A scoping review on metrics to quantify reproducibility: a multitude of questions leads to a multitude of metrics

Rachel Heyard[1*], Samuel Pawel[1], Joris Frese[2], Bernhard Voelkl[3], Hanno Würbel[3], Sarah K. McCann[4], Leonhard Held[1], Kimberley E. Wever[5], Helena Hartmann[6], Louise Townsin[7], and Stephanie Zellers[8]

[1]Center for Reproducible Science, Epidemiology, Biostatistics and Prevention Institute, University of Zurich, Zurich, Switzerland

[2]Department of Political and Social Sciences, European University Institute, Florence, Italy

[3]Animal Welfare Division, University of Bern, Bern, Switzerland

[4]Berlin Institute of Health at Charité-Universitätsmedizin Berlin, QUEST Center, Berlin, Germany

[5]Department of Anesthesiology, Pain and Palliative Care, Radboud University Medical Center, Nijmegen, The Netherlands

[6]Clinical Neurosciences, Department of Neurology and Center for Translational, Neuro- and Behavioral Sciences (C-TNBS), University Hospital Essen, Essen, Germany

[7]Research and Innovation Office, Torrens University Australia, Australia

[8]Institute for Molecular Medicine Finland, University of Helsinki, Helsinki, Finland

[*]Corresponding author: Rachel Heyard, rachel.heyard@uzh.ch

November 26, 2024

Abstract

Background: Reproducibility is recognized as essential to scientific progress and integrity. Replication studies and large-scale replication projects, aiming to quantify different aspects of reproducibility, have become more common. Since no standardized approach to measuring reproducibility exists, a diverse set of metrics has emerged and a comprehensive overview is needed.

Methods: We conducted a scoping review to identify large-scale replication projects that used metrics and methodological papers that proposed or discussed metrics. The project list was compiled by the authors. For the methodological papers, we searched Scopus, MedLine, PsycINFO and EconLit. Records were screened in duplicate against predefined inclusion criteria. Demographic information on included records and information on reproducibility metrics used, suggested or discussed was extracted.

Results: We identified 49 large-scale projects and 97 methodological papers, and extracted 50 metrics. The metrics were characterized based on type (formulas and/or statistical models, frameworks, graphical representations, studies and questionnaires, algorithms), input required, and appropriate application scenarios. Each metric addresses a distinct question.

Conclusions: Our review provides a comprehensive resource in the form of a "live", interactive table for future replication teams and meta-researchers, offering support in how to select the most appropriate metrics that are aligned with research questions and project goals.

Keywords: Reproducibility, replicability, generalizability, translatability, meta-research, literature review, metrics, quantify

1 Introduction

Reproducibility of research results is often referred to as a cornerstone of science. Historically, the idea of replication as a means to establish the trustworthiness of a reported observation can be traced back at least one thousand years to the Persian scholars al-Biruni and al-Haytham [1]. Later, Gallileo emphasized that he repeated his experiments on movement on the inclined plane a hundred times in order to give the results more credibility [2]. The first scientific society in modern Europe, the *Accademia del Cimento*, founded in Florence in 1657, considered replication to be such a fundamental concept, that it chose "*provando e riprovando*" ("to verify repeatedly"), as the society's motto. Similarly, the Royal Society London declared replication of experiments as the sole method for establishing "matters of fact" [1]. Yet, early authors were very vague regarding how they established that a replication confirmed the original observation. Even today, there is no universally accepted definition of "reproducibility", as usage of the term and suggestions for how to establish or quantify reproducibility can vary widely among researchers and disciplines [3, 4]. Acknowledging that there is an ongoing debate on the definition of different aspects of reproducibility, we will use the terms as suggested by the iRISE (improving Reproducibility In SciencE) consortium [5], for the purpose of our study. Here, replicability is defined as "*the extent to which design, implementation, analysis, and reporting of a study enable a third party to repeat the study and assess its findings*", replication as "*a study that repeats all or part of another study and allows researchers to compare their findings*", and reproducibility as "*the extent to which the results of a study agree with those of replication studies*". This definition of reproducibility immediately asks for a specification of how to quantify the extent of agreement between a study and its replication. While there is no definition of reproducibility that is universally accepted across disciplines and research types, even less is known on the metric that best captures the reproducibility of a study or finding. However, selecting the most appropriate outcome for a reproducibility study[1] is crucial to ensure the accuracy and credibility of research into the reproducibility of science.

An increasing number of articles has recently discussed the relevance of various metrics to define "successful replication" in the pairwise comparison of original-replication study pairs. Hereafter, we define a successful replication as "a replication study for which the results agree with the corresponding original study". In a rapid review of replication studies in psychology published in 2013, Anderson and Maxwell [6] investigated the decision criteria for successful replication. They concluded that the majority of published replication studies (44 of the 50 included studies) classified the replication as successful when both studies came to the same conclusions based on statistical significance. Cobey et al. [7] conducted a scoping review of replication studies published in 2018 and 2019 in economics, education, psychology, health sciences, and biomedicine to describe the epidemiological characteristics of this literature. They found large variability in how authors assessed reproducibility, although most of the included studies used a comparison of effect sizes to define success. Further, large scale reproducibility efforts, e.g. the replication projects in psychology [8], experimental economics [9], or cancer biology [10], all used a whole set of metrics based on statistical significance, effect sizes, or methodology from meta-analysis to summarise the reproducibility of a research field. This list of traditional metrics for reproducibility includes the significance criterion, where a replication is considered successful if it finds a statistically significant effect in the same direction as the original study, and effect size comparisons, where success is determined by the similarity between the effect size of the replication and the original study. To investigate whether there is one best metric for the quantification of replication success, Muradchanian et al. [11] conducted a simulation study to examine the performance of a set of metrics in terms of their classification accuracy under varying degrees of publication bias. Their findings revealed no clear "winner" across all simulation conditions, emphasizing

[1] We define a reproducibility study as any type of study investigating the reproducibility of a field, study, analysis or finding.

that the choice of the most appropriate metric may depend on the specific context or objective of the analysis. In line with this, Anderson and Maxwell [6] directly link the criteria for replication success to distinct replication goals. Existing reviews examining the usefulness and limitations of various metrics for reproducibility (including Hung and Fithian [12] and Nosek et al. [13]) typically lack a systematic search of the literature. Moreover, they tend to focus on one narrow aspect of reproducibility and scenario of application: specifically, where a replication study applies the same design, methodology or analysis as the original study to newly collected data.

In our review, we aim to gain a more comprehensive overview of metrics that have been used or suggested to quantify, assess, explain or predict different types of reproducibility. We sought to identify all metrics used in larger studies and projects, as well as those suggested in methodological literature. To achieve this we conducted a literature review of applied and methodological research. We did not restrict our comprehensive search to statistical metrics based on formulas. We addressed the following research questions: (1) Which metrics have been used or suggested to quantify, assess, explain or predict reproducibility? and (2) Which of these metrics have solely been suggested theoretically, and which have been proposed or discussed together with information on their practical implementation (e.g., clear implementation steps, ready-to-use tools, or open-source code)? We also identified the scenarios in which each metric proved most useful and associated each with a research question to guide users in interpreting the metrics. Additionally, we extracted details on any reported assumptions and limitations.

The metrics identified in our review are summarized in a table designed to inform various audiences in reproducibility research. A "live" and interactive version of the table can be found on `rachelhey.github.io/reproducibility_metrics/`. These target audiences include replication teams planning future reproducibility studies, newcomers to the field seeking a first comprehensive overview of available metrics, and the broader meta-research community, particularly those requiring outcome measures for intervention studies aimed at improving reproducibility. Additionally, our findings will support peer reviewers and researchers alike in critically evaluating the appropriateness of metrics used in reproducibility efforts, ensuring they align with the study's goal. This review is part of the work done by the iRISE (improving Reproducibility In SciencE) consortium. Since iRISE is committed to mainstreaming Equity, Diversity and Inclusion (EDI, see also the iRISE EDI statement: osf.io/b4crd), we collected data on potential EDI dimensions considered in reproducibility assessment, to perform an exploratory analysis.

We first outline our review methods, including the paper eligibility criteria, search strategy, data screening, and data extraction process in Section 2. The results are presented separately for the metrics used in large scale reproducibility efforts (Section 3.3) and the metrics suggested in methodological research (Section 3.4). We finish with a discussion of our results, limitations, and future directions in Section 4.

2 Methods

The protocol of the present study was preregistered on the Open Science Framework prior to initiating the literature screening and data extraction [14]. The protocol, as well as this manuscript, follow the PRISMA-ScR reporting guidelines for scoping reviews [15] (see the filled checklist osf.io/v7tas). Any deviations from the protocol were recorded and are discussed in Section 3.1. When referring to metrics, we include any metrics that provide a binary classifier of a study, part of a study, or results of a study being reproducible. We also include any metrics that provide a continuous quantification of reproducibility, or a level of reproducibility (for example on a numeric scale, or from "not at all" to "fully reproducible") and are interested in any tools, algorithms or models that measure, aim at explaining, or predicting reproducibility in a broader sense. Our search strategy was developed under the guidance of an information specialist from the University of Zurich and aims to identify two classes of papers: *application papers* and *methodological papers*. Therefore, the review was divided into two parts:

(1) **Application papers** – To gain an understanding of the metrics used to quantify, assess, explain, or predict a specific type of reproducibility in practice, a list of large-scale reproducibility projects[2] was compiled by the project team (available via our Zotero library[3]).

(2) **Methodological papers** – A systematic search was conducted to identify literature in which metrics to quantify, assess, explain, or predict reproducibility were proposed or discussed.

The screening and data extraction of the application papers preceded and informed the screening and data extraction of the methodological papers.

2.1 Eligibility criteria

All papers, protocols, or preprints discussing the methodology or the results of a large-scale reproducibility project were included as application papers. A reproducibility project was defined as a large-scale effort to measure the reproducibility of a field, method, type of study, or similar [for example, 16, 17]. These projects attempt to reproduce a series of previous results, to repeat a specific part of a series of previous studies, or to repeat one analysis multiple times in independent teams. They further aimed at summarising the results into a quantification of overall reproducibility. All methodological papers or preprints suggesting the use of a specific metric to quantify, assess, explain or predict a certain type of reproducibility were included as methodological papers. The aspect of reproducibility discussed in these papers had to align with the current definition of reproducibility set in the iRISE Reproducibility Glossary [5]. More specifically, papers using the same terminology but in a different, unrelated context (including translation in linguistics, image replication, sexual reproduction, cell or bacteria replications, virus reproduction ratio) were excluded. All years of publication and fields of research were included. For the systematic search of methodological papers all languages were included, while the list of application papers was compiled by the project team and is therefore limited to English literature. Commentaries, editorials and opinion pieces were excluded unless it was apparent from the abstract that a metric or measure was suggested or discussed. Single study application papers, e.g., papers discussing single replications of single original findings, were excluded, because they generally used the same set of traditional metrics, including metrics based on statistical significance and effect size comparisons [7], and the effort of assessing such papers in depth was considered disproportionate to the amount of potential information to be gained.

2.2 Search strategy, information sources, and screening

To collect the application papers, e.g., description of the methodology or the results of large-scale reproducibility efforts, two team members (RH and SP) initialized a list of projects that was complemented via a call for contributions (see osf.io/a2wrj). Once the list was finalized (mid March 2024) it was uploaded to the Systematic Review Facility (SyRF) [18], and five team members (HH, JF, LT, RH, SP) screened the titles and abstracts of the documents for final inclusion. All documents were screened in duplicate and conflicts were resolved by a third independent reviewer as automatically implemented in SyRF.
For the methodological papers, a systematic search was performed in the following databases: Scopus, MedLine (via Ebsco), PsycINFO (via Ebsco), and EconLit (via Ebsco), where the selection of discipline-specific databases was inspired by Cobey et al. [7]. The search strings can be found in Appendix A. The literature search was performed on May 13th, 2024. The search results were deduplicated in `R` (via their

[2]These were defined as larger projects where a group or a consortium of researchers attempt to reproduce a set of original studies, or the same original study several times. They do not include single efforts to reproduce part or all of an original study. To qualify as a large-scale reproducibility project, the project team should, in addition to conducting the set of replication studies, attempt to summarize the results of the set of studies.

[3]zotero.org/groups/5397531/reproducibilitymetrics/collections/HST4PWW8

digital object identifier, DOI) and imported into SyRF. A screening guide was developed, see Appendix B.1, and tested and adapted using a random sample of twenty methodological papers. Six team members (HH, JF, LT, RH, SP, SZ) screened titles and abstracts in duplicate and conflicts were resolved by a third independent reviewer as automatically implemented in SyRF. While screening, the reviewers had the option to annotate papers that were not, by definition, methodological papers, but documented an "interesting application". A paper was labelled an "interesting application paper" whenever it was apparent from the title and/or abstract that the authors applied an innovative or non-traditional reproducibility metric (i.e., other than significance criterion, meta-analysis or effect size comparison).
During data extraction of the application papers and with the flagged "interesting application papers", more potential methodological papers were retrieved. Additionally, a forward-backward reference and citation search was performed on the included methodological papers, that were not flagged "interesting application papers", via OpenAlex using the `openalexR` `R` package [19]. The titles of the papers identified via OpenAlex were subjected to a keyword search, and only those papers with at least one of the following terms were retained for screening: *quantify, measure, evaluate, assess, quantifying, measuring, evaluating, assessing, metric, score, rating, quantification, measurement, evaluation,* and *assessment*. The retained 296 potential methodological papers were pre-screened by one team member (RH). The records retained after pre-screening, as well as the potential methodological papers extracted from the application papers and the "interesting application papers", were screened by four team members (JF, LT, RH, SZ) using the screening guide in Appendix B.2. Each document was screened in duplicate and conflicts were resolved by a third independent reviewer as automatically implemented in SyRF.

2.3 Data extraction

All data extraction was performed in SyRF. For the application papers five team members (HH, JF, LT, RH, SP) extracted information on the research question or aim of the project, the type of project and, if applicable, the definition of reproducibility given by the authors, or inferred from the text. The type of project is of particular interest for application papers, as it determines what format of data is collected and what type of metrics can be used. McShane et al. [20] defined the types "Many Phenomena, One Study", where many original hypotheses are tested, each in one replication study, "One Phenomenon, Many Studies", where one original hypothesis is tested by many different teams or in many separate studies, and "Many Phenomena, Many Studies", where many original hypotheses are tested in many separate studies. Information on the metrics to quantify reproducibility was extracted using a predefined list (with traditional reproducibility metrics such as "agreement in statistical significance") and free text for less traditional metrics. If the authors mentioned other papers or documents with further information on the metrics used, their DOIs were retrieved and fed into the systematic search for methodological papers. Additionally, any text discussing limitations or assumptions related to the metrics used was extracted. Finally, text related to a discussion of EDI dimensions of the metrics was extracted (see Section 2.4 for more information). The full list of questions used for data extraction for the application papers can be found in Appendix C.1. Each document was annotated by at least two reviewers. One team member (RH) merged the individual data extraction sheets together and reconciled any differences. The "interesting application papers" which were included in the screening of the methodological papers were annotated by four team members (HH, LT, RH, SZ), notably to identify any potential methodological papers that were cited (the data extraction guide is in Appendix C.2.1).
The 97 methodological papers were annotated by six team members (not in duplicate by HH, JF, LT, RH, SP, SZ) using the extraction guide in Appendix C.2.2. In particular, details on whether the metric was designed for the purpose of quantifying reproducibility, the particular type of reproducibility or related concept the metric addresses, and the type of measure, including a formula, a model, or a metric

derived from a study or survey (see the extraction guide for examples), were extracted. We also collected information on the implementation, the required data input, any assumptions or limitations discussed, as well as mentions of EDI dimensions.
After all the information was extracted on a paper-level, one team member (RH) identified the distinct metrics that were either used in the application papers or suggested and discussed in the methodological papers, and composed a table on the level of reproducibility metric. This table was reviewed by the other team members.

2.4 Exploratory analysis on EDI dimensions considered in reproducibility assessment

Since data on EDI dimensions in the reproducibility space is limited, it is of great value to collect EDI relevant data whenever possible. We therefore collected any mention of equity, diversity or inclusion in the included records. We were specifically interested in whether authors who suggested or used a certain metric to assess or quantify reproducibility discussed its applicability or generalizability across research fields, research types or research communities. The extracted EDI content was reviewed by RH and SZ and grouped into topics for descriptive purposes, based on the EDI terms in the iRISE glossary [5]. This analysis was purely exploratory and not preregistered.

3 Results

3.1 Protocol amendments

The search string for the methodological papers was adapted to be more specific and ensure the number of records to screen was feasible for our small team. To narrow the scope of our manuscript, review questions 3 and 4 from the protocol on the interpretation, assumptions and limitations were only answered in a descriptive manner, based on the limited information extracted from the included records. A more focused discussion on the interpretation, assumptions and limitations of each metric remains to be performed. We decided against using Rayyan for screening and instead performed both the screening and data extraction in SyRF. While the forward-backward search of the references and citations was mentioned in the protocol, the exact procedure was not pre-specified. We added an exploratory data analysis on the EDI dimensions.

3.2 Included records

As outlined in Figure 1, our research team identified 54 records potentially discussing a large-scale reproducibility effort. Following screening, 50 of these papers were retained for data extraction, while one was later retracted by the journal and therefore excluded from our analysis. During data extraction of the 49 included application papers, we identified 13 potential methodological papers. The literature search for the methodological papers yielded 1,316 records of which 1,215 were excluded during the screening process. We retained 101 records of which 47 were flagged as "interesting application papers". The remaining 54 records were classified as methodological papers. Data extraction from the "interesting application papers" led to the identification of an additional 33 potential methodological papers. Subsequently, a forward and backward citation search on the 54 included methodological papers resulted in 4,346 records, with 296 of these containing relevant keywords in their title. After screening the 296 records, 42 more records were added to the list of potential methodological papers. In the final step, we screened the 88 potential methodological papers, identified through data extraction, and forward and backward citation search, after deduplication. After data extraction of the methodological papers, one record was excluded as it

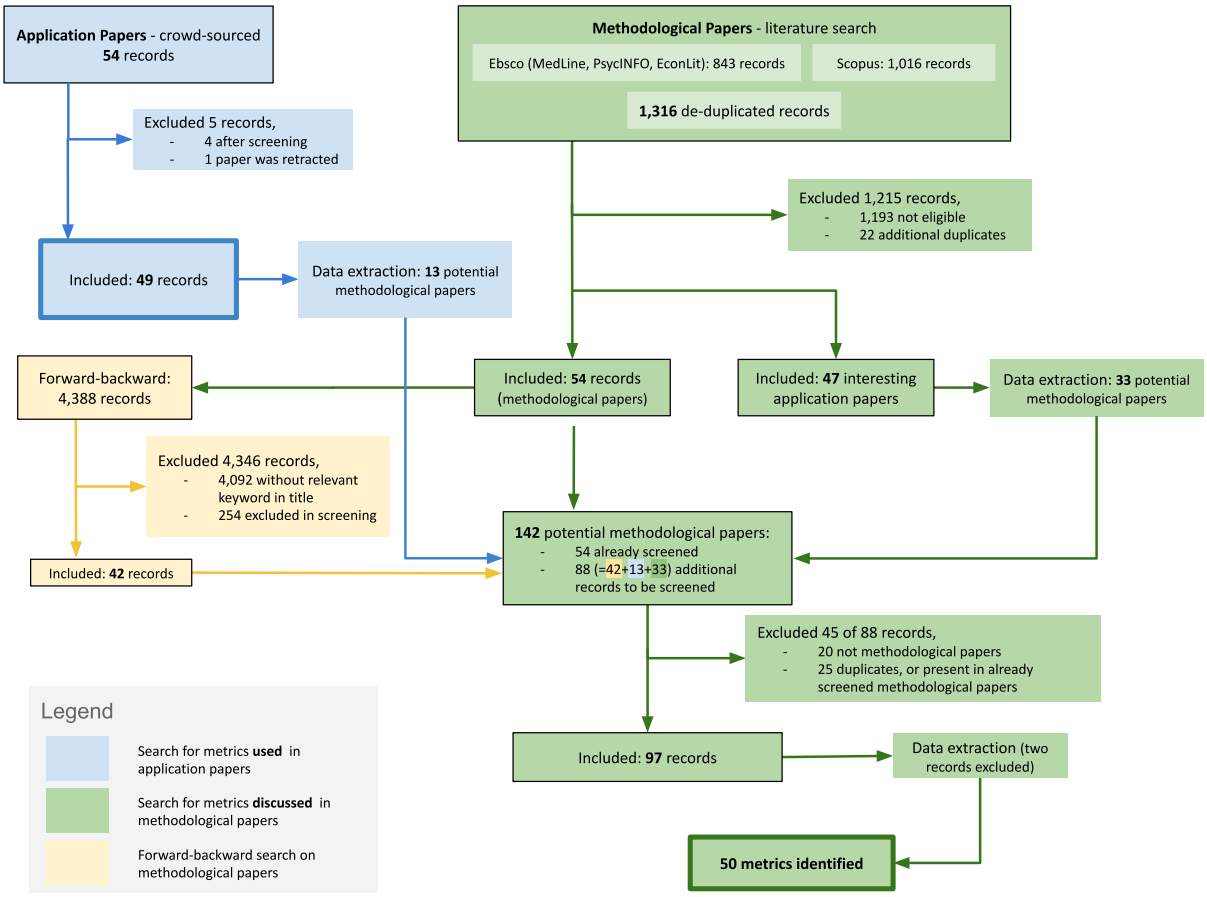

Figure 1: Flow-chart of the search strategy for both application and methodological papers.

was written in Czech and no team member was fluent in Czech. We also found one more duplicate. Ultimately, a total of 95 distinct methodological papers were included in this review. In the following sections, the results for the application papers and methodological papers will be presented separately.

3.3 Application papers

3.3.1 Characteristics of the included application papers

Table 1 gives a first impression of the characteristics of the 49 included application papers. Most large-scale reproducibility efforts were performed in the Social Sciences (67%) and only a minority in the Health and Life Sciences (20%) and Physical Sciences (12%). Less than half of the included records (23/49 = 47%) clearly defined what they meant by "reproducibility", i.e., we were able to identify a clear definition in the paper. When categorizing the aspect of reproducibility using the texts, we concluded that most records (27/49 = 55%) report that in their effort they used the same analysis on different data, defined as a form of "replication" in Voelkl et al. [5]. Among the included records, we found an equal share of project types. One of the project records presented two types of project: the protocol by Page et al. [21] presents the REPRISE project, a large effort encompassing four studies, where studies two and three were of interest in our review; one was classified as a "Many Phenomena, Many Studies" project and one a "Many Phenomena, One Study" project. Most of the included reproducibility efforts were conducted by a large team of authors (median number of project authors = 24), while some were conducted by only one or a handful of authors. The included papers were fairly recently published (median year of publication = 2020), and were, generally, already heavily cited (median number of citations = 61, September 28,

2024).

3.3.2 Characteristics of reproducibility metrics used

Eight (16%) reproducibility efforts employed only a single metric, while the remainder used at least two metrics to evaluate reproducibility (see Figure 2). A total of 12 metrics were recorded for Wang et al. [22]. The metrics used were of varying types and investigated agreement in significance or effect size, using meta-analysis methodology or subjective assessment.

Agreement in statistical significance: Thirty-two (65%) of the included application papers used at least one metric based on statistical significance. These 32 projects were equally likely to be either type of project, as seen in Figure 3, and Figure 4 shows that most of these projects repeated the same analysis on different data. Usually, "Many Phenomena, One Study" project types like Errington et al. [10] investigate whether the original and replication studies found a significant effect in the same direction. For "One Phenomenon, Many Studies" or "Many Phenomena, Many Studies" projects like Klein et al. [23], measuring reproducibility based on statistical significance means computing a proportion of samples or replications that rejected the null hypothesis in the expected direction. "Many Phenomena, Many Studies" project types, including the Brazilian Reproducibility Initiative [24], where each study was replicated three times, usually employed a pooled version of the effect sizes of the replication studies to assess reproducibility. "One Phenomenon, Many Studies" project types, on the other hand, reported rates, shares or counts of studies or analyses obtaining statistically significant results, as for example Schweinsberg et al. [25].

Agreement in effect size: Seventy-one percent ($35/49 = 71\%$) of the application papers used at least one metric based on the agreement in effect sizes. These metrics come in different forms. Irvine, Hoffman, and Wilkinson-Ryan [26], for example, informally describe how the original and replication effect sizes compare to each other in tables and figures. One of the seven reproducibility metrics used by Errington et al. [10] was to simply check that the direction of the effect was the same in the original and replication studies. Cova et al. [16] and Camerer et al. [9] used a binary measure assessing whether the 95% confidence interval (CI) of the replication effect size includes the original effect size. Since this metric does not acknowledge sampling error in both the original and the replication study, Camerer et al. [9] and Boyce, Mathur, and Frank [27] investigated whether the replication effect sizes were included in a 95% prediction interval of the original effect size, as suggested by Patil, Peng, and Leek [28]. For projects where multiple replication studies were performed for one phenomenon or original study, the effects for all replications were aggregated and then compared to the original effect (as for example in Ebersole et al. [29]). Klein et al. [23], a "Many Phenomena, Many Studies" project, investigated variation across samples and settings using intra-class correlation coefficients and the heterogeneity of effect sizes using Cochran's Q and I^2. Chang, Chilcott, and Latimer [30], who followed Wang, Schneeweiss, and RCT-DUPLICATE Initiative [31] to design their project, assessed reproducibility using standardized differences to investigate whether the effect sizes of original and replication studies (here randomized controlled trials vs real-world evidence emulations) were significantly different. In addition, they claimed successful replication (or emulation) if the effect estimates of the replication fell within the 95% CI of the original study. Ebersole et al. [32], Errington et al. [10] and Boyce, Mathur, and Frank [27] used *p-original*, defined as the p-value for the null-hypothesis that the effect sizes of the original and replication study follow the same distribution [33]. This metric can take effect size heterogeneity into account and assesses statistical consistency between original and replication studies.

Meta-analysis of study results: Only nine (18%) of the included application papers reported that they used a meta-analysis of study results to decide on successful replication or degree of reproducibility. In "Many Phenomena, One Study" projects this usually entailed performing a fixed-effect meta-analysis of the findings from the original and the corresponding replication study and flagging successful replication if the meta-analytical effect size was found to be significant in the same direction as the original effect [as in 8, 9, 24, 10]. The remaining reproducibility projects, specifically "Many Phenomena, Many Studies" and "One Phenomenon, Many Studies", performed meta-analyses, usually random-effects, of all replication effect sizes to assess and quantify reproducibility [e.g., 34, 35, 32, 36]. If there was an original study, these meta-analytical results were then compared to the original results. Ebersole et al. [32] used meta-analytical approaches to investigate whether certain interventions could improve reproducibility.

Subjective assessment: Twenty-nine percent ($14/49 = 29\%$) of the application papers reported using some form of subjective or narrative assessment of reproducibility. This often implied asking replication teams, informally or using a survey questionnaire, for their assessment on the reproducibility of a study after having performed its replication [8, 26, 16, 37]. More specifically, the replication team in Naudet et al. [38], for instance, classified papers into four categories: "fully reproduced", "not fully reproduced but same conclusion", "not reproduced and different conclusion", and "not reproduced (or partially reproduced) because of missing information". Boyce, Mathur, and Frank [27] used a subjective replication score coded on a scale from [0, 0.25, 0.5, 0.75], which allowed raters to subjectively summarize multiple important outcomes or features of reproducibility. Low et al. [39] summarized the methodology used and conclusions drawn from two independent systematic reviews in a narrative manner. Other projects used so-called "prediction markets", in which experts trade contracts on the possible outcome of the replication study, informed by the results of an original study and information on the design of a planned replication study [among others 40, 9]. The market price can then be interpreted as the predicted reproducibility of the study. Alipourfard et al. [41] explain that they will use the repliCATS platform [42], which uses a modified form of a Delphi protocol to aggregate expert reproducibility assessments. In their project where two datasets were re-analysed by four research teams, using either Bayesian or frequentist statistics, Dongen et al. [43] summarized the findings only in a subjective and narrative manner during discussions. The RepliSims project presented in Luijken et al. [44] describes the differences in the results of simulation studies in a qualitative and narrative way: "are trends in the results moving in the same direction or do the performance rankings of different simulation scenarios match those in the original study?".

Additional metrics and analyses: In addition to the metrics described above, some application papers used less traditional metrics to summarize the reproducibility of findings. Often these were secondary or complementary analyses of the results. Specifically, Milcu et al. [35], a "One Phenomenon, Many Studies" project, used Tukey's post-hoc honest significant difference test [45], to investigate "how many laboratories produced results that were statistically indistinguishable from one another". Schweinsberg et al. [25], who asked several teams of analysts to answer the same research question, examined whether independent analysts would arrive at similar analyses and statistical results, and performed a multiverse analysis using the Boba approach as suggested in Liu et al. [46]. The Boba multiverse gave the project authors an opportunity to further understand which analysis choices played a major role in creating differences in the independent analysts' results. In the Yale Open Data Access Medtronic Project [39], two independent research teams used the same data and analysis, and the project authors not only compared the final results and conclusions of the two teams, but were particularly interested in differences in inclusion criteria and statistical methodology applied on the data, which were summarized in a narrative fashion. Many replication projects summarized differences in original and/or replication studies in a descriptive manner, including percentages, counts, or number of differences and correlation coefficients

Table 1: Characteristics of the included application papers.

	N (%), unless otherwise indicated
Total records	49
Field of research (OpenAlex)	
Health and Life Sciences	10 (20.4%)
Physical Sciences	6 (12.2%)
Social Sciences	33 (67.3%)
Authors defined reproducibility?	
No	26 (53.1%)
Yes	23 (46.9%)
Aspect of reproducibility	
Combination*	2 (4.1%)
Different data - different analysis	3 (6.1%)
Different data - same analysis	27 (55.1%)
Same data - different analysis	13 (26.5%)
Same data - same analysis	4 (8.2%)
Type of project	
Many Phenomena, Many studies	16 (32.7%)
Many Phenomena, Many studies; Many Phenomena, One Study	1 (2%)
Many Phenomena, One Study	15 (30.6%)
One Phenomenon, Many Studies	17 (34.7%)
Number project authors	
Median	24
Range	1 - 260
Citation count	
Median	61
Range	0 - 6,739
Year of publication	
Median	2020
Range	2007 - 2024
Number of measures used	
Median	2
Range	1 - 12
Agreement in statistical significance	
No	17 (34.7%)
Yes	32 (65.3%)
Agreement in effect size	
No	14 (28.6%)
Yes	35 (71.4%)
Meta-analaysis of study results	
No	40 (81.6%)
Yes	9 (18.4%)
Subjective assessment	
No	35 (71.4%)
Yes	14 (28.6%)
Used none of the predefined measures	
No	45 (91.8%)
Yes	4 (8.2%)

* Theses papers presented projects with several sub-projects looking at different aspects of reproducibility

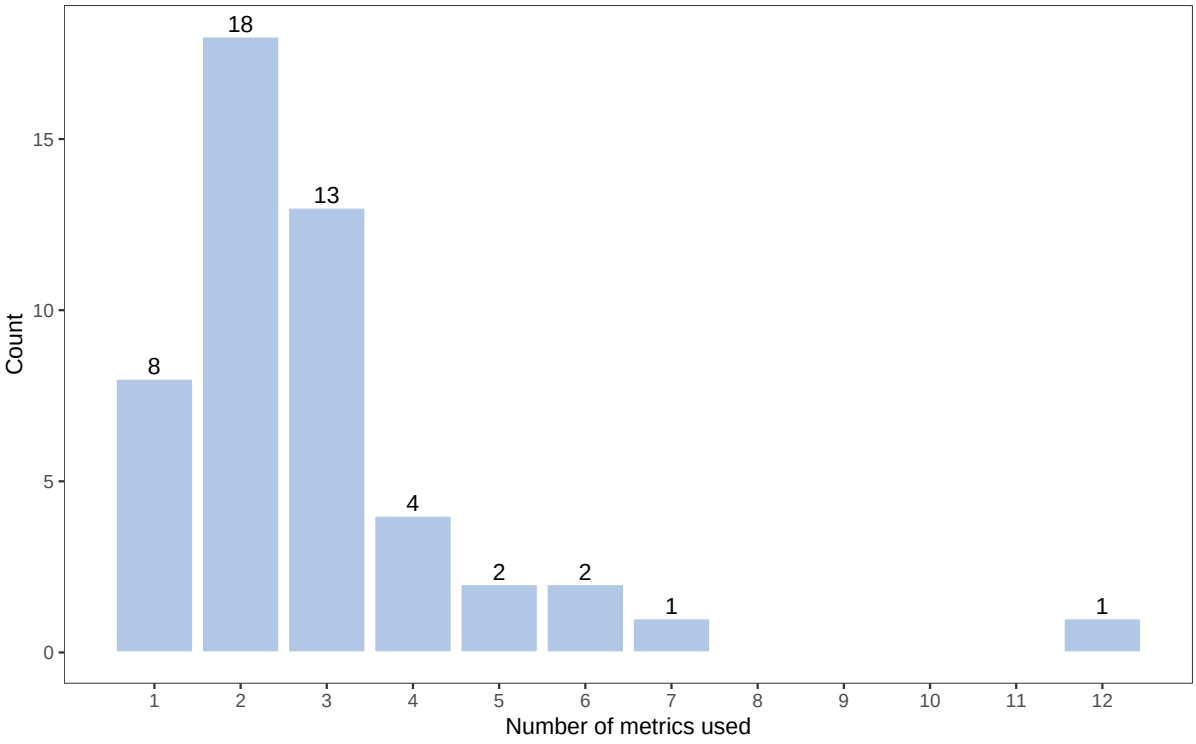

Figure 2: The total number of metrics used in the application papers to summarise reproducibility.

[e.g., 47, 48, 36]. Bastiaansen et al. [49] and Huntington-Klein et al. [47], for example, recorded differences in processing and analysis steps and decisions. Wang et al. [22] used calibration and Bland-Altman plots to represent their findings and assess agreement of original and replication results.

3.3.3 Limitations and assumptions of metrics discussed in application papers

Less than a third ($15/49 = 30\%$) of the included application papers discussed any assumptions or limitations of using specific metrics or measures to summarize or investigate reproducibility. Milcu et al. [35], for example, mentioned that using statistical significance to determine reproducibility might be "viewed as overly restrictive". They argue that they employed this approach due to the lack of a better alternative. Cova et al. [16] mentioned that the use of statistical significance as a replication success criterion for original "null" results is "especially dubious", which was recently discussed in Pawel et al. [50]. Some reproducibility projects reported that they are specifically using subjective assessment metrics because they accommodate the consideration of multiple outcomes of interest and are applicable across a diverse set of outcome measures [27], while others mention the subjectivity as a limitation [37]. Wang et al. [22] discuss that the proportion of studies with effect estimates of the same sign is imperfect as a metric for studies with small effect sizes, as the smallest implementation differences could result in a sign change in the reproduction attempt. In the next section, some of the many of the metrics used in application projects are explained in more detail.

3.4 Methodological papers

3.4.1 Characteristics of the included methodological papers

Of the 95 distinct records for which data were extracted, more than half ($57/95 = 60\%$) were categorized to the field of Social Sciences by `openalexR`. Sixty percent ($57/95 = 60\%$) were original research papers, 17% ($16/95 = 17\%$) were review papers and 15% ($14/95 = 15\%$) were classified as tutorial papers (see Table

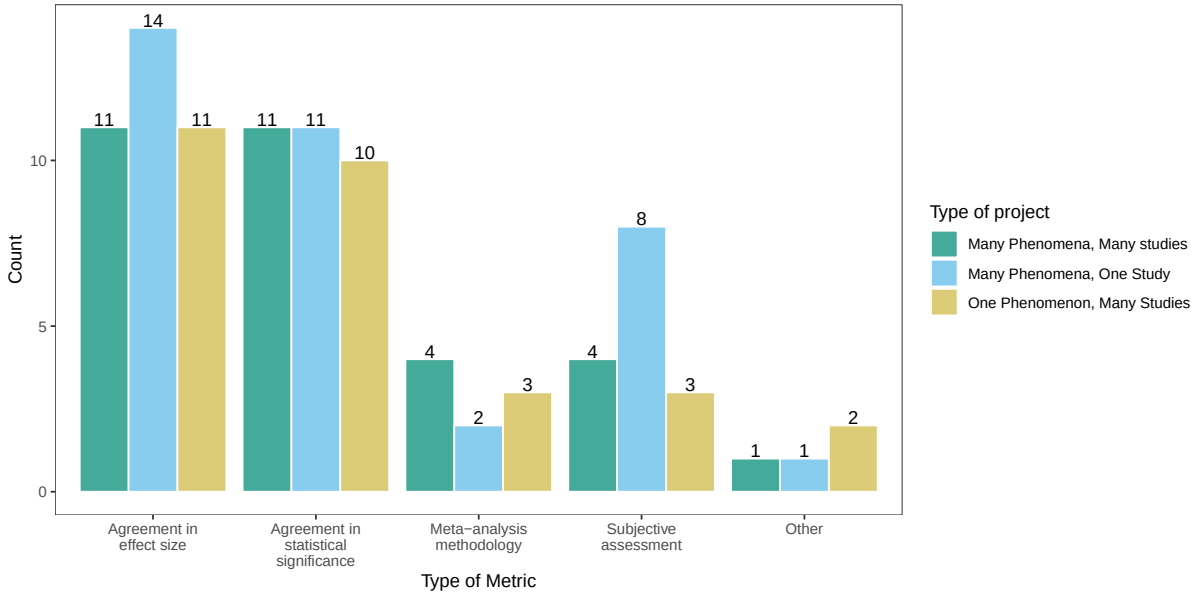

Figure 3: Count of mentions of different types of metrics depending on the type of project. Note that projects classified as more than one (combined) type were split into multiple projects.

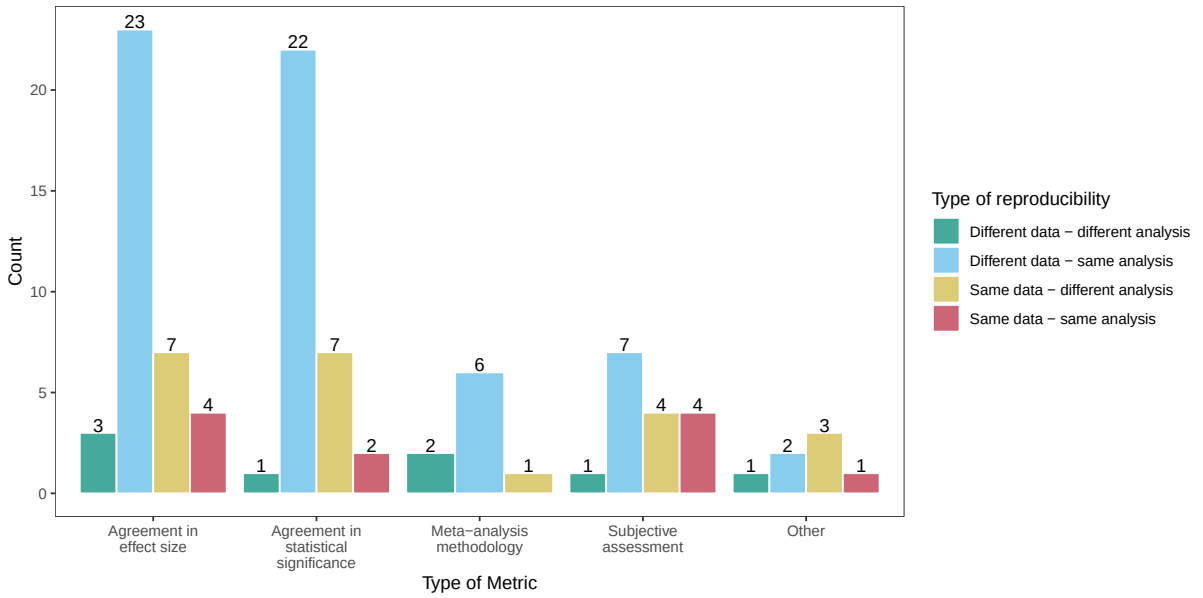

Figure 4: Count of mentions of different types of metrics by type of reproducibility. Note that projects classified as investigating several types of reproducibility were split into multiple projects.

Table 2: Summary of methodological papers included.

	N (%)
Total records	95
Field of research	
Health and Life Sciences	11 (11.6%)
Physical Sciences	27 (28.4%)
Social Sciences	57 (60%)
Type of paper	
Conference paper	1 (1.1%)
Editorial, comment, or similar	7 (7.4%)
Original research paper	57 (60%)
Review paper	16 (16.8%)
Tutorial paper	14 (14.7%)

2). We extracted a total of 50 distinct reproducibility metrics from these records. Table 3 summarizes the key attributes of the metrics. Note that all metrics used in the application papers were included, except for the Boba multiverse approach used in Schweinsberg et al. [25], and the comparison of study results using various descriptive statistics, because those methods are less suited for the quantification or classification of reproducibility.

3.4.2 Characteristics of the identified reproducibility metrics

Sixty percent (30/50 = 60%) of the metrics were specifically designed to assess reproducibility or a closely related concept, while the remaining 40% (20/50 = 40%) were initially proposed for a different context, but used or suggested to be used in reproducibility studies. We extracted 37 metrics (37/50 = 74%) that were formulas or statistical models. A type of metric we did not expect to find was "a framework". They were called "framework" by the authors and either formalize conditions or outline a standardized workflow to quantify or interpret reproducibility. Four metrics summarize the reproducibility in a graphical representation, while another four quantify reproducibility using a study, a survey or a questionnaire. Three metrics are based on an algorithm. The "Purpose of metric" column informs on whether the metric quantifies reproducibility in a continuous way or classifies it into "reproducible" vs. "not reproducible" or replication success vs. failure. Some metrics were specifically presented as being useful to explain or predict reproducibility. Most of the metrics (47/50 = 94%) can be used to quantify reproducibility in a continuous manner. Twenty-four (24/50 = 48%) were proposed or discussed together with a ready-to-use open tool or open-source software and code, while eleven metrics (11/50 = 22%) were classified as hard or costly to implement. This was mostly due to the metric relying on costly data retrieved using a study, e.g., prediction markets, or because the implementation was not clearly described. A large majority (39/50 = 78%) use results in the form of numbers and tables to quantify or assess reproducibility.
Table 4 presents the descriptions of the 50 identified metrics, including their name, a brief description, the research questions they address, application scenarios, their purposes, and relevant references (when they were first mentioned, discussed, or applied in the context of reproducibility). The metrics are organized by type: first, the 37 metrics that are based on formulas and statistical models, followed by those using frameworks, graphs, and studies, surveys or questionnaires. A more detailed version of the table, including information on their implementation, data input requirements, the extracted assumptions and limitations is available online rachelhey.github.io/reproducibility_metrics/. The assumptions and limitations listed are drawn directly from the reviewed records. All identified metrics come with some assumption or limitation, and each targets a specific research question. Thus there is no single "best" metric to quantify, classify, explain or predict reproducibility in general. Replication teams and meta-researchers should first

define the research question they seek to answer and then select the most suitable metric and project type. In the following sections, we first summarise "statistical metrics" (i.e., metrics based on formulas and statistical models), followed by a discussion of the other types of identified metrics.

Metrics based on formulas and statistical models

Of the identified metrics, 37 ($37/50 = 74\%$) were classified as being based on a formula or statistical model, making the majority "statistical metrics". These metrics typically provide a quantitative assessment of reproducibility, with one exception: the correspondence test. This test, recently introduced by Steiner, Sheehan, and Wong [51], combines both difference and equivalence testing. While the two individual tests, which are also part of the identified metrics, provide a quantitative assessment, the correspondence test categorizes their combined outcome into four levels. At a predefinded significance threshold α, it returns *equivalence* when the difference test finds no significant difference between the effect sizes of two studies and the equivalence test is significant. Alternatively, it can establish *difference*, *trivial difference*, or *indeterminacy*. This test is particularly relevant when comparing an original study to its replication, addressing the question "To what extent does the effect size from the replication study differ or is equivalent to that of the original study?". In contrast, the individual underlying tests provide more direct measures of the strength of evidence in terms of p-values.

The difference test, often referred to as Q-test (see "difference in effect size" in Table 4), has been widely used in large-scale replication projects, in some form or another. In a pairwise comparison of an original study with its replication, the research question addressed by this metric is "To which degree do the effects from a replication study mirror the original?", which can be extended to "To which degree do the effects from a set of replication studies mirror each other?" in a scenario where several replications are considered. This metric enables a direct comparison of effect sizes between two or more studies. Other related metrics for comparing effect sizes between original and replication studies include those based on 95% confidence and prediction intervals. P-original, P_{orig}, and $\hat{P}_{>0}$ have been suggested specifically for a scenario where one original study is replicated several times [52, 33]. The Z-curve methodology, related to the P-curve [53], quantifies reproducibility by predicting the success rate of direct replication studies based on the mean power after selection for significance [54] and the expected discovery rate [55]. The Z-curve is most useful when quantifying the reproducibility of a set of replication studies. As discussed in Section 3.3, meta-analysis methodology has also extensively been used in large-scale replication studies in pairwise original-replication assessment but also when several replication studies are compared with each other.

Most metrics that provide a quantitative assessment of reproducibility, can be dichotomized to classify a study as "reproducible" vs. "not reproducible". We illustrate this using one of the most commonly used metrics for reproducibility: the significance criterion. When comparing two studies, the criterion deems the replication of an original study successful if both studies report a significant effect in the same direction at a predefined level α. This creates a binary outcome of either replication success or failure. To quantify the strength of evidence that both studies found a statistically significant effect in the same direction, the maximum p-value, $\max\{p_o, p_r\}$, can be used, where p_o and p_r are the p-values from the original and replication. The binary classification is determined by checking whether $\max\{p_o, p_r\} < \alpha$. This illustrates the scenario of pairwise comparisons between an original study and its replication. However, the same criterion was employed in "Many phenomena, One study" projects, which involve multiple original-replication study pairs. In these cases, overall reproducibility was quantified by calculating the proportion of study pairs that achieve success [8, 10, 31]. Conversely, in projects of the type "One phenomenon, Many studies", where multiple replications test the same hypothesis or analyse the same data, reproducibility was quantified by determining the proportion of replications that yield

Table 3: Summary statistics of attributes of identified reproducibility metrics.

	N (%)
Total number of metrics	50
Designed for reproducibility	
No*	20 (40%)
Yes	30 (60%)
Type of reproducibility	
Different data - different analysis	1 (2%)
Different data - same analysis	27 (54%)
Different data - same/different analysis	4 (8%)
Same data - different analysis	1 (2%)
Same data - same analysis	1 (2%)
Same data - same/different analysis	1 (2%)
Same/different data - same analysis	5 (10%)
Same/different data - same/different analysis	10 (20%)
Type of metric	
A formula and/or statistical model	37 (74%)
A framework	3 (6%)
A graph	3 (6%)
A study, survey, or questionnaire	4 (8%)
An algorithm	3 (6%)
Purpose of metric	
To classify	3 (6%)
To quantify	16 (32%)
To quantify and classify	21 (42%)
To quantify and explain	4 (8%)
To quantify and predict	6 (12%)
Type of assessment	
Qualitative	3 (6%)
Qualitative and quantitative	5 (10%)
Quantitative	42 (84%)
Implementation	
Clear implementation	1 (2%)
Easy to implement	13 (26%)
Hard, costly or unclear implementation	11 (22%)
Ready-to-use closed tool provided	1 (2%)
Ready-to-use open tool provided	24 (48%)
Data Input	
Original raw data, code, and/or software	3 (6%)
Qualitative data, surveys or questionnaires	3 (6%)
Results - figures	1 (2%)
Results - figures, numbers and tables	2 (4%)
Results - number and tables	37 (74%)
Text, meta-data, and information on design	4 (8%)

* includes unclear

statistically significant outcomes in the same direction [47, 40, 25]. This also shows how the same metric can be used to assess reproducibility in different contexts, such as when different methods are applied to the same dataset (e.g. "One phenomenon, Many studies"), but also when the same methods are applied on different data (e.g. "Many phenomena, One study"). Related to this, the metric called *P interval* offers a more nuanced interpretation of the p-value of an original finding by computing a prediction interval for the p-value of a hypothetical replication study [56]. Many included methodological review papers discuss the limitations of the significance criterion. For example, the significance criterion could potentially indicate replication failure even when the effect estimates in the original and replication study are the same. This is why some authors have designed metrics that combine the comparison of effect size with an investigation of the strength of evidence in the original and replication studies (e.g., the sceptical p-value [57, 58] and the small telescopes approach [59]).

In addition to frequentist approaches for the assessment of reproducibility, some identified metrics included Bayesian methodology. For example, we identified Bayes factors (BFs) specifically designed for pairwise comparisons of original and replication studies: the equality-of-effect size BF [60], the replication BF [61] and the sceptical BF [62]. Some of the identified metrics were designed to quantify reproducibility in a specific field of research, including the quantified reproducibility assessment developed for studies in natural language processing and the Jaccard similarity coefficient applied to fMRI (functional magnetic resonance imaging) research.

Other types of metrics

We identified three metrics classified as *frameworks*. While we did not predefine what a framework entails, these three were initially classified as "Other" but later grouped as frameworks, as this is how the authors described them. All three outline how various aspects of reproducibility can be combined into a more nuanced assessment. For example, the unified framework for estimating the credibility of published research evaluates aspects such as transparency of methods and data, computational reproducibility, robustness and effect reproducibility [63]. While it does not offer a final summary across these aspects, it collects diverse evidence for a nuanced qualitative judgment on reproducibility. The framework by McIntosh et al. [64], targeted at biomedical research, includes 119 items operationalizing research transparency that are integrated in an assessment tool (RepeAT). The iRISE glossary refers to such items as proxy measures [5]. Although the authors suggest automation, its implementation remains unclear. Unlike the latter frameworks, which are useful to quantify or assess the reproducibility of one or several original studies, the causal replication framework by Steiner, Wong, and Anglin [65], is designed for use when at least one replication study is available or planned. It helps interpret and explain replication outcomes by examining the assumptions under which replication success can be expected.

Among the graphical representations identified, Bland-Altman plots have long been used in medical research to assess the agreement of two measurements. Wang et al. [22] employed this plot to assess the computational reproducibility in real-world evidence studies, while Page et al. [21] used it for agreement between original and replication effect sizes in evidence synthesis. These examples highlight the plot's potential applications to different aspects of reproducibility: computational reproducibility in the first case and conceptual replication in the second, as defined in Voelkl et al. [5]. Other graphical representations, such as *Reproducibility Maps* (specific for fMRI research) and modified Brinley plots (more broadly applicable to a setting of several replications of the same intervention study) were developed specifically for reproducibility.

Four identified metrics involved actual studies, where participants, often field experts, assess the reproducibility of studies. The participants in prediction markets, used in two of our application papers [40, 9], trade contracts which will be worth a certain amount of money based on replication outcomes. The

final price of the contracts will reflect the predicted probability of successful replication. Prediction markets are most applicable when a set of original studies are planned to be replicated. Other metrics use survey techniques to evaluate whether an original study's design, methods or reporting meet community standards of reproducible research. The RepliCATS methodology uses a modified Delphi process, where experts are asked to reach a consensus on the reproducibility of a study in several rounds before the data are aggregated into a final reproducibility assessment. In many "Many phenomena, One study" projects, replication teams are asked to assess replication outcomes using a binary scale (success/failure) or a more nuanced scale (e.g., Likert). While the implementation for these metrics was generally classified as clear, they can be labour- and cost-intensive due to the need to recruit participants, or pay participants in prediction markets.

Finally, we identified three algorithm-based metrics. Two involve checking the presence or absence of certain reproducibility-related proxy features using automated software tools. Another algorithm uses machine learning models to quantify reproducibility based on the texts and meta-data of a study. These algorithm-based metrics are useful for evaluating the reproducibility of single original studies, but might again come with substantial costs, as they are computationally extensive, or require specialized software and IT knowledge.

Table 4: Metrics table: Summary of the 50 identified metrics, ordered alphabetically and grouped by the type of metric: a formula or statistical model, a framework, a graph, a study, survey or questionnaire, or an algorithm. The name and description of the metric is followed by one or several research questions summarising the type of question the metric can answer. The scenario of application gives insights into the type of project design needed to compute or use the metric. We then collapsed all the references for further reading, where have the metrics first been mentioned in relation to reproducibility, which papers discussed them further and which application papers demonstrate how to use them.

Name (also called/related to)	Description	Research question	Scenario of application	Purpose of Metric	References
A formula and/or statistical model					
Bayes Factor: Equality-of-effect-size BF test	This test compares the null hypothesis that the effect sizes from two experiments (o and r for original and replication) are equal against an alternative hypothesis that they are not. Suppose $H_0 : \theta_o = \theta_r$ and $H_1 : \theta_o \neq \theta_r$, then the equality-of-effect-size Bayes factor is defined as $$B_{01} = \frac{f(Y_o, Y_r \mid H_0)}{f(Y_o, Y_r \mid H_1)},$$ where $f(Y_o, Y_r \mid H_i)$ is the marginal likelihood of the data under hypothesis H_i with $i \in \{0, 1\}$. B_{01} higher than 1 indicate support for H_0 and is indicative of a successful replication.	"What is the evidence for the effect size in the replication attempt being equal vs. unequal to the effect size in the original study?"	Two exchangeable studies: one original and one replication	To quantify	First mentioned in [60]. Discussed in [61, 66].
Bayes Factor: Fixed-effect meta-analysis BF Test (Meta-analytic BF)	The meta-analytic Bayes factor quantifies the evidence provided by the data of several experiments/studies for the hypothesis that the true effect is present (H_1) versus absent (H_0): $$B_{10} = \frac{f(Y_1, ..., Y_M \mid H_1)}{f(Y_1, ...Y_M \mid H_0)},$$ where $f(\dots \mid H_i)$ is the marginal likelihood of the data under hypothesis H_i with $i \in \{0, 1\}$. A high B_{10} indicates that the evidence from the pooled data supports H_1.	"When pooling all data, what is the evidence for the effect being present vs. absent?"	A series of exchangeable studies: one original and many replications; many replications without an original	To quantify	First mentioned in [67]. Discussed in [61, 66, 11]. Used in [36].

Table 4: Metrics table: Summary of the 50 identified metrics *(continued)*

Name (also called/related to)	Description	Research question	Scenario of application	Purpose of Metric	References
Bayes Factor: Independent Jeffreys-Zellner-Siow BF test (default BF)	This test compares the null hypothesis that the effect size is zero against an alternative hypothesis that the effect is not zero. Suppose $H_0 : \theta = 0$ and $H_1 : \theta \sim \text{Cauchy}(0, 1)$, then the Bayes factor is defined as $$B_{10} = \frac{f(Y \mid H_1)}{f(Y \mid H_0)},$$ where $f(Y \mid H_i)$ is the marginal likelihood of the data Y under hypothesis H_i with $i \in \{0, 1\}$. B_{10} higher than 1 indicate support for H_1, whereas lower than 1 indicate support for H_0. In the replication setting, the Bayes factor is used to test the absence or presence of an effect in the replication study. Note that the Jeffreys-Zellner-Siow prior is a prior that is specifically designed for the t-test / linear regression setting (normal data with unknown mean and variance).	"What is the evidence for the effect being present or absent in light of a replication attempt, given that we know relatively little about the expected effect size beforehand?"	Two exchangeable studies: one original and one replication	To quantify	Discussed in [61, 66, 68, 11]. Used in [69].

Table 4: Metrics table: Summary of the 50 identified metrics *(continued)*

Name (also called/related to)	Description	Research question	Scenario of application	Purpose of Metric	References
Bayesian Evidence Synthesis (variant: Meta-Analysis Model-based Assessment of replicability (MAMBA))	The approach assumes that multiple studies exist that investigate a common general theory. These studies might be so diverse in design and measurements, that the study-specific informative hypotheses reflecting the common theory can differ. First the evidence for or against the hypothesis of interest in each individual study is quantified. The evidence is then pooled over studies, providing a joint level of support for the general theory. The aggregation uses updated model probabilities, that is, the posterior odds after observing a first data set are used as the prior odds for the second study; and the posterior odds after inclusion of the second study are used as the prior odds for the third study. This process can be repeated for each additional replication study as presented in: $$\left(\frac{P(H_1 \mid Y)}{P(H_2 \mid Y)}\right)^N = \frac{P(H_1)}{P(H_2)} \prod_{n=1}^{N} (B_{12})^n,$$ where $n = 1, \ldots, N$ indicates the number of studies and Y is the denotes the data. Note that the prior odds before the first study $P(H_1)/P(H_2)$ is often set to one, reflecting no preference for either hypothesis before any data was observed. A closely linked variant of this is the MAMBA, introduced for replicability for genome data.	"Given several conceptual replications with substantial diversity in data, design and methods but investigating the same theory, what is the evidence undelying a certaing theory of interest?"	Several substantially different replications investigating the same theory of interest	To quantify	First mentioned in [70]. Discussed in Variant for genome data in [71].
Bayesian mixture model for reproducibility rate	It is a model for the p-values from the original results and the replications, in order to assess the reproducibility rate and to investigate whether some characteristics of the studies are associated with how likely they reproduce. In the mixture model each pair of p-values (original and replication) comes from a mixture distribution were one component describes the p-value behaviour under the null hypothesis and the second under the alternative. All included original studies claim a significant result, the weight given to the second component of the mixture can be seen as a reproducibility rate. As such, the model is linked to the significance criterion.	"Given the results (p-values) from a set of original and replication studies, what is the rate of reproducibility, and how is it related to certain aspects of the experiments?"	Several pairs of original and replication studies	To quantify and explain	First mentioned in [72].

Table 4: Metrics table: Summary of the 50 identified metrics *(continued)*

Name (also called/related to)	Description	Research question	Scenario of application	Purpose of Metric	References
Confidence interval: original effect in replication 95% CI (Coverage)	For an original-replication study pair, this metric entails a binary check on whether the original effect size is included in the 95% confidence interval of the replication effect size. When several original-replication study pairs are considered, coverage is calculated as the proportion of pairs in which the original effect was in the CI of the replication.	"Given an original effect size, (what is the probability that) does a repetition of the experiment, with an independent sample of participants, produce(s) a CI that overlaps with the original effect?"	One original and one replication study; or one original and many replication studies	To quantify and classify	First mentioned in [73]. Discussed in [74, 75]. Used in [10, 9, 8, 69, 16].
Confidence interval: replication effect in original 95%CI (Capture probability)	For an original-replication study pair, this metric entails a binary check on whether the replication effect size is included in the 95% confidence interval of the original effect size. When several replication studies are performed the shares of replications in that interval is captured via the capture probability, which is defined as the percentage of replication means that (will) fall within a given original CI.	"Given an effect size and 95% CI, (what is the probability that) does a repetition of the experiment, with an independent sample of participants, give(s) an effect that falls within the original CI?"	One original and one replication study; or one original and many replication studies	To quantify and classify	First mentioned in [73]. Discussed in [76, 68]. Used in [10, 30, 31].
Consistency of original with replications, P_{orig}	This metric represents the probability that the effect estimate from the original study would be as extreme or more extreme than it actually was if the original study and the replications were statistically consistent (defined here as being drawn from the same distribution)	"To what extent are the replication effect sizes consistent with the effect sizes of an original study?"	One original study and several replication studies	To quantify	First mentioned in [77]. Discussed in [52, 33]. Used in [27, 32].
Continuously cumulating meta-analytic approach	Continuously cumulating metaanalysis (CCMA) uses standard meta-analytic calculations in a continuing fashion after each new replication attempt completes. Instead of simply noting whether each individual replication attempt reached significance, CCMA combines the data from all studies that were completed so far and computes meta-analytic indexes to quantify the evidence	"Given subsequent replications that were performed to date, what is the current evidence for an effect?"	One original study and several replication studies; or several replications	To quantify	First mentioned in [78]. Discussed in [79, 6, 74].
Correlation between effects	Replication is assessed in terms of the linear relationship between effect estimates, including numerically with the Pearson or Spearman correlation as well as visually with scatterplots. For successful replications the correlation should be close to 1.	"Do the replication studies and the original studies produce effects that are correlated?"	Several pairs of original and replication studies	To quantify	Discussed in [80]. Used in [22].

Table 4: Metrics table: Summary of the 50 identified metrics *(continued)*

Name (also called/related to)	Description	Research question	Scenario of application	Purpose of Metric	References
Correspondence test	This measure combines a difference (related to the Q-test) and equivalence test in the same framework. The correspondence test allows for a more nuanced inference regarding replication success or failure based on whether the null hypothesis of either test can or cannot be rejected. The test has four possible outcomes: equivalence if the difference test is non-significant and the equivalence test is significant, difference if the difference test is significant and the equivalence test is non-significant, trivial difference if the difference test is significant and the equivalence test is significant and indeterminacy if the difference test or the equivalence test are significant .	"To what extent does the effect size from the replication study differ or is equivalent to that of the original study?"	One original study and one replication study	To classify	First mentioned in [51].
Credibility analysis (Reverse-Bayes, probability of credibility, probability of replicating an effect)	The analysis of credibility uses the results of a study (specifically the confidence interval) and uses a Reverse-Bayes approach to find the prior that is required to generate credible evidence for the existence of an effect (i.e., a posterior that excludes no effect). The prior is then compared with internal or external evidence to assess if the finding is credible or not.	"How credible are the results of a study, in a Bayesian framework?"	One original study	To quantify and classify	First mentioned in [81]. Discussed in [82, 83].
Cross-validation methods (Jackknife, bootstrap)	Internal cross-validation methodology are used to test result replicability, where the results received in one subsample of the raw data can be confirmed in the remaining data. The degree of shrinkage (validity shrinkage) is then estimated using the difference in R^2 between the subsamples providing a theoretical basis to evaluate the reproducibility of result. The closer shrinkage is estimated to be zero, the greater the degree of stability and more confidence in the replicability/generalisability of the results. Alternatively, jackknife and bootstrap validation methods can be used.	"To what extent can the stability of a result be trusted, and to what extent can the result be generalized?"	One original study	To quantify and predict	First mentioned in [84]. Discussed in [85, 86].

Table 4: Metrics table: Summary of the 50 identified metrics *(continued)*

Name (also called/related to)	Description	Research question	Scenario of application	Purpose of Metric	References
Design analysis	Given that a study was performed that yielded an estimate d with standard error s. Then a true effect-size D (the value that d would take if observed in a very large sample) has to be considered. The random variable d^{rep} is defined as the estimate that would be observed in a hypothetical replication study with a design identical to that used in the original study. A probability model for d^{rep} then gives the following three summaries: (1) The power: the probability that the replication d^{rep} is larger (in absolute value) than the critical value that is considered to define "statistical significance" in this analysis; (2) The Type S error rate: the probability that the replicated estimate has the incorrect sign, if it is statistically significantly different from zero; (3) The exaggeration ratio (expected Type M error): the expectation of the absolute value of the estimate divided by the effect size, if statistically significantly different from zero.	"Given the results of an original study and an effect of a hypothetical replication study, what is the probability of the estimate being in the wrong direction, and what is the factor by which the magnitude of the effect is overestimated?"	One original study	To quantify and explain	First mentioned in [87].

Table 4: Metrics table: Summary of the 50 identified metrics *(continued)*

Name (also called/related to)	Description	Research question	Scenario of application	Purpose of Metric	References
Difference in effect size (Q-statistic, (meta-analytic) Q-test, difference test, Tukey's post-hoc honest significant difference test)	The original and replication effect sizes can be compared by calculating their difference together with its confidence interval. They can further be compared in a significance testing paradigm using the Q-statistic or difference test. Alternatively, when there is data for several original-replication study pairs, a paired t-test and/or Wilcoxon test can be applied on the effect size estimates for the original and replication studies. Tukey's post-hoc honest significant difference test can be used to answer the question of how many replications produced results that were statistically indistinguishable from one another.	"To which degree do the effects from a replication study mirror the original"	One original and one replication study; or several replications (meta-analytic Q-test)	To quantify and classify	First mentioned in [88] (Q-statistic for reproducibility). Discussed in [52, 89, 90, 91, 68, 80, 75, 92, 74, 51]. Used in [26, 35, 38, 93, 32, 47, 10, 25, 94, 95, 96, 34, 97, 22, 8, 98, 69, 36, 30, 99, 29, 100, 101, 31, 102, 17, 24, 21, 103, 104].
Equivalence testing (TOST (two one-sided tests))	An equivalence range is constructed based on an equivalence margin, or a smallest effect size of interest. When assessing the replication of an original "null" (non-significant) finding a successful replication would reject the null hypothesis of an effect being outside the equivalence region. Alternatively, when interested in assessing whether the original and the replication study find consistent or equivalent effects, one can test whether the difference in effect size falls within a region of equivalence.	"For the replication of an original null finding, does the replication study find an effect that is equally negligeable?" - "Are the results from the replication statistically equivalent to the results of the original study?"	One original and one replication study	To quantify and classify	Discussed in [6, 68, 92, 51, 105].
Externally standardized residuals	For each $i = 1, \ldots, n$, the replication effect size i is compared to the weighted mean effect size of all replications excluding study i via a standardized difference. These residuals can than inform on a failure to replicate. They tend to be ambiguous about successful replications. This metric is related to the measure of reproducibility of the studies included in a meta-analysis introduced by [106].	"Is the original study consistent with the replication(s)?" - "Are all studies included in a meta-analysis replicable?"	One original study and one replication; or one original study and many replications	To quantify and classify	First mentioned in [90]. Discussed in [106].

Table 4: Metrics table: Summary of the 50 identified metrics *(continued)*

Name (also called/related to)	Description	Research question	Scenario of application	Purpose of Metric	References
Fragility Index (Fragility quotient)	The fragility index was proposed to quantify the robustness of statistical significance of clinical studies with binary outcomes. It is defined as the minimal event status modifications that can alter statistical significance. If the original study result is statistically significant (with $p(0,0) < \alpha$), the fragility index is defined as $$FI = \min_{p(f_0,f_1)\geq\alpha} |f_0| + |f_1|,$$ where f_0 and f_1 are the numbers of non-events changed to events in groups 0 and 1, respectively. If the original study result is non-significant (with $p(0,0) \geq \alpha$), the min is searched for all f_0 and f_1 with $p(f_0,f_1) < \alpha$. A smaller value of FI indicates a more fragile results. The FI was extended to meta-analyses and network meta-analyses. One may use the relative measure, fragility quotient (FQ), to compare the multiple studies' fragility. Specifically, $$FQ = \frac{FI}{n_0+n_1} \times 100\%$$ where $n_0 + n_1$ is the total sample size of the study. Thus, the FQ represents the minimal percentage change of event status among all participants that can alter the significance (or non-significance), and it ranges within 0 and 10%.	"Given the results of an original study were significant, what is the smallest change in the original data that is needed to deem the results non-significant? and vice-versa for original null results" - "How fragile are the original results to small changes in the underlying data?"	One original study	To quantify	First mentioned in [107]. Discussed in [108, 109].

Table 4: Metrics table: Summary of the 50 identified metrics *(continued)*

Name (also called/related to)	Description	Research question	Scenario of application	Purpose of Metric	References
I squared - I^2 (Estimation of effect variance)	I squared describes the percentage of total variation across studies (replications) that is due to heterogeneity rather than chance, and is calculated from basic results obtained from a typical meta-analysis: $$I^2 = 100\% \times (Q - \mathrm{df})/Q,$$ where Q is Cochran's heterogeneity statistic and df the degrees of freedom. Any negative values of I^2 are set to zero so that it lies between 0 and 100%. A value of 0% indicates no observed heterogeneity, and larger values show increasing heterogeneity.	"Given a set of replications, to what extent is the total variation across study results due to heterogeneity?" - "How consistent are the results across replications?"	Several replications; one original and several replications	To quantify	First mentioned in [110]. Discussed in [91]. Used in [34, 23].
Jaccard similarity coefficient (Coefficient of similarity)	The percent overlap of activation between two fMRI studies (j and l) is defined as $$w_{j,l} = \frac{V_{j,l}}{V_j + V_l - V_{j,l}},$$ where V_j and V_l are the number of voxels identified as activated in either experiment and $V_{j,l}$ is the number of voxels identified as activated in both experiments. [111] suggest using a measure that is closely related to the Jaccard coefficient to measure reproducibility in omics data analysis.	"By what extent do the results of two (or more) fMRI experiments overlap?"	One original study and one replication study; or several replications	To quantify	Discussed in [112, 111]. Used in [113].
Leave-one-out error	A model is trained on all data without the ith data point, and tested on the ith data point. The leave-one-out error is then directly related to the average loss or error over all i.	"Given a deep learning model, how generalisable are its results?"	One original study	To quantify and predict	Discussed in [114].

Table 4: Metrics table: Summary of the 50 identified metrics *(continued)*

Name (also called/related to)	Description	Research question	Scenario of application	Purpose of Metric	References
Likelihood-based approach for reproducibility (Likelihood-ratio)	The design of the original study is used to derive an estimate of a theoretically interesting effect size, d_{tie}. A likelihood ratio is then calculated to contrast the match of two models to the data from the replication attempt: a model based on the derived d_{tie}, and a null model. More specifically, a null model assuming no effect and a replication model that assumes the effect is d_{tie}. The magnitude of the likelihood ratio describes the strength of the evidence in favor of one or the other model. Very large ratios in favor of d_{tie} would be considered strong evidence for replication. Symmetrically, very large ratios in favor of the null model would be strong evidence against replication.	"Given a theoretically interesting effect size derived from the original study, what is the evidence for or against replicating this effect?"	One original study and one replication study	To quantify and classify	First mentioned in [115].
Mean relative effect size (Percentage difference in effect size)	The mean relative effect size is defined as $\nu = \sum_{j=1}^{m} \frac{\theta_{2j}/\theta_{1j}}{m}$, where θ_{2j} and θ_{1j} are the effect sizes from either the original or the replication study and m is the number of findings that were replicated. This value is usually used to assess by how much the effect size changed from original to replication study. Alternatively, the percentage difference can be used.	"What is the average ratio of replication study effects to original study effects?"	Several pairs of original and replication studies	To quantify	Discussed in [80]. Used in [32, 116, 48, 113, 22, 8].
Meta-analysis	Fixed-effect or random-effects meta-analyses can be used to combine the results from an original and a replication study, or from several replication studies. In the pairwise scenario, a replication is often considered successful if the results of the meta-analysis align with the results of the original study (significance and direction of effect). When several replications are conducted of the same phenomenon, meta-analysis methodology can be used to assess the reproducibility of the finding. To account for potential heterogeneity between studies, random-effects models are used.	"Given an original-replication study pair, does the pooled effect align with that of the original study?" - "Given a set of replications, is the effect size reproducible across studies?"	One original and one replication study; or one original and many replication studies; or several replications	To quantify and classify	Discussed in [117, 89, 68, 75, 92, 74, 11, 118]. Used in [35, 32, 10, 34, 9, 119, 69, 24].
Minimum effect testing	Based on the results of the original study, a minimal level of evidence required to support the original study is defined, as a range constituting the null hypothesis. A test is performed to see whether the replication effect size lies within the range (H_0) or outside (H_1).	"Is the replication effect size significantly different from a minimal effect size of interest, required to support the original study?"	One original and one replication study	To classify	Discussed in [68]. Used in [22].

Table 4: Metrics table: Summary of the 50 identified metrics *(continued)*

Name (also called/related to)	Description	Research question	Scenario of application	Purpose of Metric	References
Network Comparison Test, NCT	This test was proposed to statistically evaluate the similarity of network models.	"Given two network structures, how similar are they to each other?"	One original study and one replication study	To quantify and classify	Discussed in [120, 121].
P interval	The p interval, or prediction interval for p, is an interval with a specified chance (usually 80%) of including the p-value given by a replication.	"Given the results of an original study, what is the range of p-values a replication (following the same design) would lie in with 80% probability?"	One original study	To quantify and predict	First mentioned in [56].
Prediction interval: replication effect in original 95% prediction interval	Using the findings (effect size and variation) of the original study, and the expected variation of the replication study (linked to its sample size), compute the 95% prediction interval. This can be used to predict the effect size of the replication study or, for a binary criterion of replication success, check whether the replication effect size is included in the prediction interval. [75] further show how the metric based on the prediction interval is related to the Q-test.	"Do the findings from the replication study align with a reasonable expectation, given the observed variation in the original study and replication study?" - "Are the replication estimates statistically consistent with the original estimates?"	Original finding only; one original and one replication study; or one original and many replication studies	To quantify and classify	First mentioned in [28]. Discussed in [68, 80, 75, 74]. Used in [27, 10, 9], [24] checked original effect in 95% prediction interval of replications.
Proportion of population effects agreeing in direction with the original, $\hat{P}_{>0}$	This metric assesses the strength of evidence of the replication effect sizes going in the same direction as the original effect size, by estimating the proportion of population effects agreeing in direction with the original effect estimate. It can be generalized by ensuring that they do not only agree in direction but are also stronger than a chosen threshold.	"To what extent do the replication effect sizes agree with the sign found in the original study?"	One original study and several replication studies	To quantify	First mentioned in [77]. Discussed in [52, 33]. Used in [32].
Quantified reproducibility assessment, QRA	The method is based on the concepts and definitions of metrology. For QRA, the precision of measurements done in replications across varying conditions is assessed.	"After performing multiple measurements of an object, what is the precision of the measured quantity obtained?"	One original study and many replication studies	To quantify and classify	First mentioned in [122]. Discussed in [123, 124].

Table 4: Metrics table: Summary of the 50 identified metrics *(continued)*

Name (also called/related to)	Description	Research question	Scenario of application	Purpose of Metric	References
Replication Bayes factor	The replication Bayes factor tests the proponent's replication hypothesis $H_r : \theta \sim$posterior distribution from original study vs the null hypothesis $H_0 : \theta = 0$ of a skeptic who has reason to doubt the presence of an effect: $$B_{r0} = \frac{f(Y_r \mid H_r)}{f(Y_r \mid H_0)},$$ where $f(Y_r \mid H_i)$ is the marginal likelihood of the data under hypothesis H_i with $i \in \{0, 1\}$.The higher the B_{r0} the more evidence for the replication hypothesis.	"What is the evidence for the effect from the replication attempt being comparable to what was found in the original study, or absent?" - "Are the replication results more consistent with the original study or with a null effect?"	One original and one replication study	To quantify	First mentioned in [61]. Discussed in [125, 115, 68, 11, 126]. Used in [69].
Sceptical p-value (versions: nominal sceptical p-value, golden sceptical p-value, controlled sceptical p-value)	Replication success is declared if the replication study is in conflict with a sceptical prior that would make the original study non significant. The sceptical p-value quantifies the prior-data conflict. [57] introduced the nominal p-value. Two more recalibrations have been proposed since. The nominal p-value might be too stringent as it needs both original and replication study to be significant at level α. With the golden recalibration it is possible to establish replication success, original and replication study do not both necessarily need to be significant at level α, provided that the replication effect estimate does not shrink compared to the original one. The controlled p-value was introduced to guarantee overall type I error control at α^2 and is closely related to the significance criterion.	"To what extent are the results of a replication study in conflict with the beliefs of a sceptic of the original study?"	One original study and one replication study	To quantify and classify	First mentioned in [57]. Discussed in [11, 127, 58].
Sceptical Bayes Factor (Reverse-Bayes)	The sceptical Bayes factor combines reverse-Bayes analysis with Bayesian hypothesis testing. First, a sceptical prior is determined for the effect size such that the original finding is no longer convincing in terms of Bayes factors. Then, this prior is contrasted to an advocacy prior (the reference posterior of the effect size based on the original study). Replication success is flagged if the replication data favour the advocacy over the sceptical prior at a higher level than the original data favoured the sceptical prior over the null hypothesis. The highest level for which replication success would be declared is then the sceptical Bayes factor.	"In light of the replication data, at which level of evidence can an advocate of the original study convince a sceptic?"	One original study and one replication study	To quantify and classify	First mentioned in [62].

Table 4: Metrics table: Summary of the 50 identified metrics *(continued)*

Name (also called/related to)	Description	Research question	Scenario of application	Purpose of Metric	References
Significance criterion (vote counting, two-trials rule, regulatory agreement)	For an original-replication study pair, replication success is concluded when both original study and replication study find a statistically significant effect, in the same direction. This can be done either with directional two-sided hypothesis tests, or via a one-sided test. For a continuous assessment of reproducibility, $\max(p_o, p_r)$ can be used, where p_o and p_r are the p-values from the original and replication, respectively.	"Do the original and replication study both find a statistical significant effect in the same direction?"	One original and one replication study; or several original-replication study pairs, or several replications	To quantify and classify	Discussed in [6, 68, 80, 75, 92, 74, 11, 51]. Used in [26, 38, 27, 93, 40, 47, 10, 116, 25, 128, 94, 95, 39, 129, 96, 37, 9, 97, 8, 69, 30, 99, 16, 29, 23, 101, 31, 17, 24, 103, 104].
Small Telescopes	Based on the sample size and the statistical test performed in the original study, the effect that the original study has 33% power to detect, d_{33}, is computed. If the effect size of the replication study is significantly different from d_{33}, a replication failure is concluded.	"Is the replication effect size statistically significantly smaller than a small effect, defined as the effect the original study could detect if it were powered at 33%?"	One original and one replication study	To quantify and classify	First mentioned in [59]. Discussed in [125, 68, 115, 11, 130].
Snapshot hybrid (Bayesian meta-analysis)	The method combines both the original and replication effect size to evaluate the common true effect size. It is a hybrid method because it only takes the statistical significance of the original study into account, whereas it considers evidence of the replication study as unbiased. The snapshot hybrid consists of three steps. First, the likelihood of the effect sizes of the original study and replication is calculated conditional on four hypothesized effect sizes (zero, small, medium, and large). Second, the posterior model probabilities of these four effect sizes are calculated using the likelihoods of step 1 and assuming equal prior model probabilities. Equal prior model probabilities are selected by default, because this refers to an uninformative prior distribution for the encompassing model. Third, when desired, the posterior model probabilities can be recalculated for other than equal prior model probabilities.	"After replicating an original study, what is the evidence for a null, small, medium or large effect?"	One original study and one replication study	To quantify	First mentioned in [131].

Table 4: Metrics table: Summary of the 50 identified metrics *(continued)*

Name (also called/related to)	Description	Research question	Scenario of application	Purpose of Metric	References
Z-curve (Exact replication rate, p-curves)	The Z-curve methodology is a method for estimating the expected replication rate, which can be defined as the predicted success rate of exact replication studies based on the mean power after selection for significance. An extension was proposed which estimates the expected discovery rate in addition, which is the estimate of a proportion that the reported statistically significant results constitute from all conducted statistical tests and can be used to detect and quantify the amount of selection bias.	"Do all replication studies combined provide credible evidence for a phenomenon?"	Several replications	To quantify and predict	First mentioned in [54] (Z-curve), [53] (P-curve). Discussed in [55].
A framework					
Causal replication framework	The framework formalizes the conditions under which replication success can be expected, and allows for the causal interpretation of replication failures. These conditions are summarized into replication assumptions which are qualitatively or narratively assessed. Replication failure occurs when one or more of the causal replication framework assumptions are violated.	"How can a replication failure be interpreted, from a causal perspective"	One original and one replication study; or one original and many replication studies; or several replications	To quantify and explain	First mentioned in [65]. Discussed in [132].
RepeAT - Repeatability Assessment Tool	The tool was developed using a multi-phase method to determine components needed for reproducing biomedical data: a literature review generated a framework which was tested and refined. The RepeAT framework now contains 119 unique variables that were grouped into five categories which address different componnts for reproducible research: research design and aim, database and data collection methods, data mining and data cleaning, data analysis, data sharing and documentation.	"Does the presented research align with community standards of reproducible biomedical research, using electronic health records?"	One original study	To quantify	First mentioned in [64].

Table 4: Metrics table: Summary of the 50 identified metrics *(continued)*

Name (also called/related to)	Description	Research question	Scenario of application	Purpose of Metric	References
Unified framework for estimating the credibility of published research	The unified framework for estimating the credibility of published research examines four fundamental falsifiability-related dimensions: transparency of the methods and data, reproducibility of the results when the same data-processing and analytic decisions are reapplied, robustness of the results to different data-processing and analytic decisions, and reproducibility of the effect. This framework includes a standardized workflow in which the degree to which a finding has survived scrutiny is quantified along these four dimensions. More specifically, for method and data transparency: availability of design details, analytic choices, and underlying data; for analytic reproducibility: ability of reported results to be reproduced by repeating the same data processing and statistical analyses on the original data; for analytic robustness: robustness of results to different data-processing and data-analytic decisions; and for effect reproducibility: ability of the effect to be consistently observed in new samples, at a magnitude similar to that originally reported, when methodologies and conditions similar to those of the original study are used. The framework outlines the steps to investigate these four dimensions.	"For a specific published research work, what is the evidence for its credibility measured on four different dimensions: method and data transparency, analytic reproducibility, analytic robustness and effect reproducibility?"	One original study and many replication studies	To quantify and explain	First mentioned in [63].
A graph					
Bland-Altman Plot (Agreement measures)	When two measures are compared (for example replications and their original studies), the mean difference between the measures and standard deviations of the difference are used to define the limits of agreement. Then the average effect (average of replication and original effect) is plotted against the difference in effect size. The two measures can be used interchangeably if most of the points lie inside the limits of agreement. Other related agreement parameters can be used as well.	"Do the effects estimated in several original-replication study pairs agree with each other?" - "How good is the agreement between repeated measures/studies?"	Several pairs of original and replication studies	To quantify and classify	First mentioned in [133]. Discussed in [134]. Used in [22, 21].

Table 4: Metrics table: Summary of the 50 identified metrics *(continued)*

Name (also called/related to)	Description	Research question	Scenario of application	Purpose of Metric	References
Modified Brinley plot	The plot summarises the results for several replications including a comparison (A vs. B) by plotting the means of one phase (A, baseline) against the mean of the second phase (B, intervention) for each comparison. An identity line (diagonal with intercept = 0, slope = 1) is included to represent the lack of difference between means. A desired postintervention level and a desired amount of change after introducing the intervention is specified to define an area of the plot in which the dots should fall if they all meet both requirements. The share of points in the area gives the degree of replication.	"Given a pre-specified desired effect and multiple replications, what is the share of replications that, represented graphically, achieve the desired effect?"	Several replications	To quantify and classify	First mentioned in [135]. Discussed in [136].
Reproducibility Maps	The fMRI images are colored depending on whether or not the truly active voxels were stongly reproducible or not	"For fMRI research, how many and which of the truly active voxels were strongly reproduced?"	Several replications	To quantify and classify	First mentioned in [137].
A study, survey, or questionnaire					
Prediction market	Based on original results and information on the design of planned replication studies, participants in a prediction market trade contracts on the possible outcome of a replication study. The contracts pay a certain amount of money if the replication is successful. The traded contracts then allow the price to be interpreted as the predicted probability of the outcome occurring.	"What do the participants in a prediction market predict as the probability that the original findings will replicate?"	One original study with a planned replication; or several original studies with planned replications	To quantify and classify	First mentioned in [138]. Used in [40, 9].
Presence/Absence of elements ensuring reproducibilty,via proxies (Framework for evaluating rigor and reproducibility)	An original paper is checked for the presence or absence of certain design and reporting elements that are crucial for its reproducibility. This is often achieved using checklists or reporting guidelines which summarise the community standards. The elements of theses checklists or guidelines are usually integrated in a study, survey or questionnaire.	"Do the design, methods and reporting of the original paper align with community standards of reproducible and transparent research?"	One original study	To quantify and classify	Discussed in [139, 123, 140, 141].

Table 4: Metrics table: Summary of the 50 identified metrics *(continued)*

Name (also called/related to)	Description	Research question	Scenario of application	Purpose of Metric	References
RepliCATS	The process elicits expert predictions about the reproducibility of research. It is based on a modified Delphi technique and includes four steps represented in the acronym IDEA: 'Investigate', 'Discuss', 'Estimate' and 'Aggregate'. Each individual is provided a scientific claim and the original research paper to read, and provide an estimate of whether or not the claim will replicate (Investigate). They then see the group's judgements and reasoning, and can interrogate these (Discuss). Following this, each individual provides a second private assessment (Estimate). A mathematical aggregation of the individual estimates is taken as the final assessment (Aggregate).	"How reliable do experts believe the claims from an original finding are?"	One original study	To quantify and predict	First mentioned in [42]. Used in [41].
Subjective reproducibility assessment (Replication standard, assessment of feasibility)	The replication teams are surveyed/asked to answer the question "Did your results replicate the original effect?". The teams can give a binary answer, or give a more nuanced interpretation on, for example, a Likert scale. Specific fields have specified their own categories for reproducibility assessment, as for example the replication standard in agent-based modeling: "numerical identity", "distributional equivalence", and "relational alignment". For the reproducibility of simulation studies, agreement between results from the replication studies and the original studies was assessed in a qualitative manner and involved evaluating: whether numerical values from the replication studies were comparable to those in the original studies, whether trends in the results were moving in the same direction, and whether the performance rankings of different simulation scenarios matched those in the original studies [44].	"Does the replication team consider the replication as successful?" - "To what extent does the replication team trust in the reproducibility of a finding?"	One original study and one replication study	To quantify and classify	Discussed in [142, 74, 44]. Used in [26, 44, 38, 27, 40, 47, 43, 39, 44, 37, 8, 49, 16, 143, 21, 103].
An algorithm					

Table 4: Metrics table: Summary of the 50 identified metrics *(continued)*

Name (also called/related to)	Description	Research question	Scenario of application	Purpose of Metric	References
Reproducibility scale of workflow execution - Tonkaz	The metric is based on the idea of evaluating the reproducibility of results using biological feature values (e.g., number of reads, mapping rate, and variant frequency) representing their biological interpretation. The resulting reproducibility scale is a four point scale and goes from "Fully Reproduced" to "Acceptable Differences" to "Unacceptable Differences" to "Not Reproduced". The authors implemented an automated system to classify results on this scale.	"Given a certain original research paper with results based on computation, can the workflow to generate the results be executed and verified?"	One original study	To classify	First mentioned in [144].
RipetaScore	The ripetaScore combines three aspects of trust for a total of 30 points: 1. using the "Trust in Research" criteria it is determined whether a paper is a research paper. Only then will the paper continue to be scored. 2. The paper is then evaluated for the presence of reproducibility quality indicators and it can receive up to 20 points. Another 10 points come from the trust in professionalism quality indicators. For the trust in reproducibility criteria, papers are primarily evaluated with regards to their data/code sharing practices, reporting of methods, and citing software. These criteria are all assessed via natural language processing.	"Given certain trust in research, reproducibility and professionalism quality indicators, how high does a paper score?"	One original study	To quantify	First mentioned in [145].
Text-based machine learning model to estimate reproducibility	A machine learning model using an ensemble of random forest and logistic regression was trained on data from replication studies. This model can then use a paper's text and meta-data to predict its likelihood of replication, based on the significance criterion.	"Given the text of an original paper, what is the probability of replication success?"	One original study; or several original studies	To quantify and predict	First mentioned in [146]. Discussed in [147, 123]. Used in [41].

3.5 EDI considerations in reproducibility assessment

For 18 of the 49 application papers (37%) and 15 of the 95 methodological papers (16%) we extracted content related to EDI. The extracted text was grouped into five themes: *diversity in replication teams*, *diversity in replication samples*, *epistemic diversity*, *generalization of findings*, and *research culture*. Methodological papers overwhelmingly focused on epistemic diversity, defined as the diversity of knowledge production, expertise, field, method of study, epistemic values, and/or reasoning [148, 5]. This epistemic diversity was reflected in the methods papers either via encouraging future studies generalizing the metric to fields other than those initially proposed, or an explanation that the metric is only relevant for a specific field or method of study. Application papers were more likely to encourage diversity of replication teams (those conducting replication studies) or replication samples (both human and non-human samples). Several application papers highlighted the importance of generalizability and heterogeneity, noting that increased diversity and heterogeneity in replications may lead to increased generalizability when findings of multiple replications are considered in aggregate. Lastly, two papers (one application and one methodological) noted the relevance of research culture to reproducibility and reproducibility metrics, suggesting that social and cultural factors can facilitate or impede uptake of reproducible research practices and replication projects. The raw data containing the extracted texts on EDI considerations are available via osf.io/sbcy3/.

4 Discussion

In this study, we systematically searched the methodological literature on metrics to quantify, assess, explain, or predict reproducibility. This review was complemented by an investigation into the reproducibility metrics that have so far been used in large-scale replication projects. Our search included 49 replication projects and 95 distinct methodological papers. We identified 50 different metrics and summarized them in a table which organized the metrics by type – formulas or statistical models, frameworks, graphs, studies, survey, or questionnaire, and algorithms. When conceptualizing this review, we did not expect to find such a high number of metrics. The fact that they are diverse in nature and address slightly different questions and aspects of reproducibility, underpins the complexity of measuring reproducibility. Therefore, there cannot be a single, universally applicable reproducibility metric; it should be a case-by-case choice aligned with the goals of the study.

Classifying the metrics to one specific type of reproducibility was not straightforward and might not even be possible. While many metrics have been developed or applied with one aspect of reproducibility in mind, they can often be directly applied or can be extended to other aspects. Future research focusing on specific aspects of reproducibility can build on our results by selecting the metrics to apply in that context and investigate their assumption and limitations. Our reproducibility metrics table is an important contribution that provides a clear overview of available metrics, their potential applications and references for further information. We hope that it will serve as a practical tool for future replication teams to plan their projects more effectively, as it offers a way to align the type and aim of a study with the most appropriate metric(s), based on the research questions under consideration. The metrics table additionally offers opportunities for researchers to explore new metrics and make informed decisions on which metrics best fit their study design, and constraints. For those new to the field, considerations related to cost and ease of implementation of the various metrics are highlighted in the online version of our table (rachelhey.github.io/reproducibility_metrics/). Peer-reviewers can use the table to critically review reproducibility studies regarding the appropriateness of the metric(s) used. Meta-researchers can find reproducibility outcomes for future intervention studies aiming at improving reproducibility. Our table can help to align reproducibility metrics to the goals of a replication effort [6] or reproducibility

studies. Researchers who want to follow the recommendation that the design of replication efforts should be informed by the reproducibility metrics [149], may find the information in the table helpful. A noteworthy observation from our data extraction is that large-scale replication projects rarely provide a definition of reproducibility. Additionally, while these studies put a lot of effort into describing the design and methods used in the replication, they seldom outline the methods used to summarize reproducibility. Instead, they tend to only report the results in a descriptive manner in the results section. Therefore, we invite researchers to choose the metric(s) that align(s) with their research question and justify this choice. Sharing data and code could further allow for the assessment of the performance of other metrics or how they interact and complement each other in practice.
In an exploratory analysis, we extracted any mention of EDI dimensions. As expected, only a handful of papers included such considerations, but we could still find some valuable data which will be useful in the remainder of the iRISE project, which includes a work-package examining the interface of reproducibility and research culture. Our study also shows, however, that EDI dimensions are explicitly considered only in few instances, and should be given higher priority in future work.

4.1 Limitations

While our search strategy was extensive, we cannot be sure that the list of metrics is fully exhaustive. Due to the epistemic diversity in the understanding of reproducibility, it is possible that we missed relevant metrics because our keywords did not capture this diversity. Additionally, as our review only captures a snapshot in time, we hope to update the online, "live" version of our table whenever new metrics become available (as for example Held, Pawel, and Micheloud [150] which was published after our literature search). Therefore, the research community is invited to suggest the addition of other reproducibility metrics. Second, we did not critically evaluate or scrutinize the quality or effectiveness of the metrics identified but rather focused on collecting and characterising them. Future research should build on this work and involve a rigorous assessment of the metrics to better understand their strengths and weaknesses. Third, specifically for the application papers, we did not investigate the relationship between the metrics used and the outcome of the projects. For instance, different metrics might produce conflicting results, where one indicates replication success or high reproducibility while the other suggests failure or low reproducibility. Finally, due to resource constraints, we decided to exclude single study application papers from our review. While they, as described above, generally use the same set of metrics, it could be, that the way results are analyzed differs from large-scale studies (e.g., because researchers can zoom in closer, as there is only one original-replication pair). This could be another avenue for future research and complement our review of large-scale replication projects, as well as the work done by Cobey et al. [7].

4.2 Conclusion

Our review offers a comprehensive overview of various reproducibility metrics. By providing classifications of their types, their potential applications, and ease of implementation, we hope to assist future replication teams and meta-researchers to make informed research decisions. We have also paved the way for future research to critically evaluate these metrics further and explore real-world implications.

Data and software availability

All records included (after screening) in our review are organized in a Zotero library (zotero.org/groups/5397531/reproducibilitymetrics), and the methodological papers from the literature search are in included in another Zotero library

(zotero.org/groups/5630395/reproducibilitymetrics_methodsscreening). The complete set of records screened for the methodological papers is available via zotero.org/groups/5630395/reproducibilitymetrics_methodsscreening. The data files with the data extraction, of both application and methodological papers are shared via the Open Science Framework (osf.io/sbcy3/). Analysis code to produce summary statistics, Figures and Tables are available via DOI: 10.17605/OSF.IO/QWR2B.

Funding statement

RH, BV, HW, SKM, LH, KW and SZ receive funding from iRISE. iRISE receives funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101094853. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Executive Agency (ERA). Neither the European Union nor the ERA can be held responsible for them. iRISE also receives funding from the Swiss State Secretariat for Education, Research and Innovation (SERI): Direct Funding for Collaborative Projects as part of the transitional measures, and from UK Research and Innovation (UKRI). HH was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation - Project-ID 422744262 - TRR 289).

Competing interests

SP and LH have developed two of the metrics identified in this review.

Acknowledgement

We thank Robin Segerer, information specialist from University Library Zurich, for help with the search strategy, Flora Logoz, research assistant at University of Zurich, for help with the online version of our table, and Laura Caquelin and Gustav Nilsonne for valuable feedback on an earlier version of our manuscript. Additionally, we would like to thank the iRISE consortium, and specially work package 1, for continuous feedback in the conceptualization and reporting of our work, and FORRT (Framework for Open and Reproducible Research Training) for spreading our call for contribution and collaboration in their community.

Author contributions

- **Conceptualization:** RH, SP, BV, HW, SKM, LH, KW
- **Data curation:** RH, SP, JF, HH, LT, SZ
- **Formal Analysis:** RH, SP, JF, HH, LT, SZ
- **Funding acquisition:** RH, HW, SKM, LH, KW
- **Methodology:** RH, SKM, KW
- **Project administration:** RH
- **Software:** RH
- **Visualization:** RH

- **Writing – original draft:** RH
- **Writing – review & editing:** SP, JF BV, HW, SKM, LH, KW, HH, LT, SZ

References

[1] F. Steinle. "Stability and Replication of Experimental Results: A Historical Perspective". In: *Reproducibility*. John Wiley & Sons, Ltd, 2016, pp. 39–63.

[2] P. R. Dear. *Revolutionizing the sciences: European knowledge in transition, 1500-1700*. Third edition. Oxford: macmillan international, Higher Education, 2019.

[3] S. N. Goodman, D. Fanelli, and J. P. A. Ioannidis. "What does research reproducibility mean?" In: *Science Translational Medicine* 8.341 (2016). Publisher: American Association for the Advancement of Science, 341ps12–341ps12. DOI: 10.1126/scitranslmed.aaf5027.

[4] L. A. Barba. *Terminologies for Reproducible Research*. 2018. DOI: 10.48550/arXiv.1802.03311.

[5] B. Voelkl et al. "The iRISE Reproducibility Glossary". In: (2024). Publisher: Open Science Framework. DOI: https://doi.org/10.17605/OSF.IO/BR9SP.

[6] S. F. Anderson and S. E. Maxwell. "There's more than one way to conduct a replication study: Beyond statistical significance." In: *Psychological Methods* 21.1 (2016), pp. 1–12. DOI: 10.1037/met0000051.

[7] K. D. Cobey et al. "Epidemiological characteristics and prevalence rates of research reproducibility across disciplines: A scoping review of articles published in 2018-2019". In: *eLife* 12 (2023). Ed. by D. B. Allison, M. Zaidi, C. J. Vorland, A. Lupia, and J. Agley. Publisher: eLife Sciences Publications, Ltd, e78518. DOI: 10.7554/eLife.78518.

[8] Open Science Collaboration. "Estimating the reproducibility of psychological science". In: *Science* 349.6251 (2015). Publisher: American Association for the Advancement of Science, aac4716. DOI: 10.1126/science.aac4716.

[9] C. F. Camerer et al. "Evaluating replicability of laboratory experiments in economics". In: *Science* 351.6280 (2016). Publisher: American Association for the Advancement of Science, pp. 1433–1436. DOI: 10.1126/science.aaf0918.

[10] T. M. Errington et al. "Investigating the replicability of preclinical cancer biology". In: *eLife* 10 (2021). Ed. by R. Pasqualini and E. Franco. Publisher: eLife Sciences Publications, Ltd, e71601. DOI: 10.7554/eLife.71601.

[11] J. Muradchanian, R. Hoekstra, H. Kiers, and D. Van Ravenzwaaij. "How best to quantify replication success? A simulation study on the comparison of replication success metrics". In: *Royal Society Open Science* 8.5 (2021), p. 201697. DOI: 10.1098/rsos.201697.

[12] K. Hung and W. Fithian. "Statistical methods for replicability assessment". In: *The Annals of Applied Statistics* 14.3 (2020). DOI: 10.1214/20-AOAS1336.

[13] B. A. Nosek et al. "Replicability, Robustness, and Reproducibility in Psychological Science". In: *Annual Review of Psychology* 73.1 (2022), pp. 719–748. DOI: 10.1146/annurev-psych-020821-114157.

[14] R. Heyard et al. *Reproducibility Metrics - Study Protocol*. 2023. DOI: 10.17605/OSF.IO/7VC4Z.

[15] A. C. Tricco et al. "PRISMA Extension for Scoping Reviews (PRISMA-ScR): Checklist and Explanation". In: *Annals of Internal Medicine* 169.7 (2018), pp. 467–473. DOI: 10.7326/M18-0850.

[16] F. Cova et al. “Estimating the Reproducibility of Experimental Philosophy”. In: *Review of Philosophy and Psychology* 12.1 (2021), pp. 9–44. DOI: 10.1007/s13164-018-0400-9.

[17] R. A. Klein et al. “Many Labs 2: Investigating Variation in Replicability Across Samples and Settings”. In: *Advances in Methods and Practices in Psychological Science* 1.4 (2018). Publisher: SAGE Publications Inc, pp. 443–490. DOI: 10.1177/2515245918810225.

[18] Z. Bahor et al. “Development and uptake of an online systematic review platform: the early years of the CAMARADES Systematic Review Facility (SyRF)”. In: *BMJ open science* 5.1 (2021), e100103. DOI: 10.1136/bmjos-2020-100103.

[19] M. Aria, T. Le, C. Cuccurullo, A. Belfiore, and J. Choe. “openalexR: An R-Tool for Collecting Bibliometric Data from OpenAlex”. In: *The R Journal* 15.4 (2024), pp. 167–180. DOI: 10.32614/RJ-2023-089.

[20] B. B. McShane, J. L. Tackett, U. Böckenholt, and A. Gelman. “Large-Scale Replication Projects in Contemporary Psychological Research”. In: *The American Statistician* 73.sup1 (2019). Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/00031305.2018.1505655, pp. 99–105. DOI: 10.1080/00031305.2018.1505655.

[21] M. J. Page et al. “The REPRISE project: protocol for an evaluation of REProducibility and Replicability In Syntheses of Evidence”. In: *Systematic Reviews* 10.1 (2021), p. 112. DOI: 10.1186/s13643-021-01670-0.

[22] S. V. Wang et al. “Reproducibility of real-world evidence studies using clinical practice data to inform regulatory and coverage decisions”. In: *Nature Communications* 13.1 (2022), p. 5126. DOI: 10.1038/s41467-022-32310-3.

[23] R. A. Klein et al. “Investigating Variation in Replicability”. In: *Social Psychology* 45.3 (2014). Publisher: Hogrefe Publishing, pp. 142–152. DOI: 10.1027/1864-9335/a000178.

[24] O. B. Amaral, K. Neves, A. P. Wasilewska-Sampaio, and C. F. Carneiro. “The Brazilian Reproducibility Initiative”. In: *eLife* 8 (2019). Ed. by P. Rodgers, T. M. Errington, and R. Klein. Publisher: eLife Sciences Publications, Ltd, e41602. DOI: 10.7554/eLife.41602.

[25] M. Schweinsberg et al. “Same data, different conclusions: Radical dispersion in empirical results when independent analysts operationalize and test the same hypothesis”. In: *Organizational Behavior and Human Decision Processes* 165 (2021), pp. 228–249. DOI: 10.1016/j.obhdp.2021.02.003.

[26] K. Irvine, D. A. Hoffman, and T. Wilkinson-Ryan. “Law and Psychology Grows Up, Goes Online, and Replicates”. In: *Journal of Empirical Legal Studies* 15.2 (2018), pp. 320–355. DOI: 10.1111/jels.12180.

[27] V. Boyce, M. Mathur, and M. C. Frank. “Eleven years of student replication projects provide evidence on the correlates of replicability in psychology”. In: *Royal Society Open Science* 10.11 (2023). Publisher: Royal Society, p. 231240. DOI: 10.1098/rsos.231240.

[28] P. Patil, R. D. Peng, and J. T. Leek. “What Should Researchers Expect When They Replicate Studies? A Statistical View of Replicability in Psychological Science”. In: *Perspectives on Psychological Science* 11.4 (2016). Publisher: [Association for Psychological Science, Sage Publications, Inc.], pp. 539–544.

[29] C. R. Ebersole et al. “Many Labs 3: Evaluating participant pool quality across the academic semester via replication”. In: *Journal of Experimental Social Psychology*. Special Issue: Confirmatory 67 (2016), pp. 68–82. DOI: 10.1016/j.jesp.2015.10.012.

[30] J.-Y. A. Chang, J. B. Chilcott, and N. R. Latimer. *Leveraging real-world data to assess treatment sequences in health economic evaluations: a study protocol for emulating target trials using the English Cancer Registry and US Electronic Health Records-Derived Database*. Monograph. Issue: 24.01 Number: 24.01 Pages: 1-61 Publisher: Sheffield Centre for Health and Related Research, University of Sheffield. 2024.

[31] S. V. Wang, S. Schneeweiss, and RCT-DUPLICATE Initiative. "Emulation of Randomized Clinical Trials With Nonrandomized Database Analyses: Results of 32 Clinical Trials". In: *JAMA* 329.16 (2023), pp. 1376–1385. DOI: 10.1001/jama.2023.4221.

[32] C. R. Ebersole et al. "Many Labs 5: Testing Pre-Data-Collection Peer Review as an Intervention to Increase Replicability". In: *Advances in Methods and Practices in Psychological Science* 3.3 (2020). Publisher: SAGE Publications Inc, pp. 309–331. DOI: 10.1177/2515245920958687.

[33] M. B. Mathur and T. J. VanderWeele. "New Statistical Metrics for Multisite Replication Projects". In: *Journal of the Royal Statistical Society Series A: Statistics in Society* 183.3 (2020), pp. 1145–1166. DOI: 10.1111/rssa.12572.

[34] M. S. Hagger et al. "A Multilab Preregistered Replication of the Ego-Depletion Effect". In: *Perspectives on Psychological Science: A Journal of the Association for Psychological Science* 11.4 (2016), pp. 546–573. DOI: 10.1177/1745691616652873.

[35] A. Milcu et al. "Genotypic variability enhances the reproducibility of an ecological study". In: *Nature Ecology & Evolution* 2.2 (2018), pp. 279–287. DOI: 10.1038/s41559-017-0434-x.

[36] S. Coretta et al. "Multidimensional Signals and Analytic Flexibility: Estimating Degrees of Freedom in Human-Speech Analyses". In: *Advances in Methods and Practices in Psychological Science* 6.3 (2023). Publisher: SAGE Publications Inc, p. 25152459231162567. DOI: 10.1177/25152459231162567.

[37] M. Fišar et al. "Reproducibility in Management Science". In: (2024). Publisher: OSF. DOI: 10.31219/osf.io/mydzv.

[38] F. Naudet et al. "Data sharing and reanalysis of randomized controlled trials in leading biomedical journals with a full data sharing policy: survey of studies published in The BMJ and PLOS Medicine". In: *BMJ* 360 (2018). Publisher: British Medical Journal Publishing Group Section: Research, k400. DOI: 10.1136/bmj.k400.

[39] J. Low et al. "Comparison of two independent systematic reviews of trials of recombinant human bone morphogenetic protein-2 (rhBMP-2): the Yale Open Data Access Medtronic Project". In: *Systematic Reviews* 6.1 (2017), p. 28. DOI: 10.1186/s13643-017-0422-x.

[40] R. Botvinik-Nezer et al. "Variability in the analysis of a single neuroimaging dataset by many teams". In: *Nature* 582.7810 (2020). Number: 7810 Publisher: Nature Publishing Group, pp. 84–88. DOI: 10.1038/s41586-020-2314-9.

[41] N. Alipourfard et al. "Systematizing Confidence in Open Research and Evidence (SCORE)". In: (2024). Publisher: OSF. DOI: 10.31235/osf.io/46mnb.

[42] H. Fraser et al. "Predicting reliability through structured expert elicitation with the repliCATS (Collaborative Assessments for Trustworthy Science) process". In: *PLOS ONE* 18.1 (2023). Ed. by F. Catalá-López, e0274429. DOI: 10.1371/journal.pone.0274429.

[43] N. N. N. van Dongen et al. "Multiple Perspectives on Inference for Two Simple Statistical Scenarios". In: *The American Statistician* 73.1 (2019), pp. 328–339. DOI: 10.1080/00031305.2019.1565553.

[44] K. Luijken et al. "Replicability of simulation studies for the investigation of statistical methods: the RepliSims project". In: *Royal Society Open Science* 11.1 (2024), p. 231003. DOI: 10.1098/rsos.231003.

[45] J. W. Tukey. *Exploratory data analysis*. Addison-Wesley series in behavioral science. Reading (Mass.) Menlo Park (Calif.) London [etc.]: Addison-Wesley publ, 1977.

[46] Y. Liu, A. Kale, T. Althoff, and J. Heer. "Boba: Authoring and Visualizing Multiverse Analyses". In: *IEEE Transactions on Visualization and Computer Graphics* 27.2 (2021). arXiv:2007.05551 [cs], pp. 1753–1763. DOI: 10.1109/TVCG.2020.3028985.

[47] N. Huntington-Klein et al. "The influence of hidden researcher decisions in applied microeconomics". In: *Economic Inquiry* 59.3 (2021), pp. 944–960. DOI: 10.1111/ecin.12992.

[48] R. Kirkby. "Quantitative Macroeconomics: Lessons Learned from Fourteen Replications". In: *Computational Economics* 61.2 (2023), pp. 875–896. DOI: 10.1007/s10614-022-10234-w.

[49] J. A. Bastiaansen et al. "Time to get personal? The impact of researchers choices on the selection of treatment targets using the experience sampling methodology". In: *Journal of Psychosomatic Research* 137 (2020), p. 110211. DOI: 10.1016/j.jpsychores.2020.110211.

[50] S. Pawel, R. Heyard, C. Micheloud, and L. Held. "Replication of "null results" – Absence of evidence or evidence of absence?" In: *eLife* 12 (2024). Publisher: eLife Sciences Publications Limited. DOI: 10.7554/eLife.92311.2.

[51] P. M. Steiner, P. Sheehan, and V. C. Wong. "Correspondence measures for assessing replication success". In: *Psychological Methods* (2023). Place: US Publisher: American Psychological Association, No Pagination Specified–No Pagination Specified. DOI: 10.1037/met0000597.

[52] M. B. Mathur and T. J. VanderWeele. "Challenges and suggestions for defining replication "success" when effects may be heterogeneous: Comment on Hedges and Schauer (2019)." In: *Psychological Methods* 24.5 (2019), pp. 571–575. DOI: 10.1037/met0000223.

[53] U. Simonsohn, L. D. Nelson, and J. P. Simmons. "P-curve: A key to the file-drawer." In: *Journal of Experimental Psychology: General* 143.2 (2014), pp. 534–547. DOI: 10.1037/a0033242.

[54] J. Brunner and U. Schimmack. "Estimating Population Mean Power Under Conditions of Heterogeneity and Selection for Significance". In: *Meta-Psychology* 4 (2020). DOI: 10.15626/MP.2018.874.

[55] F. Bartoš and U. Schimmack. "Z-curve 2.0: Estimating Replication Rates and Discovery Rates". In: *Meta-Psychology* 6 (2022). DOI: 10.15626/MP.2021.2720.

[56] G. Cumming. "Replication and p Intervals: p Values Predict the Future Only Vaguely, but Confidence Intervals Do Much Better". In: *Perspectives on Psychological Science* 3.4 (2008), pp. 286–300. DOI: 10.1111/j.1745-6924.2008.00079.x.

[57] L. Held. "A New Standard for the Analysis and Design of Replication Studies". In: *Journal of the Royal Statistical Society Series A: Statistics in Society* 183.2 (2020), pp. 431–448. DOI: 10.1111/rssa.12493.

[58] C. Micheloud, F. Balabdaoui, and L. Held. "Assessing replicability with the sceptical p-value: Type-I error control and sample size planning". In: *Statistica Neerlandica* 77.4 (2023), pp. 573–591. DOI: 10.1111/stan.12312.

[59] U. Simonsohn. "Small Telescopes: Detectability and the Evaluation of Replication Results". In: *Psychological Science* 26.5 (2015), pp. 559–569. DOI: 10.1177/0956797614567341.

[60] M. J. Bayarri and A. M. Mayoral. “Bayesian Design of “Successful” Replications”. In: *The American Statistician* 56.3 (2002). Publisher: ASA Website _eprint: https://doi.org/10.1198/000313002155, pp. 207–214. DOI: 10.1198/000313002155.

[61] J. Verhagen and E.-J. Wagenmakers. “"Bayesian tests to quantify the result of a replication attempt": Correction to Verhagen and Wagenmakers (2014).” In: *Journal of Experimental Psychology: General* 143.6 (2014), pp. 2073–2073. DOI: 10.1037/a0038326.

[62] S. Pawel and L. Held. “The Sceptical Bayes Factor for the Assessment of Replication Success”. In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 84.3 (2022), pp. 879–911. DOI: 10.1111/rssb.12491.

[63] E. P. LeBel, R. J. McCarthy, B. D. Earp, M. Elson, and W. Vanpaemel. “A Unified Framework to Quantify the Credibility of Scientific Findings”. In: *Advances in Methods and Practices in Psychological Science* 1.3 (2018), pp. 389–402. DOI: 10.1177/2515245918787489.

[64] L. D. McIntosh et al. “Repeat: a framework to assess empirical reproducibility in biomedical research”. In: *BMC Medical Research Methodology* 17.1 (2017), p. 143. DOI: 10.1186/s12874-017-0377-6.

[65] P. M. Steiner, V. C. Wong, and K. Anglin. “A Causal Replication Framework for Designing and Assessing Replication Efforts”. In: *Zeitschrift für Psychologie* 227.4 (2019), pp. 280–292. DOI: 10.1027/2151-2604/a000385.

[66] J. Verhagen and E.-J. Wagenmakers. “Bayesian tests to quantify the result of a replication attempt.” In: *Journal of Experimental Psychology: General* 143.4 (2014), pp. 1457–1475. DOI: 10.1037/a0036731.

[67] J. N. Rouder and R. D. Morey. “Default Bayes Factors for Model Selection in Regression”. In: *Multivariate Behavioral Research* 47.6 (2012), pp. 877–903. DOI: 10.1080/00273171.2012.734737.

[68] R. M. Heirene. “A call for replications of addiction research: which studies should we replicate and what constitutes a ‘successful’ replication?” In: *Addiction Research & Theory* 29.2 (2021), pp. 89–97. DOI: 10.1080/16066359.2020.1751130.

[69] E.-J. Wagenmakers et al. “Registered Replication Report: Strack, Martin, & Stepper (1988)”. In: *Perspectives on Psychological Science* 11.6 (2016). Publisher: SAGE Publications Inc, pp. 917–928. DOI: 10.1177/1745691616674458.

[70] I. Klugkist and T. B. Volker. “Bayesian evidence synthesis for informative hypotheses: An introduction.” In: *Psychological Methods* (2023). DOI: 10.1037/met0000602.

[71] D. McGuire et al. “Model-based assessment of replicability for genome-wide association meta-analysis”. In: *Nature Communications* 12.1 (2021), p. 1964. DOI: 10.1038/s41467-021-21226-z.

[72] F. Pauli. “A Statistical Model to Investigate the Reproducibility Rate Based on Replication Experiments”. In: *International Statistical Review* 87.1 (2019), pp. 68–79. DOI: 10.1111/insr.12273.

[73] M. J. Brandt et al. “The Replication Recipe: What makes for a convincing replication?” In: *Journal of Experimental Social Psychology* 50 (2014), pp. 217–224. DOI: 10.1016/j.jesp.2013.10.005.

[74] S. C. Fletcher. “How (not) to measure replication”. In: *European Journal for Philosophy of Science* 11.2 (2021), p. 57. DOI: 10.1007/s13194-021-00377-2.

[75] J. M. Schauer and L. V. Hedges. “Reconsidering statistical methods for assessing replication.” In: *Psychological Methods* 26.1 (2021), pp. 127–139. DOI: 10.1037/met0000302.

[76] G. Cumming and R. Maillardet. “Confidence intervals and replication: Where will the next mean fall?” In: *Psychological Methods* 11.3 (2006), pp. 217–227. DOI: 10.1037/1082-989X.11.3.217.

[77] M. B. Mathur and T. J. VanderWeele. "New metrics for meta-analyses of heterogeneous effects". In: *Statistics in Medicine* 38.8 (2019), pp. 1336–1342. DOI: 10.1002/sim.8057.

[78] R. Rosenthal. "Replication in behavioral research". In: *Journal of Social Behavior & Personality* 5.4 (1990). Place: US Publisher: Select Press, pp. 1–30.

[79] S. L. Braver, F. J. Thoemmes, and R. Rosenthal. "Continuously Cumulating Meta-Analysis and Replicability". In: *Perspectives on Psychological Science* 9.3 (2014), pp. 333–342. DOI: 10.1177/1745691614529796.

[80] J. M. Schauer et al. "An evaluation of statistical methods for aggregate patterns of replication failure". In: *The Annals of Applied Statistics* 15.1 (2021). DOI: 10.1214/20-AOAS1387.

[81] R. A. J. Matthews. "Methods for Assessing the Credibility of Clinical Trial Outcomes". In: *Drug Information Journal* 35.4 (2001), pp. 1469–1478. DOI: 10.1177/009286150103500442.

[82] L. Held. "The assessment of intrinsic credibility and a new argument for $p < 0.005$". In: *Royal Society Open Science* 6.3 (2019), p. 181534. DOI: 10.1098/rsos.181534.

[83] L. Held, R. Matthews, M. Ott, and S. Pawel. "Reverse-Bayes methods for evidence assessment and research synthesis". In: *Research Synthesis Methods* 13.3 (2022), pp. 295–314. DOI: 10.1002/jrsm.1538.

[84] B. Thompson. "The Pivotal Role of Replication in Psychological Research: Empirically Evaluating the Replicability of Sample Results". In: *Journal of Personality* 62.2 (1994), pp. 157–176. DOI: 10.1111/j.1467-6494.1994.tb00289.x.

[85] J. Guan, P. Xiang, and X. D. Keating. "Evaluating the Replicability of Sample Results: A Tutorial of Double Cross-Validation Methods". In: *Measurement in Physical Education and Exercise Science* 8.4 (2004), pp. 227–241. DOI: 10.1207/s15327841mpee0804_4.

[86] Q. C. Song, C. Tang, and S. Wee. "Making Sense of Model Generalizability: A Tutorial on Cross-Validation in R and Shiny". In: *Advances in Methods and Practices in Psychological Science* 4.1 (2021), p. 251524592094706. DOI: 10.1177/2515245920947067.

[87] A. Gelman and J. Carlin. "Beyond Power Calculations: Assessing Type S (Sign) and Type M (Magnitude) Errors". In: *Perspectives on Psychological Science* 9.6 (2014), pp. 641–651. DOI: 10.1177/1745691614551642.

[88] L. V. Hedges and J. M. Schauer. "Statistical analyses for studying replication: Meta-analytic perspectives." In: *Psychological Methods* 24.5 (2019), pp. 557–570. DOI: 10.1037/met0000189.

[89] L. V. Hedges and J. M. Schauer. "More Than One Replication Study Is Needed for Unambiguous Tests of Replication". In: *Journal of Educational and Behavioral Statistics* 44.5 (2019), pp. 543–570. DOI: 10.3102/1076998619852953.

[90] J. M. Schauer and L. V. Hedges. "Assessing heterogeneity and power in replications of psychological experiments." In: *Psychological Bulletin* 146.8 (2020), pp. 701–719. DOI: 10.1037/bul0000232.

[91] L. V. Hedges and J. M. Schauer. "The Design of Replication Studies". In: *Journal of the Royal Statistical Society Series A: Statistics in Society* 184.3 (2021), pp. 868–886. DOI: 10.1111/rssa.12688.

[92] D. G. Bonett. "Design and Analysis of Replication Studies". In: *Organizational Research Methods* 24.3 (2021). Publisher: SAGE Publications Inc, pp. 513–529. DOI: 10.1177/1094428120911088.

[93] S. Hoogeveen et al. "A many-analysts approach to the relation between religiosity and well-being". In: *Religion, Brain & Behavior* 13.3 (2023), pp. 237–283. DOI: 10.1080/2153599X.2022.2070255.

[94] M. Arroyo-Araujo et al. "Systematic assessment of the replicability and generalizability of preclinical findings: Impact of protocol harmonization across laboratory sites". In: *PLOS Biology* 20.11 (2022). Publisher: Public Library of Science, e3001886. DOI: 10.1371/journal.pbio.3001886.

[95] R. Silberzahn et al. "Many Analysts, One Data Set: Making Transparent How Variations in Analytic Choices Affect Results". In: *Advances in Methods and Practices in Psychological Science* 1.3 (2018). Publisher: SAGE Publications Inc, pp. 337–356. DOI: 10.1177/2515245917747646.

[96] C. F. Camerer et al. "Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015". In: *Nature Human Behaviour* 2.9 (2018). Number: 9 Publisher: Nature Publishing Group, pp. 637–644. DOI: 10.1038/s41562-018-0399-z.

[97] I. Cheung et al. "Registered Replication Report: Study 1 From Finkel, Rusbult, Kumashiro, & Hannon (2002)". In: *Perspectives on Psychological Science* 11.5 (2016). Publisher: SAGE Publications Inc, pp. 750–764. DOI: 10.1177/1745691616664694.

[98] S. M. Amini and C. F. Parmeter. "Comparison of Model Averaging Techniques: Assessing Growth Determinants". In: *Journal of Applied Econometrics* 27.5 (2012), pp. 870–876. DOI: 10.1002/jae.2288.

[99] J. Hanousek, D. Hajkova, and R. K. Filer. "A rise by any other name? Sensitivity of growth regressions to data source". In: *Journal of Macroeconomics* 30.3 (2008), pp. 1188–1206. DOI: 10.1016/j.jmacro.2007.08.015.

[100] R. A. Klein et al. "Many Labs 4: Failure to Replicate Mortality Salience Effect With and Without Original Author Involvement". In: *Collabra: Psychology* 8.1 (2022), p. 35271. DOI: 10.1525/collabra.35271.

[101] J. Brauer. "Data, Models, Coefficients: The Case of United States Military Expenditure". In: *Conflict Management and Peace Science* 24.1 (2007). Publisher: SAGE Publications Ltd, pp. 55–64. DOI: 10.1080/07388940601102845.

[102] S. Bouwmeester et al. "Registered Replication Report: Rand, Greene, and Nowak (2012)". In: *Perspectives on Psychological Science* 12.3 (2017). Publisher: SAGE Publications Inc, pp. 527–542. DOI: 10.1177/1745691617693624.

[103] N. Breznau et al. "Observing many researchers using the same data and hypothesis reveals a hidden universe of uncertainty". In: *Proceedings of the National Academy of Sciences* 119.44 (2022). Publisher: Proceedings of the National Academy of Sciences, e2203150119. DOI: 10.1073/pnas.2203150119.

[104] A. Marcoci et al. "Predicting the replicability of social and behavioural science claims from the COVID-19 Preprint Replication Project with structured expert and novice groups". In: (2024). Publisher: OSF. DOI: 10.31222/osf.io/xdsjf.

[105] P. Mateu, B. Applegate, and C. L. Coryn. "Towards more credible conceptual replications under heteroscedasticity and unbalanced designs". In: *Quality & Quantity* 58.1 (2024), pp. 723–751. DOI: 10.1007/s11135-023-01657-0.

[106] M. Xiao, H. Chu, J. S. Hodges, and L. Lin. "Quantifying replicability of multiple studies in a meta-analysis". In: *The Annals of Applied Statistics* 18.1 (2024). DOI: 10.1214/23-AOAS1806.

[107] M. Walsh et al. "The statistical significance of randomized controlled trial results is frequently fragile: a case for a Fragility Index". In: *Journal of Clinical Epidemiology* 67.6 (2014). Publisher: Elsevier, pp. 622–628. DOI: 10.1016/j.jclinepi.2013.10.019.

[108] L. Lin and H. Chu. "Assessing and visualizing fragility of clinical results with binary outcomes in R using the fragility package". In: *PLOS ONE* 17.6 (2022). Ed. by P. A. Gagniuc, e0268754. DOI: 10.1371/journal.pone.0268754.

[109] L. Lin et al. "Assessing the robustness of results from clinical trials and meta-analyses with the fragility index". In: *American Journal of Obstetrics and Gynecology* 228.3 (2023), pp. 276–282. DOI: 10.1016/j.ajog.2022.08.053.

[110] J. P. T. Higgins. "Measuring inconsistency in meta-analyses". In: *BMJ* 327.7414 (2003), pp. 557–560. DOI: 10.1136/bmj.327.7414.557.

[111] J. Wang, H. Liang, Q. Zhang, and S. Ma. "Replicability in cancer omics data analysis: measures and empirical explorations". In: *Briefings in Bioinformatics* 23.5 (2022), bbac304. DOI: 10.1093/bib/bbac304.

[112] R. Maitra. "A re-defined and generalized percent-overlap-of-activation measure for studies of fMRI reproducibility and its use in identifying outlier activation maps". In: *NeuroImage* 50.1 (2010), pp. 124–135. DOI: 10.1016/j.neuroimage.2009.11.070.

[113] M. Veronese et al. "Reproducibility of findings in modern PET neuroimaging: insight from the NRM2018 grand challenge". In: *Journal of Cerebral Blood Flow and Metabolism: Official Journal of the International Society of Cerebral Blood Flow and Metabolism* 41.10 (2021), pp. 2778–2796. DOI: 10.1177/0271678X211015101.

[114] G. Bachmann, T. Hofmann, and A. Lucchi. *Generalization Through The Lens Of Leave-One-Out Error*. Version Number: 1. 2022. DOI: 10.48550/ARXIV.2203.03443.

[115] P. Dixon and S. Glover. "Assessing evidence for replication: A likelihood-based approach". In: *Behavior Research Methods* 52.6 (2020), pp. 2452–2459. DOI: 10.3758/s13428-020-01403-6.

[116] C. J. Soto. "How Replicable Are Links Between Personality Traits and Consequential Life Outcomes? The Life Outcomes of Personality Replication Project". In: *Psychological Science* 30.5 (2019). Publisher: SAGE Publications Inc, pp. 711–727. DOI: 10.1177/0956797619831612.

[117] L. R. Fabrigar and D. T. Wegener. "Conceptualizing and evaluating the replication of research results". In: *Journal of Experimental Social Psychology* 66 (2016), pp. 68–80. DOI: 10.1016/j.jesp.2015.07.009.

[118] B. B. McShane, U. Böckenholt, and K. T. Hansen. "Variation and Covariation in Large-Scale Replication Projects: An Evaluation of Replicability". In: *Journal of the American Statistical Association* 117.540 (2022), pp. 1605–1621. DOI: 10.1080/01621459.2022.2054816.

[119] I. Jaric et al. "Using mice from different breeding sites fails to improve replicability of results from single-laboratory studies". In: *Lab Animal* 53.1 (2024). Number: 1 Publisher: Nature Publishing Group, pp. 18–22. DOI: 10.1038/s41684-023-01307-w.

[120] D. Borsboom et al. "False alarm? A comprehensive reanalysis of "Evidence that psychopathology symptom networks have limited replicability" by Forbes, Wright, Markon, and Krueger (2017)." In: *Journal of Abnormal Psychology* 126.7 (2017), pp. 989–999. DOI: 10.1037/abn0000306.

[121] M. K. Forbes, A. G. C. Wright, K. E. Markon, and R. F. Krueger. "Quantifying the Reliability and Replicability of Psychopathology Network Characteristics". In: *Multivariate Behavioral Research* 56.2 (2021), pp. 224–242. DOI: 10.1080/00273171.2019.1616526.

[122] A. Belz, M. Popovic, and S. Mille. "Quantified Reproducibility Assessment of NLP Results". In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Dublin, Ireland: Association for Computational Linguistics, 2022, pp. 16–28. DOI: 10.18653/v1/2022.acl-long.2.

[123] T. Nordling and T. M. Peralta. *A literature review of methods for assessment of reproducibility in science*. 2022. DOI: 10.21203/rs.3.rs-2267847/v4.

[124] A. Belz. "A Metrological Perspective on Reproducibility in NLP*". In: *Computational Linguistics* 48.4 (2022), pp. 1125–1135. DOI: 10.1162/coli_a_00448.

[125] R. A. Zwaan, A. Etz, R. E. Lucas, and M. B. Donnellan. "Making replication mainstream". In: *Behavioral and Brain Sciences* 41 (2018), e120. DOI: 10.1017/S0140525X17001972.

[126] S. A. Baig. "Bayesian Inference: Evaluating Replication Attempts With Bayes Factors". In: *Nicotine & Tobacco Research* 24.4 (2022), pp. 626–629. DOI: 10.1093/ntr/ntab219.

[127] L. Held, C. Micheloud, and S. Pawel. "The assessment of replication success based on relative effect size". In: *The Annals of Applied Statistics* 16.2 (2022). DOI: 10.1214/21-AOAS1502.

[128] H. O. Balli and B. E. Sørensen. "Interaction effects in econometrics". In: *Empirical Economics* 45.1 (2013), pp. 583–603. DOI: 10.1007/s00181-012-0604-2.

[129] M. Arroyo-Araujo et al. "Reproducibility via coordinated standardization: a multi-center study in a Shank2 genetic rat model for Autism Spectrum Disorders". In: *Scientific Reports* 9.1 (2019). Number: 1 Publisher: Nature Publishing Group, p. 11602. DOI: 10.1038/s41598-019-47981-0.

[130] S. Costigan, J. Ruscio, and J. T. Crawford. "Performing Small-Telescopes Analysis by Resampling: Empirically Constructing Confidence Intervals and Estimating Statistical Power for Measures of Effect Size". In: *Advances in Methods and Practices in Psychological Science* 7.1 (2024), p. 25152459241227865. DOI: 10.1177/25152459241227865.

[131] R. C. M. Van Aert and M. A. L. M. Van Assen. "Bayesian evaluation of effect size after replicating an original study". In: *PLOS ONE* 12.4 (2017). Ed. by D. Marinazzo, e0175302. DOI: 10.1371/journal.pone.0175302.

[132] V. C. Wong, K. Anglin, and P. M. Steiner. "Design-Based Approaches to Causal Replication Studies". In: *Prevention Science* 23.5 (2022), pp. 723–738. DOI: 10.1007/s11121-021-01234-7.

[133] M. J. Bland and D. Altman. "Statistical methods for assessing agreement between two methods of clinical measurement". In: *The Lancet* 327.8476 (1986), pp. 307–310. DOI: 10.1016/S0140-6736(86)90837-8.

[134] H. C. De Vet, C. B. Terwee, D. L. Knol, and L. M. Bouter. "When to use agreement versus reliability measures". In: *Journal of Clinical Epidemiology* 59.10 (2006), pp. 1033–1039. DOI: 10.1016/j.jclinepi.2005.10.015.

[135] R. Manolov and R. Tanious. "Assessing Consistency in Single-Case Data Features Using Modified Brinley Plots". In: *Behavior Modification* 46.3 (2020), pp. 581–627. DOI: 10.1177/0145445520982969.

[136] R. Manolov, R. Tanious, and B. Fernández-Castilla. "A proposal for the assessment of replication of effects in single-case experimental designs". In: *Journal of Applied Behavior Analysis* 55.3 (2022), pp. 997–1024. DOI: 10.1002/jaba.923.

[137] M. Liou et al. "Bridging Functional MR Images and Scientific Inference: Reproducibility Maps". In: *Journal of Cognitive Neuroscience* 15.7 (2003), pp. 935–945. DOI: 10.1162/089892903770007326.

[138] A. Dreber et al. "Using prediction markets to estimate the reproducibility of scientific research". In: *Proceedings of the National Academy of Sciences* 112.50 (2015), pp. 15343–15347. DOI: 10.1073/pnas.1516179112.

[139] J. M. González-Barahona and G. Robles. "On the reproducibility of empirical software engineering studies based on data retrieved from development repositories". In: *Empirical Software Engineering* 17.1-2 (2012), pp. 75–89. DOI: 10.1007/s10664-011-9181-9.

[140] L. Belbasis and O. A. Panagiotou. "Reproducibility of prediction models in health services research". In: *BMC Research Notes* 15.1 (2022), p. 204. DOI: 10.1186/s13104-022-06082-4.

[141] T. Hildebrandt and J. M. Prenoveau. "Rigor and reproducibility for data analysis and design in the behavioral sciences". In: *Behaviour Research and Therapy* 126 (2020), p. 103552. DOI: 10.1016/j.brat.2020.103552.

[142] U. Wilensky and W. Rand. "Making Models Match: Replicating an Agent-Based Model". In: *Journal of Artificial Societies and Social Simulation* 10.4 (2007).

[143] A. C. Chang and P. Li. "Is Economics Research Replicable? Sixty Published Papers From Thirteen Journals Say "Often Not"". In: *Critical Finance Review* 11.1 (2022), pp. 185–206. DOI: 10.1561/104.00000053.

[144] H. Suetake, T. Fukusato, T. Igarashi, and T. Ohta. "A workflow reproducibility scale for automatic validation of biological interpretation results". In: *GigaScience* 12 (2022), giad031. DOI: 10.1093/gigascience/giad031.

[145] J. Q. Sumner, C. H. Vitale, and L. D. McIntosh. "RipetaScore: Measuring the Quality, Transparency, and Trustworthiness of a Scientific Work". In: *Frontiers in Research Metrics and Analytics* 6 (2022), p. 751734. DOI: 10.3389/frma.2021.751734.

[146] Y. Yang, W. Youyou, and B. Uzzi. "Estimating the deep replicability of scientific findings using human and artificial intelligence". In: *Proceedings of the National Academy of Sciences* 117.20 (2020), pp. 10762–10768. DOI: 10.1073/pnas.1909046117.

[147] W. Youyou, Y. Yang, and B. Uzzi. "A discipline-wide investigation of the replicability of Psychology papers over the past two decades". In: *Proceedings of the National Academy of Sciences* 120.6 (2023), e2208863120. DOI: 10.1073/pnas.2208863120.

[148] X. Xu. "Epistemic diversity and cross-cultural comparative research: ontology, challenges, and outcomes". In: *Globalisation, Societies and Education* 20.1 (2022). Publisher: Routledge, pp. 36–48. DOI: 10.1080/14767724.2021.1932438.

[149] S. F. Anderson and K. Kelley. "Sample size planning for replication studies: The devil is in the design". In: *Psychological Methods* (2022). Place: US Publisher: American Psychological Association. DOI: 10.1037/met0000520.

[150] L. Held, S. Pawel, and C. Micheloud. "The assessment of replicability using the sum of p-values". In: *Royal Society Open Science* 11.8 (2024). Publisher: Royal Society, p. 240149. DOI: 10.1098/rsos.240149.

[151] L. Mbuagbaw, D. O. Lawson, L. Puljak, D. B. Allison, and L. Thabane. "A tutorial on methodological studies: the what, when, how and why". In: *BMC Medical Research Methodology* 20.1 (2020), p. 226. DOI: 10.1186/s12874-020-01107-7.

A Search strings for methodological papers

A.1 Scopus

TITLE-ABS((replication* OR replicated OR reproduced OR reproduction* OR generalised OR generalisation* OR generalized OR generalization) W/1 (study* OR studies OR experiment* OR analys* OR analyz* OR estimation* OR estimate* OR result* OR finding*) OR ((reproducibility W/2 research) OR (reproducibility W/2 science) OR (replicability W/2 research) OR (replicability W/2 science) OR (generalisability W/2 research) OR (generalisability W/2 science) OR (generalizability W/2 research) OR (generalizability W/2 science) OR (translatability W/2 research) OR (translatability W/2 science))) AND TITLE-ABS ((replicable OR replication OR replicability OR reproduction OR reproducible OR reproducibility OR generalisable OR generalisability OR generalisation OR generalizable OR generalizability OR generalization OR translatable OR translation OR translatability) W/1 (quantif* OR measure* OR metric* OR evaluat* OR score* OR assess* OR rating* OR ratio* OR rate*)) AND (LIMIT-TO (DOCTYPE,"ar") OR LIMIT-TO (DOCTYPE,"cp"))

A.2 Ebsco

((((TI replication* OR AB replication*) OR (TI replicated OR AB replicated) OR (TI reproduction* OR AB reproduction*) OR (TI reproduced OR AB reproduced) OR (TI generalisation* OR AB generalisation*) OR (TI generalised OR AB generalised) OR (TI generalization* OR AB generalization*) OR (TI generalized OR AB generalized)) N1 ((TI study* OR AB study*) OR (TI studies OR AB studies) OR (TI experiment* OR AB experiment*) OR (TI analys* OR AB analys*) OR (TI analyz* OR AB analyz*) OR (TI estimation* OR AB estimation*) OR (TI estimate* OR AB estimate*) OR (TI result* OR AB result*) OR (TI finding* OR AB finding*))) OR (((TI reproducibility OR AB reproducibility) N2 (TI research OR AB research)) OR ((TI reproducibility OR AB reproducibility) N2 (TI science OR AB science)) OR ((TI replicability OR AB replicability) N2 (TI research OR AB research)) OR ((TI replicability OR AB replicability) N2 (TI science OR AB science)) OR ((TI generalisability OR AB generalisability) N2 (TI research OR AB research)) OR ((TI generalisability OR AB generalisability) N2 (TI science OR AB science)) OR ((TI generalizability OR AB generalizability) N2 (TI research OR AB research)) OR ((TI generalizability OR AB generalizability) N2 (TI science OR AB science)) OR ((TI translatability OR AB translatability) N2 (TI research OR AB research)) OR ((TI translatability OR AB translatability) N2 (TI science OR AB science)))) AND (((TI replicable OR AB replicable) OR (TI replication OR AB replication) OR (TI replicability OR AB replicability) OR (TI reproduction OR AB reproduction) OR (TI reproducible OR AB reproducible) OR (TI reproducibility OR AB reproducibility) OR (TI generalisable OR AB generalisable) OR (TI generalisability OR AB generalisability) OR (TI generalisation OR AB generalisation) OR (TI generalizable OR AB generalizable) OR (TI generalizability OR AB generalizability) OR (TI generalization OR AB generalization) OR (TI translatable OR AB translatable) OR (TI translation OR AB translation) OR (TI translatability OR AB translatability)) N1 ((TI quantif* OR AB quantif*) OR (TI measure* OR AB measure*) OR (TI metric* OR AB metric*) OR (TI evaluat* OR AB evaluat*) OR (TI score* OR AB score*) OR (TI assess* OR AB assess*) OR (TI rating* OR AB rating*) OR (TI ratio* OR AB ratio*) OR (TI rate* OR AB rate*)))

B Screening guides for methodological papers

B.1 First screening of search results for methodological papers

What papers will be included?

- Methodological papers - does title and abstract suggest that the paper presents/discusses a measure to quantify reproducibility?

  Definition of "methodological papers" in our setting (adapted from [151]) is any paper that

  * describes methods or measures to quantify the reproducibility of a field, a finding, an effect, a study, a method, . . . - Example: [33]
  * describes methods or measures to assess a successful reproduction - Example: [59]

  This includes

  * Review papers, but only if the paper reviews methods or measures to quantify reproducibility or assess successful reproductions, e.g. "methodological review papers" - Example: [11] [OPTION TO FLAG AS REVIEW PAPERS]
  * Tutorial papers, explaining or demonstrating how to use measures to quantify reproducibility. [OPTION TO FLAG AS TUTORIAL PAPERS]
  * Commentaries and editorials if it is apparent from the abstract that a new alternative measure is suggested/discussed.

- Application papers are included only if it is apparent from the abstract that they use an innovative measure to quantify reproducibility [Edge case - OPTION TO FLAG AS INTERESTING APPLICATION PAPER]
- Papers investigating any type of reproducibility can be investigated - we use reproducibility as an overarching term for aspects including computational reproducibility, replicability, translatability, and generalisability. See the iRISE glossary for more definitions [5].
- Papers discussing reproducibility in any discipline or field of study are included.
- Papers published in any year are included (until May 13th 2024).

What papers will be excluded?

- Application papers - whenever it is clear from the title and abstract that the paper presents a reproducibility study (single study or large-scale project) and only applies a certain measure we will exclude it.
- Review papers that are not "methodological review papers", reviewing methods or measures to quantify reproducibility or assess successful reproductions, are excluded.
- Papers that are off topic (while using the same terminology), e.g. translation in linguistics, image replication, sexual reproduction, cell or bacteria replications or replicability, virus reproduction ratio, etc.
- Editorials are excluded if they are not discussing a new measure
- Commentaries are excluded if if they are not discussing a new measure

B.2 Second screening of list of potential methodological papers

What papers will be included?

- Methodological papers - does title and abstract suggest that the paper presents/discusses a measure to quantify, predict or explain reproducibility? This includes more quantitative measures of reproducibility but also qualitative investigations, e.g. Delphi studies.

  Definition of "methodological papers" in our setting (adapted from [151]) is any paper that

  * describes methods or measures to quantify, predict or explain the reproducibility of a field, a finding, an effect, a study, a method, ... - Example: [33]
  * describes methods or measures to assess a successful reproduction - Example: [59]

  This includes

  * Review papers, but only if the paper reviews methods or measures to quantify reproducibility or assess successful reproductions, e.g. "methodological review papers" - Example: [11].
  * Tutorial papers, explaining or demonstrating how to use measures to quantify reproducibility.
  * Commentaries and editorials if it is apparent from the abstract that a new alternative measure is suggested/discussed.
- Application papers are included only if it is apparent from the abstract that they use an innovative measure to quantify reproducibility [Edge case - OPTION TO FLAG AS INTERESTING APPLICATION PAPER]
- Papers investigating any type of reproducibility can be investigated - we use reproducibility as an overarching term for aspects including computational reproducibility, replicability, translatability, and generalizability. See the iRISE glossary for more definitions [5].
- Papers discussing reproducibility in any discipline or field of study are included.
- Papers published in any year are included (until May 13th 2024).

What papers will be excluded?

- All types of application papers - whenever it is clear from the title and abstract that the paper presents a reproducibility study (single study or large-scale project) and only applies a certain measure we will exclude it.
- Review papers that are not "methodological review papers", reviewing methods or measures to quantify reproducibility or assess successful reproductions, are excluded.
- Papers that are off topic (while using the same terminology), e.g. translation in linguistics, image replication, sexual reproduction, cell or bacteria replications or replicability, virus reproduction ratio, etc.
- Editorials are excluded if they are not discussing a new measure
- Commentaries are excluded if if they are not discussing a new measure

C Data extraction questions

C.1 Guide for data extraction of *Application papers*

1. Field of research - as described by the authors. Select from list of the broader fields [select all that apply]:
   (a) Social Sciences and Humanities
   (b) Life Sciences
   (c) STEM, e.g. Engineering, Mathematics, Physics
   (d) N/A
2. Discipline - as described by the authors [Add up to 3 disciplines, if more than 3, write interdisciplinary]
3. Type of project - one of the following:
   (a) Many Phenomena, One Study (e.g., the Reproducibility Project Psychology): Many original hypotheses are tested. Each hypothesis is tested in one (replication) study.
   (b) One Phenomenon, Many Studies (e.g., Multilab replication studies): One original hypothesis is tested by many different teams / in many separate studies
   (c) Many Phenomena, Many Studies (e.g. FORRT Replications & Reversals, Replication Database, FORRT Replication Database): Many original hypotheses are tested, each hypothesis is tested in many separate studies
   (d) Other [Add as comment]
4. Did the authors define the type of reproducibility that is investigated? - Yes or No
   (a) (if Yes - child question of above) Aspect of reproducibility investigated (authors) - extract definition of reproducibility reported by the authors
5. Even if defined by the authors, infer the aspect of reproducibility investigated using the concept of reproducibility as it is presented in the Turing way matrix. To this end, select one or several of the following:
   (a) Same data - same analysis
   (b) Same data - different analysis
   (c) Different data - same analysis
   (d) Different data - different analysis
   (e) Other [Add as comment]
6. Did the authors measure reproducibility, i.e., summarise the results, using one of the following traditional measures (select all that apply and add details in comment cell).
   (a) Agreement in statistical significance [add details in the comment cell] - Example: Are original and replication p-values < 0.05?
   (b) Agreement in effect size [add details in the comment cell] - Examples: Do original and replication effect size go in the same direction? Is the replication effect size smaller than the original effect size? Is the replication effect size contained in the original effect size 95% confidence interval?, Is the original effect size contained in the replication effect size 95% confidence interval? Is the replication effect size contained in a 95% prediction interval based on the original effect size?
   (c) Meta-analysis of original and replication study/studies [add details in the comment cell] - Examples: is meta-analytic p-value < 0.05? How large is meta-analytic effect size? Does meta-analytic 95% confidence interval include zero? Is there evidence for heterogeneity, e.g., p-value from Q-test < 0.05?
   (d) Subjective assessment [add details in the comment cell] - Examples: Answer of replicators to "did it replicate"?, Answer of original authors to "did it replicate"?
   (e) None of the above
7. Did the authors measure reproducibility, i.e., summarise the results, using one or several measures not present in the previous list? Add all
   (a) (if Yes - child question of above) Paste description of all other measures used
8. Did the paper refer to other papers for more information on the metric(s) used? - Yes or No
   (a) (if Yes - child question of above) Paste the doi of all paper(s) and add the name of the metric it refers to (as called in previous questions) in the comment.

9. Did the authors discuss limitations or assumptions of the metric(s) used? - Yes or No
    (a) (if Yes - child question of above) Paste text on limitation/assumptions and add the name of the metric it refers to (as called in previous questions) in the comment.
10. Did the authors discuss equity, diversity, and/or inclusion (see definition below) at any point? - Yes or No
    (a) (if Yes - child question of above) Paste text
11. Research question or aim - if obvious, paste research question or aim as reported by authors, for example, "To estimate the reproducibility of field XYZ".

C.2 Guide for data extraction of *Methodological papers*

C.2.1 Interesting application papers

1. Did the authors define the type of reproducibility that is investigated? - Yes or No
    (a) (if Yes - child question of above) Aspect of reproducibility investigated (authors) - extract definition of reproducibility reported by the authors
2. Even if defined by the authors, infer the aspect of reproducibility investigated using the concept of reproducibility as it is presented in the Turing way matrix. To this end, select one or several of the following:
    (a) Same data - same analysis
    (b) Same data - different analysis
    (c) Different data - same analysis
    (d) Different data - different analysis
    (e) Other [Add as comment]
3. How did the authors measure reproducibility, i.e., summarise the results? Paste description of all measures used.
4. Did the paper refer to other papers for more information on the metric(s) used? - Yes or No
    (a) (if Yes - child question of above) Paste the doi of all paper(s).

C.2.2 Methodological papers

1. **Type of paper**
   Detailed description: What type of paper are you annotating?
    - Original research paper
    - Review paper - a review of measures/metrics to quantify reproducibility
    - Tutorial paper
    - Protocol - study protocol and alike, where the study might still be ongoing
    - Editorial, comment, or similar
    - Other [add a comment]
2. **Design purpose**
   Detailed description: Was (Were) the presented measure(s) "designed" for reproducibility? Some methods were developed for another purpose, but might have been used to quantify or assess reproducibility, in the application papers. Yes, No, Unclear [explain in comment]
3. **Name of reproducibility (or related concept)**
   Detailed description: [free text] How did the authors "call" what they are investigating? If they used more than one re-term, e.g. reproducibility and replication study, add all of them. Whenever possible use the terms from the iRISE glossary. If unclear, or you cannot find a name, leave blank.
4. **Definition of reproducibility (or related concept)**
   Detailed description: Can you find a clear definition of the type of reproducibility, or related concept, the authors are interested in? An example of a clear definition would be: Reproducibility is commonly defined as the ability to obtain "consistent results using the same input data, computational steps, methods, and conditions of analysis" (from 10.1016/j.cmpb.2023.107839). An unclear definition would be: direct replications of the original study, all following the same vetted protocol - as this text snippets only explains what has been done, but was not meant as a definition (from 10.1177/1745691616664694).

- Yes, clear definition [add text in comment]
- Yes, but unclear defined [add text in comment]
- No

5. **Type of reproducibility (or related concept)**
   Detailed description: [multiple choice] What is the type of reproducibility investigated using the discussed measure(s). Even if defined by the authors, infer the aspect of reproducibility investigated using the concept of reproducibility as it is presented in the Turing way matrix. Note that same data = using exactly the same data as the original authors, or the exact same data source and data retrieval steps; and same analysis = following a predefined set of steps, allowing for slight variations. To this end, select one or several of the following:
   - Same data - same analysis
   - Same data - different analysis
   - Different data - same analysis
   - Different data - different analysis
   - Other [Add as comment]
   - Unclear [explain in comment]
6. **Purpose of measure**
   Detailed description: [multiple choice] What is (are) the measure(s) meant to be used for? Note that one measure can be used for several (or even all) of these purposes. Select all that apply and are discussed in the paper. Use the comment field to give context, or further explanation.
   - To quantify (continuous) reproducibility or related concept. Example: a method that estimate reproducibility rates/probabilities and alike - 10.15626/MP.2021.2720
   - To classify (binary, yes or no) reproducibility or related concept. Example: a tool or similar that gives a yes-no answer to the question "is this reproducible - 10.6084/m9.figshare.c.5418242
   - To predict reproducibility or related concept. Example: a model/algorithm/tool that uses the findings or the text of one paper to predict how well the study or the findings would reproduce - 10.1073/pnas.1909046117
   - To explain reproducibility or related concept. Example: a model which tries to explain certain levels of reproducibility with covariates and alike - 10.1111/insr.12273
   - Unclear [explain in comment]
7. **Number of measures**
   Detailed description: [free text] How many distinct measures, methods or models are discussed in the paper? Usually only one, but if the paper is a review, or if several variants of a method are discussed/presented there might be more (example - this preprint discusses Edgington's method and also presents a weighted version of the same methods). If the same measure is applied in different contexts or in various studies, the number of measures is still only one. Add a comment, if the number reflects the number of variants of the same measure/method.
8. **Type of measure**
   Detailed description: [multiple choice] What type of measure(s) is (are) discussed? Note that some measures might use a combination of these types. Select all that apply, and explain in comment.
   - A formula, e.g., a percentage, a p-value, a Bayes factor
   - A statistical model, e.g., a model which relates "reproducibility" or a proxy thereof to some covariates
   - An algorithm, e.g., a tool that uses unstructured data, like text, to estimate a reproducibility rate
   - A study, e.g., a Delphi study is set up to assess the reproducibility of a study
   - A survey or questionnaire, e.g., a set of experts are asked, via a survey, whether they rate a study as fully, partially, or not at all reproducible
   - Other [explain in comment]
   - Unclear [explain in comment]
9. **Type of assessment**
   Detailed description: [multiple choice] Is (Are) the measure(s) of quantitative or qualitative nature? A quantitative measure would give a continuous result, while a qualitative measure would rather give a classification into something like "fully reproducible", "partially reproducible", "not reproducible". Some measures are deterministic and do not need any subjective input, others rely, at least in some way, on a subjective assessment. If the paper discusses several measures with some being quantitative and others qualitative, select all that apply and add a comment.

- Quantitative
- Qualitative
- Objective
- Subjective
- Unclear [explain in comment]

10. **Name of measure**
    Detailed description: [free text] How did the authors call the measure? Leave blank if they did not name the measure. If they are discussing more than one measure, add all their names separately.
11. **Implementation of measure**
    Detailed description: [multiple choice] We are interested in knowing whether the discussed measure(s) can be easily implemented by a researcher who wants to investigate reproducibility. Use the comment field to give more context. Select all that apply, especially if several measures are investigated.
    - Ready-to-use open-source tool - the authors provide a tool, a code script, or similar to use the suggested measure(s), at no added costs
    - Ready-to-use closed tool - the authors implemented or used a tool, a code script, or similar to use the suggested measure(s), but it might come at a cost or the access is restricted
    - Easy to implement - the measure(s) discussed can be easily implemented and the authors gave enough details to do so (e.g., using available software and instructions)
    - Hard to implement - the measure(s) discussed can be implemented, but it is not straightforward, labour- or time-intensive (e.g., a Delphi study is implementable, but time-consuming), or expensive
    - Unclear implementation - the authors did not give enough detail on how to implement the measure
    - Suggested only - the authors suggest a measure(s) or a general way to investigate reproducibility, but do not give guidance on how to actually use it
    - Unclear [explain in comment]
12. **Data input of measure**
    Detailed description: [multiple choice] When applying the measure to investigate reproducibility or related concept, what is the input of the measure, as in on what will the measure base its assessment on? Select all that apply. If more context or explanation is needed, use the comment field.
    - Text
    - Some demographics or meta-data
    - Code or software
    - Results - numbers and tables
    - Results - figures
    - Qualitative data, surveys or questionnaires
    - Other [add in comment]
    - Unclear [explain in comment]
13. **Assumptions or prerequisites for measure's usage**
    Detailed description: [free text] To use the measure, does the input need to be in a certain form, follow a certain distribution, or does the user need specific software and alike? Write down all assumptions and/or prerequisites (like needed software) the authors mention. If you are writing down assumptions and prerequisites of specific measures, if possible add the name of the measure you are referring to. Leave blank if the authors did not discuss anything.
14. **Limitation of measure**
    Detailed description: [free text] Did the authors discuss limitations of the measure(s)? If yes, write down all the discussed limitations you find (they might be referring back to prerequisites or assumptions - just write them down as limitations too). If you are writing down limitations of specific measures, if possible add the name of the measure you are referring to. Leave blank if the authors did not discuss anything.
15. **Equity, diversity, and/or inclusion**
    Detailed description: [free text] Did the authors discuss equity, diversity, and/or inclusion (see definition below or the iRISE glossary - second part on EDI) related to the usage of the measure, at any point? Specifically epistemic diversity (diversity of knowledge production, expertise, field of study, method of study, etc.) might be something that is discussed more often - e.g., if the measure is suggested in one specific field, can it be used in another etc? Leave blank if the authors did not discuss anything.