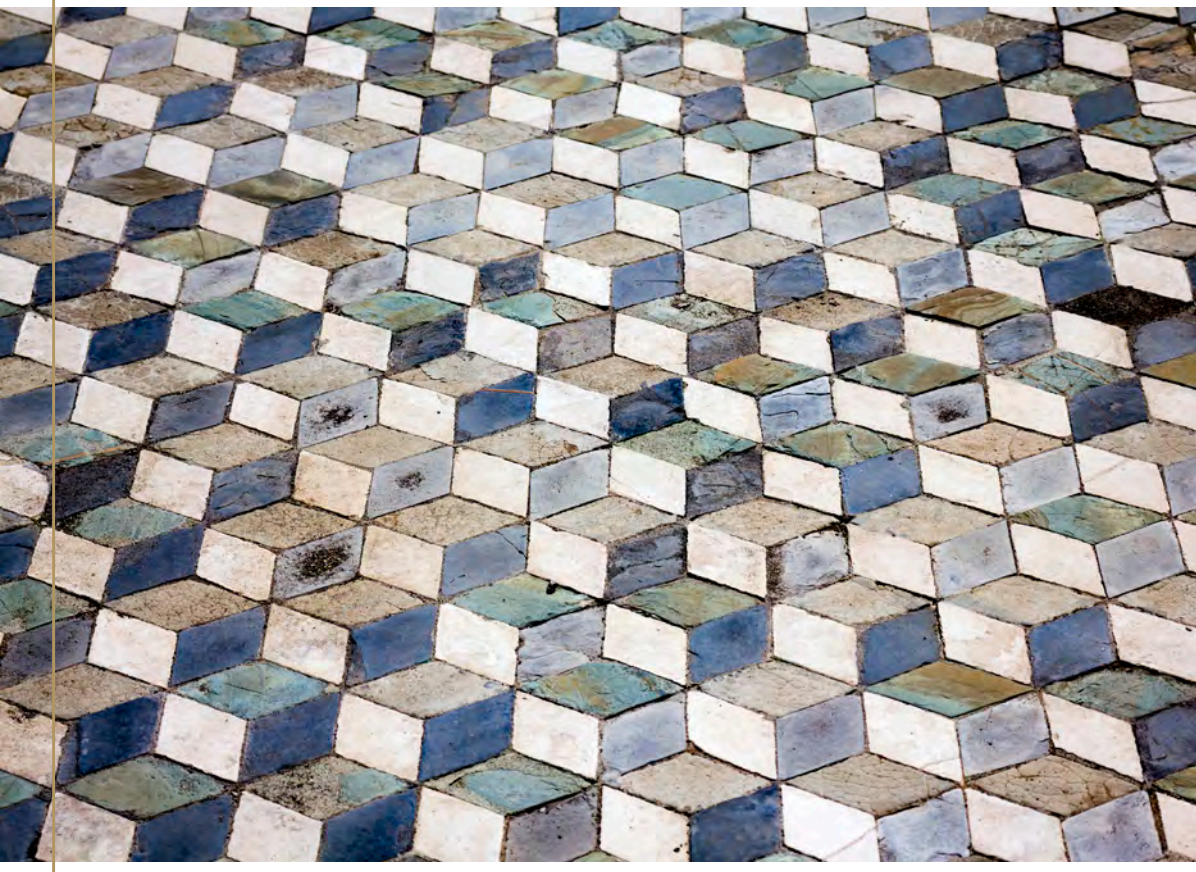
KONINKLIJKE NEDERLANDSE
AKADEMIE VAN WETENSCHAPPEN

REPLICATION STUDIES

IMPROVING REPRODUCIBILITY IN THE EMPIRICAL SCIENCES



ADVISORY REPORT

Royal Netherlands Academy of Arts and Sciences
PO Box 19121, NL-1000 GC Amsterdam
T +31 (0)20 551 0700
knaw@knaw.nl
www.knaw.nl

PDF available at www.knaw.nl
Preferred citation: KNAW (2018). *Replication studies – Improving reproducibility in the empirical sciences*, Amsterdam, KNAW.

Basic layout: Edenspiekermann, Amsterdam
Layout and index: Ellen Bouma, Alkmaar
Translation: Balance Amsterdam/Maastricht

Illustration cover: Istock by Getty Images/George-Standen, *Pompeii floor.*

ISBN: 978-90-6984-720-7

REPLICATION STUDIES

IMPROVING REPRODUCIBILITY IN THE EMPIRICAL SCIENCES

Royal Netherlands Academy of Arts and Sciences
January 2018

FOREWORD

Scientific knowledge can only grow if researchers can trust the results of earlier studies. Researchers should therefore conduct studies using systematic, rigorous methods and report in detail on how they have achieved their results. This allows other researchers to appraise results critically and to repeat a study to see whether they can reproduce its results. Being able to reproduce results is important, not only because it aids scientific progress, but also because non-reproducible results waste resources, can harm individuals and society, and may erode public trust in science.

Over the past few decades, science has become a global, multi-billion-euro enterprise with millions of researchers and other staff, churning out millions of articles a year. In addition, science is influenced by the (sometimes conflicting) interests of research institutes, funding agencies, government bodies, society, popular media, publishers, and researchers themselves. These developments have increased the pressure to present novel and important findings as quickly as possible. Meanwhile, individual researchers need to compete heavily for funding, employment, and publishing opportunities.

The pressure on research can mount to such heights that insufficient attention may go to the rigorous conduct of research and to repeating studies when necessary. Fortunately, outright fraud is rare, but questionable research practices and signs of work stress are not. Moreover, in recent years careful repetitions of earlier studies – replication studies – have been unable to reproduce many important results in various scientific disciplines. Some researchers now even speak of a 'reproducibility crisis'.

The Royal Netherlands Academy of Arts and Sciences aims to play an active role in critically reflecting on the proper conduct of science, and it is therefore issuing this

advisory report. The report shows that disciplines need better data on the degree of non-reproducibility and its causes, and that there are many promising strategies to improve reproducibility (such as higher reporting standards and proper incentives for researchers). The report, however, also concludes that replication studies should be conducted more frequently and systematically, and that this requires a joint effort from funding agencies, scientific journals, institutions and researchers.

As long as science continues to reflect on its results and methods in this manner, and on how scientific practice is organised, I am confident that we can keep contributing efficiently to society while earning society's trust in our findings.

José van Dijck,
President of the Royal Netherlands Academy of Arts and Sciences

SUMMARY

1. Introduction

In the empirical sciences, research studies aim to generate and test hypotheses through systematic observation and experimentation. Knowledge accumulates by testing increasingly specific hypotheses, building on existing results. Such scientific progress requires studies to be conducted rigorously, so that when they are repeated their results will be reasonably similar. Over the past few years, however, several systematic series of replication studies have been unable to reproduce many important results, even when applying lenient definitions of reproducibility. This has led to a debate within the scientific community about the way science is currently being conducted and the role of replication studies. This report analyses the causes of non-reproducibility, assesses the desirability of replication studies and offers recommendations for improving reproducibility and for conducting replication studies.

2. Concerns about and strategies to improve reproducibility

A replication study is a study that tries to repeat an earlier study, using similar methods and conducted under similar circumstances, to determine the reproducibility of the earlier study's results. Reproducibility is defined as the extent to which the results of a replication study agree with those of the earlier study. If the results of both studies agree, then the results of the earlier study are considered to have been 'reproduced', thereby increasing the likelihood that the results are valid. However, the results of replication studies in various empirical scientific disciplines are often not in agreement with those of the original studies, indicating that disciplines as a whole may be subject to a substantial degree of non-reproducibility. Studies with non-reproducible results can jeopardise scientific progress, waste resources, harm individuals and society, and erode public trust in science. There are various factors – for example, related to study methods, study reporting and the underlying incentive system for researchers – that can lead to non-reproducibility. These factors can and should be eliminated as much as possible.

SAMENVATTING

1. Inleiding

Binnen de empirische wetenschappen worden hypotheses opgesteld en getoetst door op een systematische wijze te observeren en te experimenteren. Kennis kan zich verder ontwikkelen doordat resultaten van eerder onderzoek de basis vormen voor het toetsen van steeds specifiekere hypotheses. Dit gaat ervan uit dat onderzoek gedegen wordt uitgevoerd zodat, wanneer een onderzoek later herhaald zou worden (in een zogenoemd replicatieonderzoek), de resultaten hetzelfde blijven. Gedurende de afgelopen jaren hebben echter diverse replicatieonderzoeken belangrijke resultaten niet kunnen reproduceren, zelfs niet als een ruim criterium voor gereproduceerde resultaten wordt gehanteerd. Dit heeft in de wetenschap geleid tot een discussie over de huidige wijze van wetenschapsbeoefening en de rol van replicatieonderzoek daarin. Dit rapport beschrijft waardoor niet-reproduceerbare resultaten kunnen ontstaan en in hoeverre replicatieonderzoek nodig is. Verder doet het rapport aanbevelingen over hoe reproduceerbaarheid en de uitvoering van replicatieonderzoek te verbeteren valt.

2. Zorgen over reproduceerbaarheid en strategieën om deze te verbeteren

Replicatieonderzoek probeert vast te stellen of de resultaten van eerder onderzoek reproduceerbaar zijn door het eerdere onderzoek te herhalen met vergelijkbare methodes en in vergelijkbare omstandigheden. Reproduceerbaarheid is daarbij de mate waarin de resultaten van het replicatieonderzoek overeenkomen met die van het eerdere onderzoek. Als de resultaten goed overeenkomen, zijn de resultaten ‘gereproduceerd’ en is het ook aannemelijker dat ze valide zijn. In diverse wetenschappelijke disciplines is echter gebleken dat de resultaten van veel replicatieonderzoek niet overeenkomen met die van eerder onderzoek. Dit kan betekenen dat disciplines over de hele breedte te kampen hebben met gebrekkige reproduceerbaarheid. Niet-reproduceerbaar onderzoek kan wetenschappelijke vooruitgang belemmeren, leiden tot het verspillen van onderzoeksmiddelen,

3. The desirability of replication studies

Replication studies can benefit research in two major ways. First, replication of individual studies can help to allay doubts about their results or their proper execution. This is especially important if these results have, or could have, a major impact on scientific progress or on meeting societal goals, or if incorrect results will lead to a waste of research resources. Whether a replication study is the best strategy in such cases also depends on the feasibility and costs of such a study compared to alternative strategies, such as conducting another, original study. Second, systematic series of replication studies are necessary to identify the extent to which results in a particular field are reproducible, the underlying causes of non-reproducibility, and the effectiveness of measures taken to improve reproducibility. The desirability of replication series depends on how much is already known about reproducibility and on the extent to which improving reproducibility is desirable compared to other investment targets for research funds.

4. Replication studies in practice

Replication studies appear to account for a small fraction of all published literature, but reliable data are lacking. Several disciplines (including preclinical animal research, clinical research, experimental psychology, genetic epidemiology and biochemistry) have taken important steps towards improving reproducibility and developing good replication practices. This has required a significant effort on the part of the research community with proper incentives from stakeholders such as scientific journals, institutions and funding agencies.

5. Barriers to and strategies for conducting more replication studies

Researchers currently face multiple barriers to conducting replication studies: studies are often not reported in sufficient detail, making it impossible for other researchers to design a proper replication study; researchers may be unsure about the right design for a replication study and the interpretation of its results; and researchers do not always appreciate the value of replication studies and may find it difficult to get them funded and published. Three broad strategies can help establish the right conditions for replication studies: improving information-sharing about original and replication studies; improving knowledge about when and how to perform replication studies; and creating better incentives for replication studies.

individuen en de samenleving schaden en het vertrouwen in de wetenschap aantasten. Diverse factoren kunnen ertoe leiden dat onderzoek niet reproduceerbaar is. Veel van deze factoren kunnen, en moeten, zo veel mogelijk worden uitgeschakeld om de reproduceerbaarheid van onderzoek te verbeteren.

3. De wenselijkheid van replicatieonderzoek

Replicatieonderzoek kan de wetenschap op twee manieren ten goede komen. Ten eerste kan het herhalen van een specifiek onderzoek helpen om twijfels over het resultaat of de juiste uitvoering van dat onderzoek weg te nemen. Dit is met name relevant in gevallen waarin resultaten veel impact (kunnen) hebben op wetenschappelijke vooruitgang of de maatschappij, of wanneer verspilling van onderzoeksmiddelen voorkomen kan worden. Of in dergelijke gevallen replicatieonderzoek de beste oplossing is, hangt ook af van de haalbaarheid en kosten van het replicatieonderzoek ten opzichte van andere strategieën, zoals een geheel nieuw onderzoek uitvoeren. De tweede manier waarop replicatieonderzoek de wetenschap kan helpen, is door het uitvoeren van systematische reeksen van replicatieonderzoeken. Dergelijke reeksen kunnen inzichtelijk maken in hoeverre de onderzoeksresultaten van een bepaald wetenschapsgebied in zijn algemeenheid reproduceerbaar zijn, wat hiervan de onderliggende factoren zijn, en of eventuele maatregelen om reproduceerbaarheid te bevorderen effectief zijn geweest. Of dergelijke replicatiereeksen wenselijk zijn, zal afhangen van wat al bekend is over reproduceerbaarheid en de vergelijking met andere wijzen om onderzoeksgelden te investeren.

4. Replicatieonderzoek in de praktijk

Replicatieonderzoeken lijken maar een klein deel te vormen van alle wetenschappelijke publicaties, maar goede gegevens ontbreken hierover. Diverse disciplines (zoals preklinisch proefdieronderzoek, experimentele psychologie, genetische epidemiologie en biochemie) hebben inmiddels stappen gezet om de reproduceerbaarheid te verbeteren en goede replicatiepraktijken te ontwikkelen. Dit heeft wel een aanzienlijke inspanning van de wetenschappelijke gemeenschap gevergd, waarbij belanghebbende partijen (zoals wetenschappelijke tijdschriften, instellingen en onderzoeksfinanciers) voor de nodige prikkels hebben gezorgd.

5. Belemmeringen en strategieën voor het uitvoeren van meer replicatieonderzoek

Onderzoekers lopen op dit moment aan tegen diverse belemmeringen voor het uitvoeren van replicatieonderzoek. Zo is het eerdere onderzoek vaak niet in

6. Conclusions and recommendations

The first step towards improving reproducibility is for empirical disciplines to assess the degree of non-reproducibility within their field and its underlying causes. The Academy is of the opinion that improving reproducibility, wherever it is found to be unsatisfactory, is extremely important. The Academy therefore recommends that researchers, funding agencies, journals and institutions should adopt the following measures:

- *Improve study methods.* Researchers should conduct research more rigorously by strengthening standardisation, quality control, evidence-based guidelines and checklists, validation studies and internal replications. Institutions should provide researchers with more training and support for rigorous study design, research practices that improve reproducibility, and the appropriate analysis and interpretation of the results of studies.
- *Improve study reporting.* Funding agencies and journals should require preregistration of hypothesis-testing studies. Journals should issue detailed evidence-based guidelines and checklists for reporting studies and ensure compliance with them. Journals and funding agencies should require storage of study data and methods in accessible repositories.
- *Create proper incentives.* Journals should be more open to publishing studies with null results and incentivise researchers to report such results. Rather than reward researchers mainly for 'high-impact' publications, 'innovative' studies and inflated claims, institutions, funding agencies and journals should also offer them incentives for conducting rigorous studies and producing reproducible research results.

The Academy also concludes that replication studies are a normal and essential part of science. Replication studies are an important tool for improving scientific knowledge, scientific methods and the functioning of scientific disciplines, and they should be conducted more frequently and systematically than is currently the case. Researchers should make careful assessments of the desirability of replication studies and consider the expected costs and benefits of conducting such studies compared to alternative approaches. To allow researchers to conduct replication studies when indicated, we recommend the following measures:

- *Improve information-sharing.* The above recommendations on study reporting also hold for replication studies: funding agencies should require preregistration of hypothesis-testing studies, and journals should issue reporting guidelines and require repositories for data and methods.
- *Improve know-how.* Researchers should share best replication practices and the

voldoende detail gerapporteerd, wat het moeilijk maakt om een goed vergelijkbaar replicatieonderzoek op te zetten. Daarnaast kan het voor onderzoekers onduidelijk zijn hoe ze een replicatieonderzoek moeten opzetten en de resultaten dienen te interpreteren. Ten slotte wordt replicatieonderzoek niet altijd op waarde geschat door onderzoekers zelf, en is het moeilijk om replicatieonderzoek gefinancierd en gepubliceerd te krijgen. Drie strategieën kunnen helpen de juiste randvoorwaarden voor replicatieonderzoek te creëren: het verbeteren van informatie-uitwisseling over oorspronkelijk onderzoek en replicatieonderzoek; het versterken van kennis over wanneer en hoe replicatieonderzoek moet worden uitgevoerd; en het creëren van betere prikkels voor replicatieonderzoek.

6. Conclusies en aanbevelingen

De eerste stap om de reproduceerbaarheid van onderzoek te verbeteren, is dat empirische disciplines bepalen in hoeverre hun resultaten niet-reproduceerbaar zijn en wat de oorzaken daarvan zijn. De KNAW acht het van zeer groot belang dat in gevallen waarin de reproduceerbaarheid als onvoldoende wordt bestempeld, deze wordt verbeterd. De KNAW beveelt aan dat onderzoekers, onderzoeksfinanciers, wetenschappelijke tijdschriften en instellingen de volgende maatregelen nemen om de reproduceerbaarheid te verbeteren.

- *Verbeter onderzoeksmethodes.* Onderzoekers dienen onderzoek op een meer gedegen wijze uit te voeren door meer aandacht te besteden aan standaardisatie, kwaliteitscontroles, op evidentie gebaseerde richtlijnen en checklists, validatieonderzoek en interne replicatie. Instellingen dienen onderzoekers beter op te leiden en te begeleiden met betrekking tot: het opzetten van gedegen onderzoek; onderzoekspraktijken die reproduceerbaarheid bevorderen; en het op de juiste wijze analyseren en interpreteren van resultaten.
- *Verbeter verslaglegging.* Financiers en tijdschriften dienen preregistratie van hypothese-toetsend onderzoek verplicht stellen. Tijdschriften dienen gedetailleerde richtlijnen en checklists voor het rapporteren van onderzoek uit te vaardigen en dienen ook te zorgen voor de naleving ervan. Tijdschriften en financiers dienen te eisen dat onderzoeksgegevens en -methodes worden opgeslagen in toegankelijke databanken (repository's).
- *Creëer de juiste prikkels.* Tijdschriften dienen onderzoek met negatieve bevindingen vaker te publiceren en dienen onderzoekers te stimuleren om dergelijke resultaten te rapporteren. Instellingen, financiers en tijdschriften dienen onderzoekers ruimhartiger te belonen voor het uitvoeren van gedegen onderzoek en reproduceerbare onderzoeksresultaten in plaats van het vooral belonen van 'high-impact'-publicaties, 'innovatief' onderzoek en overtrokken conclusies.

resources (e.g. methods, software, materials, samples, detailed analysis plans) required to conduct a particular replication study. Institutions should teach researchers how to design replication studies and assess reproducibility.

- *Create better incentives.* Funding agencies should increase funding for replication studies (e.g. by setting up programmes that allocate money specifically to replication studies and/or by requiring researchers to include replication activities in their individual proposals). Journals should encourage the submission of replication studies. Institutions should properly credit replication studies in career evaluations.

De KNAW concludeert verder dat replicatieonderzoek een normaal en essentieel onderdeel is van de wetenschap. Replicatieonderzoek is een belangrijk hulpmiddel bij het verbeteren van wetenschappelijke kennis en het functioneren van wetenschappelijke disciplines. Replicatieonderzoek zou vaker en systematischer moeten worden toegepast dan nu het geval is. Onderzoekers dienen hierbij wel per geval een zorgvuldige afweging te blijven maken in hoeverre replicatieonderzoek wenselijk is, op basis van de te verwachte opbrengsten en kosten ten opzichte van een andere aanpak. Om onderzoekers in staat te stellen replicatieonderzoek uit te voeren in gevallen dat dat wenselijk is, adviseert de KNAW:

- *Verbeter informatie-uitwisseling. De aanbevelingen in de voorgaande paragraaf over het verbeteren van verslaglegging zijn ook van direct belang voor replicatieonderzoek*: financiers dienen preregistratie van hypothese-toetsend onderzoek verplicht stellen, en tijdschriften dienen richtlijnen voor het rapporteren van onderzoek uit te vaardigen en te eisen dat onderzoeksgegevens en -methodes worden opgeslagen in toegankelijke databanken.
- *Versterk deskundigheid.* Onderzoekers dienen goede praktijken op het gebied van replicatieonderzoek en de benodigde onderzoeksmiddelen (bijv. methodes, software, materialen, monsters, gedetailleerde analyseplannen) uit te wisselen. Instellingen dienen onderzoekers te leren hoe zij goed replicatieonderzoek kunnen opzetten en de reproduceerbaarheid van onderzoek kunnen vaststellen.
- *Creëer prikkels.* Financiers dienen meer geld vrij te maken voor replicatieonderzoek (bijv. door programma's op te zetten speciaal voor replicatieonderzoek en/of door te vereisen dat replicatie deel uitmaakt van individuele onderzoeksvoorstellen). Tijdschriften dienen onderzoekers te stimuleren om manuscripten van replicatieonderzoek in te dienen. Instellingen dienen bij het beoordelen van personeel meer waarde te hechten aan replicatieonderzoek.

CONTENTS

1. INTRODUCTION

In the empirical sciences, knowledge accumulates through systematic observation and experimentation, by sharing research results within the scientific community through publication, and by generating and testing increasingly specific hypotheses, thereby building on existing results. Such scientific progress requires studies to be conducted rigorously, so that when they are repeated under similar conditions their results can be reproduced. Over the past few years, however, replication studies have been unable to reproduce many important research results in various scientific disciplines.[1] This has led to a debate within the scientific community about the way science is currently being conducted and how this may jeopardise scientific progress.[2] Some observers even speak of a replication or reproducibility 'crisis'.[3] The debate revolves around the question of whether there is indeed a problem and, if so, what its causes are and what role replication studies should play in its solution.

This report by the Royal Netherlands Academy of Arts and Sciences (KNAW) aims to:

- analyse the causes of non-reproducibility
- assess the desirability of replication studies
- make recommendations for preventing non-reproducibility and conducting replication studies.

Because the discussion about reproducibility has been most prominent in the medical sciences, life sciences and psychology, the analyses in this report are based mainly on experiences within these disciplines. The question is to what extent these analyses apply to other scientific disciplines. On the one hand, it could be argued that *all* scientific disciplines based on systematic observations (be it of elementary particles, mouse brains, human behaviour, historical events or poetry, and whether by

quantitative or qualitative methods) should, in principle, aim to generate reproducible results. On the other hand, the importance attributed to reproducibility, and certainly the scope of what can be replicated, may differ significantly between disciplines. For example, mouse brains and human behaviour can be manipulated experimentally, but unique phenomena such as historical events and poems cannot be repeated (although even in the case of unique phenomena, it is possible to repeat parts of scientific studies to test their reproducibility). We therefore invite empirical disciplines beyond the medical sciences, life sciences and psychology to consider the relevance of this report's conclusions and recommendations for their field.

The report is based on an analysis of scientific literature, recent reports by other advisory bodies, interviews with experts, a workshop with invited experts from the Netherlands and abroad, and deliberations within the Academy's Replication Studies Committee. The intended readership for the report includes the research community and individual researchers, but also policymakers and decision-makers at research institutions, funding agencies and scientific journals.

Chapter 2 analyses the current concerns about reproducibility, the causes of non-reproducibility and strategies to improve reproducibility. Chapter 3 describes what replication studies are and in what circumstances they are needed. Chapter 4 describes replication practices in various disciplines and what lessons can be drawn from them. Chapter 5 describes barriers to replication studies and strategies for overcoming these barriers. Chapter 6 summarises the Academy's conclusions and makes recommendations.

2. CONCERNS ABOUT AND STRATEGIES TO IMPROVE REPRODUCIBILITY

2.1 Defining reproducibility and replication studies

The definition of replication studies and the definition of reproducibility in this report are closely linked:

> *A replication study is a study that is an independent repetition of an earlier, published study, using similar methods and conducted under similar circumstances.* [4]

There are many ways in which researchers can repeat parts of a study within the study itself (for example, by performing multiple measurements or setting up control experiments) and either report the outcomes together with the other results or not publish them at all. However, we define a replication study as a study that is carried out *independently* of an earlier, previously published study and whose results are published separately. This report will refer to repetition *within* a study as 'internal replication'. Because a replication study is similar to (a 'repetition' of) an earlier study, its results can be compared directly with the results of that earlier study to determine whether the former have been 'reproduced'.[5]

> *Reproducibility concerns the extent to which the results of a replication study agree with those of the earlier study.*[6]

If the results of both studies agree, then the results of the earlier study are considered to have been 'reproduced'. For example, in medicine it is now common practice to repeat clinical trials on new medicines several times before physicians use these medicines to treat patients. These repetitive trials are usually highly similar to the

original trial (making them 'replication studies'), allowing a direct comparison of their results to assess 'reproducibility', for instance regarding the extent to which the new medicine is more effective in alleviating the symptoms of a disease than current standard therapy.

Unfortunately, there is currently no consensus in the literature regarding the definitions of terms such as replication, reproduction, replicability, reproducibility, or how they relate to other terms such as robustness and reliability. To add to the complexity, several typologies for replication studies and reproducibility have been proposed, none of which are widely accepted.[7] Nevertheless, we acknowledge that it may be meaningful to also consider the 'reproducibility/replicability of a *study*' (i.e. is it possible to repeat the methodology?) and the 'reproducibility/replicability of *inferences*' (i.e. are inferences based on the results of different studies consistent?), in addition to the 'reproducibility of study results'.[8] We also note that concepts such as 'robustness' of conclusions (i.e. to what extent do conclusions depend on minor changes in the procedures and assumptions?),[9] 'reliability' of measurements (i.e. what is the measurement error due to variation?)[10] and 'verifiability' of results (i.e. does the study documentation provide enough information on how results have been attained to assess compliance with relevant standards?)[11] are related to (and may even partly determine) the 'reproducibility' of study results. Some of these other aspects and concepts will be addressed in Chapters 2 and 3. However, to avoid unnecessary complexity we restrict ourselves in this report to the terms 'reproducibility of *results*' and 'replication *studies*'. In the following sections, we describe in more detail how reproducibility can be determined, what the causes of non-reproducibility are, and how reproducibility can be improved. Chapter 3 looks more closely at how replication studies can be conducted.

2.2 Determining the reproducibility of results

The reproducibility of results can be determined in various ways, requiring choices in three areas.[12] First, study results can be compared on different levels, ranging from the raw data that have been collected to the final outcomes of the analyses. Second, different methods can be used for the comparison, ranging from rigorous statistical methods to more qualitative approaches. One can also take a narrow view and only compare the original and the replication study, or a broader view and integrate the assessment into the total body of (*a priori*) knowledge about a subject. Results of a replication study are more informative if they are interpreted in the light of existing knowledge or integrated with the results of earlier studies, e.g. into a 'meta-analysis'.[13] Third, for studies with a continuous outcome (owing to underlying variation in the system being measured), it is important to define how much similarity there must be for a result to have been 'reproduced'.[14] Moreover, results of empirical studies are often surrounded by a measure of statistical uncertainty because of an inherent

variability in the samples studied as well as in measurements and other variables for which researchers could or did not adequately control (i.e. 'reliability' of results). How similar a result must be will depend on many factors, including the topic of research, the methods, samples and conditions, and the implications for knowledge. For many studies, there will be no expectation that their results can be reproduced exactly in a replication study; instead, reproducibility will be expressed in degrees. Exploratory studies may generally not be expected to generate reproducible results; this will, after all, be the focus of subsequent hypothesis-testing studies. Nevertheless, although it is impossible to identify a single, universal approach to determining reproducibility, the bottom line is that researchers expect the results of empirical studies to be reproducible.[15]

It is very important for study results to be reproducible because reproducibility increases the likelihood that they are true. We must note, however, that reproducible results are not necessarily true, nor are non-reproducible results necessarily false (see Box 1). Whenever results are considered not to have been 'reproduced', an explanation should be sought, e.g. in terms of background assumptions, variability in materials and methods, or biases in the way the studies have been conducted.[16] Sometimes this may lead to a retraction or correction of one of the studies, but more importantly, understanding *why* results have diverged can help to improve knowledge and the way studies are conducted.

BOX 1 REPRODUCIBILITY AND TRUTH

Even when reproducibility has been assessed, the fact that a result is considered to be reproducible does not immediately mean that any conclusions derived from this result are valid, generalisable, applicable or in any other way related to truth: reproducibility is simply a measure of the degree to which the results of similar studies agree. Only when results are consequently *interpreted* in the light of existing theories and hypotheses can they be considered to be contributing to truth and knowledge. Nevertheless, when results prove to be reproducible, we may feel more confident that any conclusions derived from these results are true. Similarly, when a result cannot be reproduced, it does not automatically follow that conclusions derived from the result are false. However, it does mean that previous conclusions and/or the methods that were used should be seriously reconsidered: a non-reproducible result requires a good explanation of what underlying factor in one (or both) of the studies has caused the results to diverge. Such an explanation can involve both the methods (including the analysis, study conditions, samples and reporting) and the interpretation of the results in the light of existing theories.

2.3 Occurrence and potential impact of non-reproducibility

Bearing in mind these caveats concerning the assessment of reproducibility, data on how often published study results can or cannot be reproduced are limited across the empirical sciences. Several replication series have been carried out in various disciplines (and subdisciplines) within the broad fields of medicine, life sciences and psychology (Table 1). This overview shows that a large proportion of previous studies could not be reproduced, but also suggests that there might be considerable variation between disciplines. Additional evidence concerning the study designs and settings being used to test hypotheses has also led to concerns about non-reproducibility. Using simulations, some authors have even claimed that in the case of many popular study designs and settings, it is more likely for studies to reach false conclusions than true ones, making it unlikely that the underlying results of these studies can be reproduced.[17] Moreover, studies have shown that the published literature in many disciplines is strongly biased towards 'significant, positive' results, decreasing the likelihood that published results represent actual effects and thus can be reproduced.[18]

However, we must not conclude that the disciplines in which reproducibility actually has been investigated are *more* susceptible to reproducibility problems than disciplines that lack data on the phenomenon; rather, the latter disciplines should be challenged to show how reproducible their results are. Moreover, disciplines and investigators who have been self-critical about their reproducibility performance should be applauded for setting high scientific standards by questioning and improving what they do, rather than be harassed when non-reproducibility is demonstrated.

In any case, the results of many replication studies in various empirical scientific disciplines are not in agreement with those of the original studies, indicating that disciplines as a whole may be subject to a substantial degree of non-reproducibility. In some fields, more than fifty percent of results could not be reproduced. However, good data on reproducibility are lacking in many disciplines and the degree of non-reproducibility may vary significantly across disciplines.

Because scientific knowledge accumulates by building on existing results, non-reproducible results will often – in time – be uncovered and corrected, or may simply be ignored. In other words, science is to some extent 'self-correcting'. However, this does not mean that non-reproducible results are unproblematic, because *the pace* of scientific progress will still be affected by a high proportion of non-reproducible results. A high degree of non-reproducibility can hamper scientific progress by steering research in the wrong direction and can also lead to a waste of scientific resources (harming scientific careers, human research participants and test animals).[27] Furthermore, when non-reproducible research results are used to infer

Table 1. Occurrence of non-reproducibility

Field Approach *Outcome*
Preclinical animal studies, general biology[19] Researchers from Bayer HealthCare attempted to validate data on potential drug targets obtained in 67 projects by copying models exactly or by adapting them to internal needs. *In 20% to 25% of cases, published data were completely in line with the results of the validation studies*
Preclinical studies, oncology[20] Amgen team attempted to reproduce the results of 53 'landmark' studies *Scientific results of 11% of the studies were confirmed*
Preclinical studies, genetics[21] Replication of data analyses provided in 18 articles on microarray-based gene expression studies *Two analyses (11%) were reproduced and six were partially reproduced or showed some discrepancies in results; ten could not be reproduced*
Preclinical animal studies, neurology[22] Retesting of nine potential drugs in rigorous animal tests that had been reported to slow down disease in a mouse model for ALS *None (0%) of the drugs was found to slow down ALS*
Observational and randomised studies in clinical medicine[23] A retrospective analysis of the most highly-cited articles reporting on observational and randomised studies on postulated effective medical interventions *The conclusions of 16% of articles were contradicted by subsequent studies; in a further 16%, effects in subsequent research were weaker than initially found*
Experimental psychology[24] Direct replications of 13 psychological phenomena across 36 independent samples *77% of phenomena were reproduced consistently*
Experimental psychology[25] The Open Science Collaboration attempted to independently replicate selected results from 100 studies in psychology *36% of the replication studies produced significant results, compared to 97% of the original studies. The mean effect sizes were halved*
Experimental economics[26] Replication of 18 published studies in economics *A significant effect in the same direction as in the original study was found for 11 replications (61%); on average, the replicated effect size was 66% of the original*

incorrect conclusions with practical implications, they can harm individuals (e.g. by inspiring ineffective patient treatments) and society (e.g. by inspiring ineffective policies) long before they are corrected at some later point in time. Although the

potential impact of non-reproducible results is obvious for some disciplines (for example in clinical research there can be a direct effect on the well-being of patients), such results can have important negative effects in all disciplines, albeit possibly on longer timelines or through less well-defined causal chains. A final way in which a high degree of non-reproducibility can affect scientific progress is by eroding the reputation of science and the public's trust in its findings, with negative long-term effects on both science and society.

However, what should be considered *too* high a degree of non-reproducibility will depend not only on the (potential) negative impact of non-reproducible results, but also on the positive impact on science and society of lowering their occurrence. Raising the bar towards 100% reproducibility is probably not the most efficient way to conduct research and may even cause scientific progress to grind to a halt because it could stifle creativity. We should also not expect the optimal degree of reproducibility to be the same in every discipline, and it may even depend on the specific type of research within a sub-discipline. In conclusion, studies with non-reproducible results carry a significant risk for both science and society and efforts should be made to achieve a high degree of reproducibility.

2.4 Causes of non-reproducibility

Many different decisions and circumstances during the research process can lead to non-reproducibility.[28] In Table 2 we summarise important causes of non-reproducibility related to study methods, study reporting and the underlying incentive system of scientific research. Researchers should try to eliminate or correct for the factors that threaten reproducibility.[29] Although fraud – fortunately – appears not to be a major cause of non-reproducibility, many researchers appear to engage in 'questionable research practices' that threaten reproducibility. For example, although a lack of rigour in statistical analysis is now considered a 'questionable research practice', it still occurs with considerable frequency.[30] It should be noted that some of the other factors that can threaten reproducibility, such as unexpected technical or human error or unknown changes in conditions or samples, are inherent to the pursuit of science and can never be eliminated entirely. These factors can therefore be considered 'legitimate' causes of non-reproducibility. It also bears noting that many of the causes of non-reproducibility are heavily influenced by the underlying incentive system; rewarding positive results, for example, will stimulate p-hacking[31]. If research careers depend on results that can be reported in high-impact journals, young researchers face heavy pressure to produce spectacular findings, for which they may have to cut corners. So, non-reproducibility is also very much a consequence of the research system, with its incentives, organisation and culture, and in many cases it would be unfair to blame the individual researcher for reporting a result that turns out to be non-reproducible.

Table 2. Causes of non-reproducibility

Area	Cause
Study methods	Weak experimental design/failure to control for biases
	Small sample size, increasing the risk of chance findings or inflated discoveries of otherwise true but weak signals (type I error)
	Low statistical power to detect the effect (type II error)
	Technical/human error in executing the study and poor quality control
	Fraud and fabrication of data
	Unknown variables that influence study outcomes
	Lack of rigour in statistical analyses
	Inappropriate statistical analyses
	Failure to conduct 'internal replication' (e.g. performing multiple measurements, cross-validation within a dataset, setting up control experiments)
Study reporting	Omitting null results (non-reporting and selective reporting) and selective analyses that make null results seem spuriously positive
	No sharing of data or methodology details
	Fishing (*post hoc* choices of dependent/independent variables based on results)
	Presenting *post hoc* hypotheses as tested hypotheses (HARKing)
	Outcome switching (a discrepancy between registered primary outcome and published primary outcome)
	Lack of adequate peer review
Incentive system	Rewarding many and 'high-impact' publications
	Rewarding positive and novel/'breakthrough' results
	Highly competitive funding systems
	Not rewarding open and reproducible practices
	Belief that a rigorous research process hampers discovery

Many scientific disciplines outside medicine, life sciences and psychology are also subject to these causes, so the problem of non-reproducibility is likely to occur in these other disciplines as well. For example, sociological studies can also be subject to human bias and a lack of rigour in methods and analysis, and chemistry studies can be subject to publication bias (a distortion of the scientific literature due to non-reporting or selective reporting of null results). Conceivably, differences in how strongly disciplines are affected by various factors may lead to differences in the occurrence of non-reproducibility[32]. Furthermore, although there are no data comparing the degree of reproducibility between different countries directly, we do know, for example, that countries differ with respect to the degree of publication bias, one of the factors related to non-reproducibility.[33] It is therefore possible that the degree of non-reproducibility also varies between countries. We acknowledge that even rigorously conducted studies may yield results that cannot be fully reproduced and that a certain

degree of non-reproducibility is inherent to the pursuit of science. Nevertheless, we maintain that avoidable factors that cause non-reproducibility can and should be addressed. As a first step, empirical disciplines should therefore assess the degree of non-reproducibility and its underlying causes.

2.5 Strategies to improve reproducibility

Because non-reproducibility has many causes, there are also many strategies that can help to prevent it. They fall into three broad categories: improving study methods, improving study reporting, and improving the incentive system (Table 3).[34]

Table 3. Strategies to improve reproducibility

Area	Strategy	Proposal
Study methods	Improving study design	• Comply with guidelines for designing and executing studies
	Improving methodological skills	• Train future and current researchers in statistics and research methods
	Methodological support and oversight	• Conduct an independent review of study protocols • Involve methodologists in studies
	Collaboration	• Multi-site studies • Team-science consortia[35]
	Standardisation	• Standardise research activities with technologies/ automation
	Quality control	• Set up control mechanisms: checklists, audits
	Study preregistration	• Facilitate study registries • Require preregistration of hypothesis-testing studies[36]
	Internal replication and validation[37]	• Conduct internal replication (e.g. repeat analyses in other datasets, repeat experiments) and other forms of internal validation (e.g. bootstrapping)
Study reporting	Improving reporting	• Use journal guidelines and checklists • Distinguish hypothesis-generating from hypothesis-testing studies • Use mechanisms for correcting articles ('versioning')[38]
	Transparency	• Issue guidelines for storing and providing access to data and methods
	Diversifying peer review	• Make peer review open • Base peer review on study quality, not on study results and inflated claims

Incentive system	Rewarding null results	• Publish 'negative' studies
	Less competition	• Work through long-term contracts and funding • Reward collaboration • Reward mentoring and training
	Rewarding open practices	• Reward sharing of methods and data[39]
	Rewarding reproducible practices	• Reward peer review[40] • Reward efforts to improve methods • Reward high-quality, rigorous research rather than publications in 'high-impact' journals
	Stimulating research on research	• Fund studies that monitor effects of proposed measures

Many of these measures are included in the Transparency and Openness Promotion (TOP) guidelines that have been signed by numerous scientific journals and organisations, and several disciplines have taken steps to put some of these measures into practice.[41] If these kinds of 'preventive' measures are implemented comprehensively and across the board, then the reproducibility of study results is likely to improve substantially, boosting our confidence in the validity of studies.

3. THE DESIRABILITY OF REPLICATION STUDIES

3.1 Goals of replication studies

In the previous chapter, we concluded that a certain degree of non-reproducibility is inherent to the pursuit of science and that replication studies can help to determine the reproducibility of results. However, replication studies may also serve various underlying goals.

First of all, replication studies can improve scientific knowledge. Replication of an individual study can help to allay doubts about specific results or about the proper execution of a previous study. Doubts can spring from a bad fit between results and existing knowledge, suspected flaws in the methods (i.e. study design) used, or worries about how adequately researchers have employed the methods. Box 2 illustrates the essential role that a replication study can play in scientific progress by recounting a hallmark case in the history of science: the case of Robert Boyle's air pump in 17th-century England.[42] A replication study may thus be especially important when results have (or could have) a major impact on scientific progress. However, reproducibility is a core element of the scientific method and we should not always wait for a major 'red flag' to justify replication. Replication studies may be called for when specific results could have a significant impact on society, regardless of whether serious doubts have been raised. For example, study results concerning the safety and effectiveness of new medicines may benefit or harm patients directly, justifying replication studies to determine their reproducibility. Replication studies are also very valuable in helping to correct specific results of earlier studies that are considered a 'breakthrough' and that attract generous funding. In such cases, replication studies may help to avoid wasting important research resources on what may eventually turn out to be a scientific dead end. Finally, taking the results of the original study and

ensuing replication studies together (especially if several replication studies have been conducted) can lead to more precise conclusions (e.g. by improving the 'signal to noise ratio'), or to more nuanced conclusions (e.g. by specifying the precise circumstances under which the original conclusions may be considered valid).[43]

BOX 2 REPLICATION IN THE HISTORY OF SCIENCE[44]

The first scientist to stress the importance of replication was the 17th-century chemist Robert Boyle. Boyle's air pump was designed to generate and study vacuum, which at the time was a very controversial concept. Indeed, distinguished philosophers such as René Descartes and Thomas Hobbes denied the very possibility that a vacuum could exist. In their 1985 book *Leviathan and the Air-Pump*, historians of science Steven Shapin and Simon Schaffer describe the debate between Boyle and Hobbes as fundamentally an argument about how scientific knowledge should be acquired. Boyle, a pioneer of the experimental method, maintained that the foundations of knowledge should be constituted by experimentally produced facts, which can be made believable to a scientific community by their reproducibility. By repeating the same experiment over and over again, Boyle argued, the certainty of fact will emerge.

Boyle's air pump, which was then a complicated and expensive apparatus to build, led to one of the first documented disputes over the reproducibility of a particular scientific phenomenon. In the 1660s, the Dutch scientist Christiaan Huygens had built his own air pump in Amsterdam, the first one outside the direct management of Boyle in England. Huygens reported that he saw 'anomalous suspension': water appeared to levitate in a glass jar inside his air pump (in fact suspended over an air bubble), but Boyle could not reproduce this phenomenon in his own pumps. It became clear that unless the phenomenon could be reproduced in England with one of the two pumps available, no one in England would accept the claims Huygens had made, or his competence in working the pump. Huygens was then invited to England, and under his personal guidance the phenomenon of anomalous suspension of water was reproduced. Following this, Huygens was elected a Foreign Member of the Royal Society.

At a somewhat higher level, replication studies can help to evaluate and improve scientific methods. By systematically comparing results from essentially similar studies – the original study and its replications – or by deliberately varying conditions and methods, researchers can assess in what way results depend on the specific conditions and methods (i.e. whether results are 'robust'), and critically appraise background assumptions (e.g. about what are and are not relevant parameters in a study).[45]

Finally, replication *series* can be used to gain insight into how a scientific discipline as a whole is functioning. This requires a systematic approach in which a representative cross-section of studies is targeted for replication. This can, for example, be done by replicating a random sample of a group of studies,[46] but also by using information

on risk factors for non-reproducibility (see Chapter 2) to target high-risk areas. The outcomes of such an endeavour can generate insights into the extent to which results in a particular field are reproducible, identify underlying factors related to the reproducibility or non-reproducibility of study results, and pinpoint possible interventions and evaluate their effects. We conclude that replication studies are a normal and essential part of science. Replication studies are an important tool for improving scientific knowledge, scientific methods, and the functioning of scientific disciplines.

3.2 Kinds of replication studies

In the previous chapter, we defined a replication study as a study that repeats an earlier study, using similar methods and conducted under similar circumstances. Although various analytical distinctions have been made between different kinds of replication studies, we have deliberately opted for a broad definition to allow for a comprehensive account of the phenomenon. However, three simple questions may help researchers to distinguish and choose between different approaches to replication: (1) Who will carry out the study? (2) What parts of the original study are being replicated? (3) How similar are those replicated parts to the original study?

1. *Who will carry out the replication study?* This may be the original team of investigators, an independent team, or a collaborative effort. Independence from the original team can help to avoid certain biases and can increase credibility,[47] but can also make it more difficult to ensure that the replication study is similar to the original study. Furthermore, some studies require very specific skills and considerable experience from the researchers that may be difficult for the replicating researchers to acquire. A collaborative effort can help to transfer skills and tacit knowledge about the methodology to new investigators and also to interpret any divergent results.
2. *What parts of the original study are being replicated?* A researcher may replicate a complete study or only specific parts. Studying new samples can help to generalise results, but can also make it harder to assess whether divergent results are caused by differences between samples or by some other aspect of the methodology. In some cases, it may even be impossible to collect new samples because the original study was based on a unique historical event. However, even then it should still be possible to replicate other parts of that study, such as the measurements or analyses. When new data are generated (e.g. by taking new measurements), questions may arise about how much data ('power') is needed to adequately compare the studies and whether the data from both studies can be pooled. And finally, researchers can choose whether or not to repeat the analysis (repeating only the analysis, based on existing data, is called 'reanalysis').[48]

3. *How similar are those replicated parts to the original study?* As mentioned earlier, a meaningful comparison between two studies requires them to have a considerable degree of similarity. Studies that follow the original study as closely as possible are also known as 'direct replications'. However, in practice it is impossible to produce a perfect copy of a study due to minor differences in samples, instruments, measurement conditions, and researcher skills.[49] When trying to reproduce a result, researchers thus need to consider how relevant certain study parameters are and how much variation is acceptable, based on background theories about the methods and the object of study. Furthermore, in some situations a perfect copy of the original study is undesirable, for example because a researcher wants to study precisely how parameters influence study outcomes and assess the generalisability or validity of the study, or because altering an aspect of the methodology is considered an obvious improvement.[50]

We conclude that proper replication studies can be designed in various ways.[51] Investigators should therefore offer sound arguments for their approach.

3.3 Criteria for when to conduct replication studies

We concluded above that replication studies can be important for scientific and societal progress and for improving scientific methods and the functioning of scientific disciplines. However, the frequency with which a discipline undertakes replication studies should depend not only on whether such studies help it meet these goals but also on whether they constitute an efficient use of research funds compared to alternatives, such as conducting innovative studies or taking preventive measures to improve reproducibility (see Chapter 2).[52] Table 4 summarises these considerations.

Currently, we lack sound data on the degree of non-reproducibility in various disciplines and the extent to which reproducibility could be improved. The data that are available indicate that reproducibility might vary considerably between disciplines (see Chapter 2.3). We recommend that disciplines systematically conduct replication series to generate better data on the occurrence of non-reproducibility and its causes and to monitor the effectiveness of measures aimed at improving non-reproducibility (see also Chapter 2.4).

Even if an overall desired rate of replication can be established within a discipline, this does not answer the question of which individual studies should be replicated. Replicating all studies (several times) or selecting studies for replication at random is not the most efficient way to contribute to scientific progress. We therefore recommend that researchers carefully assess the desirability of a replication study in individual cases based on its expected costs and benefits. In essence, this is no

different from the way the desirability of any (innovative) research proposal should be assessed. As described above, the potential benefits of a replication study include the impact it may have on the progress of scientific knowledge, on improving research methods, on meeting societal goals and on avoiding a waste of research resources. Potential costs of a replication study include the resources and time invested by researchers and the burden on human and animal test subjects (Table 4). These potential costs and benefits should consequently be weighed against other strategies, such as conducting another, innovative study.

Table 4. Assessment of the desirability of replication studies

Criteria	The desirability of a replication study:
Knowledge	• is higher when results from a previous study seem more implausible • is higher when there are more doubts about the validity of the methods or the proper execution of a previous study • is higher when its results may have a major impact on scientific knowledge • is higher when it may help improve research methods
Impact	• is higher when its results may have a major societal impact • is higher when it may help avoid wasting research resources on a scientific dead end • is higher when it may improve the functioning of a whole discipline (replication series)
Cost	• is lower when it requires more resources and time investment by researchers • is lower when it places a heavier burden on human and animal test subjects
Alternatives	• must be weighed against performing innovative studies • must be weighed against taking other measures to improve reproducibility

4. REPLICATION STUDIES IN PRACTICE

4.1 Data on replication studies

Data on replication studies are scarce but suggest that, at most, they account for only a few percent of published studies in most disciplines.[53] To complicate matters, those data also depend on how replication studies are defined and how they are identified (Box 3).

BOX 3 MEASURING THE OCCURRENCE OF REPLICATION STUDIES

Measurements indicating the occurrence of replication studies depend on how such studies are defined and how they are identified. For example, a survey of the psychological literature identified studies as replications if their authors had used the word 'replication' in their report.[54] This means that the actual number of replications may have been higher, since not all authors will have used this word.[55] Studies may, for example, be 'accidental' replications because researchers did not know about the previous study, but researchers may also have deliberately avoided the word 'replication' to improve the likelihood of publication.

Furthermore, we have defined a replication study as a study that is a repetition of an earlier study, using similar methods and conducted under similar circumstances. The necessary degree of similarity can be debated, however (see Chapter 3.2). Obviously, applying a more lenient definition means that more studies will qualify as a replication. In fact, the popular 'meta-analyses' in the medical sciences compile the results of many, more or less similar studies. Applying a lenient definition, all these studies may be considered replication studies, but even if we apply a strict definition, many of the underlying studies should qualify as such.

A final point to address when considering the occurrence of replication studies is that the number of published replication studies might not adequately reflect actual efforts to replicate studies. First of all, replication studies often fail to reproduce the results of

an earlier study; due to publication bias, they are then less likely to be published than the original study. Second, in-house replication in preparation for further studies will often not be published (see also Chapter 4.6). For example, companies that aim to develop a product based on the findings of a scientific study often first try to replicate the study before investing in further development. These kinds of replication studies tend not to be published, with the notable exception of the replication series by Amgen and Bayer (Table 1).[56]

Even allowing for these caveats, replication studies appear to account for only a small fraction of all publications in many disciplines. It appears that many replication studies are primarily intended to determine whether a specific result of a previous study can be reproduced when there are doubts about its validity. A more recent development concerns large-scale replication series meant to examine the overall reproducibility of scientific disciplines.[57] Although these efforts have attracted considerable attention in the scientific and popular media, these series seem to represent only a minor proportion of replication studies. Across disciplines, a 'typical' replication study: a) is carried out by a team of independent investigators; b) generates new data; c) follows the original protocol closely and justifies any deviations; and d) attempts to explain the resulting degree of reproducibility. To draw lessons from how various disciplines approach non-reproducibility and replication studies, we describe five disciplines: preclinical animal research, clinical research, experimental psychology, genetic epidemiology and biochemistry. However, a deeper understanding of the kinds of replication studies that are being conducted in various disciplines is hampered by a lack of data. We conclude that better data on replication studies are needed.

4.2 Reproducibility and replication practices in preclinical animal research

Researchers from Bayer Healthcare and Amgen brought the problem of reproducibility of preclinical animal studies into the spotlight by conducting a series of replication studies, concluding that less than a quarter of previous results could be reproduced.[58] In addition, analyses of the literature on neurological preclinical animal studies concluded that it suffered from serious publication bias, endangering the reproducibility of the reported results.[59] As explained above, a high degree of non-reproducible results can negatively impact the translation of results to human studies, leading to a waste of resources and unnecessary risks for human test subjects and slowing down the development of new medicines.

The causes of non-reproducibility of preclinical animal studies fall into two groups.[60] First, weak experimental designs are common: low power, a lack of randomisation,

non-blinded study execution, and non-blinded outcome assessment. Second, studies are often reported inadequately: essential design features such as blinding, randomisation and details about the test objects and conditions are not described, outcomes are selectively reported, studies with null results are not reported, and results are inappropriately interpreted as providing 'proof'.[61]

In an effort to improve the quality of preclinical animal studies and prevent non-reproducibility, the discipline has developed guidelines that recommend conducting hypothesis-testing research with a prospective, rigorous research plan (included blinding of outcome assessment, randomised allocation of interventions, power calculation to determine sample size, use of positive and negative controls, determination of a dose-response relationship), replication in different models, independent replication, and improved training in the use of statistics.[62] In addition, guidelines for the reporting of preclinical animal studies have been developed setting minimum information standards for reporting on experiments, the general elements of the study design (such as randomisation), and the relevant details of specific models.[63]

Although these kinds of guidelines have been widely endorsed,[64] they have not yet been implemented broadly by researchers,[65] suggesting that voluntary efforts are not enough and that ways to ensure or incentivise compliance with these guidelines are needed. In addition to drawing attention to the lack of reproducibility of preclinical animal studies, replication studies can also be used to monitor the progress that is being made on improving reproducibility.

4.3 Reproducibility and replication practices in clinical research

Traditionally, clinical research is divided into five subcategories: aetiology, diagnosis, prevention, treatment and prognosis. In treatment research, the best progress has been made in studies that take the form of clinical trials.

Collaboration among journal editors, professional organisations and scientists has resulted in a string of requirements for executing, analysing and reporting on clinical trials. First, a trial has to be 'preregistered' before recruitment commences, meaning that the study protocol is registered in a publicly accessible database with details about the intervention, selection of patients and outcome measures.[66] Sometimes this leads to the publication of a design paper with details on the rationale for selecting patients, the primary outcomes, execution details and interventions. This requirement can counteract non-publication by allowing third parties to scrutinise these registered trials for unpublished results. Second, during the registration process

or at the latest before the database is locked, a detailed statistical analysis plan may be published stipulating the planned analyses and statistical techniques. This can help to counteract selective analyses and selective publication by allowing for credible and transparent review of the manuscript. These requirements also make it possible to conduct a proper replication study because the information necessary to repeat the study becomes available. Major medical journals now adhere to this process and do not allow the publication of studies that have not complied with these requirements. This practice has increased the reproducibility of results considerably. Unfortunately, lower-impact journals are less stringent. Overall, only about half of randomised trials are preregistered, and the proportion is probably higher for trials of non-regulated interventions.[67] Furthermore, selective outcome and analysis reporting continues to be a major problem even for registered trials and even in major journals.[68]

Besides improving the quality of single trials, clinical research has also adopted a philosophy of 'one study is no study', meaning that a single trial is considered insufficient evidence for changing medical practice. New medical interventions should in principle be based on more than one trial. This has been adopted as a policy by the European Medicines Agency and the United States Food and Drug Administration, which requires at least two trials before granting marketing authorisation for a new medicine. It has also become common practice for independent but highly similar studies to be conducted simultaneously and then published together in the same journal issue.[69]

However, to improve the likelihood of publication, similar studies are often deliberately not presented as a replication but as an original study, either by not referring to the original study or by highlighting a relatively minor variation, for example in the methods or subjects. Although these studies will often be combined later in a meta-analysis (to address overall outcomes and outcomes in subgroups), this practice can lead to redundant replication.

Whereas in the past, clinical studies were performed and reported without any prior documentation, the current practice has changed markedly. For high-impact journals, authors must take well-defined steps to make their protocol and statistical approach available. This process has increased the credibility of clinical trials and at the same time provides opportunities for replication studies. Other subcategories of clinical research are improving their requirements and practice too. For example, preregistration of a protocol, including a detailed analysis plan, has been advocated for diagnostic and prognostic studies and in fact sometimes takes place,[70] although beyond randomised trials, the adoption rate for these practices is low in most areas of observational research.[71]

4.4 Reproducibility and replication practices in empirical psychology

The replication practices in psychological research have recently undergone a metamorphosis. Prior to 2011, replication research occurred mostly in the context of a series of experiments conducted by the same research team and the replication study would often include additional conditions. Replication studies were not preregistered. In 2011, two events conspired to shake psychologists' trust in the reproducibility of their results. First, the psychologist Diederik Stapel admitted to having committed fraud in more than fifty academic articles. The fraud was uncovered neither by the usual peer-review process nor by failed follow-up studies, but by whistle-blowers. Moreover, replication studies had also not been performed.[72] This showed the research community that prevailing research practices could easily allow non-reproducible results to go uncorrected. Second, psychologist Daryl Bem published a controversial article in a flagship journal in the field. In this article, 'Feeling the Future', Bem used the standard methodological toolkit and presented nine studies that showed that people have extra-sensory perception.[73] Rather than accepting Bem's claim, many researchers concluded that there was something amiss with the standard toolkit for psychological research.

Soon after 2011, several individual replication studies reported disappointing results[74] and Nobel laureate Daniel Kahneman issued a passionate plea for more replication research.[75] These events coalesced to produce the 'Open Science Movement', whose main goal is to increase the reproducibility of psychological research, in part by encouraging replication studies. It did not take long for the field of psychology to embrace the idea that replication studies were worthy of more attention. For example, the journal *Perspectives on Psychological Science* initiated a section for preregistered replications of influential results, the 'Registered Replication Reports';[76] the journal *Social Psychology* published a special issue featuring only replication studies;[77] the newly founded Center for Open Science initiated a replication project featuring 100 studies;[78] and a series of 'ManyLabs' replication projects was designed and carried out.[79] More importantly perhaps, several high-impact journals changed their guidelines and started to encourage replication research explicitly, e.g. by adhering to one of the levels in the Transparency and Openness Promotion guidelines.[80] It is now widely recognised that replication research can be highly influential, and that it can be published in high-impact journals.

Although discipline-wide data on the occurrence of replication studies are scarce, it is clear that because of these and other initiatives, within just a few years replication research has undergone an upswing in status and popularity. However, the current replication studies differ from those conducted pre-2011 in several important ways. Specifically, current replication studies are generally designed with high power, have

been preregistered, are not necessarily part of a longer series of experiments, and are conducted by independent labs.[81] Although it is difficult to pinpoint what the impact of these changes in replication practices will be, it appears that the research culture in psychology has become more rigorous and more cautious with its claims. At the very least, replication studies have become easier to publish now than ever before, implying that journals play a vital role in shaping a field's research culture.[82]

4.5 Reproducibility and replication practices in genetic epidemiology

In the second half of the previous century, genetic epidemiology evolved from estimating the risk of disease in families with hereditary disorders into providing statistical know-how and computational toolboxes required for linking rare early-onset forms of diseases to their underlying genetic mutations. Unfortunately, how some of these methods were used also threatened the reproducibility of results.

At that time, candidate genes were typically examined in low-powered studies, leading to a report on the 'discovery' of the gene for disease 'X', which was then more often than not followed by replication studies that failed to confirm the initial result.[83] The field responded to this situation by developing a large-scale meta-analysis that aggregated the published data of multiple genetic studies – but excluded the 'discovery' study – into a single analysis. A further development that threatened the reproducibility of results was the advent of high-throughput genotyping. This led to the introduction of genome-wide association studies, which allowed the testing – without any specific prior hypotheses – of many genes at once. Because of the number of possible correlations in the human genome, studies would simultaneously test millions of associations. However, based on prior experience with candidate gene studies, the strength of associations (for the risk of disease or other phenotypes of interest) to individual variants was expected to be small, with a very low prior probability of finding an actual association. Genome-wide association thus could have been the perfect stage for producing numerous non-reproducible results, because even when strong associations were present, the posterior probability of a true association would remain low owing to the low prior probability. Fortunately, partly due to its earlier experience of non-reproducibility, the field of genetic epidemiology self-imposed restraints to prevent these false positive results by:

1. increasing the statistical power of studies by forming ad hoc consortia in which data are pooled into a joint meta-analysis before publication
2. replicating results of studies before publication rather than after – as had been common practice in the previous decade. This 'internal replication' was done at little additional cost by collaborating with other consortia, sharing results and publishing results back-to-back

3. increasing the posterior probability of a true positive result in situations of multiple testing by increasing the significance threshold to genome-wide significance.[84]

Because high-quality imputations are possible in genetic research (replacing missing data with substituted values), it has become easier to combine data across studies and consortia than it is in other epidemiological disciplines, including nutrition, medicine and other -omics research. However, the general gist of these three measures could be adapted and applied in these and other fields of research as well.

4.6 Reproducibility and replication practices in biochemistry

Common practices to improve reproducibility of results in the life sciences and, more specifically, in biochemistry are to use internal controls, references and standards and to take into account previously published or investigated samples. Experiments aimed at testing a certain hypothesis are normally repeated several times ('internal replications') by a study team before results are published. These measures are intended to make studies more rigorous and can thus help prevent non-reproducibility (see also Chapter 2.5).

Another way reproducibility is addressed within biochemistry is when other researchers want to build on the results of a particular study. Researchers will then repeat parts of the experiments underlying the previous study in their own labs to prepare for their own follow-up studies. These experiments can consequently become the standard methodology in the field. However, the results of these preparatory experiments will often remain unpublished: if successful, they are considered not worth mentioning, and if unsuccessful, researchers often abandon the particular experiment and look for a different approach. In addition, a significant share of studies will not draw the interest of other researchers and will consequently never be repeated in follow-up studies.

Replication studies as defined in this report (a study that repeats a significant part of an earlier study by using similar methods and circumstances and which will be published separately) are very rare in biochemistry. Because systematic replication series have not yet been carried out in biochemistry, the reproducibility of the results of biochemistry studies remains unknown.

4.7 Lessons from replication practices in different disciplines

We can draw a number of conclusions from the practices of the disciplines described above.

- The practices show that replication in one form or another is becoming standard in many disciplines, but also that these practices vary considerably across disciplines and are still evolving.
- Developments in preclinical animal research and psychology show that a series of replication studies can cause a discipline to reflect on its research practices with regard to reproducibility and replication studies.
- As the case of genetic epidemiology illustrates, taking statistical measures, such as increasing the power of studies, performing meta-analysis across independent cohorts and increasing the threshold for significance testing, can help prevent non-reproducible results in the case of multiple testing.
- In preclinical animal research and in clinical research, the development of guidelines for study design and reporting has been an important starting point for improving reproducibility. However, the case of preclinical animal research also illustrates that voluntary efforts may not be enough to implement guidelines. Clinical research shows that journals can help ensure compliance with guidelines.
- Developments in clinical research illustrate that, although it is possible to improve reproducibility considerably by introducing rigorous research and reporting practices, this may require a coordinated and long-term effort on the part of multiple stakeholders and it remains challenging to change practices across the whole discipline.
- The case of clinical research also illustrates that a habit of replication can become an integral part of research practice and relevant regulations.
- Psychology has experienced the rapid development of a best practice for replication: replication studies are designed with high power, follow the original protocol closely, are preregistered, and are conducted by one or more independent labs.
- The case of psychology also illustrates how such a development can be made possible by a change in the research culture, fuelled by collaboration and open discussion among researchers in the field.
- Genetic epidemiology illustrates that collaborating within large consortia to replicate results before publication and then publish results back-to-back can become common practice.
- Both clinical research and psychology illustrate how high-quality journals can play a crucial role in improving reproducibility: by requiring that protocols and analysis plans are made transparent, allowing more critical review and replication if needed; by adopting guidelines aimed at improving reporting of (original) studies, allowing them to be closely replicated; and by stimulating publication of replication studies.

We conclude that disciplines are taking important steps towards improving reproducibility and the development of good replication practices, and that they can learn from one another's progress. However, it is also clear that these developments have required a significant effort on the part of the research community and proper incentives from stakeholders such as journals, institutions and funding agencies.

5. BARRIERS TO AND STRATEGIES FOR CONDUCTING MORE REPLICATION STUDIES

Although more replication studies may be desirable in many disciplines, researchers currently face a number of barriers to performing them.[85] Below, we analyse these barriers and propose how to create the right conditions for replication studies in three areas: (1) information-sharing about original and replication studies, (2) knowing when and how to perform replication studies, and (3) incentives for replication studies. Table 5 summarises the strategies and concrete proposals that can help create proper conditions for replication studies.

5.1 Information-sharing about original and replication studies

To perform replication studies, researchers, at the very least, need to have adequate information about the previous study and its results so that they can assess whether a replication study is needed and how it should be performed. However, the literature in many disciplines suffers from publication bias owing to non-publication or selective publication of null results.[86] Publication bias can generate an inappropriate degree of confidence in published results, which then leads to an underestimation of the need for replication studies. Doubts about the quality of the methods that have been used – or their proper execution – can also be a reason to perform a replication study (see Chapter 3.1 and 3.3). A general requirement in scientific research is that the methods must be described in sufficient detail to allow scrutiny (this is also referred to as 'method reproducibility' or 'verifiablity').[87] However, study reports do not always describe essential design features, partly owing to the length restrictions imposed by journals (Chapter 4.2). This makes it impossible for other researchers to design a replication study with the desired degree of similarity (Chapter 3.2).[88] When the

report provides only sketchy details on the methods and data of the original study, researchers will need to attain additional information from the original study team, who may not always be willing to provide it.[89]

To overcome these barriers and ensure that researchers share adequate information about original and replication studies, we recommend three complementary strategies. All three help improve the reproducibility of the results of original studies (as described in Chapter 2.5) *and* make it possible to conduct proper replication studies.

1. The first strategy is **preregistration** of hypothesis-testing studies. Preregistration means that the protocol of a study is registered in a database with relevant details about the methodology and analysis before data collection begins.[90] For clinical trials, this has already become common practice (Chapter 4.3).[91] Preregistration counteracts non-publication, selective analyses and selective publication, and helps to ensure that sufficient information is available to set up a replication study.[92] Although researchers are the main actor here, other stakeholders can provide an important push: funding agencies can make preregistration a requirement for receiving funding, institutions and research ethics committees can include preregistration in their guidelines for proper research conduct, and journals can also require preregistration.
2. A second strategy is to improve the **quality of study reporting**. Journals and the research community can together develop evidence-based reporting guidelines and checklists that provide a structured format indicating which details of studies should be reported,[93] with journals further lifting length restrictions for methods sections to allow more detailed descriptions either in the published paper or online.[94] Journals can also make compliance with these guidelines part of the review process. So far, most guidelines refer to reports on studies that have already been concluded, but a new wave of guidelines focuses increasingly on research design and researcher conduct. Streamlining of various guidelines is needed to avoid bureaucracy for researchers.
3. A third strategy is to provide more detailed information about methods and data of a study than is currently possible through journal publications. Lifting length restrictions for methods sections can already help, but storing information in **repositories** that are accessible to other researchers is a more comprehensive solution.[95] Again, whereas researchers are the main actor here, institutions, funding agencies and journals can help implement this strategy by facilitating repositories and by making the storage of study information in accessible repositories a requirement. However, such efforts should be centralised and processes streamlined so that researchers are not obliged to reinvent the wheel for each and every study.

5.2 Knowing when and how to perform replication studies

Although having adequate information about previous studies is crucial for replication studies, it is not sufficient in itself. Researchers should also know which cases require a replication study and how to perform them.

Researchers who work in an environment with a strong focus on innovation, competition and speediness may fail to appreciate that replication studies are an essential ingredient of scientific progress.[96] In addition, even when researchers appreciate the value of replication, they may find it difficult to decide when and when not to undertake a replication study. Researchers may also be unsure about how to set up a replication study in terms of who should carry it out, what parts should be replicated, how similar the replication should be to the original study, and what methods to use to interpret the results in terms of reproducibility (Chapters 3.2 and 2.2). And finally, an inadequate understanding of the limitations of particular study designs and misconceptions about the interpretation of statistical outcomes are barriers to choosing the right design for a replication study and to adequately interpreting its results.[97]

To address these barriers and ensure that researchers know when and how to perform a replication study, four – again complementary – strategies are available.

1. The first strategy is to assess the desired overall rate of replication by conducting systematic replication series (Chapter 3.1). This will not only generate **data** on reproducibility and its causes but also help to monitor the effectiveness of measures aimed at improving reproducibility.
2. The second strategy is to promote a more deliberate choice between replication and innovative studies. Decisions should be based on a careful **assessment of the expected costs and benefits** of replication studies, as compared to those of innovative studies (Chapter 3.3). This may well lead to a higher frequency of replication studies.
3. The third strategy concerns **education**. Institutions should pay more attention to the history and philosophy of science and the essential role of replications in them. Researchers should understand that science has been able to improve our knowledge of the world by means of a systematic and rigorously executed iterative process in which each step is checked carefully before going on to the next. This means that questions are addressed multiple times and that results are assessed for their reproducibility (Box 2). Institutions can also boost researchers' ability to assess study designs and interpret results, especially with respect to reproducibility.[98] Familiarising young researchers with replication studies can take place through on-the-job learning, for example by conducting a replication study as part of their PhD training.[99]

4. The fourth strategy concerns the exchange of **best practices**. Disciplines have developed various practices for ensuring reproducibility and conducting appropriate replication studies (Chapter 4.2-4.7). Disciplines can share these practices with other disciplines by publishing about them and discussing them, e.g. within scientific societies.[100]

5.3 Incentives for replication studies

Even when researchers have adequate information and know when and how to perform a replication study, colleagues, funding agencies, journals and institutions do not always offer them the right incentives to actually do so.

Researchers tend to be curious and to value creativity and freedom. This means that many researchers prefer performing original studies to following an existing protocol in a replication study.[101] Some researchers even view replications as demonstrating a lack of original ideas, or as an attack on peers.[102] Furthermore, when a replication study has not reproduced a result, heated arguments may ensue and credibility may be lost by the team that conducted the original study, by the replication team, or even by both parties.[103] Another set of disincentives for researchers is the way research is funded. Many funding agencies focus primarily on 'innovative' research, making it harder to get funding for replication studies. [104] In addition, the highly competitive nature of current funding can incentivise researchers to submit proposals for innovative studies instead of conservative approaches such as replications, in order to help sway the reviewers. Current publication practices also form disincentives for conducting replication studies. Researchers have (or expect to have) difficulty publishing the results of replication studies in 'high-impact' journals because such journals often maintain a criterion of 'originality' for accepting manuscripts[105] and want to send out unambiguous messages, not muddied discussions about replications. Furthermore, researchers may assume that the lack of reproducibility in their replication study has been caused by human error or by an uninteresting confounding variable and therefore choose not to publish their study. [106] A final disincentive for researchers is that career evaluations are based largely on having many and 'high-impact' publications and on having acquired funding, and that they are given less credit for the rigorousness of their research and for replication studies. Moreover, in their external communication institutions focus on presenting new results to attract funding, students and personnel and may even consider reporting the results of replication studies as bad publicity.

In order to encourage researchers to perform replication studies when needed, colleagues, funding agencies, journals and institutions must turn these disincentives around.

1. The research community can create a **positive culture** of replication by publicly appreciating its value. It could do this, for example, through awards, prizes and editorial comments in journals that commend replication activities, but simply making replication studies a more mainstream activity by reporting on them regularly can also help to turn the negative culture around.[107] In addition, the research community could frame an unwillingness to assist with replication studies when needed as a questionable research practice. Institutions can play a role by encouraging researchers to share any specific materials, instruments, skills and other resources required to conduct a replication study.
2. Funding agencies can provide more **funding opportunities** for replication studies.[108] Because the desirability of replication studies in disciplines is unknown (Chapter 3.3), it is not possible to specify what portion of research funding should be made available for replication. And even if the overall portion were known in a specific discipline, it would still beg the question of precisely which replication studies should be funded. However, in view of the high degree of non-reproducible results found in some disciplines, the funding of replication studies should represent a sizable proportion of research funding overall. Conducting systematic replication series within disciplines would be the first step to determining the desired level of funding for replication studies. This funding can be provided by setting up dedicated programmes that allocate money specifically to proposals for replication studies or by requiring a certain part of each research programme or individual proposal to be dedicated to replication activities. Funding agencies could also include researchers' efforts to improve reproducibility and conduct replication studies as a criterion when reviewing grant applications. Funding agencies could, at the very least, monitor what proportion of funding goes to replication studies, allowing them to reflect on whether this is the most efficient allocation of funds.
3. Journals can also play a key role in incentivising replication studies. This has become clear in psychology, where several high-impact journals changed their guidelines and started to encourage replication research explicitly (Chapter 4.4). For a start, journals can develop an editorial policy that communicates clearly that replication studies have a fair chance of being published.[109] An initiative that goes a step further is the Registered Replication Reports format, in which researchers can have a research protocol for a replication study accepted by a journal before actually collecting data, on the promise that the journal will later publish the results regardless of the outcome.[110] Other journals have taken a different approach and require independent replication *before* they will publish a result.[111] Another measure that could encourage the **publication** of replications is to have journals commit to publishing at least one high-quality replication study of a previous report from their own annals.[112] And finally, other outlets for publishing replication studies, such as blogs and posts on preprint servers, have sprung up.[113]

4. Finally, institutions can create proper incentives for replication studies by giving researchers more credit in **career evaluations** for their efforts to improve the reproducibility of research and for conducting replication studies. Data-sharing can, for example, be incentivised by routinely linking project initiators and data collectors to the data and citing their basic work. Moreover, it should be considered a significant mark of esteem if the data researchers have gathered are used worldwide in the service of science and society at large. In addition, analogous to trial registries, data sources can be registered and linked to those who established them.

*Table 5. Strategies to stimulate replication studies**

Area	Strategy	Proposal	Researchers	Funding agencies	Journals	Institutions
Information-sharing	Preregistration	• Register study protocols in a database with relevant details about the methodology and analysis • Make preregistration a requirement for funding • Include preregistration in guidelines for proper research conduct • Require preregistration for publication	X	X	X	X
	Quality of reporting	• Develop reporting guidelines and checklists for a structured reporting format with relevant details of the methodology • Lift length restrictions for methods sections • Make adherence to reporting guidelines part of the review process	X x		X X X	
	Repositories	• Store detailed information about study data and methods in repositories that are accessible to other researchers • Make storage of study information in a repository mandatory	X	X	X	X
Know-how	Data	• Conduct replication series to generate data on reproducibility and its causes and monitor the effectiveness of measures to improve reproducibility	X	x		x
	Assessment of benefits and costs	• Assess the overall desirability of replication studies in a discipline based on expected costs and benefits compared to alternatives • Make a careful assessment of the desirability of a replication study in individual cases	X X	x		
	Education	• Educate researchers in the history and philosophy of science and the essential role of replications • Train researchers in assessing (replication) study designs and interpreting results, especially with respect to reproducibility • Familiarise young researchers with replication studies through on-the-job learning, for example by conducting a replication study as part of their PhD training	x x X			X X x
	Best practices	• Share best practices with other disciplines by publishing about them and discussing them, e.g. within scientific societies	X		x	x
Incentives	Positive culture	• Commend replication activities through awards, prizes, and editorial comments in journals • Report regularly on replication attempts, making replication studies a mainstream activity • Frame willingness to help with replication studies when needed as a good research practice • Have researchers share any specific materials, instruments, skills and other resources required to conduct a replication study	X X X X	X	X X	X x
	Funding opportunities	• See that replication studies comprise a sizable fraction of funding • Set up programmes that allocate money specifically to proposals for replication studies • Require replication to be part of individual research proposals • Monitor what proportion of funding goes towards replication efforts		X X X X		
	Publication	• Amend editorial policies to state that replication studies have a fair chance of being published • Encourage replication studies explicitly, e.g. through a tailored format • Publish at least one replication study of a previous report in that same journal			X X X	
	Career evaluations	• Credit the efforts of researchers to improve reproducibility of research and conduct replication studies		x		X

** An uppercase X indicates the main stakeholder for carrying out a proposal and a lowercase x indicates the involvement of additional stakeholders*

6. CONCLUSIONS AND RECOMMENDATIONS

6.1 Improve reproducibility

The results of many replication studies in various empirical scientific disciplines do not agree with those of the original studies. This non-reproducibility of results carries a significant risk for both science and society. The degree of non-reproducibility may vary across disciplines. However, it should not be concluded that the disciplines in which reproducibility has been investigated are *more* susceptible to reproducibility problems than disciplines that lack data on the phenomenon; rather, the latter disciplines should be challenged to show how reproducible their results are. The Academy concludes that, as a first step, empirical disciplines should assess the degree of non-reproducibility in their field and its underlying causes by systematically conducting series of replication studies. The Academy is of the opinion that it is extremely important to improve reproducibility wherever it is found to be unsatisfactory. Non-reproducible results can and should be prevented as much as possible by addressing factors that cause non-reproducibility, such as suboptimal study design, execution, analysis and reporting, and the underlying incentive system. The Academy therefore recommends that researchers, funding agencies, journals and institutions should take adequate measures to improve reproducibility wherever necessary. The Academy particularly recommends the following strategies for improving methods, reporting and incentives (see Table 3, Chapter 2, for more detailed proposals):

IMPROVE STUDY METHODS

Researchers should conduct research more rigorously by strengthening standardisation, quality control, evidence-based guidelines and checklists, validation studies and internal replications. Institutions should provide researchers with more training and support for rigorous study design, research practices that improve reproducibility, and the appropriate analysis and interpretation of the results of studies.

IMPROVE STUDY REPORTING

Funding agencies and journals should require preregistration of hypothesis-testing studies. Journals should issue detailed evidence-based guidelines and checklists for reporting studies and ensure compliance with them. Journals and funding agencies should require storage of study data and methods in accessible repositories.

CREATE PROPER INCENTIVES

Journals should be more open to publishing studies with null results and incentivise researchers to report such results. Rather than reward researchers mainly for 'high-impact' publications, 'innovative' studies and inflated claims, institutions, funding agencies and journals should also offer them incentives for conducting rigorous studies and producing reproducible research results.

6.2 Conduct more replication studies

The Academy believes that if these 'preventive' measures are implemented comprehensively and across the board, the reproducibility of study results can be improved substantially. The Academy, however, is also of the opinion that replication studies will always be a normal and essential part of science. Replication studies are an important tool for improving scientific knowledge, scientific methods, and the functioning of scientific disciplines. Researchers should carefully assess the desirability of replication studies based on the likelihood of results being non-reproducible and the expected costs and benefits of conducting such studies compared to alternative approaches. When replication studies are indicated, researchers have a responsibility to perform them. Considering that replication studies appear to account for only a few percent of published studies in most disciplines, the Academy expects that overall, more replication studies are desirable. Several disciplines have shown that reproducibility can be improved considerably in time and have developed good replication practices, thanks to a coordinated effort on the part of the research community supported by proper incentives from journals, institutions and funding agencies.

The Academy recommends that researchers, funding agencies, journals and institutions adopt the following measures to ensure that researchers exchange relevant information appropriately, have the proper know-how and are given the right incentives for replication studies (see Table 5, Chapter 5 for a summary of more detailed proposals):

IMPROVE INFORMATION-SHARING

The above recommendations on study reporting also hold for replication studies: funding agencies should require preregistration of hypothesis-testing studies, and journals should issue reporting guidelines and require repositories for data and methods.

IMPROVE KNOW-HOW

Researchers should share best replication practices and the resources (e.g. methods, software, materials, samples, detailed analysis plans) required to conduct a particular replication study. Institutions should teach researchers how to design replication studies and assess reproducibility.

CREATE BETTER INCENTIVES

Funding agencies should increase funding for replication studies (e.g. by setting up programmes that allocate money specifically to replication studies and by requiring researchers to include replication in their individual proposals). Journals should encourage the submission of replication studies. Institutions should properly credit replication studies in career evaluations.

ANNEXES

1. Resolution inaugurating the Replication Research Committee

Having regard to Article 8 of the Academy's Regulations, the Board of the Royal Netherlands Academy of Arts and Sciences has decided to install the Replication Research Committee (hereinafter: 'the Committee').

Preamble

- Both the scientific community and society in general consider the reliability of research outcomes a significant point of concern.
- Replication studies can test the reproducibility of research findings and improve their reliability.
- The Academy wishes to actively promote the reliability of research and, therefore, the correct use of replication studies.
- The Committee will focus on the medical sciences and (for purposes of comparison) a number of related areas of research. Reasons to do so are the scope of the research area, its direct relationship to human lives, and society's interest in medical progress.

Article 1. The Committee's tasks

The Committee's tasks are as follows:

1. to survey the desirability of replication studies in the medical sciences, psychology and biochemistry/biophysics
2. to identify opportunities for promoting replication studies
3. to propose action that the Academy can take and make recommendations to relevant parties based on the foregoing points

4. to draft an advisory report in Dutch.[1]

In addition to the medical sciences, the advisory report will focus on two related areas of research for comparison purposes, i.e. psychology and biochemistry/biophysics, in order to identify lessons learned and best practices and to formulate research-wide recommendations.

The Committee will present a draft of the report to the Board in the spring of 2017.

Article 2. Composition and term

The following persons are appointed to the Committee in a private capacity:

Chairperson

- Prof. Johan Mackenbach (Professor of Public Health, Erasmus MC)

Members

- Prof. Cock van Duijn (Professor of Genetic Epidemiology, Erasmus MC)
- Prof. Harry Büller (Professor of Vascular Medicine, Academic Medical Center Amsterdam)
- Prof. Aad van der Vaart (Professor of Stochastics, Leiden University)
- Prof. Eric-Jan Wagenmakers (Professor of Neurocognitive Modelling, University of Amsterdam)
- Dr Patricia Dankers (Associate Professor of Biomaterials, Eindhoven University of Technology)
- Prof. Lex Bouter (Professor of Methodology and Integrity, VU University Amsterdam)

The Committee's term will run until the summer of 2017.

Prof. Philip Scheltens will serve as an agenda member for the Academy Board.

The Committee will be assisted by the Academy's Bureau in accordance with the Director General's instructions. Dr Jean Philippe de Jong of the Academy Bureau will serve as the official secretary

Article 3. Quality management

Prior to their appointment, the members of the Committee familiarised themselves with the *Code ter voorkoming van oneigenlijke beïnvloeding door belangenverstrengeling* [Code of conduct to prevent inappropriate influence owing to conflicts of interests] and filled in and returned the declaration contained therein before the Committee's first meeting.

1 The Academy will also publish a translation of the report into English.

The Committee members have familiarised themselves with the Academy's *Handleiding adviezen KNAW* [Manual concerning Academy Advisory Reports] as adopted by the Academy Board on 21 May 2013. The peer review policy is described in Appendix B to this Manual. There will be no deviation from that policy.

Article 4. Follow-up and communication

The Committee will address follow up and communication concerning its findings.

Article 5. Costs and remuneration

The Committee members will be reimbursed for their travel expenses in accordance with Article 18(2) of the Academy Regulations.

Article 6. Confidentiality

The members of the Committee will observe confidentiality in respect of all information that becomes known to them in the context of the implementation of this resolution and that can be considered to be of a confidential nature.

Adopted in Amsterdam on 9 May 2016 by the Board of the Royal Netherlands Academy of Arts and Sciences.

On behalf of the Academy Board,

M. Zaanen, LLM
Director General of the Royal Netherlands Academy of Arts and Sciences

2. Individuals consulted

- Prof. R. Bernards, Netherlands Cancer Institute, University Medical Center Utrecht
- Dr P. Borst, Professor Emeritus, University of Amsterdam, Netherlands Cancer Institute
- Prof. A.F. Cohen, Centre for Human Drug Research, Leiden University Medical Center
- Prof. H.W. van den Doel, Leiden University
- Dr B.D. Earp, Yale University, University of Oxford
- Dr J.M. Fenterer van Vlissingen, Erasmus University Medical Center
- Prof. J.N.C. de Geus, VU University Medical Center
- Prof. C.C.A.M. Gielen, Netherlands Organisation for Scientific Research, Radboud University
- Prof. S.N. Goodman, Stanford University
- Prof. W.A. van Gool, Health Council of the Netherlands, Academic Medical Center Amsterdam
- Prof. M.H. van Ijzendoorn, Leiden University
- Prof. J.P.A. Ioannidis, Stanford University
- Prof. L.A.L.M. Kiemeney, Radboud University Medical Center
- Prof. J.A. Knottnerus, Maastricht University
- Dr D. Lakens, Eindhoven University of Technology
- I.C. Lether, Dutch Arthritis Foundation
- Prof. M. Macleod, University of Edinburgh, Forth Valley Royal Hospital
- Dr R. Manna, Netherlands Organisation for Health Research and Development
- Dr E. Marcus, Cell Press/Elsevier
- Dr J.W.M van der Meer, Professor Emeritus, Radboud University Medical Center
- Prof. K.G.M. Moons, University Medical Center Utrecht, Julius Center for Health Sciences and Primary Care
- Prof. B.A. Nosek, University of Virginia
- Prof. J.J. van Os, University Medical Center Utrecht
- Prof. S. Repping, Academic Medical Center Amsterdam
- Prof. F.R. Rosendaal, Leiden University Medical Center
- Prof. P. Scheltens, VU University Medical Center
- Dr A.G.J. van de Schoot, Utrecht University
- Prof. T.K. Sixma, Netherlands Cancer Institute, Erasmus University Medical Center
- Prof. E.W. Steyerberg, Leiden University Medical Center, Erasmus University Medical Center
- Dr C.H. Vinkers, University Medical Center Utrecht
- Prof. J.M. Wicherts, Tilburg University

3. Review

At the request of the Academy's Board, a draft of this report was reviewed by the following reviewers:

- Prof. J.N.C. de Geus, VU University Medical Center
- Prof. J.A. Knottnerus, Maastricht University
- Prof. J.P.A. Ioannidis, Stanford University

In addition, the report was reviewed by:

- Prof. B.A. Nosek, University of Virginia
- Dr B.D. Earp, Yale University, University of Oxford
- The Academy's Council for Medical Sciences
- The Academy's Social Sciences Council

The reviewers are not responsible for the final report.

BIBLIOGRAPHY

AMS 2015. The Academy of Medical Sciences. (2015). *Reproducibility and reliability of biomedical research: improving research practice. Symposium report.*

Baker 2014. Baker, D., Lidster, K., Sottomayor, A., & Amor, S. (2014). Two years later: journals are not yet enforcing the ARRIVE guidelines on reporting standards for pre-clinical animal studies. *PLoS biology*, *12*(1), e1001756.

Baker 2016. Baker, M. (2016). Is there a reproducibility crisis? A Nature survey lifts the lid on how researchers view the 'crisis rocking science and what they think will help. *Nature*, *533*(7604), 452-455.

Baker 2017. Baker, M., & Dolgin, E. (2017). Reproducibility project yields muddy results. *Nature*, *541*(7637), 269-270.

Begley 2012. Begley, C. G., & Ellis, L. M. (2012). Drug development: Raise standards for preclinical cancer research. *Nature*, *483*(7391), 531-533.

Begley 2015. Begley, C. G., & Ioannidis, J. P. (2015). Reproducibility in science. *Circulation research*, *116*(1), 116-126.

Bem 2011. Bem, D. J. (2011). Feeling the future: Experimental evidence for anomalous retroactive influences on human cognition and affect. *Journal of Personality and Social Psychology,*100(3), 407-25.

Boccia 2016. Boccia, S., Rothman, K. J., Panic, N., Flacco, M. E., Rosso, A., Pastorino, R., Manzoli, L., La Vecchia, C., Villari, P., Boffetta, P., Ricciardi, W., & Ioannidis, J. P. (2016). Registration practices for observational studies on ClinicalTrials.gov indicated low adherence. *Journal of Clinical Epidemiology*, 70, 176-182.

Brandt 2014. Brandt, M. J., IJzerman, H., Dijksterhuis, A., Farach, F. J., Geller, J., Giner-Sorolla, R., Grange, J.A. Peruginig, M., Spiesh, J.R., & Van 't Veer, A. (2014). The replication recipe: What makes for a convincing replication? *Journal of Experimental Social Psychology*, *50*, 217-224.

Camerer, C. F., Dreber, A., Forsell, E., Ho, T. H., Huber, J., Johannesson, M., ... & Wu, H. (2016). Evaluating replicability of laboratory experiments in economics. *Science*, 351(6280), 1433-1436.

Chambers 2013. Chambers, C. D. (2013). Registered reports: a new publishing initiative at Cortex. *Cortex*, *49*(3), 609-610.

Chambers 2017. Chambers, C. D. (2017). Chambers, C. (2017). *The seven deadly sins of psychology: A manifesto for reforming the culture of scientific practice*. Princeton University Press.

Cook 2016. Cook, B. G., Collins, L. W., Cook, S. C., & Cook, L. (2016). A replication by any other name: A systematic review of replicative intervention studies. *Remedial and Special Education*, *37*(4), 223-234.

Cutcher-Gershenfeld 2017. Cutcher-Gershenfeld, J., Baker, K. S., Berente, N., Flint, C., Gershenfeld, G., Grant, B., Haberman, M, King, J. L., Kirkpatrick, C., Lawrence, B., Lewis, S., Lenhardt, W. C., Mayernik, M., McElroy, C., Mittleman, B., Shin, N., Stall, S., Winter, S., & Zaslavsky, I. (2017). Five ways consortia can catalyse open science. *Nature*, *543*(7647), 615.

Dal-Ré 2014. Dal-Ré, R., Ioannidis, J. P., Bracken, M. B., Buffler, P. A., Chan, A. W., Franco, E. L., La Vecchia, C., &Weiderpass, E. (2014). Making prospective registration of observational research a reality. Science Translational Medicine, 6(224), 224cm1.

Dal Ré 2015. Dal-Ré R., Bracken, M. B., & Ioannidis, J. P. (2015). Call to improve transparency of trials of non-regulated interventions. *BMJ*, 350, h1323.

DFG 2017. Deutsche Forschungsgemeinschaft (German Research Foundation). (2017). *Statement on the Replicability of Research Results.*

Doyen 2012. Doyen, S., Klein, O., Pichon, C. L., & Cleeremans, A. (2012). Behavioral priming: it's all in the mind, but whose mind? *PloS one*, *7*(1), e29081.

Earp 2015. Earp, B. D., & Trafimow, D. (2015). Replication, falsification, and the crisis of confidence in social psychology. *Frontiers in psychology*, *6*, 621.

Everett 2015. Everett, J. A., & Earp, B. D. (2015). A tragedy of the (academic) commons: interpreting the replication crisis in psychology as a social dilemma for early-career researchers. *Frontiers in psychology*, *6*, 1152.

Fanelli 2009. Fanelli, D. (2009). How many scientists fabricate and falsify research? A systematic review and meta-analysis of survey data. *PloS one*, *4*(5), e5738.

Fanelli 2011. Fanelli, D. (2011). Negative results are disappearing from most disciplines and countries. *Scientometrics*, *90*(3), 891-904.

Fanelli 2017. Fanelli, D., Costas, R., & Ioannidis, J. P. (2017). Meta-assessment of bias in science. *Proceedings of the National Academy of Sciences*, 114(14), 3714–3719.

Freedman 2015. Freedman, L. P., Cockburn, I. M., & Simcoe, T. S. (2015). The economics of reproducibility in preclinical research. *PLoS biology*, *13*(6), e1002165.

Gomez 2010. Gómez, O.S., Juristo, N., & Vegas, S. (2010). Replication, Reproduction and Re-analysis: Three ways for verifying experimental findings. *RESER Cape Town, South Africa*, 35.

Goodman 2016. Goodman, S. N., Fanelli, D., & Ioannidis, J. P. (2016). What does research reproducibility mean? *Science translational medicine*, *8*(341), 341ps12.

Greenland 2016. Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., & Altman, D. G. (2016). Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. *European journal of epidemiology*, *31*(4), 337-350.

Harris 2013. Harris, C. R., Coburn, N., Rohrer, D., & Pashler, H. (2013). Two failures to replicate high-performance-goal priming effects. *PloS one*, *8*(8), e72467.

Harris 2017. Harris, R. (2017). *Rigor Mortis: How Sloppy Science Creates Worthless Cures, Crushes Hope, and Wastes Billions*. Hachette UK.

Hüffmeier 2016. Hüffmeier, J., Mazei, J., & Schultze, T. (2016). Reconceptualizing replication as a sequence of different studies: A replication typology. *Journal of Experimental Social Psychology*, *66*, 81-92.

IAP 2016. Interacademy Partnership for Health. (2016). *A call for action to improve the reproducibility of biomedical research.*

Ijzendoorn 1994. Ijzendoorn, M. H. van. (1994). *Chapter 3. A Process Model Of Replication Studies: On The Relation Between Different Types Of Replication*. p 57-70. Leiden University Library

Ioannidis 2005a. Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS medicine*, *2*(8), e124.

Ioannidis 2005b. Ioannidis, J. P. (2005). Contradicted and initially stronger effects in highly cited clinical research. *Journal of the American Medical Association*, *294*(2), 218-228.

Ioannidis 2009. Ioannidis, J. P., Allison, D. B., Ball, C. A., Coulibaly, I., Cui, X., Culhane, A. C., Falchi, M., Furlanello, C., Game, L., Jurman, G., Mangion, J., Mehta, T., Nitzberg, M., Page, G. P., Petretto, E., & van Noort, V. (2009). Repeatability of published microarray gene expression analyses. *Nature genetics*, *41*(2), 149-155.

Ioannidis 2014a. I Ioannidis, J. P., Greenland, S., Hlatky, M. A., Khoury, M. J., Macleod, M. R., Moher, D., Schulz, K.F., & Tibshirani, R. (2014). Increasing value and reducing waste in research design, conduct, and analysis. *The Lancet*, *383*(9912), 166-175.

Ioanndis 2014b. Ioannidis, J. P. (2014). How to make more published research true. *PLoS medicine*, *11*(10), e1001747.

Ioannidis 2017. Ioannidis, J. P., Caplan, A. L., & Dal-Ré, R. (2017). Outcome reporting bias in clinical trials: why monitoring matters. *BMJ,*356, j408.

Iqbal 2016. Iqbal, S. A., Wallach, J. D., Khoury, M. J., Schully, S. D., & Ioannidis, J. P. (2016). Reproducible research practices and transparency across the biomedical literature. *PLoS biology*, *14*(1), e1002333.

Jonas 2016. Jonas, K. J., & Cesario, J. (2016). How can preregistration contribute to research in our field? *Comprehensive Results in Social Psychology,* 1(1-3), 1-7.

Kaplan 2015. Kaplan, R. M., & Irvin, V. L. (2015). Likelihood of null effects of large NHLBI clinical trials has increased over time. *PLoS One*, *10*(8), e0132382.

Kelly 2006. Kelly, C. D. (2006). Replicating empirical research in behavioral ecology: how and why it should be done but rarely ever is. *The Quarterly Review of Biology*, *81*(3), 221-236.

Kerr 1998. Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review*, *2*(3), 196-217.

Kilkenny 2010. Kilkenny, C., Browne, W. J., Cuthill, I. C., Emerson, M., & Altman, D. G. (2010). Improving bioscience research reporting: the ARRIVE guidelines for reporting animal research. *PLoS biology*, *8*(6), e1000412.

Gary King. 1995. King, G. (1995). Replication, Replication. *PS: Political Science and Politics*, 28, 444-452.

Klein 2014. Klein, R. A., Ratliff, K. A., Vianello, M., Adams Jr, R. B., Bahník, ⍰., Bernstein, M. J., ... & Nosek, B.A. (2014). Investigating variation in replicability. *Social psychology*, 45, 142-152.

Knottnerus 2016. Knottnerus, J. A., (2016). Research data as a global public good. *Journal of Clinical Epidemiology,* 70, 270-271.

Kunert 2016. Kunert, R. (2016). Internal conceptual replications do not increase independent replication success. *Psychonomic bulletin & review*, *23*(5), 1631-1638.

Leek 2017. Leek, J. T., & Jager, L. R. (2017). Is most published research really false? *Annual Review of Statistics and Its Application*, *4*, 109-122.

Levelt 2012. Commissie Levelt, Commissie Noort, Commissie Drenth. (2012). *Falende wetenschap: De frauduleuze onderzoekspraktijken van sociaal-psycholoog Diederik Stapel.*

Makel 2012. Makel, M. C., Plucker, J. A., & Hegarty, B. (2012). Replications in psychology research: How often do they really occur? *Perspectives on Psychological Science*, *7*(6), 537-542.

Makel 2016. Makel, M. C., Plucker, J. A., Freeman, J., Lombardi, A., Simonsen, B., & Coyne, M. (2016). Replication of special education research: Necessary but far too rare. *Remedial and Special Education*, *37*(4), 205-212.

Martin 2017. Martin, G. N., & Clarke, R. M. (2017). Are psychology journals anti-replication? A snapshot of editorial practices. *Frontiers in Psychology*, *8, 523*.

Mogil 2017. Mogil, J. S., & Macleod, M. R. (2017). No publication without confirmation. *Nature*, *542*(7642), 409-411.

Munafò 2017. Munafò, M. R., Nosek, B. A., Bishop, D. V., Button, K. S., Chambers, C. D., du Sert, N. P., Simonsohn, U., Wagenmakers, E. J., Ware, J. J., & Ioannidis, J. P. (2017). A manifesto for reproducible science. *Nature Human Behaviour*, *1*, 0021.

Mueller-Langer 2017. Mueller-Langer, F., Fecher, B., Harhoff, D., & Wagner, G. G. (2017). The economics of replication. *Max Planck Institute for Innovation & Competition Research Paper No. 17-03.*

NAS 2016. National Academies of Sciences, Engineering, and Medicine. (2016). *Statistical challenges in assessing and fostering the reproducibility of scientific results: Summary of a workshop*. National Academies Press.

Nature 2016. Go forth and replicate! Editorial. *Nature*, 536(7617), 373

Nosek 2014. Nosek, B. A., & Lakens, D. (2016). Registered Reports: A Method to Increase the Credibility of Published Reports. *Social Psychology*, 45(3), 137-141.

Nosek 2012. Nosek, B. A., Spies, J. R., & Motyl, M. (2012). Scientific utopia: II. Restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science*, *7*(6), 615-631.

Nosek 2015. Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., … & Yarkoni, T. (2015). Promoting an open research culture. *Science*, 348, 6242, 1422-1425

Nosek 2017. Nosek, B. A., & Errington, T. M. (2017). Reproducibility in cancer biology: making sense of replications. *Elife*, *6*, e23383.

NSF 2015. Cacioppo, J. T., Kaplan, R. M., Krosnick, J. A., Olds, J. L., & Dean, H. (2015). *Social, Behavioral, and Economic Sciences Perspectives on Robust and Reliable Science. Report of the Subcommittee on Replicability in Science Advisory Committee to the National Science Foundation Directorate for Social, Behavioral, and Economic Sciences*

Nuzzo 2015. Nuzzo, R. (2015). How scientists fool themselves-and how they can stop. *Nature*, *526*(7572), 182.

OSC 2015. Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716.

Parker 2016. Parker, T. H., Forstmeier, W., Koricheva, J., Fidler, F., Hadfield, J. D., Chee, Y. E., Kelly, C. D., Gurevitch, J., & Nakagawa, S. (2016). Transparency in ecology and evolution: real problems, real solutions. *Trends in ecology & evolution*, *31*(9), 711-719.

Pashler 2012. Pashler, H., & Wagenmakers, E. J. (2012). Editors' introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspectives on Psychological Science*, *7*(6), 528-530.

Perrin 2014. Perrin, S. (2014). Preclinical research: Make mouse studies work. *Nature*, *507*(7493), 423-425.

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

Popper 1959. Popper, K. (2005). *The logic of scientific discovery*. Routledge.

Prinz 2011. Prinz, F., Schlange, T., & Asadullah, K. (2011). Believe it or not: how much can we rely on published data on potential drug targets? *Nature reviews Drug discovery*, *10*(9), 712-712.

Ravenzwaaij 2017. Ravenzwaaij, D. van, Ioannidis, J. P., (2017). A simulation study of the strength of evidence in the recommendation of medications based on two trials with statistically significant results. *PLoS One,* 12(3), e0173184.

Resnik 2017. Resnik, D. B., & Shamoo, A. E. (2017). Reproducibility and Research Integrity. *Accountability in research*, *24*(2), 116-123.

Rifai 2014. Rifai, N., Bossuyt, P. M., Ioannidis, J. P., Bray, K. R., McShane, L. M., Golub, R. M., Hooft, L. (2014). Registering diagnostic and prognostic trials of tests: is it the right thing to do? *Clinical Chemistry*, 60(9), 1146-1152.

Sena 2010. Sena, E. S., Van Der Worp, H. B., Bath, P. M., Howells, D. W., & Macleod, M. R. (2010). Publication bias in reports of animal stroke studies leads to major overstatement of efficacy. *PLoS biology*, *8*(3), e1000344.

Shapin 1985. Shapin, S., & Schaffer, S. (2011). *Leviathan and the Air-Pump: Hobbes, Boyle, and the Experimental Life*. Princeton University Press.

Simons 2014. Simons, D. J. (2014). The value of direct replication. *Perspectives on Psychological Science*, *9*(1), 76-80.

Spellman 2015. Spellman, B. A. (2015). A short (personal) future history of revolution 2.0. *Perspectives on Psychological Science*, 10(6), 886-899.

Tsilidis 2013. Tsilidis, K. K., Panagiotou, O. A., Sena, E. S., Aretouli, E., Evangelou, E., Howells, D. W., ... & Ioannidis, J. P. (2013). Evaluation of excess significance bias in animal studies of neurological diseases. *PLoS biology*, *11*(7), e1001609.

Srivastava 2012. Srivastava, S. (2012). A Pottery Barn rule for scientific journals. *September*, *27*, 2012. https://hardsci.wordpress.com/2012/09/27/a-pottery-barn-rule-for-scientific-journals/

VSNU 2014. VSNU – Association of universities in the Netherlands. (2014). *The Netherlands Code of Conduct for Academic Practice. Principles of good academic teaching and research.*

Wicherts 2016. Wicherts, J. M., Veldkamp, C. L., Augusteijn, H. E., Bakker, M., Van Aert, R. C., & Van Assen, M. A. (2016). Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid p-hacking. *Frontiers in psychology*, *7, 1832*.

Yong 2012. Yong, E. (2012). Bad copy. *Nature*, *485*(7398), 298.

NOTES

1 See Table 1

2 Several academic societies have issued statements or reports regarding reproducibility (NAS 2016, NSF 2015, AMS 2015, IAP 2016)

3 The feeling that there is a 'crisis' in science is not new. In the field of psychology, concerns were voiced as far back as the seventies about the quality of reporting, the proper use of statistical analysis and problems with replication. However, in an era of fast and worldwide communication, it has become a more prominent topic, including in the popular media (Spellman 2015, Pashler 2012). According to a survey by *Nature* in 2016, 90% of respondents think there is a 'reproducibility crisis' (Baker 2016). For an overview of the current debate about reproducibility in the life sciences, see Harris 2017

4 This definition is in line with the report issued by the United Kingdom's Academy of Medical Science (AMS 2015)

5 Psychology in particular has distinguished between 'direct' replications, which closely follow the approach of the original study, and 'indirect' or 'conceptual' replications, which take a wholly different approach to answer the same scientific question (Earp 2015). These indirect replications can indeed contribute to scientific knowledge by corroborating (or questioning) conclusions or by helping to assess whether the methods used are valid. Some authors call the extent to which the same conclusions can be drawn from results 'inferential reproducibility' (Goodman 2016). However, indirect replications cannot help to assess whether the results of the previous study can be *reproduced* as defined in this report, i.e. to what extent the results of a study are similar to those of an earlier, similar study (Earp 2015). To avoid muddling the discussion, we will not call these studies 'replication studies'. We elaborate on how to assess reproducibility and the need for similarity in replications studies in Chapters 2.2 and 3.2

6 This definition is in line with the report of the United Kingdom's Academy of Medical Science (AMS 2015)

7 Several of these typologies have provided helpful analyses for this report, e.g.: 'direct versus indirect replication' (Earp 2015); 'methods/results/inferential reproducibility' (Goodman 2016); 're-analysis/replication/reproduction' (Gomez 2010); or even more complex typologies (Huffmeier 2016)

8 Other authors have used reproducibility to denote the extent to which the *methods* of a study can be reproduced (referred to as 'method reproducibility') and the extent to which the same conclusions can be drawn from results (referred to as 'inferential reproducibility') (Goodman 2016).

9 NAS 2016

10 AMS 2015

11 VSNU 2014

12 OSC 2015, Begley 2015, Nosek 2017
13 IJzendoorn 1994
14 OSC 2015, Begley 2015
15 Begley 2015
16 IJzendoorn 1994, Goodman 2016, AMS 2015
17 Ioannidis 2005a. Ioannidis has proposed an analytical framework in which a research finding is less likely to be true when: the studies conducted in a field are smaller; when effect sizes are smaller; when there is a greater number and lesser preselection of tested relationships; where there is greater flexibility in designs, definitions, outcomes, and analytical modes; when there is greater financial and other interest and prejudice; and when more teams are involved in a scientific field chasing statistical significance. Applying this framework to various scientific fields, he argued that for most study designs and settings, it is more likely for a research claim to be false than true. His argument was based on the 2x2 table of statistical hypothesis testing, as in Notes Table 1. Here α and β are the probabilities of the errors of reporting a discovery (i.e. rejecting the null) if there is no effect, or not reporting a discovery if the effect does exist, and π is the prior probability that the effect exists (i.e. that the null is false). From the table, one can compute the probability that the effect exists given that it is reported as $(1-\beta)\pi/((1-\beta)\pi+\alpha(1-\pi))$, the relative weight within the first row of the table that the effect exists. Ioannidis' argument is that in many situations this number is small, even for given small *level* α. Specifically, small studies and small effect sizes both lead to small *power* 1-β, while selecting many hypotheses to study will lead to small π. Furthermore, this situation is exacerbated by study bias and by many researchers performing the same study with only success stories being reported. These two effects lead to Notes Table 2, where *u* is the bias towards reporting and *n* is the number of researchers studying the same hypothesis. The number *u* is larger in studies that have flexible, less clearly defined outcomes or use less well-established, more exploratory techniques, and *n* tends to be large in popular research areas. The probability that an effect exists given that it is reported, computed as before, but now from the first row of Notes Table 2, increases with *u* and *n*. Ioannidis suggests values ranging from 0.85 (π=0.5, *u*=0.1, α=0.05, β=0.2, *n*=1) in an adequately powered random clinical trial to 0.001 (π=1/1001, α=0.05, β=0.2, *u*=0.8, *n*=1) in discovery-oriented exploratory research with massive testing. (The effect of *n>1* is to replace β by β^n and 1-α by $(1-\alpha)^n$, which can be dramatically smaller numbers.) He suggests that even in well-performed, but underpowered randomised phase I/II clinical trials, the probability of a finding being correct does not exceed 0.23, and this would indeed prove the assertion that in this setting most published research findings are false. The available empirical evidence is perhaps too small for such precise calculations, and statistical practice nowadays corrects for multiplicity and/or in terms of false discovery rates, for example by choosing a lower threshold α than the typical 0.05. Nevertheless, Ioannidis' assertion that 'most published research is false' has been a key driver for the much-needed reflection on statistical approaches and the proper interpretation of the results of statistical tests

Notes Table 1: probabilities of true and false decisions. π is the probability that the effect exists, β the probability that it is found given that it exists, α the probability that it is found given that it does not exist

	effect exists	effect does not exist
report discovery	(1-β)π	α(1-π)
do not report	βπ	(1-α)(1-π)

Notes Table 2: probabilities of true and false decisions. π, β and α are as in Notes Table 1. u is the fraction of discoveries that would not be reported in an unbiased experiment but are reported. n is the number of independent attempts to make the discovery

	effect exists	effect does not exist
report discovery	$(1-\beta^n+u\beta^n)\pi$	$(1-(1-\alpha)^n(1-u))(1-\pi)$
do not report	$(1-u)\beta^n\pi$	$(1-u)(1-\alpha)^n(1-\pi)$

18 Fanelli 2011

19 Prinz, 2011

20 Begley 2012

21 Ioannidis 2009

22 Perrin 2014

23 Ioannidis 2005b

24 Klein 2014

25 OSC 2015. Another project from the Open Science Collaboration is ongoing. The Reproducibility Project: Cancer Biology is scrutinising key results in 29 cancer papers published in *Nature*, *Science*, *Cell* and other high-impact journals by independently replicating experiments (Nosek 2017)

26 Camerer 2016

Begley 2012

27 Ioannidis 2014a, Freedman 2015

28 AMS 2015, Ioannides 2005a, Begley 2015, Munafò 2017, Wicherts 2016, Chambers 2017

29 Resnik 2017, Fanelli 2009

30 Examples of a lack of rigour in statistical analyses include statistically testing many correlations between variables without first devising a specific hypothesis and then presenting the statistically significant outcomes of such analyses as the result of a hypothesis-testing design. The technical terms for these questionable research practices are 'data dredging'/'p-hacking' in combination with 'HARKing' (Hypothesising After Results are Known). Wicherts 2016, Kerr 1998

31 Wicherts 2016

32 Fanelli 2017

33 The possibility of differences between reproducibility across countries is supported by the finding that countries differ with respect to publication bias. The United States had published significantly fewer positive results over the years than Asian countries (and particularly Japan) but more than European countries (especially the United Kingdom). Fanelli 2011

34 AMS 2015, Begley 2015, Munafò 2017, Ioannidis 2014b, IAP 2016

35 Cutcher-Gershenfeld 2017

36 For example, when the National Heart, Lung, and Blood Institute required preregistration of primary outcomes (the main outcome against which success should be judged) in clinical trials, the proportion of studies reporting a benefit fell from 57% to 8%. This indicates that preregistration can reduce false positive results substantially (Kaplan 2015)

37 We consider 'internal replications' as another way to *prevent* the non-reproducibility of published results and have not included them in our definition of replication *studies.* The reason is that internal replications are activities carried out by a researcher to replicate parts of his or her own study (for example, by performing multiple measurements, cross-validation within a dataset, or setting up control experiments), and these results will normally either be published together with the other, 'primary' results or not at all. We

concede that this is a somewhat arbitrary criterion. For example, if a researcher were to conduct multiple identical experiments and publish the results together in one journal article, this would count as an internal replication, whereas if he were to publish the results in separate journal articles, it should count as a replication study. Nevertheless, the important point here is that besides *preventing* non-reproducible results from becoming part of the literature – and this would include most internal replications – there is a need for replication studies to assess whether results that have already become part of the literature can be reproduced – and this would not include reports about internal replications. In Chapter 4 we describe replication practices, including internal replications and replication studies, in various disciplines

38 Versioning' is a publication practice in which published articles are not considered 'final' but can be improved transparently afterwards

39 Data-sharing can, for example, be incentivised by linking project initiators and data collectors routinely to the data and citing their basic work. Moreover, it should be considered a significant mark of esteem if the data researchers have gathered are used worldwide in the service of science and society at large. In addition, analogous to trial registries, data sources can be registered and linked to those who established them (Knottnerus 2016)

40 For example, researchers can gain recognition by registering peer-review activities on the website https://publons.com

41 Munafò 2017, https://cos.io/our-services/top-guidelines/, Nosek 2015

42 Popper 1959

43 NAS 2016

44 This case has been described by Steven Shapin and Simon Schaffer in *Leviathan and the Air-Pump* (Shapin 1985), Princeton University Press, Princeton, New Jersey (1985). The text in the box is based on en.wikipedia.org/wiki/Reproducibility, retrieved 11-2-2017

45 Earp uses the term 'auxiliary assumptions' for this (Earp 2015

46 This has been the approach of the Open Science Collaboration (OSC 2015)

47 Brandt 2014. Kunert et al. have even suggested that internal replications do not increase the likelihood that results will be reproduced in independent replication studies. They attribute this failure to questionable research practices by the researcher (Kunert 2016)

48 However, Leek calls reanalysis reproducibility, and what we call reproducibility, he calls replicability (Leek 2017)

49 This is clearly evident in research areas that investigate unique events, such as climate change, supernovas, volcanic eruptions or past events, or which focus on the observation of contingent phenomena (e.g. in the earth system sciences or in astrophysics) (DFG 2017). However, even in the laboratory there will always be parameters that vary from one experiment to the next

50 Whether the results of a study are considered to have been reproduced is based on many auxiliary assumptions about which study parameters are and are not relevant. Some parameters can be altered without negatively affecting the outcome of the study (Earp 2015). Indeed, altering a parameter of a study can be an obvious improvement and may increase its statistical power to detect an effect. In addition to assessing reproducibility, researchers may want to assess the 'robustness' of a study: the stability of experimental conclusions when there are variations in either baseline assumptions or experimental procedures (Goodman 2016). To do so, Ijzendoorn has proposed a 'process model' of replication in which aspects of a study are altered systematically (Ijzendoorn 1994)

51 This contrasts with Brandt, who has proposed a 'replication recipe' for replications,

including the 'ingredients' that the study is to be conducted by independent researchers and that it should follow the original study exactly (Brandt 2014)

52 Various indicators can serve to quantify the extent to which replication studies should be carried out, e.g. a proportion of funding, of published studies, or of a researcher's time investment. Because funding comes from many different sources, that percentage can be hard to pin down exactly, and because publication bias possibly affects replication studies more than original studies, we suggest that time investment by researchers is the best indicator

53 Many authors have concluded that replications are rare (see Makel 2012, Kelly 2006, Parker 2016, Iqbal 2016, Poldrack 2017, Makel 2016). Only two reports have produced quantitative estimates. Makel et al. concluded that 1.07% of studies in the 100 most frequently cited psychology journals had been presented as a replication study, and that there had been an increase in recent decades (Makel 2012). However, the studies that were sampled included 34% 'internal' replications (replication efforts as part of a larger study, the results of which were presented together in the same report). So according to our definition of a replication study, the percentage of replication studies would be 0.7%. Iqbal et al. surveyed a random sample (n=259) of the biomedical literature between 2000 and 2014. 1.5% of articles claimed or were inferred to be replication efforts trying to validate previous knowledge, 51.7% claimed to present some novel findings and (1.5%) had clear statements of both study novelty and some form of replication. In 45.2% of cases, it was unclear whether they reported novel findings or replication efforts (Iqbal 2016)

54 Makel 2012

55 Cook did indeed find that replication studies do not always use the word replication (Cook 2016)

56 Begley 2012, Prinz 2011

57 For example: Begley 2012, Prinz 2011, OSC 2015

58 Begley 2012, Prinz 2011

59 Sena 2010, Tsilidis 2013

60 Begley 2015

61 Begley 2015

62 Mogil 2017

63 For example, the ARRIVE guidelines (Kilkenny 2010). https://www.nc3rs.org.uk/arrive-guidelines

64 For example, the International Committee of Medical Journal Editors has provided recommendations for the conduct, reporting, editing, and publication of scholarly work in medical journals

65 Baker 2014

66 Preregistration of clinical trials is now standard practice in many countries, for example in the Unites States' Clinicaltrials.gov database or the European Union EudraCT database

67 Dal-Ré 2015

68 Ioannidis 2017

69 However, even with two studies, the chance of having weak evidence for new treatments is still very substantial (Ravenzwaaij 2017)

70 Rifai 2014

71 Boccia 2016, Dal-Ré 2014

72 Levelt 2012

73 Bem, 2011

74 Doyen 2012, Harris 2013

75 Yong 2012
76 *Perspectives on Psychological Science* has begun publishing a new type of article, tailored to reporting replication studies. It asks psychologists to nominate an influential study for replication and draw up a plan. The original author is then invited to offer suggestions on the protocol, multiple labs volunteer to collect data. Consequently, the results — whatever they may be — are published as a registered replication report (RRR) (https://www.psychologicalscience.org/publications/replication)
77 Social Psychology, May 19, 2014
78 OSC 2015
79 Klein 2014
80 Nosek 2015
81 Jonas 2016
82 Bem 2011
83 This is known as the 'winner's curse': the phenomenon that the first study overestimates the association due to low statistical power and publication bias, resulting in the failure of low-powered replication studies to corroborate the association
84 ($p<5x10-8$)
85 Baker 2016, Nature 2016, Munafò 2017
86 AMS 2015, Ioannides 2005, Begley 2015, Munafò 2017
87 Goodman 2016, VSNU 2014. King (1995) has, for example, argued that empirical political scientists need to have access to the body of data necessary to replicate existing studies so as to understand, evaluate, and especially build on this work. He has proposed archiving studies in the form of 'replication data sets' that include all information necessary to replicate empirical results
88 NAS 2016, Baker 2017
89 Referred to as 'method reproducibility' (Goodman 2016)
90 The website of the Open Science Framework lists more than 7000 registrations (http://osf.io/), showing the take-up of this practice in basic/pre-clinical sciences. It also provides a guided workflow to help researchers learn about and participate in preregistration (http://cos.io/prereg/)
91 Preregistration of clinical trials is now standard practice in many countries, for example in the Unites States' Clinicaltrials.gov database or the European Union EudraCT database. However, in Clinicaltrials.gov 45.2% of completed trials have not published their results (https://trialstracker.ebmdatalab.net/#/). An initiative that takes preregistration a step further is the Registered Reports format (Chambers 2013)
92 Kaplan 2015
93 Examples include the ARRIVE guidelines (https://www.nc3rs.org.uk/arrive-guidelines), STAR methods (http://www.cell.com/star-methods) and the EQUATOR network (https://www.equator-network.org/reporting-guidelines/)
94 For example, the journal *Cell* has lifted length restrictions for methods sections
95 The European Open Science Cloud initiative by the European Commission aims to provide cloud-based services and a data infrastructure to store, share and re-use scientific data across disciplines and borders (https://ec.europa.eu/research/openscience/index.cfm?pg=open-science-cloud)
96 Nosek 2012
97 Simons 2014, Greenland 2016, Munafò 2017
98 Begley 2015
99 Everett 2015

100 For example, the field of neuroimaging can draw lessons from genetics (Poldrack 2017)
101 Ioannidis 2014a
102 Spellman 2015, Baker2016, Nuzzo 2015
103 Ijzendoorn 1994
104 Examples include the European Research Council, which focuses on 'frontier-research' (https://erc.europa.eu/about-erc/mission), the public-private funded Innovative Medicines Initiative (http://www.imi.europa.eu/content/mission), Horizon 2020, the 'EU Framework Programme for Research and Innovation' (https://ec.europa.eu/programmes/horizon2020/en) and national agencies such as the Dutch NWO with its 'Innovational Research Incentives Schemes'
105 For example, 'the criteria for publication of scientific papers (Articles and Letters) in *Nature* are that they: report original scientific research' (http://www.nature.com/nature/authors/get_published/#a1). The *British Medical Journal* states that 'If your research is novel, ethical, and methodologically robust, and it deals with questions that are directly related to clinical care, public health, or healthcare policy, we invite you to submit it to the BMJ' (http://www.bmj.com/content/346/bmj.f2433). Moreover, only 3% of psychology journals state in their aims or instructions to authors that they accept replications. (Martin 2017)
106 Baker 2016
107 *Nature* 2016
108 For example, the Dutch funding agency NWO and the German DFG have committed themselves to funding replication studies (https://www.nwo.nl/en/research-and-results/programmes/replication+studies, DFG 2017)
109 This recommendation is in line with the Transparency and Openness Promotion (TOP) guidelines that have been adopted by journals such as *Science*, *Nature* and *PLoS ONE* (Nosek 2015). The guideline includes a standard on replication, ranging from encouraging replication studies to using the 'registered reports' format as a submission option for replication studies with peer review prior to observing the study outcomes
110 *Perspectives on Psychological Science* has launched such an initiative, see Chapter 4.4
111 This has become common practice for genetic epidemiology studies, see Chapter 4.5
112 Srivastava 2012. For example, the journal *Royal Society Open Science* commits to publishing any methodologically sound attempt to replicate any study that has been published within this and a number of other journals (http://neurochambers.blogspot.nl/2016/11/an-accountable-replication-policy-at.html)
113 For example, the online platform F1000 launched the dedicated Preclinical Reproducibility and Robustness channel for refutations, confirmations or more nuanced replication studies