TIER2

Enhancing Trust, Integrity, and Efficiency in Research through Next-Level Reproducibility Impact Pathways

Deliverable D3.1 - State-of-play on methods, tools, practices to increase reproducibility across diverse epistemic contexts

31/12/2023

Lead Beneficiary: Know-Center GmbH

Author/s: Tony Ross-Hellauer, Sven Ulpts, Eva Kormann, Nicki Lisa Cole, Simone Kopeinik, Dominik Kowald, Christopher Osborne, Jesper Schneider

Reviewer/s: Joeri Tijdink, Eleni Adamidi

Funded by
the European Union

Views and opinions expressed are those of the author(s) only and do not necessarily reflect those of the European Union or the European Commission. Neither the EU nor the EC can be held responsible for them.

Prepared under contract from the European Commission
Grant agreement No. 101094817

EU Horizon Europe Research and Innovation action

Project acronym:	**TIER2**
Project full title:	**Enhancing Trust, Integrity, and Efficiency in Research through Next-Level Reproducibility Impact Pathways**
Start of the project:	January 2023
Duration:	36 months
Project coordinator:	Dr. Tony Ross-Hellauer

Deliverable title:	State-of-play on methods, tools, practices to increase reproducibility across diverse epistemic contexts
Deliverable n°:	D3.1
Version n°:	1
Nature of the deliverable:	Report
Dissemination level:	Public

WP responsible:	WP3
Lead beneficiary:	Know-Center GmbH

TIER2 Project, Grant agreement No. 101094817

Due date of deliverable:	Month 12
Actual submission date:	Month 12

Deliverable status:

Version	Status	Date	Author(s)
0.9	Draft	11 Dec 2023	Tony Ross-Hellauer, Sven Ulpts, Eva Kormann, Nicki Lisa Cole, Simone Kopeinik, Dominik Kowald, Christopher Osborne, Jesper Schneider Know-Center GmbH, Aarhus University, University of Oxford
0.91	Review	20 Dec 2023	Joeri Tijdink, Eleni Adamidi VUMC, ARC
0.92	Revised according to reviewer feedback	30 Dec 2023	Tony Ross-Hellauer, Sven Ulpts, Jesper Schneider
1	Final (incl. final editing/formatting)	31 Dec 2023	Tony Ross-Hellauer

The content of this deliverable does not necessarily reflect the official opinions of the European Commission or other institutions of the European Union.

Table of contents

Executive Summary

TIER2, over the course of its three-year duration (2023-2025), aims to contribute to improving this situation in various ways. Key to our approach is to centre “epistemic diversity” (defined below) by selecting three broad research areas — social, life, and computer sciences, and two cross-disciplinary stakeholder groups of research publishers and funders — to systematically investigate the roles, nature, and meanings of reproducibility across contexts. Through coordinated co-creation with these communities, TIER2 aims to boost knowledge on reproducibility, create tools, engage communities, implement interventions and policy across different contexts to increase reproducibility where it is relevant.

This Deliverable details work to provide the theoretical, evidential and strategic framework for the project. The aim is to capture the complexity in the meaning(s) of reproducibility across contexts, provide a conceptual framework that systematically relates epistemic diversity to reproducibility by identifying key research characteristics affecting the relevance and feasibility of different types of reproducibility, establish current levels of knowledge on which interventions work in which contexts (including in two specific cross-cutting research methods (qualitative and Machine Learning-driven research), and devise a strategic intervention logic for designing and implementing interventions that aim at sustainable behavioural change towards increased reproducibility.

This work has been addressed through seven ambitious individual studies:

- “Definitions of reproducibility” (Section 2)
- “Epistemic diversity and Knowledge Production Modes” (Sec. 3)
- “Scoping review and evidence mapping of interventions aimed at improving reproducible and replicable science” (Sec. 4)
- “Review of conceptions and facilitators of and barriers to reproducibility of qualitative research” (Sec. 5)
- “Review of conceptions and practices regarding reproducibility in Machine Learning (ML)-driven research” (Sec. 6)
- “Changing behaviour in the academy: A strategy for improving research culture and practice” (Sec. 7)

This work hence fills knowledge gaps to enable the mapping of “impact pathways”, i.e., the possible paths that connect input to output, outcome and impact (including linkages of causal mechanisms and drivers/barriers), to elucidate the routes to increased reproducibility across diverse contexts. This work is crucial to inform the future stages of TIER2, especially to design, implement and test a series of new tools and instruments (the “pilots”) conducted within TIER2 Work Packages 4 and 5.

List of Abbreviations

BCW – Behaviour Change Wheel DMP – Data Management Plan

CCS – Culture Change Strategy

CORDIS - Community Research and Development Information Service (EC)

EU – European Union

EUA – European University Association

KPM – Knowledge Production Mode

ML – Machine Learning

OSF – Open Science Framework

PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses: extension for Scoping Reviews

UKRI - UK Research and Innovation

UNESCO - United Nations Educational, Scientific and Cultural Organization

1. Introduction

Although definitions and usages of the term reproducibility vary widely, for many it refers, in a broad sense, to the possibility for the scientific community to obtain the same or similar results as the originators of a specific finding through repetition of research methods or analyses (Barba, 2018). Recently, concerns about supposedly low rates of reproducibility have come up in a variety of disciplines, but mainly in the behavioural and medical sciences. Proposed drivers of poor reproducibility include lack of transparency in reporting, data, and analysis, lack of replication studies, publication bias towards reporting of positive results, and questionable research practices (Atmanspacher & Maase, 2016; Munafo et al., 2017; Nosek et al., 2022). While poor levels of reproducibility are seen by some as a serious threat to scientific self-correction, efficiency of research processes, and societal trust in research results, the applicability of reproducibility is highly contested depending on the nature of the research. Suggested research characteristics that affect the applicability of reproducibility are, for instance, the cost, effort, and time it takes to be reproducible, the degree of standardization in the research domain, the reliance of inferential statistics, the epistemology, the nature of the subject of investigation, the type of reproducibility (replication), and ethical and legal constraints to sharing and transparency (Leonelli, 2018).

TIER2, over the course of its three-year duration (2023-2025), aims to contribute to improving this situation in various ways. Key to our approach is to centre “epistemic diversity” (defined below) by selecting three broad research areas — social, life, and computer sciences, and two cross-disciplinary stakeholder groups of research publishers and funders — to systematically investigate the roles, nature, and meanings of reproducibility across contexts. Through coordinated co-creation with these communities, TIER2 aims to boost knowledge on reproducibility, create tools, engage communities, implement interventions and policy across different contexts to increase reproducibility where it is relevant.

Figure 1, below, shows the methodological steps whereby these aims will be achieved. In the later stages we aim to design, implement and assess a series of novel pilot interventions, tools and practices addressing various stakeholders (researchers, funders and publishers), and then consolidate findings into a cohesive vision that evaluates gains and savings, and produces a roadmap for future action.

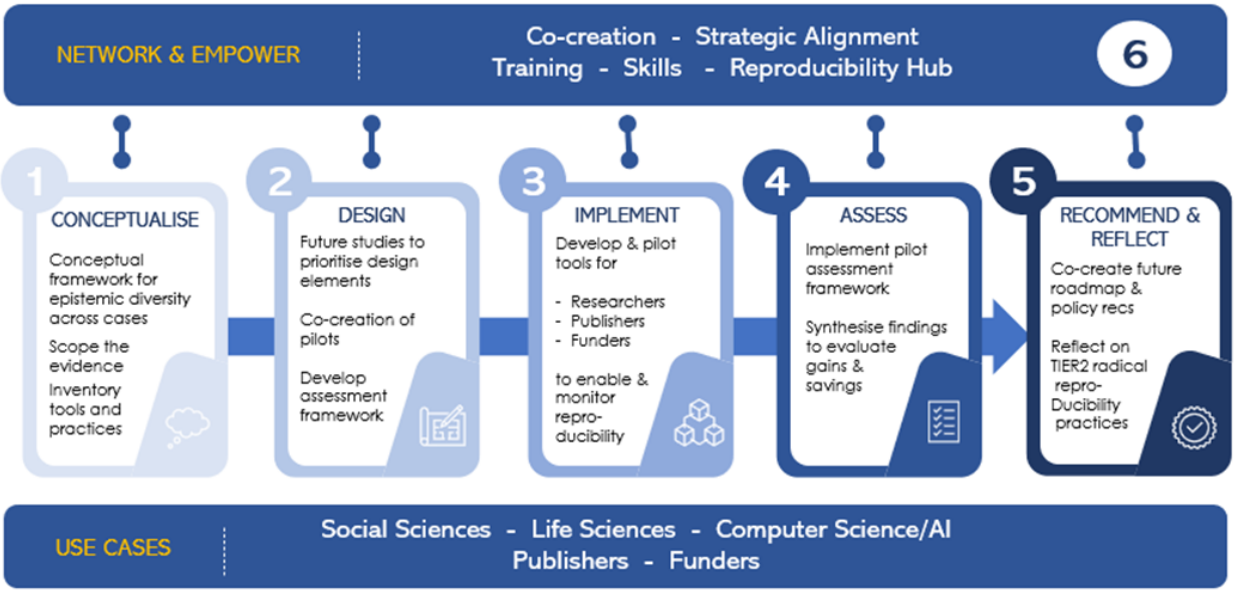

NETWORK & EMPOWER
Co-creation - Strategic Alignment
Training - Skills - Reproducibility Hub
6
1
CONCEPTUALISE
Conceptual framework for epistemic diversity across cases
Scope the evidence
Inventory tools and practices
2
DESIGN
Future studies to prioritise design elements
Co-creation of pilots
Develop assessment framework
3
IMPLEMENT
Develop & pilot tools for
- Researchers
- Publishers
- Funders
to enable & monitor repro-ducibility
4
ASSESS
Implement pilot assessment framework
Synthesise findings to evaluate gains & savings
5
RECOMMEND & REFLECT
Co-create future roadmap & policy recs
Reflect on TIER2 radical repro-Ducibility practices
USE CASES
Social Sciences - Life Sciences - Computer Science/AI
Publishers - Funders

Figure 1. TIER2 methodological steps

As seen in Figure 1, the essential first step on this journey (“Conceptualise”) is to capture the complexity in the meaning(s) of reproducibility across contexts, provide a conceptual framework that systematically relates epistemic diversity to reproducibility by identifying key research characteristics affecting the relevance and feasibility of different types of reproducibility, and establish current levels of knowledge on which interventions work in which contexts. Doing so aims to fill knowledge gaps to enable the mapping of “impact pathways”, i.e., the possible paths that connect input to output, outcome and impact (including linkages of causal mechanisms and drivers/barriers), to elucidate the routes to increased reproducibility across diverse contexts. This work is crucial to inform the future stages of TIER2, especially to design, implement and test a series of new tools and instruments (the “pilots”) conducted within TIER2 Work Packages 4 and 5.

These steps were operationalised via two tasks within Work Package 3 (“Concept, evidence, synthesis and recommendations”) of TIER2:

- **Task 3.1. Conceptual framework for reproducibility across contexts:** In our proposal, we identified a lack of clarity on meanings, limits, implications, but also relevance of reproducibility across modes of knowledge production as a key issue. Other scholars (Devezer et al., 2019; Guttinger, 2020; Leonelli, 2018; Penders et al., 2019; Tuval-Mashiach, 2021) have already highlighted the effect of “epistemic diversity”[1] on the meanings, implications, and applicability of different types of reproducibility across diverse epistemic contexts. Since epistemic diversity is a key component of TIER2’s aim to understand the role of ‘reproducibility’ across epistemic contexts while taking into account the prevailing conceptual confusion about ‘reproducibility’ building and extending on the

[1] Epistemic diversity is defined by (Leonelli, 2022) as "the condition or fact of being different or varied, which affects the development and/or understanding of knowledge”. Sources of such variance, for Leonelli, can be conceptual, material, methodological, infrastructural, socio-cultural, and institutional.

existing conceptual work led to the formulation of the knowledge production modes (KPM) framework as the conceptual foundation of TIER2. Grasping the conceptual confusion required a snowballing of the literature based on existing reviews about the meanings of reproducibility and related terms (see, e.g., Barba, 2018; Gómez et al., 2014; Gundersen, 2021; Köhler & Cortina, 2021; Matarese, 2022; Plesser, 2018). The synthesis and critical appraisal of the existing literature on the relation between epistemic diversity and 'reproducibility' also included discussions with some original authors about their views on the issue. In a separate strand of work, in order to distil basic principles for the design of interventions to improve reproducibility, we also aimed to further investigate theories of behaviour change. This builds especially upon the Culture Change Strategy from the Center for Open Science (Nosek, 2019) and the Behaviour-Change Wheel (Michie et al., 2011). The resultant general strategy for improving research culture and practice provides accessible guidance for designing and incrementally improving and assessing interventions that aim at behavioural change towards improvement of research, including reproducibility.

- **Task 3.2. Evidence-base and inventory of reproducibility tools and practices:** Acknowledging that reproducibility has very different meanings and consequences across epistemic contexts, it is essential that any analysis be rooted in an understanding of how interventions vary according to these contexts. As a first step, a rigorous accounting of current knowledge on which interventions work, in which circumstances, is essential. Hence, this task aimed to consolidate and valorise current knowledge on practices and tools for reproducibility by systematically scoping, critically appraising, and synthesising the literature. It included a comprehensive scoping review of tested reproducibility interventions, as well as two smaller reviews of reproducibility as it relates to qualitative research and research employing methods of Machine Learning (ML). Qualitative research and ML-driven research were chosen as two cross-cutting methods which are relevant across TIER2's focus disciplines (social, life and computer sciences), as well as areas of particular interest. Within qualitative research, reproducibility is often claimed to be an inadequate epistemic criterion for such research (Leonelli, 2018), and within ML, their "black-box" nature and inherent indeterminism provides interesting challenges for conceptions of "computational reproducibility".

This Deliverable reports the outcomes of this work. It includes seven individual studies addressing different aspects of these issues to provide a fundamental conceptual and evidential basis for the subsequent stages of TIER2. Our aim here is to report these findings as briefly as possible, linking to the full presentation of results available via the various pre-prints we have already made publicly available.

Note that full results for the studies described in Sections 4 and 5 are unfortunately not yet available, due to unforeseen delays. It is our intention to update this report once those results are available (in Spring 2024).

The deliverable proceeds as follows:

- Section 2, “Definitions of reproducibility”, presents work to capture the meanings of reproducibility and related terms like replication, replicability, reproduction, repeatability and repetition across different contexts, diverse research approaches and a wide range of disciplines.
- Section 3, “Epistemic diversity and Knowledge Production Modes”, examines how the appropriateness of ‘reproducibility’ relates to specific aspects of epistemic diversity. It proposes a knowledge production modes (KPM) framework, an analytical construct essential for identifying characteristics of the research that affect the appropriateness of ‘reproducibility’ and for contextualizing diverse research activities within and across different fields as well as research situations.
- Section 4, “Scoping review and evidence mapping of interventions aimed at improving reproducible and replicable science”, presents ongoing work to scope (using the PRISMA-ScR Scoping Review methodology) formal evidence of interventions that aim to boost reproducibility. This has involved screening of over 36,000 initial records. (Note that due to unforeseen delays this work is unfortunately incomplete, with data extraction still underway. Full results will be ready to share in Spring 2024).
- Section 5, “Review of conceptions and facilitators of and barriers to reproducibility of qualitative research”, presents work to review the conceptual framing and definitions of reproducibility with regard to qualitative research, and key facilitators of it, including Open Science practices, and barriers to it.
- Section 6, “Review of conceptions and practices regarding reproducibility in Machine Learning (ML)-driven research”, presents work to examine the situation of ML reproducibility in different research fields, which issues arise across research contexts, as well as their barriers and drivers (including tools, practices, and interventions).
- Section 7, “Changing behaviour in the academy: A strategy for improving research culture and practice”, presents work to develop a comprehensive, theoretically informed strategy for maximising the adoption of research improving behaviours, including those contributing to greater reproducibility.
- Section 8, “Discussion and conclusion”, makes clear the relevance of this work to the research community at large, as well as for the future work in TIER2

2. Definitions of reproducibility

2.1.Research questions

The main goal of this component of work task 3.1. is to capture the meanings of reproducibility and related terms like replication, replicability, reproduction, repeatability and repetition across different contexts, diverse research approaches and a wide range of disciplines. Hence, the overarching research question is:

- **RQ1:** What do reproducibility and related terms mean?

Importantly, this question is further divided into two aspects of the meanings of these terms, namely the practices involved, and the purposes associated (intended) with them. Therefore, we seek to answer

- **RQ1a:** What practices are referred to by these terms?
- **RQ1b:** What functions (purposes) are stated to be achieved or intended in relation to these terms?

2.2.Background

Since eminent failures to replicate in some biomedical and social sciences in the early 2010s, there is an increased (re)focus on the practice of replication and reproducibility as well as their role in research (see e.g., Open Science Collaboration, 2015). The current review is by no means the first attempt at capturing the replication and reproducibility talk (e.g., Nelson et al., 2021). Previous reviews about the usage of these terms indicate that there is conceptual confusion concerning the terminology of replication and reproducibility across the research landscape. There are numerous conflicting definitions of kinds of reproducibility and replication circulating in the literature across and within disciplines. As Renee Borges put it: “There is therefore confusion in the definition of the terms themselves, although everyone believes that they know what is being said” (Borges, 2022 p.1).

Numerous scholars have tried to capture the meanings and confusions around the replication and reproducibility terminology across disciplines and over time (e.g., Barba, 2018; Matarese, 2022). Barba (2018) for instance, provides an overview of some conceptualizations of the terms “replicate” and “reproduce” and locates their use across scientific disciplines. She distinguishes between three different kinds of usages of these terms in the literature. The first, which one might call the ignorant or careless usage, is that they are used indistinguishably without acknowledging any kinds of differences in meaning or understanding. The second is that reproduce refers to the reuse of the originator’s code and data to obtain the same outcome, while replicate describes the regeneration of the same result using different analyses on different (new) data. The third kind of usage is the inverse of the second kind. Matarese (2022) also distinguishes three kinds of classifications regarding reproducibility, replicability, repeatability, and replication. Matarese (2022) differentiates between Method-Result-Inferential Reproducibility (MRIR) from Goodman et al (2016), Exact-Direct-Conceptual Replicability (EDCR), and Repeatability-Replicability-Reproducibility (RRR), before presenting their own definition. Furthermore, in recent years there has been a discussion, pushed to the foreground due, in part, to claims about a so-called replication crisis in some disciplines and policy changes in line with a replication drive, about how universal or general the practices and criteria of replication and reproducibility should be across the sciences and humanities (the research landscape) (e.g., Guttinger, 2020; Peels & Bouter, 2018; Penders et al., 2019; Sikorski & Andreoletti, 2023).

2.3.Methods

In an approach comparable to Albertoni et al. (2023) used for Machine Learning, but more elaborate, we have analysed an extensive corpus of literature starting with existing reviews about the meaning and usage of replication, replicability, reproduction and reproducibility (see e.g., Barba, 2018; Matarese, 2022). This broad set of pivotal documents was used as seed for a subsequent snowball sampling of primary documents containing definitions across all scientific fields until a saturation point was achieved resulting in over 400 definitions of these and related terms like repetition, repeatability and reanalysis. Saturation means a sufficient pool of definitions for analysing the underlying dimensions of these terms.

Next up, we set out to analyse the underlying dimensions of terms. We reviewed definitions from a wide range of over 40 disciplines over the last 55 years. The sampling and synthesis approaches are fully described in the respective preprints (see Section 2.5).

2.4.Results

Our analyses provide two main findings:

1. When examined across all fields, the terminology surrounding ‘reproducibility’ is actually even more complex, conflated and confused than we anticipated based on the existing reviews. We conclude that a new broad definition based on the current confused terminology and its many qualifiers is futile (RQ 1).
2. Two underlying dimensions when taken together can clarify what type(s) of ‘reproducibility’ or ‘replication’ one is talking about: the actual practice and the perceived function (purpose) (RQ 1a, b).

Ad 1. There are conceptual confusions in multiple ways. There are higher level distinctions, with huge variation in meanings and implications, between replicability, reproducibility, replication, reproduction, and repeatability (Barba, 2018; Plesser, 2018), but there are also confusions at lower levels between different kinds of replication and reproducibility that are supposed to be distinguished by qualifiers like exact, conceptual, direct, computational, and operational, such as computational reproducibility, inferential reproducibility, direct and conceptual or exact and inexact replication, etc. We see diverse and contradicting meanings of these terms. More precisely, we found the same term with multiple different meanings and multiple different terms with the same meanings across and within disciplines, as well as approaches over a period of the last 55 years on both levels. Worsening the immense confusion is the circumstance that ever-new terms and definitions are introduced in the literature. One of the reasons why new definitions continue to emerge and why there is redundancy and overlap in terms (Albtertoni et al., 2023) is that diverse kinds of research and research situations relate differently to the notion of ‘reproducibility’ (Leonelli, 2018; Penders et al., 2019). Likewise, numerous scholars attempt to provide definitions of reproducibility or replication that are appropriate for the nature of their research (e.g., Huijnen & Huistra, 2022; Schöch, 2023; TalkadSukumar & Metoyer, 2019). Hence, any new terminology or typology might be short lived and inevitably controversial and contestable according to the local, epistemological, ontological, disciplinary, methodological, theoretical and institutional context (see e.g., Penders et al., 2019). Therefore, any new typology or terminology that is based on the terms replication or reproducibility might just make the situation even more confusing and less practicable.

Ad 2. As a constructive solution, and somewhat similar to Matarese's (2022) distinction between enabling and a replication test, and Peels' and Bouter's (2018) distinction between a replication (study) and replicability, we propose to use two generic terms, **redoing,** inspired by Schickore (2011) and **enabling** to indicate the fundamental practices underlying the 'reproducibility' terminology thereby discarding the confused terminology. Redoing is the actual redoing of the whole or parts of a study (similar/same or varied). Enabling are practices that either enable redoing, or other types of intersubjective accountability usually via forms of sharing or reporting (transparency). This decoupling of enabling from redoing was also inspired by Pratt et al's. (2020) proposed decoupling of transparency from replication in qualitative research.

The practices and outcomes of redoing are mostly differentiable based on what components of a study are supposed to be kept the same, similar or vary. Among these aspects of a study are investigators, methods, procedures, analyses, instruments, measurements, technologies, software, lab, data, hypotheses, theories, operationalizations, results, protocols, interpretations conclusions /inferences, environments, cultures, samples, populations, variables, and ancillary assumptions. Importantly, redoing needs to be qualified by what aspects of a study are to be kept the same, similar or vary (Albertoni et al., 2023 separate between the *team* and the *workflow component)*.

The function of 'reproducibility' (see Matarese, 2022; Schmidt, 2009 for functional accounts) is the actual purpose(s) of the overall practice (what Albertoni et al., 2023 call *reason*). Among the many functions we identified are: Enable replication (redoing), identify findings likely to be wrong, determine effect sizes, detect, rule out and understand bias, increase certainty, assess and establish the integrity of scientific knowledge or science, empirically justify/ qualify results, test theory, ensure reproducibility, generalizability, robustness, reliability, falsification, confirmation/ verification/ corroboration, various kinds of validity, check scientific merit, demonstrate flexibility in methods and variability, trustworthiness/ increase trust, rigor, credibility, estimate, ruling out or reducing errors of various kinds, theory development & refinement, improve understanding/ knowledge, learning/ training, confidence in findings, understanding or conclusions, likelihood of findings being true/ facts (getting closer to “the” truth), prevent waste, calibration, traceability/ transparency, knowledge accumulation, fraud detection or elimination, identify and reduce questionable research practices , establish stability/ consistency, objectivity, establish existence, improve sensitivity of analysis or measurement, improve design/ methodology, accuracy, alleviate different kinds of underdetermination, reveal uncertainties, scientific progress, and precision. Unfortunately, the literature does not indicate the slightest agreement about what practices of redoing or enabling precisely fulfil what function. Therefore, the many identified functions further complicate the conceptual confusion of 'reproducibility' and 'replication'.

Our review suggests that due to the conceptual confusion about practices and function as well as the fact that there are so many types of ‘reproducibility’ across the research landscape any new typology or terminology that is based on the terms replication or reproducibility will just exacerbate the confusion. Despite the conceptual confusion, everyone seems to think to know what is meant with the reproducibility terminology (Borges, 2022). Therefore, it becomes practically meaningless to speak of the (ir)relevance, (in)feasibility, importance, or necessity of replication or reproducibility without specifying the type(s) of ‘reproducibility’’ and replication’ one is talking about. Using our distinction between enabling and redoing in addition to the associated functions, this boils down to detailing what is to be done (what is the practice?) and for what purpose (what is the function?).

If we inspect the example case of *conceptual replication*, which is relatively common in the social sciences (especially psychology), we identified 31 definitions for that concept alone. There is huge variation in how that concept is defined. Meanings vary, for instance, in what is the intended function, ranging from generalizability and robustness to accuracy and validity. Definitions also diverge about what combination of changed and identical components are supposed to be (re)done in comparison to an original study. Part of this confusion about what components of a study should vary or be the same (similar) is also what is supposed to be tested with conceptual replication, with claims including, among others, the same hypothesis, different hypotheses from the same theory, a specific research question or a research idea. Furthermore, some authors even distinguish between different types of conceptual replication. What is more, as pointed out by Barba (2018), for some the terms replicability or reproducibility themselves also mean what others understand by the concept conceptual replication. And finally, there are also different terms referring to the same or similar practices as conceptual replication that have the same or similar functions, like triangulation or multiple determination.

2.5.Next steps and implications for TIER2

The full draft of this work is available via the Open Science Framework:

- Ulpts, Sven. 2023 Unravelling the conceptual and epistemic confusion surrounding the replications and reproducibility discourse. Open Science Framework. https://osf.io/4gsp6/

The work is currently being prepared for submission to a relevant peer-reviewed academic journal.

This work on clarifying definitions also leads into future work in TIER2. The key distinction between enabling and redoing feeds into the construction of the Knowledge Production Modes framework (Section 3) and into the development of a new pilot activity to design and test a “decision support aid” for researchers interested to investigate the relevance and feasibility of reproducibility for their own work (pilot outlined in Sec. 8 “Discussion and conclusions”). In addition, it aids clarity on the intervention logic for all other pilots by offering clarity on which aspect(s) of reproducibility each pilot aims to address.

3. Epistemic diversity and Knowledge Production Modes

3.1.Research questions

The main purpose of the work task 3.1. is to establish a conceptual framework for reproducibility in the realm of epistemic diversity and across contexts. This work is divided into two components. The first component examines definitions and understandings of reproducibility across research areas (described above in Sec. 2); and the second component examines how the appropriateness of ‘reproducibility’ relates to specific aspects of epistemic diversity. Therefore, the central element of this component is the knowledge production modes (KPM) framework. The concept of KPM is developed as an analytical construct essential for identifying characteristics of the research that affect the appropriateness of ‘reproducibility’ and for contextualizing diverse research activities within and across different fields as well as research situations. Hence, ultimately this framework will allow for the assessment of the relevance and feasibility of ‘reproducibility’ across different modes of knowledge production. Due to the conceptual confusion and existence of plural types of ‘reproducibility’ (see Sec. 2) the KPM framework was formulated independent of any specific conceptualization of ‘reproducibility’ because it intends to be compatible with all of them. The KPM framework is subsequently used as the foundation for the development of a decision aid that enables researchers, funders, and publishers to make informed decisions in accordance with the appropriateness of ‘reproducibility’. Based on the conceptual analyses from the meanings of reproducibility review, in the decision aid we supplant the confused synonymous reproducibility terms with redoing and enabling to ensure simplicity and diversity. The main research aim of this component is:

- **RQ1:** The development of the knowledge production modes (KPM) framework for the assessment of the relevance and feasibility of reproducibility (the different activities of redoing and enabling redoing or general intersubjectivity (accountability) in the face of epistemic diversity.

3.2.Background

The so-called ‘reproducibility crisis' has fostered normative discussions in some fields about the supposed paramount importance of reproducibility for the quality, integrity, and trustworthiness of research leading to demands for mandates or incentives for reproducibility of all kinds of research. Such claims of general appropriateness and necessity of reproducibility across all research contexts are not without its critics. Qualitative and humanistic scholars have debated the issue for some time resulting in an important debate about epistemic diversity and how traditional positivistic notions of reproducibility may not necessarily fit well into other ways of knowing. Epistemic diversity stresses that there are a variety of ways to produce and justify knowledge. It is clear from these discussions that prior to any mandates or incentives for reproducibility, it has to be assessed whether such mandates and incentives are appropriate.

A few attempts have been proposed to frame reproducibility according to types of research based on approaches, methods or fields (see e.g., Guttinger, 2020; Leonelli, 2018; Penders et al., 2019; Tuval-Mashiach, 2021). While these attempts are interesting, they are also wanting as they narrowly focus on methods design and neglect amongst other things the important epistemological dimension of knowledge production (explained below). The latter is essential for understanding to what extent reproducibility may be relevant for a certain knowledge production mode. Consequently, the relevance of ‘reproducibility’ is first and foremost linked to epistemic traditions that can differ within fields and methods. They are therefore not suitable entities when it comes to assessing to what extent ‘reproducibility’ may be relevant to a particular kind of research in a given context. Here, we therefore suggest knowledge production modes (KPM) as a more suitable entity that captures crucial aspects of epistemic diversity that influence the appropriateness of reproducibility for diverse kinds of research. KPM encapsulates both the epistemic and social aspects of knowledge production and are local in the sense that they are organized around a subject matter, an epistemic position, and preferred methodologies within a specific research situation. Moreover, KPMs can form parts of research specialties.

3.3.Methods

The work in component 2 is based upon conceptual analyses combined with ongoing online discussions with subject experts (including authors of key previous studies) where we presented current versions of the framework and received constructive feedback which could then be used to adjust the framework. The basic idea behind KPMs is to join both epistemic and social dimensions when characterizing knowledge production, something which is often neglected. We do this by analysing and rethinking existing work from a wide range backgrounds like philosophy of science, sociology of science, and science and technology studies to establish the analytical entity of KPMs.

3.4.Results

We constructed the KPM framework that allows for the assessment of the appropriateness of different types of ‘reproducibility’ for diverse kinds of research in their specific research situations. We distinguish between two components of appropriateness. The relevance and the feasibility of ‘reproducibility’. The relevance is mainly assessed based on epistemic factors and the feasibility is assessed based on practical aspects of the research. The aspects for the assessment of the relevance are the epistemology, which is simply put what people want to know or the way of knowing, the criteria and practices that establish the quality and/ or trustworthiness of the research, and the research goal. The research goal can be non-epistemic and override epistemic considerations, for instance, in cases where commercial or proprietary interests are interfering or even motivating the research.

The feasibility is determined by the nature of the subject of investigation, the uncertainty related to the research, how much the research depends on certain resources, and therefore also on the availability and costs of such resources. The nature (complexity) of the subject of investigation concerns, for instance, whether it is stable or changing over time and whether it is interacting with its environment or rather indifferent. The uncertainty has two components. The theoretical uncertainty is related to how well the subject of investigation is understood and to what degree such understanding guides the investigation. In contrast, the methodological uncertainty is about how well the researchers understand the methods and procedures, for example, regarding how to use and justify them.

We provide two examples, one for relevance and one for feasibility.

Relevance: Many believe that the appropriateness of ‘reproducibility’ is linked to whether methods are quantitative or qualitative. However, the quantitative-qualitative distinction is not helpful because even within qualitative research the relevance of ‘reproducibility’ varies. For example, in the social sciences, like in International Relations, several ways of knowing are present within quantitative and qualitative research. Most pronounced are marked differences between positivist and interpretivist ways of knowing underlying qualitive research. While they can study the same phenomena with the same method, they can still base the research on different assumptions about what the nature the knowledge is. Positivist qualitative research usually assumes one underlying ‘truth’ that is supposed to be approximated or discovered and where the achievement of consensus in findings can provide epistemic value for the quality of the research and the resulting knowledge claims. In contrast, interpretivist qualitative research usually does not assume one underlying ‘truth’ and it is not about discovering facts, but about capturing interpretations or opinions, and where the diversity of findings can even be the goal. Hence, for such positivist qualitative research, *redoing* studies with the aim to obtain the same or similar findings, can be of value, while for such interpretivist qualitative research, *redoing* with the aim of obtaining the same or similar findings, would be counter to the intention behind the research.

Feasibility: Even though in cases where some form of ‘reproducibility’ is relevant, the feasibility, and therefore also what can be expected of ‘reproducibility’, varies depending on the type of reproducibility wanted, the research set-up and the context. For example, genome research is well situated within positivist notions of research where basic elements such as reliability, validity and robustness are fundamental epistemic values. The aim is to discover facts; hence some form of ‘reproducibility’ might be relevant for this way of knowing. However, while relevant, the feasibility of ‘reproducibility’ can still vary to a high degree. Guttinger (2020) provides the following example from genome research. The subject of investigation, i.e., the gene, contrary to popular belief, is actually highly reactive to its history and environment. This reactivity of the subject significantly affects how stable it is in different studies, and therefore also how feasible it is to establish ‘reproducible’ settings, and how feasible it is to expect similar findings. According to Guttinger (2020), this has important consequences for fields such as molecular biology, genomics, and biochemistry, where 'reproducibility’ in some form seems an ideal.

3.5.Next steps and implications for TIER2

This work has been published as a preprint:

- Ulpts, S., & Schneider, J. W. (2023). Knowledge Production Modes: The Relevance and Feasibility of ‘Reproducibility’. https://doi.org/10.31222/osf.io/ujnd9

The manuscript is currently being prepared for submission to a relevant peer-reviewed academic journal.

The KPM framework builds the conceptual foundation for a prototypical tool aimed at assisting in decisions about relevance and feasibility of redoing and/or enabling research across different modes of knowledge production which will be piloted later in TIER 2 (described in Section 8

“Discussion”). Further, it also provides a theoretical framework that the other work tasks of TIER2 can draw upon. Finally, the KPM framework will also be informing policy briefs for funders and publishers about reproducibility guidelines and policies in the face of epistemic diversity.

4. Scoping review and evidence mapping of interventions aimed at improving reproducible and replicable science

4.1. Research questions

A key aim of TIER2 Task 3.2 was to scope interventions, tools and practices to boost reproducibility. Dialogue with our sister project OSIRIS[2] in early 2023 identified that both projects planned similar efforts (c.f., OSIRIS Task 2.1). Given the wide scope of such interventions across contexts, as well as the divergent and complementary disciplinary expertise across the two projects, the decision was made to pool resources and collaborate. Given our belief that reforms and interventions should be based on evidence wherever possible, as well as awareness of the huge extent of the literature from pilot searching, the focus was on formal evidence of interventions. The main RQ was hence:

- **RQ1:** Which interventions have been formally investigated regarding their influence on reproducibility and replicability?

4.2. Background

Causes of poor reproducibility are diverse, including lack of transparency (e.g., poor reporting/publishing of methods/data/code/analysis), lack of reproduction/replication studies, publication bias towards reporting of positive results, and questionable research practices (Atmanspacher & Maase, 2016). In response, a range of interventions have been proposed and tested within the scientific literature. Strategies to increase reproducibility and replicability have primarily focused on improving research transparency through various open science practices. These aim to ensure that the research process is documented and widely accessible so that it can be checked, critiqued, re-used, and built upon in future research.

However, no systematic analysis of these interventions has yet been conducted. We know that although interventions to improve reproducibility are currently often targeted broadly, much of the work to understand issues of reproducibility have been predominantly led by select disciplines, especially psychology and clinical medicine (Cobey et al., 2022). However, these disciplines are only part of the funding landscape. There is hence a clear need to systematize knowledge of which interventions are appropriate in which contexts to determine the impact pathways whereby interventions result in increased reproducibility (and the extent to which this is desirable in different contexts). In addition, acknowledging that reproducibility has very different meanings and consequences across epistemic, social, and technical contexts, it is essential that any analysis of gains and savings be rooted in an understanding of how interventions vary according to these contexts.

This study hence uses scoping review and evidence mapping methodology to examine which interventions have been empirically investigated, and their actual impact on reproducibility and

[2] https://osiris4r.eu/

replicability. This will allow us to understand the evidence base for interventions for increasing the reproducibility and replicability of research and the documented barriers and facilitators in the process of creating more reproducible and replicable research using scoping review and evidence mapping methodology.

4.3.Methods

The protocol for the scoping review was developed prior to the search and has since been published via *Open Research Europe* (Dudda et al., 2023). Further methodological materials (e.g., data extraction sheets, data analysis plans), will be added to the Open Science Framework folder of this project[3]. The study follows the Preferred Reporting Items for Systematic Reviews and Meta-Analyses: extension for Scoping Reviews (PRISMA-ScR) guidelines (Tricco et al., 2018).

The scoping review methodology was applied to create an overview of the evidence on the effectiveness of interventions in improving reproducibility and replicability in science applied on various levels. Definitions of replicability and reproducibility were derived from Nosek et al. (2022) (2022) and the European Commission's scoping and final report on reproducibility in research (European Commission. Directorate General for Research and Innovation., 2020; European Commission. Directorate General for Research and Innovation. et al., 2022).

Eligible for inclusion were studies evaluating the effectiveness of an intervention which aimed to either aimed to increase reproducibility or increase practices associated with improved reproducibility (e.g., data sharing). Potential interventions were collected prior to the search, but additional practices were also included. A list of reproducibility outcomes was also compiled, however, studies were included independent of the specific approach applied to measure the outcome variable(s), such as using proxies or direct indicators. Proxies of reproducibility/replicability as outcomes eligible for inclusion were defined through the project team's synthesis and through referencing widely cited prescriptive texts that provide a conceptual framework for the selection made (European Commission. Directorate General for Research and Innovation., 2020; European Commission. Directorate General for Research and Innovation. et al., 2022; Munafò et al., 2017; Stodden et al., 2016). Various study designs were considered for inclusion, as long as an intervention was investigated. Studies were excluded that only reported prevalence of a practice or made claims about effectiveness without empirical evidence. Reviews were collected to serve as a basis for the snowball search.

The search strategy was developed based on the list of practices and outcomes and through collection and citation coupling of manually identified key articles. Due to the complexity of the search objectives, the search string was revised throughout the title/abstract screening in order to improve its quality. An example search string for the Medline database) is below in Box 1.

Box 1. Medline Search string.

(((data or code or workflow or practices or materials or notebook) adj2 (open or share or shared or sharing or preservation or stewardship)) or "open science" or ((computational or data or open or research or conclusion* or inferential or analytic or conceptual or direct or exact or statistical) adj3 (reproducib* or

[33] https://osf.io/d65ys/

replicability or replicable)) or (research adj5 (transparen* or credib*))).ti,ab. or (reporting adj3 guideline*).ti.

(Clinical study/ or Case control study/ or Family study/ or Longitudinal study/ or Retrospective study/ or Prospective study/ or Cohort analysis/ or Comparative Study/ or (Cohort adj (study or studies)).mp. or (Case control adj (study or studies)).tw. or (follow up adj (study or studies)).tw. or (observational adj (study or studies)).tw. or (epidemiologic$ adj (study or studies)).tw. or (cross sectional adj (study or studies)).tw. or (comparative adj stud*).mp. or (("randomized controlled trial" or "controlled clinical trial" or "multicenter study" or "pragmatic clinical trial").pt. or non- randomized controlled trials as topic/ or interrupted time series analysis/ or controlled before-after studies/ or random*.ti,ab. or groups.ab. or (trial or multicenter or "multi center" or multicentre or "multi centre").ti. or (intervention? or effect? or impact? or controlled or control group? or (before adj12 after) or (pre adj5 post) or ((pretest or "pre test") and (posttest or "post test")) or quasiexperiment* or quasi experiment* or pseudo-experiment* or pseudoexperiment* or evaluat* or "time series" or time-point? or "repeated measur*" or ((experimental or empirical or qualitative) adj5 (study or studies))).ti,ab.)) not ((news or comment or editorial).pt. or comment on.cm.)

Searched databases were: Medline, Embase, Web of Science, PsycINFO, Scopus, CAB Direct, Agris, PubAg, AGRICOLA and Eric. For grey literature the project consortia of TIER2, OSIRIS and iRISE were contacted and requested to contribute relevant literature they are aware of. The collected reviews will serve as a basis for a snowball search.

For screening and data-charting, the EPPI Reviewer software[4] was used. Given the very large number of initial results, Title/Abstracts were screened by one reviewer (but with liberal inclusion criteria applied). Full-text screening occurred in duplicate, with a third person (senior researcher) adjudicating in case of disagreements.

The data extraction sheet encompasses data on description of the publication, type of publication, study design, description of the sample, description of the intervention, outcome variables, disciplinary scope, and results.

All included studies will be summarized in a narrative synthesis. The evidence will be mapped according to the intervention (rows) and reproducibility/replicability (columns) using the EPPI Reviewer software.

4.4.Results

Search across all databases, once deduplicated, gave 36,063 initial records. Title/Abstract screening of these records left 1,643 included studies, and screening of full-texts left 172 included studies. Data extraction of these results is currently underway.

4.5.Next steps and implications for TIER2

Once data-charting is completed, the team will complete snowball and grey literature search. The review will then be written-up, shared via preprint (expected March 2024), and submitted to a relevant peer-reviewed journal for publication. The results will then also be showcased via the TIER2 “Reproducibility Hub” (to be developed in 2024, and hosted via the Embassy of Good Science[5]). This work will aid all the TIER2 pilot activities identify relevant evidence related to

[4] https://eppi.ioe.ac.uk/cms/Default.aspx?tabid=2914

[5] https://embassy.science/wiki/Main_Page

specific interventions (see Sec. 8), and will also spotlight areas where enough evidence exists for potential follow-up systematic review activities. In addition, the results from literature screening and data-charting will be provided to our sister project iRISE[6]. IRISE is creating an algorithmic approach to automatically curating evidence on such interventions. The data from this study will be rich training data for those activities.

[6] https://www.irise-project.eu/

5. Review of conceptions and facilitators of and barriers to reproducibility of qualitative research

5.1. Research questions

This integrative review took a targeted approach to reviewing the conceptual framing and definitions of reproducibility with regard to qualitative research, and key facilitators of it, including Open Science practices, and barriers to it. The work hence builds upon the work on definitions of reproducibility (Section 2) and its relevance and feasibility across Knowledge Production Modes by drilling down into an area of research where the applicability of reproducibility as an epistemic criterion is especially contested. An integrative review method, as described by (Toronto & Remington, 2020; Torraco, 2016; Whittemore & Knafl, 2005) was selected for this study because we include both conceptual/theoretical literature and empirical research, and this type of review method accounts for such a duality of literatures. This review was conducted in a targeted and systematic fashion – with focused search terms limited to title and keywords (where possible) – because it was carried out as a complement to the much larger cross-project scoping review described in Chapter 4 of this document. This review was preregistered through the OSF on July 13, 2023 (Cole, Ulpts, et al., 2023).

The research questions for this review are:

- **RQ1:** How is reproducibility conceptualized and discussed in relation to qualitative research?
- **RQ2:** Which factors and practices enable, and which undermine, the potential reproducibility of qualitative research?

5.2. Background

In response to what is viewed as a "reproducibility crisis" within some fields, many consider Open Science to offer solutions by fostering transparency of the research process. There has been, therefore, a normative shift towards evaluation, assessment, and reward in accordance with a demand for reproducibility (though open practices) from researchers by funders, institutes, and publishers (see e.g., Bissell, 2013; Guttinger, 2020; Penders et al., 2019). Effectively, such a shift changes who gets access to resources that are necessary to conduct certain kinds of research and who is excluded from conducting research. However, as made clear in Chapter 2 of this deliverable, research is not one unified entity, but there is a diverse landscape of different kinds of research relying on varying kinds of quality criteria with diverging epistemological and ontological positions, as well as purposes for which the research is conducted. The constellation of these factors affects not only what kind of reproducibility might be relevant, but also whether there is even any relevant place for it at all.

In response to both the perception of a "reproducibility crisis" and recognition that responses to it may foster epistemic injustices, TIER2 aims to contribute to increasing the re-use and overall quality of research results while centering epistemic diversity to ensure that definitions of reproducibility (and replicability) and expectations for them reflect the diversity of academic disciplines, research fields, and research practices that constitute scientific research.

We began this work by developing a conceptual framework for reproducibility across contexts, which has already highlighted the tensions between quantitatively driven definitions of and expectations for reproducibility and the values, norms, ethics and practices of qualitative research

(Section 3, c.f., Ulpts & Schneider 2023). Building on this, as part of broader work to construct an evidence-base and inventory of reproducibility tools and practices (Task 3.2), we focus in this integrative review on reproducibility as it relates to qualitative research.

Importantly, even qualitative research is not one unified entity, but rather a loose constellation of diverse approaches (Pownall, 2022; Pratt et al., 2020) that include established social science methods like ethnography, interviews, focus groups, discourse and content analysis, and case studies, as well as methodological approaches common to humanities including archival and comparative research, among others. Numerous scholars argue that the call for higher appreciation and application of replication and reproducibility stem from and are based on quantitative (post)positivist approaches to research. Applying foreign research quality criteria and practices to communities and approaches has the potential of harming them by pushing more appropriate and already established practices and criteria out, as well as putting a burden on them and/or even practically preventing them from conducting their research, due to ethical, practical and epistemic dependencies of qualitative approaches (see e.g., Bazzoli, 2022; Bennett, 2021). Therefore, prior to a widespread adoption of such practices, we must better understand how ‘reproducibility’ relates to qualitative research approaches and whether they are appropriate and applicable to prevent epistemic injustices (Penders et al., 2019).

There have already been some small reviews uncovering some of the facilitators and barriers to Open Science and kinds of reproducibility and replication in qualitative approaches (see e.g., Field et al., 2021). Some have attempted to identify or define reproducibility in ways relevant to and feasible for qualitative research (e.g., TalkadSukumar & Metoyer, 2019; Tuval-Mashiach, 2021). Others have conducted investigations into researchers’ perceived applicability of reproducibility to qualitative approaches (Reischer & Cowan, 2020). However, this literature appears quite scattered and somewhat underappreciated by reform movements. Accordingly, in this study we aimed to review the literature to identify, evaluate and synthesize conceptualizations of reproducibility in qualitative research, as well as identify barriers to and enablers of it within this set of research practices. We further aimed to provide insight into the relevancy and feasibility of reproducibility, and Open Science practices that support and enable it, in diverse qualitative research approaches.

5.3.Methods

This review was designed in keeping with the guidance for integrative reviews developed by Whittemore and Knafl (2005) and elaborated by Torraco (2016) and Toronto and Remington (2020). An integrative review is ideal for these purposes because reproducibility (and replicability) of qualitative research is an emerging topic with a growing body of literature surrounding it that is not uniform in stance nor conclusions. As Torraco observes, conducting an integrative review can lead to “an initial or preliminary conceptualization of the topic rather than a reconceptualization of existing models.”

The top-level keywords used for the search include: reproducibility, replicability, open data, open science, accountability, transparency, preregistration, qualitative research, mixed methods research.

Search strings based on these keywords were run in Databases: Scopus, Web of Science Core Collection, Dimensions, JSTOR, PubMed, and APA PsycInfo.

Additionally, a search for grey literature was conducted using these keywords in the following sources: CORDIS, EU Publications, Science Europe, EUA, National Academy of Sciences, JISC, Center for Open Science, OSF Preprint Archive, Open Research Funders Group, UKRI, and UNESCO.

The inclusion criteria for the search were defined by the research questions, and facilitated a search focused on literature that discusses the intersection of reproducibility with qualitative research/methods. We included the concept of “mixed methods” to capture literature that bridges qualitative and quantitative methods. Additionally, we included literature that addresses transparency and accountability in relation to qualitative research/methods, as these terms appear to be conceptually linked to discussions of reproducibility of qualitative research.

To capture discussions and evidence of the use of Open Science practices to support reproducibility of qualitative research, we focused our search on those practices that are known to be in use and/or possible for qualitative research, including Open Science generally, open data, open methods, and pre-registration. We did not include Open Science practices that are not known to be relevant to fostering the reproducibility of *qualitative* research, namely open code/software/tools, open evaluation, and Open Access publishing.

We limited our search to English-language texts due to the language capacities of our research team and imposed no fixed time span on the publication date of literature included.

A snowball search procedure was conducted following full-text screening and data charting of academic literature. The results of this search were also subjected to full-text screening and data charting.

The screening and data charting of search results were carried out in the SyRF online review platform in the following phases:

1. Title and abstract screening of initial academic search results
2. Full-text screening and data charting of included academic and grey literature
3. Full-text screening and data charting of snowballed literature

For a full account of our research methods, and inclusion and exclusion criteria used throughout the review process, see our published protocol (Cole et al. 2023).

To date, we conducted preliminary coding in NVivo and analysis on the papers included from the original academic and grey literature searches to develop a code system and pilot our data analysis procedure. We have since synthesized all data from included papers and finalised a code system and data analysis plan for collaborative qualitative analysis of our data. We will analyse quantitative data using Python. Here, we share some preliminary findings derived from our preliminary analysis.

5.4.Preliminary results

The initial search of academic databases resulted in 3215 unique papers (after deduplication) that were put forward for abstract screening. The abstract screening process resulted in 289 included papers that were moved forward to full-text screening and data charting. Simultaneous to this process, we conducted the search for grey literature which yielded 276 results; 31 of which were included after abstract screening. Therefore, we screened and charted 294 total papers, including academic (deduplicated, with full-text available) and grey literature, which resulted in 170 included papers.

We conducted a snowball citation search of included academic papers flagged for this purpose which resulted in 124 additional papers identified for full-text screening and data charting (in progress as of this draft). Of these, we included and charted 82 papers for a total number of included papers of 254.

In response to RQ1, *How is reproducibility conceptualized and discussed in relation to qualitative research?*, preliminary analysis indicates that a minority of papers frame reproducibility (and/or replicability) positively in the context of qualitative research, meaning that they frame it is favorable and/or possible. The authors that frame reproducibility this way typically recognize that it is neither desirable nor possible for all types of qualitative research, but that for some, it is in fact both, when open research practices are implemented to provide enough information and context about the study to make shared data or materials usable for this purpose.

In contrast, most papers either take a critical approach to reproducibility (and/or replicability) as it relates to qualitative research, or do not mention it at all, instead focusing on other research values that are more relevant and central to the diverse epistemologies of qualitative research. These include transparency (the most discussed), quality, rigor, and credibility, rather than reproducibility and replicability, noting that these are longstanding values in qualitative research epistemologies. These values may be associated with the pursuit of reproducibility or replicability, but they are also framed as stand-ins for these within qualitative research.

In response to RQ2, *Which factors and practices enable, and which undermine, the potential reproducibility of qualitative research?*, preliminary analysis indicates that the longstanding pursuit of transparency, rigor and credibility are enablers of the (possibility) of reproducibility and replicability, and that established Open Science practices like data sharing, preregistration, open methods and open analysis are useful in achieving them throughout the research lifecycle. Some point out that traditional practices within qualitative research, like field notes, analytic memos, research diaries, and researcher reflexivity are already designed to produce transparent, rigorous and credible research, and that they overlap with and expand upon established Open Science practices.

Key barriers referenced include the role of the researcher (as often an instrument) in qualitative research methods and the interpretive nature of much qualitative research, which would make it very difficult, if not impossible, for another researcher to reproduce or replicate such a study. Another oft cited barrier is that qualitative research is (for the most part) not generalizable due to its context-specific nature, and therefore not replicable. Other frequently referenced barriers are tied to the concept of Open Data or data sharing more broadly. Common in the literature are concerns about the implications for research participant privacy and confidentiality, the process and scope of informed consent, data anonymization processes and time costs (and implications for data usability), and the challenge of documenting community research context and researcher positionality and standpoint well enough to enable reproducibility or replicability. Some also discuss data ownership issues particular to qualitative research, given that researchers and participants often co-create the data together.

5.5.Next steps and implications for TIER2

Full analysis and synthesis of the data has begun, and reporting and publication will be completed during the first quarter of 2024. The preliminary results from this study indicate that reproducibility (and replicability) is not viewed as a priority, nor even possible, by most of the qualitative research community. While it might be possible for some qualitative research projects, credibility, rigor, transparency and overall quality are given primacy of place among research values. This is a clear sign to TIER2 that there is no good use in attempting to force reproducibility as a value on qualitative research. However, the emphasis on transparency within the literature, and its connection to Open Science values and practices, is something that TIER2 could celebrate and support. One could frame this as the *pursuit* of reproducibility as a largely unattainable ideal, in which practices of transparent research foster increased research quality and *transferability* of findings. In fact, as some authors in our study observe, the transparent and reflexive practices of qualitative research can both broaden and deepen Open Science practices in ways that can, when adopted, help foster reproducible research across quantitative and mixed-method research, and for some qualitative research.

6. Review of conceptions and practices regarding reproducibility in Machine Learning (ML)-driven research

6.1.Research questions

As in many other scientific fields, research in artificial intelligence (AI) in general, and machine learning (ML) in particular, has been argued to be facing a reproducibility crisis (Hutson, 2018a). This has raised doubts about the reliability and validity of many scientific findings, directing to a need for more confidence in the overall body of scientific knowledge. Misleading or unreproducible results can lead to wasted resources, hinder scientific progress, and impact decision-making in various fields.

ML introduces many new barriers to reproducibility, mainly unpublished source code and sensitivity to ML training conditions, making it nearly impossible to reproduce existing ML publications and very hard to verify the claims and findings stated in the publications. Moreover, in a recent study, Gundersen et al. (2022) found that the same experiment executed on different ML platforms leads to different results. This finding suggests that much research is still needed until out-of-the-box reproducibility can be provided. Apart from that, ML is nowadays used in various research fields, but it is still unclear in which research fields also reproducibility aspects of ML are discussed. Therefore, we raise the following research question:

- **RQ1:** *What is the situation of ML reproducibility in different research fields?*

By building on the results of this research question, we are further interested in studying the particular reasons that hinder the reproducibility of studies employing ML. Thus, we formulate the second research question of this work:

- **RQ2:** *Which reproducibility issues exist in research fields applying ML, and what are the barriers that cause these issues?*

Having identified the different barriers of ML reproducibility in research, we want to identify potential solutions for this problem by stating the third research question of this work:

- **RQ3:** Which drivers (tools, practices, and interventions) exist to support ML reproducibility?

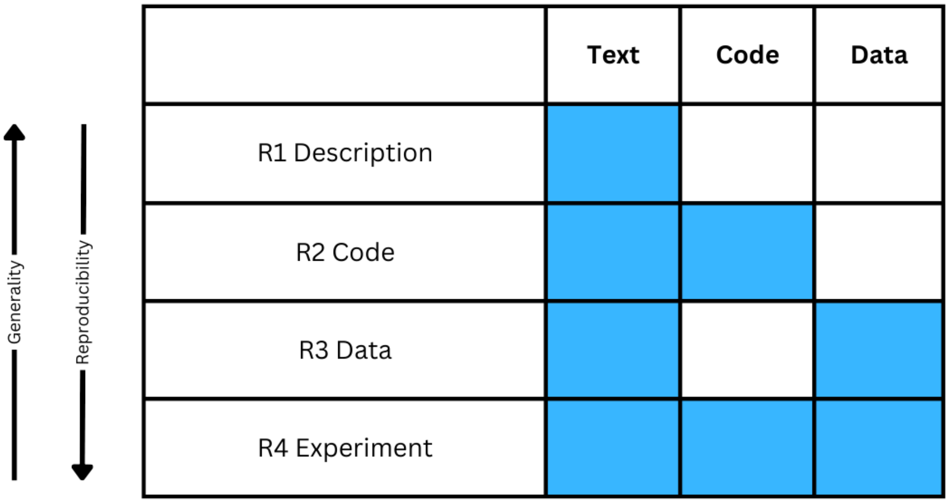
The following subsections briefly summarize the findings of a pre-print that we have recently submitted to arXiv (Semmelrock et al., 2023).

6.2.Background

Different definitions of reproducibility exist in ML, but the one by Gundersen (2021) is the most agreeable to researchers in the field. This definition aligns with four types or levels of ML reproducibility, that are also visualized in Figure 1:

- **R1 (Description):** Only a textual description of the original experiment is used for reproducing this experiment. The text could describe the experimental procedure, the target system and its behaviour, the implementation of the target system, for example, in the form of pseudocode, the data collection procedure, the data, the outcome, the analysis and so on.
- **R2 (Code):** The code and the textual description of the experiment are used as reference when conducting the reproducibility experiment. The code could cover the target system, the workflow, data pre-processing, experiment configurations, visualization and analyses.
- **R3 (Data):** Data and the textual description of the experiment are used as reference for conducting the reproducibility experiment. The data could include training, validation and test sets, as well as the outcome produced in the experiment.
- **R4 (Experiment):** The complete documentation of the experiment, including data and code in addition to the textual description shared by the original investigators, is used as a reference for the reproducibility experiment.

As seen in Figure 2, below, R4 provides the highest degree of reproducibility, while R1 provides the highest degree of generality. Thus, there is a clear trade-off between these dimensions.



	Text	Code	Data
R1 Description	🟦		
R2 Code	🟦	🟦	
R3 Data	🟦		🟦
R4 Experiment	🟦	🟦	🟦

Generality
Reproducibility

Figure 2. The four levels/degrees of ML reproducibility according to Gundersen (2021).

6.3.Methods

To conduct this review, we performed literature searches in Google Scholar and Scopus, taking advantage of the benefits of the two different search engines. To find relevant literature, we used keywords such as "machine learning", "artificial intelligence", "barriers", "challenges", "drivers", "tools", "FAIR", "reproducibility", and "replicability". In addition, we searched for these keywords in combination within naming the different scientific disciplines (e.g., health) to retrieve the domain-specific literature. We manually reviewed the retrieved papers for relevance and continued to find relevant literature by scanning the reference lists of the papers we used (i.e., snowballing). A sample query used in Scopus is given as follows: "TITLE-ABS-KEY ((reproduci* OR replica*)

AND ("machine learning" OR "artificial intelligence" OR "AI") AND (tool* OR practice* OR intervention* OR barrier*))".

In order to deal with the large number of results, we first reviewed the literature with the highest number of citations and adjusted the search queries for more fine-grained results. We also restricted our search to more recent literature, i.e., all the literature we used was published after 2005. This way, the literature is more relevant to the modern reproducibility crisis, while it has been an issue since the 2010s when it became a big topic in psychology.

6.4.Results

RQ1: *Research fields discussing ML reproducibility*

ML reproducibility is discussed in several research fields, including Chemistry, Material Science, Neuroimaging, Computer Science, and Health and Life Science. However, for the sake of space, in the following, we focus on Computer Science as well as Health and Life Science:

Computer Science

In Computer Science, reproducibility is mainly discussed in two specific sub-fields of ML, i.e., Deep learning and Reinforcement learning. Additional literature can be found in research fields such as recommender systems (Cremonesi & Jannach, 2021).

Deep learning (DL) is a subfield of ML which uses neural networks, specifically deep neural networks with many hidden layers, to solve problems. These neural networks are widely used in ML and can be applied to supervised and unsupervised learning tasks. Neural networks, however, are known to be inherently non-deterministic and produce different computational outcomes on multiple reruns due to the many sources of randomness during training (Ahmed & Lofstead, 2022).

Reinforcement learning (RL) is a subfield of ML where reproducibility issues are discussed extensively. RL is especially susceptible to reproducibility issues, partially because of additional sources of non-determinism in the learning process, which other areas are not subjected to, such as the environment or policy (Nagarajan et al., 2018). In general, RL differs from traditional supervised and unsupervised learning techniques, since the model is not simply trained on data. Instead, RL models usually learn by trial and error.

Health and Life Science

The uptake of ML in health and life science has heavily increased in recent years. For example, in a systematic review, Rahimi et al. (2022) observed ML models to be very promising in the clinical management and prediction of diabetes. In the review, however, the authors also uncovered a clear need for better reporting and validation of these models. Otherwise, integration into routine clinical care would not be possible. Provenzano et al. (2021) drew the same conclusion in a comparison between the performance of ML models and radiologists in cancer imaging. While ML models seem to consistently outperform radiologists, the publications presenting these ML models do not allow for reproduction of the algorithms.

RQ2: *Barriers of ML reproducibility*

The barriers towards ML reproducibility can be categorized using the four levels of reproducibility described beforehand (**R1 – R4**). Within **R1** (Description), poor reporting practices are mainly discussed in the literature (Andaur Navarro et al., 2023). Within **R2** (Code), a study by Hutson (2018b) found that published ML research is often not accompanied by available data and code. Only one-third of researchers share the data, and even fewer share the source code. This phenomenon can have a lot of reasons, such as private data or code that itself is based on unpublished code. Furthermore, the problem may also be attributed to the increasing pressure on researchers to publish quickly, which in turn does not allow them to invest in polishing and preparing the code and decreases the willingness to release the code.

Concerning **R3** (Data), the issue of data leakage is discussed in the literature. Data leakage happens when some form of data, which the ML model should not be trained on, leaks into the training process (Kapoor & Narayanan, 2023). Data leakage has become a widespread issue of research applying ML recently, which can partly be attributed to the increasing number of non-experts using ML for various research fields. The ease of application of ML libraries and no-code off-the-shelf AI tools accelerate this development. Finally, for **R4** (Experiment), studies have shown that both hardware differences, such as different GPUs or CPUs, and different compiler settings, can result in varying computation outcomes (Hong et al., 2013). Furthermore, comparing the same ML algorithm with fixed random seeds executed using different ML frameworks also resulted in different performances (Pouchard et al., 2020).

RQ3: *Drivers for ML reproducibility*

We can mainly distinguish between technology-based drivers and procedural drivers for ML reproducibility.

Technology-based drivers

Concerning technology-based drivers, reproducing the exact same environment as the original author of a study is not trivial, as it also means replicating the operating system, software versions and dependencies. Virtualization in the form of virtual environments, called virtual machines, can help overcome this problem. If the original author runs the experiment in a virtual machine, the environment can easily be shared with other researchers. However, for these solutions to be adopted by researchers, they must be easy to use and integrate into current workflows (Boettiger, 2015). Here, virtualization tools such as Docker or ReproZip (Chirigati et al., 2016) are discussed in the literature. In cases where datasets cannot be shared due to privacy concerns (e.g., in the health field), the creation of synthetic data, which captures the same information as the original data, is proposed in the literature (Walonoski et al., 2018).

Procedural drivers

With respect to procedural drivers, the idea of checklists and guidelines has been discussed in the literature. On top of that, Pineau et al. (2021) proposed an ML reproducibility checklist, which should ensure the inclusion of necessary information for reproducing the work, and has been suggested as best practice by researchers of different fields (Artrith et al., 2021). Similar to checklists, model info sheets are questionnaires specifically tailored towards handling data leakage, i.e., the detection and prevention of it.

6.5.Next steps and implications for TIER2

A preprint reporting this work more fully is available via arXiv:

- Semmelrock, H., Kopeinik, S., Theiler, D., Ross-Hellauer, T., & Kowald, D. (2023). Reproducibility in Machine Learning-Driven Research (arXiv:2307.10320). arXiv. https://doi.org/10.48550/arXiv.2307.10320

The manuscript is currently being prepared for submission to a relevant peer-reviewed academic journal or conference. As a next step, we aim to map the various drivers and barriers. Based on this mapping, we will be able to contribute to the understanding of potential solutions (i.e., drivers) for reproducibility barriers identified in TIER2. This will also contribute to the development and implementation of the TIER2 Pilots (see Sec, 8), especially Pilot 1 (Decision aid for researchers, to be pilot-tested with ML researchers) and Pilots 3 and 4 (tools for reproducible workflows).

7. Changing behaviour in the academy: A strategy for improving research culture and practice

7.1.Research questions

This study aims to develop a comprehensive, theoretically informed strategy for maximising the adoption of research improving behaviours, including those contributing to greater reproducibility. Our inspiration for this study comes from the Culture Change Strategy (henceforth CCS; Nosek, 2019). Strengths of the strategy include its simplicity, but limitations include its lack of a clear theoretical justification for itself and no clear route to implementation. We therefore aim to address these limitations and construct a strategy that offers theoretical justification for itself (via the COM-B model of behaviour) as well as links to a prominent framework of behaviour change (the Behaviour Change Wheel, henceforth BCW; Michie et al., 2011).

- **RQ1:** How can the COM-B model (Michie et al., 2011) be used to develop a comprehensive strategy for maximising the adoption of research improving behaviours?
- **RQ2:** How can this strategy be linked to the BCW (Michie et al., 2011) to provide a clear route from strategy to implementation?

7.2.Background

Improving reproducibility is a substantial and complex problem, involving a multiplicity of actors, behaviours, and influencing factors. Fundamentally, the collective behaviour of researchers and other research stakeholders will need to change (Corker, 2018; Norris & O’Connor, 2019; Osborne & Norris, 2022). However, for some, cultural influences prevent behaviour change, and so the behaviour of some will need to be leveraged to change research culture and so facilitate mass behaviour change.

The Culture Change Strategy (Nosek 2019) is perhaps the most well-known strategy aimed at maximising the adoption of research improving behaviours. It was used to inform the TIER2 proposal (Ross-Hellauer et al., 2022), has been a focus of several recent papers (Armeni et al., 2021; Cole, Reichmann, et al., 2023; Mellor, 2021; Robson et al., 2021; Shaw et al., 2022) and has become ubiquitous at Open Science related conferences and presentations. The CCS (see Figure 3) instructs us to first, provide the infrastructure for the behaviour to make it possible; second, improve the user experience to make it easy; third, foster communities of practice to make it normative; fourth, provide incentives to make it rewarding; and, last, enact policy to make it required. The CCS therefore informs us of who to target, what to do, when to do it, and why.

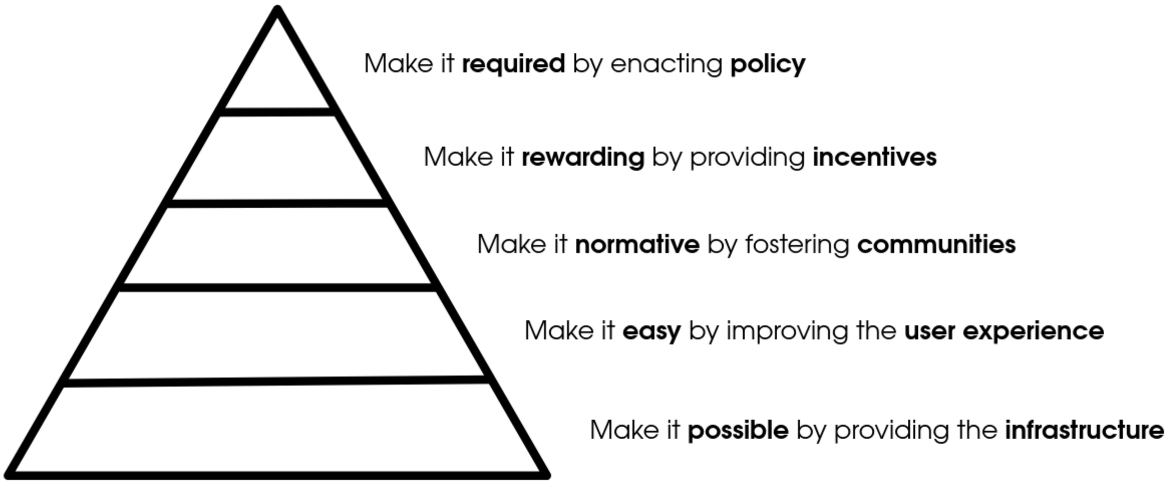

Make it **required** by enacting **policy**
Make it **rewarding** by providing **incentives**
Make it **normative** by fostering **communities**
Make it **easy** by improving the **user experience**
Make it **possible** by providing the **infrastructure**

Figure 3. The Culture Change Strategy pyramid, adapted from (Nosek, 2019)

We suspect that part of the CCS's success has been its intuitive simplicity as well as it being based on the well-established diffusion of innovations theory (Rogers, 2003). Briefly, the DOI explains how innovations (ideas, technologies, and behaviours perceived as new) spread within a population and how this may be used to influence their rate of adoption. However, whilst the Strategy is based on and extends (Rogers, 2003), it does not offer a theoretical justification for doing so and so its credibility is unclear. Furthermore, it offers no guidance on how it should be implemented: using 'common-sense' approaches or recognised behaviour change methods? Unfortunately, our 'common-sense' often leads us astray, resulting in ineffective interventions at best or harmful and costly ones at worst (Michie et al., 2014; West & Gould, 2022) and so recognised behaviour change approaches should be preferred.

Within this study, we aim to address the limitations of the CCS and so construct a new strategy that provides theoretical justification for itself as well as a clear path from strategy to implementation. In doing so, we aim to provide present and potential future users of the CCS with a more credible strategy and one that offers clear guidance on its implementation.

7.3.Methods

The interdisciplinary research process (Repko & Szostak, 2020) was used to inform our methods. Step 1 involved identifying insights from the DOI and BCW framework (including COM-B) that would be relevant to constructing a strategy to maximise the adoption of research improving behaviours. Step 2 involved identifying common ground between insights (what the DOI and BCW agreed upon). Step 3 involved identifying conflict between insights and to resolve this. The final step involved integrating all these insights to construct our strategy. We then linked our strategy to the BCW.

7.4.Results

Integrating insights from the CCS (Nosek, 2019) and BCW (Michie et al., 2011) first informed us that the Strategy should: be based on the COM-B model of behaviour (see Figure 4) and feature a limited number of comprehensive and coherent levels that instruct who to target, when to target them, what to do, how to do it, and why.

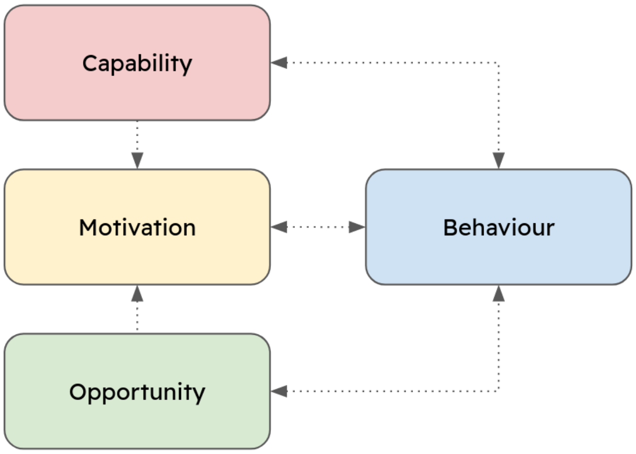


Figure 4. The COM-B model of behaviour, adapted from (Michie et al., 2011)

The levels of our strategy were developed from the COM-B model of behaviour, which we expanded using insights from the BCW (Michie et al., 2014; West et al., 2020; West & Gould, 2022) and DOI (Rogers, 2003) to identify a list of factors influential to behaviour. Figure 5 presents this extended model.

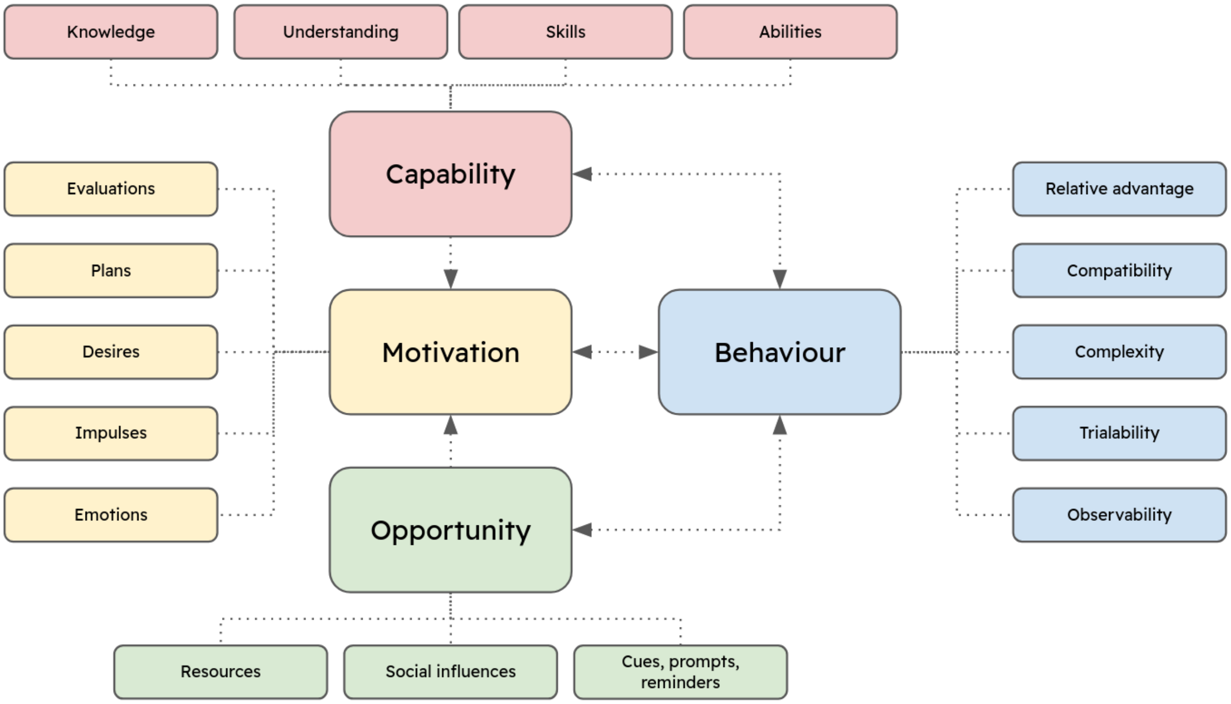
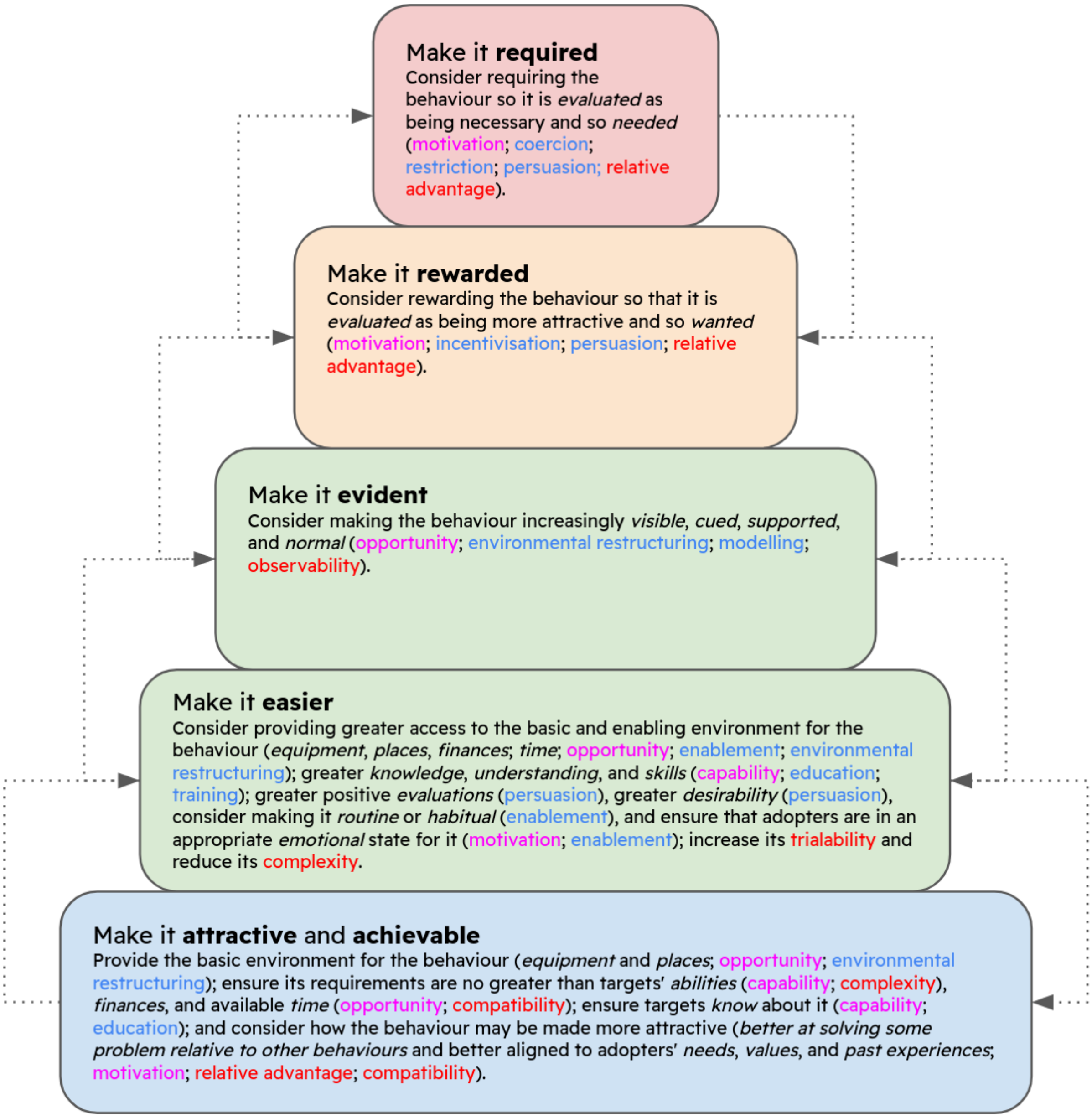


Figure 5. The extended COM-B model, based on insights from (Rogers 2003, Michie et al., 2014, and West et al. 2020).

These factors were then grouped into a limited number of coherent stages: (a) make it attractive and achievable; (b) make it easier; (c) make it evident; (d) make it rewarded; and (e) make it required. These factors and stages were then linked to the types of intervention listed within the BCW (Michie et al., 2014). All factors and types of intervention were included within the stages, and so the Strategy was considered to be comprehensive. Figure 6 presents the Strategy for Improving Research Culture and Practice.



Make it **required**
Consider requiring the behaviour so it is *evaluated* as being necessary and so *needed* (motivation; coercion; restriction; persuasion; relative advantage).

Make it **rewarded**
Consider rewarding the behaviour so that it is *evaluated* as being more attractive and so *wanted* (motivation; incentivisation; persuasion; relative advantage).

Make it **evident**
Consider making the behaviour increasingly *visible*, *cued*, *supported*, and *normal* (opportunity; environmental restructuring; modelling; observability).

Make it **easier**
Consider providing greater access to the basic and enabling environment for the behaviour (*equipment*, *places*, *finances*, *time*; opportunity; enablement; environmental restructuring); greater *knowledge*, *understanding*, and *skills* (capability; education; training); greater positive *evaluations* (persuasion), greater *desirability* (persuasion), consider making it *routine* or *habitual* (enablement), and ensure that adopters are in an appropriate *emotional* state for it (motivation; enablement); increase its trialability and reduce its complexity.

Make it **attractive** and **achievable**
Provide the basic environment for the behaviour (*equipment* and *places*; opportunity; environmental restructuring); ensure its requirements are no greater than targets' *abilities* (capability; complexity), *finances*, and available *time* (opportunity; compatibility); ensure targets *know* about it (capability; education); and consider how the behaviour may be made more attractive (*better at solving some problem relative to other behaviours* and better aligned to adopters' *needs*, *values*, and *past experiences*; motivation; relative advantage; compatibility).

Figure 6. The Strategy for Improving Research Culture and Practice, based on insights from Rogers 2003; Michie, Atkins, and West 2014; West et al. 2020, and West and Gould, 2022.

7.1.Next steps and implications for TIER2

The full draft of this work is publicly available via the Open Science Framework:

- Osborne, Christopher et al. 2023. "Changing Behaviour in the Academy: A Strategy for Improving Research Culture and Practice". Open Science Framework. https://osf.io/g8a5j/

We will continue to refine this work and then submit for publication. In creating an accessible entry-point for those concerned to design interventions for improving research cultures and practices, our insights will feed directly into future work in TIER2, especially the design, implementation and assessment of the Pilots (described in Sec. 8).

8. Discussion and conclusions

This Deliverable reports work on seven individual studies conducted within the first year of TIER2 that provide the theoretical, evidential, and strategic basis for the project. Some of this work (especially the scoping review of interventions) remains ongoing due to unforeseen delays. We intend to update this Deliverable in the first quarter of 2024 once these results are available.

The general need for this work has been underlined by TIER2's co-creation activities with stakeholders, and especially the work in Task 4.1 "Future studies to identify priorities from the stakeholder community to predict future of reproducibility and identify actionable steps". As reported in the Deliverable report from that work, D4.1 The Future(s) of Reproducibility in Research (Tijdink et al., 2023), our stakeholders were keenly aware that lack of clarity on the meaning and relevance of reproducibility across contexts can negatively impact decision-making at all levels[7], and could risk marginalising research areas where reproducibility is less applicable if translated into policies which treat reproducibility as universally applicable. The need for consolidation of evidence on which interventions to increase reproducibility are effective, in which contexts, was also noted. Our approach of conducting a broad review of the evidence, complemented by specific reviews on qualitative methods and Machine Learning, add both breadth and depth to our understanding in this regard.

In addition to their usefulness to the research community in general, these studies also play a crucial role for future activities within TIER2. We will incorporate these findings into our future training and awareness-raising activities and resources (to be curated via the upcoming "Reproducibility Hub"), and they will underpin our final synthesis and recommendations at the end of the project (Task 3.3). The findings also feed directly into the design of our Pilot activities. The TIER2 proposal indicated eight provisional Pilot activities, with the aim of refinement or amendment of these plans through co-creation with stakeholders. The current plans for the Pilots are listed in Table 1.

Table 1. TIER2 Pilots to be developed and implemented in the next stages of the project

Nr	Pilot title	Pilot aims	Target stakeholder
1	Reproducibility decision aid	Will aid clarity on meaning, relevance, and feasibility of 'reproducibility' for researchers to aid them in identifying what type of reproducibility is relevant for their research and indicate what they must consider regarding how feasible such 'reproducibility' would be for them. Tool will be piloted with two	Researchers

[7] As one publisher told us: "I almost feel it's so critical because it affects everything that is discussed subsequently, including the costs, et cetera , is this issue of what exactly we mean. I know there's some different definitions, but certainly when I speak with scientists, I don't even think there's like replicability and reproducibility. I think there's about five different things here. And, you know, the ways in which we frame the issue. And again, as you alluded to, this differs by discipline. You can get very different answers as to what the amount of effort that you're willing to spend, the amount of cost and whether you need to worry about this at all..." ((Tijdink et al., 2023, p.34)

		researcher groups (qualitative and Machine Learning researchers)	
2	Reproducibility management plans (RMPs)	Will extend Data Management Plans (DMPs) to include planning for elements related to reproducibility more broadly. Will extend the Argos DMP tool through user-testing/surveys for definition. Core functionalities will then be tested through prototyping and piloting with researchers and funders.	Researchers, Funders
3	Reproducible workflows (Life science / Computer Science)	Will customise and evaluate tools/practices for reproducible workflows in Life science and Computer Science. The goal is to enrich Schema platform to support researchers to facilitate the reproducibility of their analyses by providing them a Virtual Research environment (VRE) to run experiments, save in RO-crates and easily reproduce them.	Researchers
4	Reproducible workflows (Social science)	Will facilitate social science data analysis and its reproducibility by creating a community-supported catalogue of peer-reviewed computational methods (GESIS Methods Hub). The main aim in TIER2 is to design and test review workflows and researcher checklists for reproducible computational social science.	Researchers
5	Reproducibility Promotion Plans	Will produce and test practical advice for funders on how to create a plan to boost the reproducibility of their funded-results, and how to include reproducibility in their assessments and monitoring of proposals	Funders
6	Reproducibility monitoring dashboard	Will develop and test tools to enable funding agencies in tracking and monitoring reusability of research artefacts (datasets, software, tools/systems, etc) created in funded projects	Funders
7	Reproducibility Handbook for publishers	Will co-create and test (via intervention), with interested publishers, training material for editors. This handbook will be both educational and operational guidance in support of reproducibility and FAIRness.	Publishers
8	Data Availability Statements	Will provide journal editors with straightforward workflows to improve data availability of the research they publish. Data availability statements (DAS) related to the data underlying publications are currently not effective at promoting data sharing and reproducibility. Will design and test an intervention study where editors request enhanced information on reasons for not sharing and provide advice to authors on how issues may be overcome.	Publishers

The work reported here informs and supports the design of these Pilots in multiple ways.

Knowledge gained from the work on definitions (Section 2), including clarity on the different functions of reproducibility, and the work on Knowledge Production Modes (Section 3) to highlight aspects of research methods, epistemologies, and situational factors that shape the relevance and feasibility of reproducibility, it will be crucial for Pilots 2 (RMPs), 4 (reproducibility promotion plans), 5 (monitoring dashboard), and 7 (editorial handbook), which all attempt to create tools, policies or workflows which aim at addressing the full range of research contexts. Ensuring that any such tools are sensitive enough to the contextual factors outlined in this work is essential to ensure that wrong standards are not applied to the wrong kinds of research. For example, any Reproducibility Promotion Plan for funders will have to ensure that areas for which reproducibility is less applicable (e.g., qualitative research) are not penalised or marginalised through such efforts. This work also leads directly into Pilot 1 (Decision aid), which has been conceived as a direct continuation of this work (see Box 2).

Box 2. Description of Pilot 1 (Reproducibility decision aid)

Based on the conceptual analyses of reproducibility and knowledge production modes, we have constructed a prototype analytical tool. The tool will help the different stakeholders (i.e., researchers, publishers and funders) in three ways. First, it will help them see through the aforementioned conceptual confusion and understand the meaning of ‘reproducibility’ in question for their specific case. Second it will provide clarity about whether the specific type of reproducibility is relevant for the case at hand. Lastly it will illuminate whether the specific type of reproducibility is feasible in the specific situation.

The tool has been split into a policy brief for publishers and funders (gatekeepers) and an online decision aid for researchers. The policy brief provides overall recommendations for gatekeepers about how relevance and feasibility of reproducibility relate to diverse research characteristics across the research landscape emphasising epistemic diversity and how this should be taken into consideration if gatekeepers are planning to mandate or incentivize ‘reproducibility’. The intended purpose for the online tool is more concrete, to aid researchers in identifying what type of reproducibility is relevant for their intend and indicate what they have to consider regarding how feasible such ‘reproducibility’ would be for them.

The prototype of the online tool will be piloted during 2024. First, a cognitive test will be implemented locally at Aarhus University where researchers together with TIER2 members will systematically go through the tool to assess and then adjust it. Subsequently, an updated prototype will be piloted internationally in two research communities: Among groups of qualitative researchers, and groups machine learning researchers. Evaluation of the tool will be a combination of a short survey and interviews/focus groups where individual features, usability and overall relevance will be assessed. If needed, a second round of piloting will be performed on an updated prototype. If unsuccessful, the decision aid can still function as an analytical tool be transforming it into an information sheet guiding the stakeholders through considerations important for the relevance and feasibility of reproducibility given specific knowledge production modes.

The outcomes of our review activities will also inform the Pilots. The scoping review of interventions (Sec. 4) will (once complete) provide rich background on which types of interventions have demonstrated effectiveness in which contexts. The two topical reviews (on qualitative and Machine Learning research, Sec. 5 and 6), meanwhile, provide essential background on the relevance and feasibility of reproducibility in specific contexts and as such will be central to the development of Pilot 1 (Decision aid) which will be piloted on researchers using these methods. In addition, the topical reviews provide further context that will be relevant as edge cases for our pilots (RMPs, decision tool, dashboard, workflows, DAS, handbook).

Finally, our work to develop an accessible entry-point for those concerned to design interventions for improving research cultures and practices (Sec. 7), our insights will feed directly into future work in TIER2, especially the design, implementation and assessment of the Pilots.

8.1.Limitations

Note that full results for the studies described in Sections 4 and 5 are unfortunately not yet available, due to unforeseen delays. It is our intention to update this report once those results are available (in Spring 2024).

References

Ahmed, H., & Lofstead, J. (2022). Managing Randomness to Enable Reproducible Machine Learning. *Proceedings of the 5th International Workshop on Practical Reproducible Evaluation of Computer Systems*, 15–20. https://doi.org/10.1145/3526062.3536353

Albertoni, R., Colantonio, S., Skrzypczyński, P., & Stefanowski, J. (2023, February 24). *Reproducibility of Machine Learning: Terminology, Recommendations and Open Issues*. arXiv.Org. https://arxiv.org/abs/2302.12691v1

Andaur Navarro, C. L., Damen, J. A. A., Takada, T., Nijman, S. W. J., Dhiman, P., Ma, J., Collins, G. S., Bajpai, R., Riley, R. D., Moons, K. G. M., & Hooft, L. (2023). Systematic review finds "spin" practices and poor reporting standards in studies on machine learning-based prediction models. *Journal of Clinical Epidemiology*, *158*, 99–110. https://doi.org/10.1016/j.jclinepi.2023.03.024

Armeni, K., Brinkman, L., Carlsson, R., Eerland, A., Fijten, R., Fondberg, R., Heininga, V. E., Heunis, S., Koh, W. Q., Masselink, M., Moran, N., Baoill, A. Ó., Sarafoglou, A., Schettino, A., Schwamm, H., Sjoerds, Z., Teperek, M., van den Akker, O. R., van't Veer, A., & Zurita-Milla, R. (2021). Towards wide-scale adoption of open science practices: The role of open science communities. *Science and Public Policy*, *48*(5), 605–611. https://doi.org/10.1093/scipol/scab039

Artrith, N., Butler, K. T., Coudert, F.-X., Han, S., Isayev, O., Jain, A., & Walsh, A. (2021). Best practices in machine learning for chemistry. *Nature Chemistry*, *13*(6), Article 6. https://doi.org/10.1038/s41557-021-00716-z

Atmanspacher, H., & Maase, S. (Eds.). (2016). *Reproducibility: Principles, Problems, Practices, and Prospects*. Wiley.

Barba, L. A. (2018, February 9). *Terminologies for Reproducible Research*. arXiv.Org. https://arxiv.org/abs/1802.03311v1

Bazzoli, A. (2022). Open science and epistemic pluralism: A tale of many perils and some opportunities. *Industrial and Organizational Psychology*, *15*(4), 525–528. https://doi.org/10.1017/iop.2022.67

Bennett, E. A. (2021). Open Science From a Qualitative, Feminist Perspective: Epistemological Dogmas and a Call for Critical Examination. *Psychology of Women Quarterly*, *45*(4), 448–456. https://doi.org/10.1177/03616843211036460

Bissell, M. (2013). Reproducibility: The risks of the replication drive. *Nature*, *503*(7476), 333–334. Scopus. https://doi.org/10.1038/503333a

Boettiger, C. (2015). An introduction to Docker for reproducible research. *ACM SIGOPS Operating Systems Review*, *49*(1), 71–79. https://doi.org/10.1145/2723872.2723882

Borges, R. M. (2022). Reproducibility and replicability in science: A Sisyphean task. *Journal of Biosciences*, *47*(1), 15. https://doi.org/10.1007/s12038-022-00259-6

Chirigati, F., Rampin, R., Shasha, D., & Freire, J. (2016). ReproZip: Computational Reproducibility With Ease. *Proceedings of the 2016 International Conference on Management of Data*, 2085–2088. https://doi.org/10.1145/2882903.2899401

Cobey, K., Fehlmann, C. A., Franco, M. C., Ayala, A. P., Sikora, L., Rice, D. B., Xu, C., Ioannidis, J. P. A., Lalu, M., Menard, A., Neitzel, A., Nguyen, B., Tsertsvadze, N., & Moher, D. (2022). *Epidemiological characteristics and prevalence rates of research reproducibility across disciplines: A scoping review*. OSF Preprints. https://doi.org/10.31219/osf.io/k6nf4

Cole, N. L., Reichmann, S., & Ross-Hellauer, T. (2023). Toward equitable open research: Stakeholder co-created recommendations for research institutions, funders and researchers. *Royal Society Open Science*, *10*(2), 221460. https://doi.org/10.1098/rsos.221460

Cole, N. L., Ulpts, S., Ross-Hellauer, T., Bochynska, A., & Klebel, T. (2023). *Integrative review of conceptions and facilitators of and barriers to reproducibility of qualitative research*. Open Science Framework. https://doi.org/10.17605/OSF.IO/Q4XWK

Corker, K. (2018). Open Science is a Behavior. *Center for Open Science Blog*. https://www.cos.io/blog/open-science-is-a-behavior

Cremonesi, P., & Jannach, D. (2021). Progress in Recommender Systems Research: Crisis? What Crisis? *AI Magazine*, *42*(3), Article 3. https://doi.org/10.1609/aimag.v42i3.18145

Devezer, B., Nardin, L. G., Baumgaertner, B., & Buzbas, E. O. (2019). Scientific discovery in a model-centric framework: Reproducibility, innovation, and epistemic diversity. *PLOS ONE*, *14*(5), e0216125. https://doi.org/10.1371/journal.pone.0216125

Dudda, L., Kozula, M., Ross-Hellauer, T., & Kormann, E. (2023). Scoping review and evidence mapping of interventions aimed at improving reproducible and replicable science: Protocol [version 1; peer review: 1 approved with reservations]. *Open Research Europe*, *3*(179). https://doi.org/10.12688/openreseurope.16567.1

European Commission. Directorate General for Research and Innovation. (2020). *Reproducibility of scientific results in the EU: Scoping report.* Publications Office. https://data.europa.eu/doi/10.2777/341654

European Commission. Directorate General for Research and Innovation., PPMI., Know Center., & Athena RC. (2022). *Assessing the reproducibility of research results in EU Framework Programmes for Research: Final report.* Publications Office. https://data.europa.eu/doi/10.2777/186782

Field, S. M., Van Ravenzwaaij, D., Pittelkow, M.-M., Hoek, J. M., & Derksen, M. (2021). *Qualitative Open Science – Pain Points and Perspectives* [Preprint]. Open Science Framework. https://doi.org/10.31219/osf.io/e3cq4

Gómez, O. S., Juristo, N., & Vegas, S. (2014). Understanding replication of experiments in software engineering: A classification. *Information and Software Technology*, *56*(8), 1033–1048. https://doi.org/10.1016/j.infsof.2014.04.004

Goodman, S. N., Fanelli, D., & Ioannidis, J. P. A. (2016). What does research reproducibility mean? *Science Translational Medicine*, *8*(341), 341ps12-341ps12. https://doi.org/10.1126/scitranslmed.aaf5027

Gundersen, O. E. (2021). The fundamental principles of reproducibility. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *379*(2197), 20200210. https://doi.org/10.1098/rsta.2020.0210

Gundersen, O. E., Shamsaliei, S., & Isdahl, R. J. (2022). Do machine learning platforms provide out-of-the-box reproducibility? *Future Generation Computer Systems*, *126*, 34–47. https://doi.org/10.1016/j.future.2021.06.014

Guttinger, S. (2020). The limits of replicability. *European Journal for Philosophy of Science*, *10*(2), 10. https://doi.org/10.1007/s13194-019-0269-1

Hong, S.-Y., Koo, M.-S., Jang, J., Kim, J.-E. E., Park, H., Joh, M.-S., Kang, J.-H., & Oh, T.-J. (2013). An Evaluation of the Software System Dependency of a Global Atmospheric Model. *Monthly Weather Review*, *141*(11), 4165–4172. https://doi.org/10.1175/MWR-D-12-00352.1

Huijnen, P., & Huistra, P. (2022). *On the Use of Replications in History*. Zenodo. https://doi.org/10.5281/zenodo.7037401

Hutson, M. (2018a). Artificial intelligence faces reproducibility crisis. *Science*, *359*(6377), 725–726. https://doi.org/10.1126/science.359.6377.725

Hutson, M. (2018b). Missing data hinder replication of artificial intelligence studies. *Science*. https://doi.org/10.1126/science.aat3298

Kamel Rahimi, A., Canfell, O. J., Chan, W., Sly, B., Pole, J. D., Sullivan, C., & Shrapnel, S. (2022). Machine learning models for diabetes management in acute care using electronic medical records: A systematic review. *International Journal of Medical Informatics*, *162*, 104758. https://doi.org/10.1016/j.ijmedinf.2022.104758

Kapoor, S., & Narayanan, A. (2023). Leakage and the reproducibility crisis in machine-learning-based science. *Patterns*, *4*(9). https://doi.org/10.1016/j.patter.2023.100804

Köhler, T., & Cortina, J. M. (2021). Play It Again, Sam! An Analysis of Constructive Replication in the Organizational Sciences. *Journal of Management*, *47*(2), 488–518. https://doi.org/10.1177/0149206319843985

Leonelli, S. (2018). Rethinking Reproducibility as a Criterion for Research Quality. In L. Fiorito, S. Scheall, & C. E. Suprinyak (Eds.), *Research in the History of Economic Thought and Methodology* (Vol. 36, pp. 129–146). Emerald Publishing Limited. https://doi.org/10.1108/S0743-41542018000036B009

Leonelli, S. (2022). Open Science and Epistemic Diversity: Friends or Foes? *Philosophy of Science*, 1–21. https://doi.org/10.1017/psa.2022.45

Matarese, V. (2022). Kinds of Replicability: Different Terms and Different Functions. *Axiomathes*, *32*(2), 647–670. https://doi.org/10.1007/s10516-021-09610-2

Mellor, D. (2021). Improving norms in research culture to incentivize transparency and rigor. *Educational Psychologist*, *56*(2), 122–131. https://doi.org/10.1080/00461520.2021.1902329

Michie, S., van Stralen, M. M., & West, R. (2011). The behaviour change wheel: A new method for characterising and designing behaviour change interventions. *Implementation Science*, *6*(1), 42. https://doi.org/10.1186/1748-5908-6-42

Michie, S., West, R., Campbell, R., Brown, J., & Gainforth, H. (2014). *ABC of Behaviour Change Theories*. Silverback Publishing.

Munafo, M. R., Nosek, B. A., Bishop, D. V. M., Button, K. S., Chambers, C. D., du Sert, N. P., Simonsohn, U., Wagenmakers, E. J., Ware, J. J., & Ioannidis, J. P. A. (2017). A manifesto for reproducible science. *Nat Hum Behav*, *1*, 0021. PubMed-not-MEDLINE. https://doi.org/10.1038/s41562-016-0021

Munafò, M. R., Nosek, B. A., Bishop, D. V. M., Button, K. S., Chambers, C. D., Percie du Sert, N., Simonsohn, U., Wagenmakers, E.-J., Ware, J. J., & Ioannidis, J. P. A. (2017). A manifesto for reproducible science. *Nature Human Behaviour*, *1*(1), 0021. https://doi.org/10.1038/s41562-016-0021

Nagarajan, P., Warnell, G., & Stone, P. (2018). *The impact of nondeterminism on reproducibility in deep reinforcement learning*. https://prabhatnagarajan.com/publications/2018/nagarajanrml2018.pdf

Nelson, N. C., Ichikawa, K., Chung, J., & Malik, M. M. (2021). Mapping the discursive dimensions of the reproducibility crisis: A mixed methods analysis. *PLOS ONE*, *16*(7), e0254090. https://doi.org/10.1371/journal.pone.0254090

Norris, E., & O'Connor, D. B. (2019). Science as behaviour: Using a behaviour change approach to increase uptake of open science. *Psychology & Health*, *34*(12), 1397–1406. https://doi.org/10.1080/08870446.2019.1679373

Nosek, B. A. (2019). Strategy for Culture Change. *Center for Open Science Blog*. https://www.cos.io/blog/strategy-for-culture-change

Nosek, B. A., Hardwicke, T. E., Moshontz, H., Allard, A., Corker, K. S., Dreber, A., Fidler, F., Hilgard, J., Kline Struhl, M., Nuijten, M. B., Rohrer, J. M., Romero, F., Scheel, A. M., Scherer, L. D., Schönbrodt, F. D., & Vazire, S. (2022). Replicability, Robustness, and Reproducibility in Psychological Science. *Annual Review of Psychology*, *73*(1), 719–748. https://doi.org/10.1146/annurev-psych-020821-114157

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

Osborne, C., & Norris, E. (2022). Pre-registration as behaviour: Developing an evidence-based intervention specification to increase pre-registration uptake by researchers using the Behaviour Change Wheel. *Cogent Psychology*, *9*(1), 2066304. https://doi.org/10.1080/23311908.2022.2066304

Peels, R., & Bouter, L. (2018). The possibility and desirability of replication in the humanities. *Palgrave Communications*, *4*(1), 95. https://doi.org/10.1057/s41599-018-0149-x

Penders, Holbrook, & de Rijcke. (2019). Rinse and Repeat: Understanding the Value of Replication across Different Ways of Knowing. *Publications*, *7*(3), 52. https://doi.org/10.3390/publications7030052

Pineau, J., Vincent-Lamarre, P., Sinha, K., Lariviere, V., Beygelzimer, A., d'Alche-Buc, F., Fox, E., & Larochelle, H. (2021). Improving Reproducibility in Machine Learning Research(A Report from the NeurIPS 2019 Reproducibility Program). *Journal of Machine Learning Research*, *22*(164), 1–20.

Plesser, H. E. (2018). Reproducibility vs. Replicability: A Brief History of a Confused Terminology. *Frontiers in Neuroinformatics*, *11*. https://doi.org/10.3389/fninf.2017.00076

Pouchard, L., Lin, Y., & Van Dam, H. (2020). Replicating Machine Learning Experiments in Materials Science. In I. Foster, G. R. Joubert, L. Kučera, W. E. Nagel, & F. Peters (Eds.), *Advances in Parallel Computing*. IOS Press. https://doi.org/10.3233/APC200105

Pownall, M. (2022). *Is replication possible for qualitative research?* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/dwxeg

Pratt, M. G., Kaplan, S., & Whittington, R. (2020). Editorial Essay: The Tumult over Transparency: Decoupling Transparency from Replication in Establishing Trustworthy Qualitative Research. *Administrative Science Quarterly*, *65*(1), 1–19. https://doi.org/10.1177/0001839219887663

Provenzano, D., Rao, Y. J., Goyal, S., Haji-Momenian, S., Lichtenberger, J., & Loew, M. (2021). Radiologist vs Machine Learning: A Comparison of Performance in Cancer Imaging. *2021 IEEE Applied Imagery Pattern Recognition Workshop (AIPR)*, 1–10. https://doi.org/10.1109/AIPR52630.2021.9762211

Reischer, H. N., & Cowan, H. R. (2020). Quantity Over Quality? Reproducible Psychological Science from a Mixed Methods Perspective. *Collabra: Psychology*, *6*(1), 26. https://doi.org/10.1525/collabra.284

Repko, A. F., & Szostak, R. (2020). *Interdisciplinary Research: Process and Theory*. SAGE Publications.

Robson, S. G., Baum, M. A., Beaudry, J. L., Beitner, J., Brohmer, H., Chin, J. M., Jasko, K., Kouros, C. D., Laukkonen, R. E., Moreau, D., Searston, R. A., Slagter, H. A., Steffens, N. K., Tangen, J. M., & Thomas, A. (2021). Promoting Open Science: A Holistic Approach to Changing Behaviour. *Collabra: Psychology*, *7*(1), 30137. https://doi.org/10.1525/collabra.30137

Rogers, E. M. (2003). *Diffusion of innovations* (5th ed). Free Press.

Ross-Hellauer, T., Klebel, T., Bannach-Brown, A., Horbach, S. P. J. M., Jabeen, H., Manola, N., Metodiev, T., Papageorgiou, H., Reczko, M., Sansone, S.-A., Schneider, J., Tijdink, J., & Vergoulis, T. (2022). TIER2: Enhancing Trust, Integrity and Efficiency in Research through next-level Reproducibility. *Research Ideas and Outcomes*, *8*, e98457. https://doi.org/10.3897/rio.8.e98457

Schickore, J. (2011). The Significance of Re-Doing Experiments: A Contribution to Historically Informed Methodology. *Erkenntnis*, *75*(3), 325–347. https://doi.org/10.1007/s10670-011-9332-9

Schmidt, S. (2009). Shall we Really do it Again? The Powerful Concept of Replication is Neglected in the Social Sciences. *Review of General Psychology*, *13*(2), 90–100. https://doi.org/10.1037/a0015108

Schöch, C. (2023). Repetitive research: A conceptual space and terminology of replication, reproduction, revision, reanalysis, reinvestigation and reuse in digital humanities. *International Journal of Digital Humanities*, *5*(2), 373–403. https://doi.org/10.1007/s42803-023-00073-y

Semmelrock, H., Kopeinik, S., Theiler, D., Ross-Hellauer, T., & Kowald, D. (2023). *Reproducibility in Machine Learning-Driven Research* (arXiv:2307.10320). arXiv. https://doi.org/10.48550/arXiv.2307.10320

Shaw, L. C., Errington, T. M., & Mellor, D. T. (2022). Toward Open Science: Contributing to Research Culture Change. *Science Editor*, *45*(1), 14–17. https://doi.org/10.36591/SE-D-4501-14

Sikorski, M., & Andreoletti, M. (2023). Epistemic Functions of Replicability in Experimental Sciences: Defending the Orthodox View. *Foundations of Science*. https://doi.org/10.1007/s10699-023-09901-4

Stodden, V., McNutt, M., Bailey, D. H., Deelman, E., Gil, Y., Hanson, B., Heroux, M. A., Ioannidis, J. P. A., & Taufer, M. (2016). Enhancing reproducibility for computational methods. *Science*, *354*(6317), 1240–1241. https://doi.org/10.1126/science.aah6168

TalkadSukumar, P., & Metoyer, R. (2019). *Replication and Transparency of Qualitative Research from a Constructivist Perspective*. OSF Preprints. https://doi.org/10.31219/osf.io/6efvp

Tijdink, J. K., Leitner, B., Cole, N. L., Horbach, S., Kopeinik, S., & Ross-Hellauer, T. (2023). *TIER2 D4.1 The Future(s) of Reproducibility in Research*. https://doi.org/10.17605/OSF.IO/DZQ9E

Toronto, C. E., & Remington, R. (Eds.). (2020). *A Step-by-Step Guide to Conducting an Integrative Review*. Springer International Publishing. https://doi.org/10.1007/978-3-030-37504-1

Torraco, R. J. (2016). Writing Integrative Literature Reviews: Using the Past and Present to Explore the Future. *Human Resource Development Review*, *15*(4), 404–428. https://doi.org/10.1177/1534484316671606

Tricco, A. C., Lillie, E., Zarin, W., O'Brien, K. K., Colquhoun, H., Levac, D., Moher, D., Peters, M. D. J., Horsley, T., Weeks, L., Hempel, S., Akl, E. A., Chang, C., McGowan, J., Stewart, L., Hartling, L., Aldcroft, A., Wilson, M. G., Garritty, C., … Straus, S. E. (2018). PRISMA Extension for Scoping Reviews (PRISMA-ScR): Checklist and Explanation. *Annals of Internal Medicine*, *169*(7), 467–473. https://doi.org/10.7326/M18-0850

Tuval-Mashiach, R. (2021). Is replication relevant for qualitative research? *Qualitative Psychology*, *8*(3), 365–377. https://doi.org/10.1037/qup0000217

Walonoski, J., Kramer, M., Nichols, J., Quina, A., Moesel, C., Hall, D., Duffett, C., Dube, K., Gallagher, T., & McLachlan, S. (2018). Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record. *Journal of the American Medical Informatics Association*, *25*(3), 230–238. https://doi.org/10.1093/jamia/ocx079

West, R., & Gould, A. (2022). *Improving health and wellbeing: A guide to using behavioural science in policy and practice*. World Health Organization Collaborating Centre on Investment for Health and Well-being. https://phwwhocc.co.uk/resources/improving-health-and-wellbeing-a-guide-to-using-behavioural-science-in-policy-and-practice/

West, R., Michie, S., Chadwick, P., Atkins, L., & Lorencatto, F. (2020). *Achieving behaviour change: A guide for local government and partners*. https://assets.publishing.service.gov.uk/media/5fa537c7d3bf7f03b249aa12/UFG_National_Guide_v04.00__1___1_.pdf

Whittemore, R., & Knafl, K. (2005). The integrative review: Updated methodology. *Journal of Advanced Nursing*, *52*(5), 546–553. https://doi.org/10.1111/j.1365-2648.2005.03621.x