: Psychology*, 5(1):46.

Finkel, E. J., Eastwick, P. W., and 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.

Flake, J. K. and Fried, E. I. (2020). Measurement Schmeasurement: Questionable Measurement Practices and How to Avoid Them. *Advances in Methods and Practices in Psychological Science*, 3(4):456–465. Publisher: SAGE Publications Inc.

Forscher, P. S., Paris, B., Primbs, M., and Coles, N. A. (2020). PSACR: The Psychological Science Accelerator's COVID-19 Rapid-Response Project. Technical report, PsyArXiv. type: article.

Forsell, E., Viganola, D., Pfeiffer, T., Almenberg, J., Wilson, B., Chen, Y., Nosek, B. A., Johannesson, M., and Dreber, A. (2019). Predicting replication outcomes in the Many Labs 2 study. *Journal of Economic Psychology*, 75:102117.

Fraley, R. C. and 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).

Franco, A., Malhotra, N., and Simonovits, G. (2014). Publication bias in the social sciences: Unlocking the file drawer. *Science*, 345(6203):1502–1505.

Frank, M. C. and Ben, R. (2021). The ManyBabies Project. Retrieved November 1, 2021, from https://manybabies.github.io/.

Frank, M. C., Bergelson, E., Bergmann, C., Cristia, A., Floccia, C., Gervain, J., Hamlin, J. K., Hannon, E. E., Kline, M., Levelt, C., Lew-Williams, C., Nazzi, T., Panneton, R., Rabagliati, H., Soderstrom, M., Sullivan, J., Waxman, S., and Yurovsky, D. (2017). A Collaborative Approach to Infant Research: Promoting Reproducibility, Best Practices, and Theory-Building. *Infancy*, 22(4):421–435.

Fried, E. I. (2020). Lack of Theory Building and Testing Impedes Progress in The Factor and Network Literature. *Psychological Inquiry*, 31(4):271–288. Publisher: Routledge _eprint: https://doi.org/10.1080/1047840X.2020.1853461.

Furukawa, T. A., Barbui, C., Cipriani, A., Brambilla, P., and Watanabe, N. (2006). Imputing missing standard deviations in meta-analyses can provide accurate results. *Journal of Clinical Epidemiology*, 59(1):7–10.

Garfield, E. (2006). The History and Meaning of the Journal Impact Factor. *JAMA*, 295(1):90–93.

Gerring, J. (1999). What Makes a Concept Good? A Criterial Framework for Understanding Concept Formation in the Social Sciences. *Polity*, 31(3):357–393. Publisher: The University of Chicago Press.

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

Glasziou, P., Meats, E., Heneghan, C., and Shepperd, S. (2008). What is missing from descriptions of treatment in trials and reviews? *BMJ*, 336(7659):1472–1474. Publisher: British Medical Journal Publishing Group Section: Analysis.

Goodhart, C. A. E. (1984). Problems of Monetary Management: The UK Experience. In Goodhart, C. A. E., editor, *Monetary Theory and Practice: The UK Experience*, pages 91–121. Macmillan Education UK, London.

Goodman, S. (2008). A Dirty Dozen: Twelve P-Value Misconceptions. *Seminars in Hematology*, 45(3):135–140.

Gopalakrishna, G., Riet, G. t., Cruyff, M. J. L. F., Vink, G., Stoop, I., Wicherts, J., and Bouter, L. (2021). Prevalence of questionable research practices, research misconduct and their potential explanatory factors: a survey among academic researchers in The Netherlands. Technical report, MetaArXiv. type: article.

Greenland, S. and Pearl, J. (2014). Causal Diagrams. In Balakrishnan, N., Colton, T., Everitt, B., Piegorsch, W., Ruggeri, F., and Teugels, J. L., editors, *Wiley StatsRef: Statistics Reference Online*, page stat03732. John Wiley & Sons, Ltd, Chichester, UK.

Guilford, J. P. (1946). New Standards For Test Evaluation. *Educational and Psychological Measurement*, 6(4):427–438. Publisher: SAGE Publications Inc.

Hagger, M. S., Chatzisarantis, N. L. D., Alberts, H., Anggono, C. O., Batailler, C., Birt, A. R., Brand, R., Brandt, M. J., Brewer, G., Bruyneel, S., Calvillo, D. P., Campbell, W. K., Cannon, P. R., Carlucci, M., Carruth, N. P., Cheung, T., Crowell, A., De Ridder, D. T. D., Dewitte, S., Elson, M., Evans, J. R., Fay, B. A., Fennis, B. M., Finley, A., Francis, Z., Heise, E., Hoemann, H., Inzlicht, M., Koole, S. L., Koppel, L., Kroese, F., Lange, F., Lau, K., Lynch, B. P., Martijn, C., Merckelbach, H., Mills, N. V., Michirev, A., Miyake, A., Mosser, A. E., Muise, M., Muller, D., Muzi, M., Nalis, D., Nurwanti, R., Otgaar, H., Philipp, M. C., Primoceri, P., Rentzsch, K., Ringos, L., Schlinkert, C., Schmeichel, B. J., Schoch, S. F., Schrama, M., Schütz, A., Stamos, A., Tinghög, G., Ullrich, J., vanDellen, M., Wimbarti, S., Wolff, W., Yusainy, C., Zerhouni, O., and Zwienenberg, M. (2016). A Multilab Preregistered Replication of the Ego-Depletion Effect. *Perspectives on Psychological Science*, 11(4):546–573.

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

Heath, A., Kunst, N., Jackson, C., Strong, M., Alarid-Escudero, F., Goldhaber-Fiebert, J. D., Baio, G., Menzies, N. A., and Jalal, H. (2020). Calculating the Expected Value of Sample Information in Practice: Considerations from 3 Case Studies. *Medical Decision Making*, 40(3):314–326. Publisher: SAGE Publications Inc STM.

Heirene, R. (2020). A call for replications of addiction research: Which studies should we replicate & what constitutes a "successful" replication? preprint, PsyArXiv.

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. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/16066359.2020.1751130.

Hernán, M. and Robins, J. (2020). *Causal Inference: What If*. Chapman & Hall/CRC, Boca Raton.

Hervé, M. (2021). RVAideMemoire: Testing and Plotting Procedures for Biostatistics.

Huber, D. E., Potter, K. W., and Huszar, L. D. (2019). Less "Story" and more "Reliability" in cognitive neuroscience. *Cortex; a journal devoted to the study of the nervous system and behavior*, 113:347–349.

Ioannidis, J. P. A., Boyack, K. W., Small, H., Sorensen, A. A., and Klavans, R. (2014). Bibliometrics: Is your most cited work your best? *Nature News*, 514(7524):561. Section: Comment.

Isager, P. M. (2018). What To Replicate? Justifications Of Study Choice From 85 Replication Studies. Publisher: Zenodo.

Isager, P. M. (2020). Test validity defined as d-connection between target and measured attribute: Expanding the causal definition of Borsboom et al. (2004). Technical report, PsyArXiv. type: article.

Isager, P. M., Aert, R. C. M. v., Bahník, ., Brandt, M., DeSoto, K. A., Giner-Sorolla, R., Krueger, J., Perugini, M., Ropovik, I., Veer, A. v. t., Vranka, M. A., and Lakens, D. (2020). Deciding what to replicate: A formal definition of "replication value" and a decision model for replication study selection. Technical report, MetaArXiv. type: article.

Isager, P. M., van t' Veer, A., Nosten, T., Janson, E., and Lakens, D. (2019). Quantifying Replication Value: A guide in the decision of what to replicate.

Isager, P. M., Veer, A. v. t., and Lakens, D. (2021). Replication value as a function of citation impact and sample size. Technical report, MetaArXiv. type: article.

Javadi, A.-H., Emo, B., Howard, L. R., Zisch, F. E., Yu, Y., Knight, R., Pinelo Silva, J., and Spiers, H. J. (2017). Hippocampal and prefrontal processing of network topology to simulate the future. *Nature Communications*, 8(1):14652. Number: 1 Publisher: Nature Publishing Group.

JESP (2018). JESP Registered Reports Guidelines.

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

Kassam, K. S., Markey, A. R., Cherkassky, V. L., Loewenstein, G., and Just, M. A. (2013). Identifying Emotions on the Basis of Neural Activation. *PLoS ONE*, 8(6):e66032.

Ke, Q., Ferrara, E., Radicchi, F., and Flammini, A. (2015). Defining and identifying Sleeping Beauties in science. *Proceedings of the National Academy of Sciences*, 112(24):7426–7431. Publisher: National Academy of Sciences Section: Physical Sciences.

Kim, J. H. and Choi, I. (2021). Choosing the Level of Significance: A Decision-theoretic Approach. *Abacus*, 57(1):27–71. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/abac.12172.

Kitcher, P. (1995). *The advancement of science: science without legend, objectivity without illusions*. Oxford Univ. Press, New York, 1. pb. publ edition. OCLC: 33278791.

Klautzer, L., Hanney, S., Nason, E., Rubin, J., Grant, J., and 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):201–209.

Klein, R. A., Ratliff, K. A., Vianello, M., Adams Jr., R. B., Bahník, ., Bernstein, M. J., Bocian, K., Brandt, M. J., Brooks, B., Brumbaugh, C. C., Cemalcilar, Z., Chandler, J., Cheong, W., Davis, W. E., Devos, T., Eisner, M., Frankowska, N., Furrow, D., Galliani, E. M., Hasselman, F., Hicks, J. A., Hovermale, J. F., Hunt, S. J., Huntsinger, J. R., IJzerman, H., John, M.-S., Joy-Gaba, J. A., Barry Kappes, H., Krueger, L. E., Kurtz, J., Levitan, C. A., Mallett, R. K., Morris, W. L., Nelson, A. J., Nier, J. A., Packard, G., Pilati, R., Rutchick, A. M., Schmidt, K., Skorinko, J. L., Smith, R., Steiner, T. G., Storbeck, J., Van Swol, L. M., Thompson, D., van 't Veer, A. E., Vaughn, L. A., Vranka, M., Wichman, A. L., Woodzicka, J. A., and Nosek, B. A. (2014). Investigating variation in replicability: A "many labs" replication project. *Social Psychology*, 45(3):142–152.

Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Adams, R. B., Alper, S., Aveyard, M., Axt, J. R., Babalola, M. T., Bahník, ., Batra, R., Berkics, M., Bernstein, M. J., Berry, D. R., Bialobrzeska, O., Binan, E. D., Bocian, K., Brandt, M. J., Busching, R., Rédei, A. C., Cai, H., Cambier, F., Cantarero, K., Carmichael, C. L., Ceric, F., Chandler, J., Chang, J.-H., Chatard, A., Chen, E. E., Cheong, W., Cicero, D. C., Coen, S., Coleman, J. A., Collisson, B., Conway, M. A., Corker, K. S., Curran, P. G., Cushman, F., Dagona, Z. K., Dalgar, I., Dalla Rosa, A., Davis, W. E., de Bruijn, M., De Schutter, L., Devos, T., de Vries, M., Doğulu, C., Dozo, N., Dukes, K. N., Dunham, Y., Durrheim, K., Ebersole, C. R., Edlund, J. E., Eller, A., English, A. S., Finck, C., Frankowska, N., Freyre, M.- ., Friedman, M., Galliani, E. M., Gandi, J. C., Ghoshal, T., Giessner, S. R., Gill, T., Gnambs, T., Gómez, ., González, R., Graham, J., Grahe, J. E., Grahek, I., Green, E. G. T., Hai, K., Haigh, M., Haines, E. L., Hall, M. P., Heffernan, M. E., Hicks, J. A., Houdek, P., Huntsinger, J. R., Huynh, H. P., IJzerman, H., Inbar, Y., Innes-Ker, . H., Jiménez-Leal, W., John, M.-S., Joy-Gaba, J. A., Kamiloğlu, R. G., Kappes, H. B., Karabati, S., Karick, H., Keller, V. N., Kende, A., Kervyn, N., Knežević, G., Kovacs, C., Krueger, L. E., Kurapov, G., Kurtz, J., Lakens, D., Lazarević, L. B., Levitan, C. A., Lewis, N. A., Lins, S., Lipsey, N. P., Losee, J. E., Maassen, E., Maitner,

A. T., Malingumu, W., Mallett, R. K., Marotta, S. A., Međedović, J., Mena-Pacheco, F., Milfont, T. L., Morris, W. L., Murphy, S. C., Myachykov, A., Neave, N., Neijenhuijs, K., Nelson, A. J., Neto, F., Lee Nichols, A., Ocampo, A., O'Donnell, S. L., Oikawa, H., Oikawa, M., Ong, E., Orosz, G., Osowiecka, M., Packard, G., Pérez-Sánchez, R., Petrović, B., Pilati, R., Pinter, B., Podesta, L., Pogge, G., Pollmann, M. M. H., Rutchick, A. M., Saavedra, P., Saeri, A. K., Salomon, E., Schmidt, K., Schönbrodt, F. D., Sekerdej, M. B., Sirlopú, D., Skorinko, J. L. M., Smith, M. A., Smith-Castro, V., Smolders, K. C. H. J., Sobkow, A., Sowden, W., Spachtholz, P., Srivastava, M., Steiner, T. G., Stouten, J., Street, C. N. H., Sundfelt, O. K., Szeto, S., Szumowska, E., Tang, A. C. W., Tanzer, N., Tear, M. J., Theriault, J., Thomae, M., Torres, D., Traczyk, J., Tybur, J. M., Ujhelyi, A., van Aert, R. C. M., van Assen, M. A. L. M., van der Hulst, M., van Lange, P. A. M., van 't Veer, A. E., Vásquez-Echeverría, A., Ann Vaughn, L., Vázquez, A., Vega, L. D., Verniers, C., Verschoor, M., Voermans, I. P. J., Vranka, M. A., Welch, C., Wichman, A. L., Williams, L. A., Wood, M., Woodzicka, J. A., Wronska, M. K., Young, L., Zelenski, J. M., Zhijia, Z., and Nosek, B. A. (2018). Many Labs 2: Investigating Variation in Replicability Across Samples and Settings. *Advances in Methods and Practices in Psychological Science*, 1(4):443–490.

KNAW (2018). *Replication studies – Improving reproducibility in the empirical sciences*. KNAW, Amsterdam.

Koo, T. K. and Li, M. Y. (2016). A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *Journal of Chiropractic Medicine*, 15(2):155–163.

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

Kuehberger, A. and Schulte-Mecklenbeck, M. (2018). Selecting target papers for replication. *Behavioral and Brain Sciences*, 41.

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. *Frontiers in Psychology*, 4.

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

Lakens, D. (2016b). Why Within-Subject Designs Require Fewer Participants than Between-Subject Designs.

Lakens, D. (2019). The value of preregistration for psychological science: A conceptual analysis. *Japanese Psychological Review*, 62(3):221–230.

Lakens, D. (2021). Sample Size Justification. Technical report, PsyArXiv. type: article.

Lakens, D., Adolfi, F. G., Albers, C. J., Anvari, F., Apps, M. A. J., Argamon, S. E., Baguley, T., Becker, R. B., Benning, S. D., Bradford, D. E., Buchanan, E. M., Caldwell, A. R., Van Calster, B., Carlsson, R., Chen, S.-C., Chung, B., Colling, L. J., Collins, G. S., Crook, Z., Cross, E. S., Daniels, S., Danielsson, H., DeBruine, L., Dunleavy, D. J., Earp, B. D., Feist, M. I., Ferrell, J. D., Field, J. G., Fox, N. W., Friesen, A., Gomes, C., Gonzalez-Marquez, M., Grange, J. A., Grieve, A. P., Guggenberger, R., Grist, J., van Harmelen, A.-L., Hasselman, F., Hochard, K. D., Hoffarth, M. R., Holmes, N. P., Ingre, M., Isager, P. M., Isotalus, H. K., Johansson, C., Juszczyk, K., Kenny, D. A., Khalil, A. A., Konat, B., Lao, J., Larsen, E. G., Lodder, G. M. A., Lukavský, J., Madan, C. R., Manheim, D., Martin, S. R., Martin, A. E., Mayo, D. G., McCarthy, R. J., McConway, K., McFarland, C., Nio, A. Q. X., Nilsonne, G., de Oliveira, C. L., de Xivry, J.-J. O., Parsons, S., Pfuhl, G., Quinn, K. A., Sakon, J. J., Saribay, S. A., Schneider, I. K., Selvaraju, M., Sjoerds, Z., Smith, S. G., Smits, T., Spies, J. R., Sreekumar, V., Steltenpohl, C. N., Stenhouse, N., Świątkowski, W., Vadillo, M. A., Van Assen, M. A. L. M., Williams, M. N., Williams, S. E., Williams, D. R., Yarkoni, T., Ziano, I., and Zwaan, R. A. (2018a). Justify your alpha. *Nature Human Behaviour*, 2(3):168–171.

Lakens, D. and 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. Publisher: SAGE Publications Inc.

Lakens, D., Scheel, A. M., and Isager, P. M. (2018b). Equivalence Testing for Psychological Research: A Tutorial. *Advances in Methods and Practices in Psychological Science*, 1(2):259–269.

Landreth, A. and Silva, A. J. (2013). The need for research maps to navigate published work and inform experiment planning. *Neuron*, 79(3):411–415.

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

LeBel, E. P., McCarthy, R. J., Earp, B. D., Elson, M., and 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.

LeBel, E. P. and 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.

Lenth, R. V. (2001). Some Practical Guidelines for Effective Sample Size Determination. *The American Statistician*, 55(3):187–193. Publisher: Taylor & Francis _eprint: https://doi.org/10.1198/000313001317098149.

Lewandowsky, S. and Oberauer, K. (2020). Low replicability can support robust and efficient science. *Nature Communications*, 11(1):358.

Li, W., Mai, X., and Liu, C. (2014). The default mode network and social understanding of others: what do brain connectivity studies tell us. *Frontiers in Human Neuroscience*, 0. Publisher: Frontiers.

Lindsay, D. S. (2015). Replication in Psychological Science. *Psychological Science*, 26(12):1827–1832.

Lindsay, D. S. (2017). Preregistered Direct Replications in Psychological Science. *Psychological Science*, 28(9):1191–1192.

Lodder, P., Ong, H. H., Grasman, R. P. P. P., and Wicherts, J. M. (2019). A comprehensive meta-analysis of money priming. *Journal of Experimental Psychology: General*, 148(4):688–712. Place: US Publisher: American Psychological Association.

Lundberg, I., Johnson, R., and Stewart, B. (2020). What is Your Estimand? Defining the Target Quantity Connects Statistical Evidence to Theory. Technical report, SocArXiv. type: article.

Machery, E. (2020). What Is a Replication? *Philosophy of Science*, 87(4):545–567. Publisher: The University of Chicago Press.

Mackey, A. and Porte, G. (2012). Why (or why not), when and how to replicate research. *Replication research in applied linguistics*, 2146. Publisher: Cambridge University Press Cambridge, UK.

MacLeod, C. M. (1991). Half a century of research on the Stroop effect: An integrative review. *Psychological Bulletin*, 109(2):163–203.

Maier, M. and Lakens, D. (2021). Justify Your Alpha: A Primer on Two Practical Approaches. Technical report, PsyArXiv. type: article.

Makel, M. C., Plucker, J. A., and Hegarty, B. (2012). Replications in Psychology Research: How Often Do They Really Occur? *Perspectives on Psychological Science*, 7(6):537–542.

Martin, B. R. (2011). The Research Excellence Framework and the 'impact agenda': are we creating a Frankenstein monster? *Research Evaluation*, 20(3):247–254.

Martin, G. N. and Clarke, R. M. (2017). Are Psychology Journals Anti-replication? A Snapshot of Editorial Practices. *Frontiers in Psychology*, 8.

Martín-Martín, A., Orduna-Malea, E., Thelwall, M., and 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.

Matiasz, N. J., Wood, J., Doshi, P., Speier, W., Beckemeyer, B., Wang, W., Hsu, W., and Silva, A. J. (2018). ResearchMaps.org for integrating and planning research. *PloS One*, 13(5):e0195271.

Matiasz, N. J., Wood, J., Wang, W., Silva, A. J., and Hsu, W. (2017). Computer-Aided Experiment Planning toward Causal Discovery in Neuroscience. *Frontiers in Neuroinformatics*, 11(February):1–8.

Maxwell, S. E. and Delaney, H. D. (2004). *Designing experiments and analyzing data: a model comparison perspective*. Lawrence Erlbaum Associates, Mahwah, N.J, 2nd ed edition.

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

Mayo, D. G. (2018). *Statistical Inference as Severe Testing: How to Get Beyond the Statistics Wars*. Cambridge University Press, 1 edition.

McElreath, R. and Smaldino, P. E. (2015). Replication, Communication, and the Population Dynamics of Scientific Discovery. *PLOS ONE*, 10(8):e0136088. Publisher: Public Library of Science.

McKenna, H. P. (1994). The Delphi technique: a worthwhile research approach for nursing? *Journal of Advanced Nursing*, 19(6):1221–1225. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1365-2648.1994.tb01207.x.

Meadows, D. H. (2008). *Thinking in Systems: A Primer*. Chelsea Green Publishing. Google-Books-ID: CpbLAgAAQBAJ.

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology*, 46(4):806–834.

Meehl, P. E. (1990). Appraising and Amending Theories: The Strategy of Lakatosian Defense and Two Principles that Warrant It. *Psychological Inquiry*, 1(2):108–141.

Meehl, P. E. (1992). Cliometric metatheory: The actuarial approach to empirical, history-based philosophy of science. *Psychological Reports*, 71(2):339–467. Place: US Publisher: Psychological Reports.

Meehl, P. E. (1997). The Problem Is Epistemology, Not Statistics: Replace Significance Tests by Confidence Intervals and Quantify Accuracy of Risky Numerical Predictions. In Harlow, L. L., Mulaik, S. A., and Steiger, J. H., editors, *What If There Were No Significance Tests?*

Michell, J. (2003). Epistemology of Measurement: The Relevance of its History for Quantification in the Social Sciences. *Social Science Information*, 42(4):515–534.

Moonesinghe, R., Khoury, M. J., and Janssens, A. C. J. W. (2007). Most Published Research Findings Are False—But a Little Replication Goes a Long Way. *PLoS Medicine*, 4(2):e28.

Morey, R. and Lakens, D. (2016). Why Most Of Psychology Is Statistically Unfalsifiable. Publisher: Zenodo.

Moshontz, H., Campbell, L., Ebersole, C. R., IJzerman, H., Urry, H. L., Forscher, P. S., Grahe, J. E., McCarthy, R. J., Musser, E. D., Protzko, J., Flake, J. K., Forero, D. A., Janssen, S. M., Keene, J., Aczel, B., Ansari, D., Antfolk, J., Baskin, E., Batres, C., Lucia, M., Brick, C., Castille, C. M., Chandel, P., Chopik, W. J., Clarance, D., Corker, K. S., Dixson, B., Dranseika, V., Dunham, Y., Evans, T. R., Fiedler, S., Fu, C. H., Gardiner, G., Garrison, S. M., Gill, T., Hahn, A., Jaeger, B., Kačmár, P., Gwenaël, K., Kanske, P., Kekecs, Z., Kline, M., Koehn, M. A., Kujur, P., Levitan, C., Miller, J. K., Okan, C., Olsen, J., Oviedo-Trespalacios, O., gru, A. A. . U., Pande, B., Parganiha, A., Parveen, N., Pfuhl, G., Pradhan, S., Ropovik, I., Rule, N., Saunders, B., Schei, V., Schmidt, K., Singh, M. M., Sirota, M., Solas, S. ., Steltenpohl, C. N., Stieger, S., Storage, D., Sullivan, G. B., Szabelska, A., Tamnes, C. K., Vadillo, M. A., vanpaemel, w., Vergauwe, E., Verschoor, M., Vianello, M., Voracek, M., Williams, G. P., Wilson, J. P., Zickfeld, J. H., Awlia, D., Chatard, A., Fernandez, A. M., Kapucu, A., Mensink, M. C., and Chartier, C. R. (2018). Psychological Science Accelerator: Advancing Psychology through a Distributed Collaborative Network. *Advances in Methods and Practices in Psychological Science*.

Mueller-Langer, F., Fecher, B., Harhoff, D., and Wagner, G. G. (2019). Replication studies in economics—How many and which papers are chosen for replication, and why? *Research Policy*, 48(1):62–83.

Munafò, M. R., Nosek, B. A., Bishop, D. V. M., Button, K. S., Chambers, C. D., Percie du Sert, N., Simonsohn, U., Wagenmakers, E.-J., Ware, J. J., and Ioannidis, J. P. A. (2017). A manifesto for reproducible science. *Nature Human Behaviour*, 1(1):1–9. Bandiera_abtest: a Cc_license_type: cc_by Cg_type: Nature Research Journals Number: 1 Primary_atype: Reviews Publisher: Nature Publishing Group Subject_term: Social sciences Subject_term_id: social-sciences.

Muradchanian, J., Hoekstra, R., Kiers, H., and van Ravenzwaaij, D. (2021). How best to quantify replication success? A simulation study on the comparison of replication success metrics. *Royal Society Open Science*, 8(5):201697. Publisher: Royal Society.

Murphy, J., Mesquida, C., Caldwell, A. R., Earp, B. D., and Warne, J. (2021). Selection Protocol for Replication in Sports and Exercise Science. Technical report, OSF Preprints. type: article.

Muschelli, J. (2019). rscopus: Scopus Database 'API' Interface.

Muthukrishna, M. and Henrich, J. (2019). A problem in theory. *Nature Human Behaviour*, 3(3):221–229.

Navarro, D. J. (2021). If Mathematical Psychology Did Not Exist We Might Need to Invent It: A Comment on Theory Building in Psychology. *Perspectives on Psychological Science*, page 1745691620974769. Publisher: SAGE Publications Inc.

Nicholson, J. M., Mordaunt, M., Lopez, P., Uppala, A., Rosati, D., Rodrigues, N. P., Grabitz, P., and Rife, S. C. (2021). scite: a smart citation index that displays the context of citations and classifies their intent using deep learning. *bioRxiv*, page 2021.03.15.435418. Publisher: Cold Spring Harbor Laboratory Section: New Results.

Noah, T., Schul, Y., and Mayo, R. (2018). When both the original study and its failed replication are correct: Feeling observed eliminates the facial-feedback effect. *Journal of Personality and Social Psychology*, 114(5):657–664.

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., Buck, S., Chambers, C., Chin, G., Christensen, G., Contestabile, M., Dafoe, A., Eich, E., Freese, J., Glennerster, R., Goroff, D. L., Green, D., Hesse, B. W., Humphreys, M., Ishiyama, J., Karlan, D., Kraut, A., Lupia, A., Mabry, P. L., Madon, T., malhotra, n., Mayo-Wilson, E., McNutt, M., Miguel, E., Paluck, E. L., Simonsohn, U., Soderberg, C. K., Spellman, B., Turitto, J., VandenBos, G., Vazire, S., Wagenmakers, E.-J., Wilson, R. K., Yarkoni, T., Stodden, V., and DeHaven, A. C. (2016). Transparency and Openness Promotion (TOP) Guidelines. preprint, Open Science Framework.

Nosek, B. A. and Errington, T. M. (2020). What is replication? *PLOS Biology*, 18(3):e3000691. Publisher: Public Library of Science.

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

Nuijten, M. B., van Assen, M. A. L. M., Hartgerink, C. H. J., Epskamp, S., and Wicherts, J. M. (2017). The Validity of the Tool "statcheck" in Discovering Statistical Reporting Inconsistencies. preprint, PsyArXiv.

NWO (2019). Replication Studies. Retrieved August 28, 2019, from https://bit.ly/3EJWnX8.

Oberauer, K. and Lewandowsky, S. (2019). Addressing the theory crisis in psychology. *Psychonomic Bulletin & Review*, 26(5):1596–1618.

of Arts and Sciences, R. N. A. (2018). Replication studies – Improving reproducibility in the empirical sciences,. Technical report.

Oh, H. C. and 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.

Orben, A. and Lakens, D. (2020). Crud (Re)Defined. *Advances in Methods and Practices in Psychological Science*, 3(2):238–247.

Paris, B., IJzerman, H., and Forscher, P. S. (2020). PSA 2020-2021 study capacity report. preprint, PsyArXiv.

Park, D. C., Smith, A. D., Lautenschlager, G., Earles, J. L., Frieske, D., Zwahr, M., and Gaines, C. L. (1996). Mediators of long-term memory performance across the life span. *Psychology and Aging*, 11(4):621–637.

Parolo, P. D. B., Pan, R. K., Ghosh, R., Huberman, B. A., Kaski, K., and Fortunato, S. (2015). Attention decay in science. *Journal of Informetrics*, 9(4):734–745.

Pashler, H. and 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.

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

Pearl, J., Glymour, M., and Jewell, N. P. (2016). *Causal inference in statistics: a primer*. John Wiley & Sons Ltd, Chichester, West Sussex, UK.

Peirce, C. S. (1967). Note on the Theory of the Economy of Research. *Operations Research*, 15(4):643–648. Publisher: INFORMS.

Peterson, D. and Panofsky, A. (2021). Arguments against efficiency in science. *Social Science Information*, 60(3):350–355. Publisher: SAGE Publications Ltd.

Pittelkow, M.-M., Field, S. M., Isager, P. M., Veer, A. v. t., and Ravenzwaaij, D. v. (2021). The Process of Replication Target Selection: What to Consider? Publisher: OSF.

Pittelkow, M.-M., Hoekstra, R., Karsten, J., and van Ravenzwaaij, D. (2020). Replication Crisis in Clinical Psychology: A Bayesian and Qualitative Re-evaluation. preprint, PsyArXiv.

Plucker, J. A. and Makel, M. C. (2021). Replication is important for educational psychology: Recent developments and key issues. *Educational Psychologist*, 0(0):1–11. Publisher: Routledge _eprint: https://doi.org/10.1080/00461520.2021.1895796.

Poldrack, R. A., Baker, C. I., Durnez, J., Gorgolewski, K. J., Matthews, P. M., Munafò, M. R., Nichols, T. E., Poline, J.-B., Vul, E., and Yarkoni, T. (2017). Scanning the horizon: towards transparent and reproducible neuroimaging research. *Nature Reviews Neuroscience*, 18(2):115–126.

Psychfiledrawer (2014). Top-20 List of Studies Users Would Like to see Replicated!

Purkayastha, A., Palmaro, E., Falk-Krzesinski, H. J., and 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.

Radicchi, F., Weissman, A., and Bollen, J. (2017). Quantifying perceived impact of scientific publications. *Journal of Informetrics*, 11(3):704–712.

Raiffa, H., Raiffa, F. P. R. P. o. M. E. E. H., and Schlaifer, R. (1961). *Applied Statistical Decision Theory*. Division of Research, Graduate School of Business Adminitration, Harvard University. Google-Books-ID: wPBLAAAAMAAJ.

Raiffa, H. and Schlaifer, R. (1974). *Applied statistical decision theory*. Studies in managerial economics. Div. of Research, Graduate School of Business Administration, Harvard Univ, Boston, 6. print edition. OCLC: 256126576.

Ram, K. (2017). rAltmetric: Retrieves Altmerics Data for Any Published Paper from 'Altmetric.com'.

Ranehill, E., Dreber, A., Johannesson, M., Leiberg, S., Sul, S., and 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.

Revelle, W. (2021). psych: Procedures for Psychological, Psychometric, and Personality Research.

Ritchie, S. J., Wiseman, R., and French, C. C. (2012). Failing the Future: Three Unsuccessful Attempts to Replicate Bem's 'Retroactive Facilitation of Recall' Effect. *PLoS ONE*, 7(3):e33423.

Robinaugh, D. J., Haslbeck, J. M. B., Ryan, O., Fried, E. I., and Waldorp, L. J. (2021). Invisible Hands and Fine Calipers: A Call to Use Formal Theory as a Toolkit for Theory Construction. *Perspectives on Psychological Science*, page 1745691620974697. Publisher: SAGE Publications Inc.

Rohrer, J. M. (2018). Thinking Clearly About Correlations and Causation: Graphical Causal Models for Observational Data. *Advances in Methods and Practices in Psychological Science*, 1(1):27–42.

Rouder, J. N. and 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.

Royal Society Open Science (2020). Replication studies.

Sale, C. and Mellor, D. (2018). A call for replication studies in Nutrition and Health. *Nutrition and Health*, 24(4):201–201. Publisher: SAGE Publications Ltd.

Scheel, A. M., Schijen, M., and 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.

Scott, S. (2013). Pre-registration would put science in chains.

Serra-Garcia, M. and Gneezy, U. (2021). Nonreplicable publications are cited more than replicable ones. *Science Advances*, 7(21):eabd1705. Publisher: American Association for the Advancement of Science Section: Research Article.

Sharpe, D. (1997). Of apples and oranges, file drawers and garbage: Why validity issues in meta-analysis will not go away. *Clinical Psychology Review*, 17(8):881–901.

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

Silva, A. J. and 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.

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

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

Smaldino, P. (2016). Models Are Stupid, and We Need More of Them.

Smaldino, P. (2019). Better methods can't make up for mediocre theory. *Nature*, 575(7781):9–9. Number: 7781 Publisher: Nature Publishing Group.

Smaldino, P. E., Turner, M. A., and Contreras Kallens, P. A. (2019). Open science and modified funding lotteries can impede the natural selection of bad science. *Royal Society Open Science*, 6(7):190194.

Soderberg, C. K., Errington, T., Schiavone, S. R., Bottesini, J. G., Thorn, F. S., Vazire, S., Esterling, K. M., and Nosek, B. A. (2020). Initial Evidence of Research Quality of Registered Reports Compared to the Traditional Publishing Model. Technical report, MetaArXiv. type: article.

Strathern, M. (1997). 'Improving ratings': audit in the British University system. *European Review*, 5(3):305–321. Publisher: Cambridge University Press.

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

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

Sullivan, G. M. and Feinn, R. (2012). Using Effect Size—or Why the P Value Is Not Enough. *Journal of Graduate Medical Education*, 4(3):279–282.

Szpunar, K. K., St Jacques, P. L., Robbins, C. A., Wig, G. S., and Schacter, D. L. (2014). Repetition-related reductions in neural activity reveal component processes of mental simulation. *Social Cognitive and Affective Neuroscience*, 9(5):712–722.

Szucs, D. and Ioannidis, J. P. A. (2017). Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. *PLOS Biology*, 15(3):e2000797. Publisher: Public Library of Science.

Tamir, D. I. and Mitchell, J. P. (2012). Disclosing information about the self is intrinsically rewarding. *Proceedings of the National Academy of Sciences*, 109(21):8038–8043.

Tay, A., Kramer, B., and Waltman, L. (2020). Why openly available abstracts are important - overview of the current state of affairs.

Torgerson, D., Ryan, M., and Ratcliffe, J. (1995). Economics in sample size determination for clinical trials. *QJM: An International Journal of Medicine*, 88(7):517–521.

Tunç, D. U. and Tunç, M. N. (2020). A Falsificationist Treatment of Auxiliary Hypotheses in Social and Behavioral Sciences: Systematic Replications Framework. Technical report, PsyArXiv. type: article.

Tversky, A. and Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin*, 76(2):105–110.

van Eck, N. J. and Waltman, L. (2010). Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics*, 2(84):523–538.

van Eck, N. J. and Waltman, L. (2014). Visualizing Bibliometric Networks. In Ding, Y., Rousseau, R., and Wolfram, D., editors, *Measuring Scholarly Impact: Methods and Practice*, pages 285–320. Springer International Publishing, Cham.

van Eck, N. J., Waltman, L., van Raan, A. F. J., Klautz, R. J. M., and Peul, W. C. (2013). Citation Analysis May Severely Underestimate the Impact of Clinical Research as Compared to Basic Research. *PLoS ONE*, 8(4):e62395.

Vargha, A. and 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. Publisher: American Educational Research Association.

Verhaeghen, P. and De Meersman, L. (1998). Aging and the Stroop effect: A meta-analysis. *Psychology and Aging*, 13(1):120–126. Place: US Publisher: American Psychological Association.

Wagenmakers, E.-J., Beek, T., Dijkhoff, L., Gronau, Q. F., Acosta, A., Adams, R. B., Albohn, D. N., Allard, E. S., Benning, S. D., Blouin-Hudon, E.-M., Bulnes, L. C., Caldwell, T. L., Calin-Jageman, R. J., Capaldi, C. A., Carfagno, N. S., Chasten, K. T., Cleeremans, A., Connell, L., DeCicco, J. M., Dijkstra, K., Fischer, A. H., Foroni, F., Hess, U., Holmes, K. J., Jones, J. L. H., Klein, O., Koch, C., Korb, S., Lewinski, P., Liao, J. D., Lund, S., Lupianez, J., Lynott, D., Nance, C. N., Oosterwijk, S., Ozdoğru, A. A., Pacheco-Unguetti, A. P., Pearson, B., Powis, C., Riding, S., Roberts, T.-A., Rumiati, R. I., Senden, M., Shea-Shumsky, N. B., Sobocko, K., Soto, J. A., Steiner, T. G., Talarico, J. M., van Allen, Z. M., Vandekerckhove, M., Wainwright, B., Wayand, J. F., Zeelenberg, R., Zetzer, E. E., and Zwaan, R. A. (2016). Registered Replication Report: Strack, Martin, & Stepper (1988). *Perspectives on Psychological Science*, 11(6):917–928.

Wagge, J. R., Brandt, M. J., Lazarevic, L. B., Legate, N., Christopherson, C., Wiggins, B., and Grahe, J. E. (2019). Publishing Research With Undergraduate Students via Replication Work: The Collaborative Replications and Education Project. *Frontiers in Psychology*, 10. Publisher: Frontiers.

Waltman, L. and van Eck, N. J. (2013). A systematic empirical comparison of different approaches for normalizing citation impact indicators. *Journal of Informetrics*, 7(4):833–849.

Waltman, L. and van Eck, N. J. (2019). Field Normalization of Scientometric Indicators. In Glänzel, W., Moed, H. F., Schmoch, U., and Thelwall, M., editors, *Springer Handbook of Science and Technology Indicators*, Springer Handbooks, pages 281–300. Springer International Publishing, Cham.

Waltman, L., van Eck, N. J., van Leeuwen, T. N., Visser, M. S., and van Raan, A. F. J. (2011). Towards a new crown indicator: Some theoretical considerations. *Journal of Informetrics*, 5(1):37–47.

Wang, D., Song, C., and Barabási, A.-L. (2013). Quantifying Long-Term Scientific Impact. *Science*, 342(6154):127–132. Publisher: American Association for the Advancement of Science Section: Report.

Wang, M., Ren, J., Li, S., and Chen, G. (2019). Quantifying a Paper's Academic Impact by Distinguishing the Unequal Intensities and Contributions of Citations. *IEEE Access*, 7:96198–96214. Conference Name: IEEE Access.

Wen, H., Wang, H.-Y., He, X., and Wu, C.-I. (2018). On the low reproducibility of cancer studies. *National science review*, 5(5):619–624.

Westfall, J. (2016). Designing multi-lab replication projects: Number of labs matters more than number of participants. Library Catalog: jakewestfall.org.

Westfall, J., Kenny, D. A., and 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.

Whitehead, S. J. and Ali, S. (2010). Health outcomes in economic evaluation: the QALY and utilities. *British Medical Bulletin*, 96(1):5–21.

Wible, J. R. (1994). Charles Sanders Peirce's economy of research. *Journal of Economic Methodology*, 1(1):135–160.

Wilson, E. C. F. (2015). A Practical Guide to Value of Information Analysis. *PharmacoEconomics*, 33(2):105–121.

Yang, Y., Youyou, W., and 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.

Yarkoni, T. (2020). The generalizability crisis. *Behavioral and Brain Sciences*, pages 1–37.

Yuan, S., Tang, J., Zhang, Y., Wang, Y., and Xiao, T. (2018). Modeling and Predicting Citation Count via Recurrent Neural Network with Long Short-Term Memory. *arXiv:1811.02129 [physics]*. arXiv: 1811.02129.

Zwaan, R. A., Etz, A., Lucas, R. E., and Donnellan, M. B. (2018). Making replication mainstream. *Behavioral and Brain Sciences*, 41.