ECONSTOR
Make Your Publications Visible.

A Service of

ZBW Leibniz-Informationszentrum Wirtschaft
Leibniz Information Centre for Economics

Dreber, Anna; Johannesson, Magnus

Working Paper

A framework for evaluating reproducibility and replicability in economics

I4R Discussion Paper Series, No. 38

Provided in Cooperation with:
The Institute for Replication (I4R)

Suggested Citation: Dreber, Anna; Johannesson, Magnus (2023) : A framework for evaluating reproducibility and replicability in economics, I4R Discussion Paper Series, No. 38, Institute for Replication (I4R), s.l.

This Version is available at:
https://hdl.handle.net/10419/271678

Standard-Nutzungsbedingungen:

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

Terms of use:

Documents in EconStor may be saved and copied for your personal and scholarly purposes.

You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.

If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.

WWW.ECONSTOR.EU

Mitglied der Leibniz-Gemeinschaft

INSTITUTE for REPLICATION

No. 38
I4R DISCUSSION PAPER SERIES

A Framework for Evaluating Reproducibility and Replicability in Economics

Anna Dreber

Magnus Johannesson

June 2023

INSTITUTE for REPLICATION

I4R DISCUSSION PAPER SERIES

I4R DP No. 38

A Framework for Evaluating Reproducibility and Replicability in Economics

Anna Dreber[1], Magnus Johannesson[2]

[1]Stockholm School of Economics, Dept. of Economics, Stockholm/Sweden and University of Innsbruck, Dept. of Economics, Innsbruck/Austria
[2]Stockholm School of Economics, Dept. of Economics

JUNE 2023

Any opinions in this paper are those of the author(s) and not those of the Institute for Replication (I4R). Research published in this series may include views on policy, but I4R takes no institutional policy positions.

I4R Discussion Papers are research papers of the Institute for Replication which are widely circulated to promote replications and meta-scientific work in the social sciences. Provided in cooperation with EconStor, a service of the ZBW – Leibniz Information Centre for Economics, and RWI – Leibniz Institute for Economic Research, I4R Discussion Papers are among others listed in RePEc (see IDEAS, EconPapers). Complete list of all I4R DPs - downloadable for free at the I4R website.

I4R Discussion Papers often represent preliminary work and are circulated to encourage discussion. Citation of such a paper should account for its provisional character. A revised version may be available directly from the author.

Editors

Abel Brodeur
University of Ottawa

Anna Dreber
Stockholm School of Economics

Jörg Ankel-Peters
RWI – Leibniz Institute for Economic Research

E-Mail: joerg.peters@rwi-essen.de
RWI – Leibniz Institute for Economic Research
Hohenzollernstraße 1-3
45128 Essen/Germany
www.i4replication.org

ISSN: 2752-1931

A framework for evaluating reproducibility and replicability in economics

Anna Dreber and Magnus Johannesson*

Abstract

A fundamental question to the scientific enterprise is to what extent published scientific findings are credible. This question is related to the reproducibility and replicability of scientific findings where reproducibility is defined as testing if the results of an original study can be reproduced using the same data and replicability is defined as testing if the results of an original study hold in new data. We provide a framework for evaluating reproducibility and replicability in economics and divide reproducibility and replicability studies into five types: computational reproducibility, recreate reproducibility, robustness reproducibility, direct replicability and conceptual replicability, and we propose indicators to be reported for each type.

*Dreber: Department of Economics, Stockholm School of Economics (e-mail: anna.dreber@hhs.se) and Department of Economics, University of Innsbruck, Innsbruck, Austria; Johannesson: Department of Economics, Stockholm School of Economics (e-mail: magnus.johannesson@hhs.se). For financial support, we thank Jan Wallander and Tom Hedelius Foundation (grant P21-0091 to A.D.), Knut and Alice Wallenberg Foundation (grant KAW 2018.0134 to A.D.), Marianne and Marcus Wallenberg Foundation (grant KAW 2019.0434; to A.D.), and Riksbankens Jubileumsfond (grant P21-0168 to M.J.). We thank Jörg Ankel-Peters, Abel Brodeur and Fernando Hoces de la Guardia for helpful comments.

1. Introduction

Can we trust scientific findings? This question has been brought to the forefront of research in the social sciences in recent years with the movement towards open science practises and pre-registration. The single most important event for this development in the social sciences was probably the publication of the reproducibility project psychology (RPP) in 2015 (Open Science Collaboration 2015) replicating 100 studies published in three top psychology journals in 2008. While 97 of the 100 original studies reported a statistically significant result, only 35 of the replications could replicate a statistically significant result in the same direction. Although this question has only gained momentum in recent years, it is a question that has been raised many times before with some well-known contributions being Ioannidis (2005) claiming that most published research findings are false and Leamer (1983) with the classic article title "Let's take the con out of econometrics".[1]

While conducting independent replications is crucial for accumulating scientific knowledge, direct replications were relatively rare in the social sciences until the publication of RPP (Mueller-Langer et al. 2019; Ryan & Tipu 2022). After RPP the interest in replications have increased and several additional systematic replication studies have been published (Klein et al. 2014, 2018; Camerer et al. 2016, 2018; Ebersole et al. 2016). Taken together these studies suggest a replication rate of about 50% for experimental studies in the social sciences both in terms of the fraction of replications with a statistically significant effect in the same direction as the original study and in terms of the effect sizes in the replications relative to the effect sizes of the original study. Several potential explanations for these low replication rates have been offered such as "researcher degrees of freedom" including p-hacking (Simmons et al. 2011; John et al. 2012; Gelman & Loken 2014; Brodeur et al. 2016, 2020; Nelson et al. 2018), low statistical power (Button et al. 2013; Ioannidis et al. 2017), testing hypotheses with low priors (Maniadis et al. 2014; Dreber et al. 2015; Johnson et al 2017), and publication bias (Hedges 1992; Stern & Simes 1997; Franco et al. 2014, 2015).

The systematic replication projects referred to above are based on what is often termed "direct replications", which implies that the hypothesis tested in the original article is tested again in new data using the same research design and analysis as the original article. Several other types of tests of the validity and reliability of research findings are possible such as testing if the posted data and code reproduce the results in a published paper or testing if a published result

[1] See also the overview article for economics research by Christensen and Miguel (2018).

is robust to alternative equally plausible specifications to test the hypothesis. Tests based on using the same data as in the original article are often referred to as tests of reproducibility to distinguish those tests from tests based on new data (referred to as replicability).

In this article we propose a framework for evaluating reproducibility and replicability in economics.[2] We divide reproducibility and replicability studies into five types: computational reproducibility, recreate reproducibility, robustness reproducibility, direct replicability and conceptual replicability and we propose indicators to be reported for each type.

2. Typology of reproducibility and replicability studies

Our proposed typology is provided in Table 1. We define reproducibility as testing if results and conclusions of original studies can be reproduced based on the same data as used in the original studies, and replicability as testing if results and conclusions of original studies can be repeated using new data (i.e. different data than in the original studies). We furthermore divide reproducibility into computational reproducibility, recreate reproducibility and robustness reproducibility and replicability into direct and conceptual replicability. The definitions of direct replicability, conceptual replicability, and computational reproducibility are in line with how these terms are typically used in the literature (although we distinguish between sub-groups depending on the sample used), whereas robustness reproducibility and recreate reproducibility are not yet established terms. We use the term original study for the study that is reproduced or replicated.

Table 1. Types of reproducibility and replicability.

Types of reproducibility:	Definition	Sub-groups
Computational reproducibility	To what extent results in original studies can be reproduced based on data and code posted or provided by the original authors.	
Recreate reproducibility	To what extent results in original studies can be reproduced based on the information in the papers and access to the same raw data or data source, but without having access to the analysis code of the original study and/or the data set it was applied to.	A. Having access to the data set that the analysis code of the original study was applied to, but not the analysis code. B. Having access to the analysis code of the original study, but not the data set the analysis code was applied to. C. Not having access to the analysis code of the original

[2] We believe this can be applied also to other quantitative fields in the social sciences.

		study or the data set the analysis code was applied to.
Robustness reproducibility	To what extent results in original studies are robust to alternative plausible analytical decisions on the same data.	
Types of replicability:		
Direct replicability	To what extent results in original studies can be repeated on new data using the same research design and analysis as the original study.	A. Data from the same population. B. Data from a similar population. C. Data from a different population.
Conceptual replicability	To what extent results in original studies can be repeated on new data using an alternative research design and/or analysis to test the same hypothesis.	A. Data from the same population. B. Data from a similar population. C. Data from a different population.

2.1. Computational reproducibility

Computational reproducibility implies testing to what extent the data and code of a published paper yield the results reported in the paper.[3] One would expect computational reproducibility to be very high as it essentially implies testing for errors in running the original code on the original data in original studies, though software availability and software obsolescence can complicate this. However, several studies suggest that there are substantive computational reproducibility problems. Already in 1986, Dewald, Thursby and Anderson (1986) published a paper about the computational reproducibility of macroeconomics papers published in the Journal of Money, Credit and Banking. They tried to collect analysis code and data for 54 papers to test if they could reproduce the results of these papers, but only managed to reproduce the results of two (4%) papers. This fraction increased to 22% if estimated based on the 9 papers that they had data and code for, thus achieving a computational reproducibility rate of 22%. Several additional studies on computational reproducibility in economics and finance have been published since then typically yielding meagre reproducibility rates (e.g. McCullough et al. (2006, 2008), Glandon (2011), Chang & Li (2017), Gertler et al. (2018), Herbert et al. (2021), and Perignon et al. (2022)). Some economics journals, such as the journals of the

[3] See Berkeley Initiative for Transparency in the Social Sciences (2020) for guidelines on conducting computational reproducibility studies.

American Economic Association, now use Data Editors to check that the data and code yield the results in the paper prior to publication. With the increased use of Data Editors, computational reproducibility will likely improve.

2.2. Recreate reproducibility

Recreate reproducibility implies trying to reanalyze the results of an original study as closely as possible without having access to the analysis code and/or the exact data the code was applied to. It can be divided into three sub-groups (A-C) and the most challenging case is C when the original study has posted neither the data nor the analysis code; A and B border on computational reproducibility and could alternatively have been included as sub-groups of computational reproducibility. It is still rare with systematic studies of recreate reproducibility in economics, but one recent example is the study by Black et al (2022) that examined the reproducibility of four papers using the same randomized field experiment on short-sale restrictions for identifying causal effects. This field experiment did not find any effects on the directly studied outcomes related to short-sale restrictions, but a sizeable literature has tested for various other "indirect effects" and over 60 papers have been published in finance, economics and accounting reporting evidence of various indirect effects (on effects like earnings management and workplace safety). Black et al. (2022) selected four prominent papers from this literature and tried to reproduce the results of each paper, but only between 0% and 9% of the results could be reproduced based on the "statistical significance indicator" (this indicator for replication is discussed further below). The multi-analyst study by Huntington-Klein et al. (2021) on two economics papers is also in the intersection between recreate reproducibility and robustness reproducibility.[4]

2.3. Robustness reproducibility

Testing a hypothesis in a data set involves making many analytical decisions, and a published paper reports the results for a specific combination of such choices and possibly some robustness tests. Robustness reproducibility implies using the same data and testing if the results are robust to various alternative plausible ways of testing the hypothesis. A test of robustness reproducibility could in principle be anything from testing a few alternative analytical decisions to a full-blown multiverse analysis that explores all combinations of plausible analytical decisions (Steegen et al. 2016; Simonsohn et al. 2020). The ideal test of

[4] Bergh et al. (2017) and Delios et al. (2022) are also two examples of systematic recreate reproducibility studies in management.

robustness reproducibility would be moving towards conducting multiverse analysis; but systematically testing if a literature (a group of papers) is robust to a specific analytical decision is also an interesting possibility.

There are many studies testing the robustness of individual papers in economics, often published as comments. But it is difficult to draw general conclusions about robustness reproducibility from such studies as they are likely to be selected based on results being non-robust and there is little published systematic evidence on robustness reproducibility. However, we know of some more systematic studies that are currently being conducted (see, e.g., the work from the Institute for Replication) and we expect more systematic work to be published on robustness reproducibility in the coming years.[5]

2.4. Direct replicability

For experimental studies a direct replication as closely as possible uses the same experimental design as the original study and the same analysis to test the same hypothesis (ideally using the experimental instructions, software and analysis code used in the original study, which should ideally be publicly posted for all experimental studies).

It could be argued that a direct replication should be carried out in a sample drawn from an as similar population as possible as the population in the original study. However, completely ruling out any systematic differences in the samples of the original study and the replication study would involve randomly drawing the sample from the same population. In most cases that is not possible as it would imply that both the original study sample and the replication sample would have to be randomly drawn from the same population at the same time (to control for that the population has not changed over time). For most direct replications we therefore cannot rule out systematic differences in the sample included in the original study and the replication. In terms of terminology, we recommend using the term direct replication even if the sample differs and to distinguish between three possible types of direct replications. Direct replications based on samples from the same population, direct replications based on samples from similar populations (for instance university students at a Western university), and direct

[5] There is an element of robustness reproducibility in eight papers published in a special section of the Journal of Development Studies summarised by Brown & Wood (2019) in an introduction to the special section. These reproducibility studies are mainly about reproducing the original results using the same raw data and code, and borders on both computational reproducibility (if both data and code is available) and recreate reproducibility (when not all data and code are available), but there is also an element of robustness reproducibility as some alternative specifications are also typically tested.

replication based on samples from different populations (such as a university student population in the original study and a general population in the direct replication).

Direct replications of studies using observational data can use the same terminology and use the term direct replications when using the same research design and analysis as the original study applied to another sample than in the original study. A distinction can be made here as well between direct replications with a new sample from the same population, a new sample from a similar population, or a new sample from a different population.

The systematic replication projects on experimental economics and experimental social sciences by Camerer et al. (2016, 2018) are examples of systematic direct replication project in economics. There are also several examples of individual direct replication studies in economics.

2.5. Conceptual replicability

Conceptual replications imply using a different research design and/or analysis than the original study to test the same hypothesis in a new sample. For experiments this could be a different experimental design or a different analysis than used in the original study and for observational data studies the research design or analysis could differ from the original study. As above one can distinguish between conceptual replications based on new samples from the same population, a similar population, or a different population.

One can think of conceptual replications as all studies testing the same hypothesis, and how broad literature this implies depends on exactly how the hypothesis is defined. Studies testing the same hypothesis pooled in a meta-analysis can thus be viewed as conceptual replications (with the exception of the first study that was conducted, which would be the original study). This would imply a sizeable and growing literature on conceptual replications in economics as meta-analysis is increasing in popularity. The recent literature on replicating anomalies in finance also has elements of conceptual replications; see for instance the studies by Hou et al. (2020) and Jensen et al. (2022).

3. Reproducibility indicators

For both reproducibility and replicability, we propose two indicators for whether original studies are systematically biased. The first of these is the statistical significance indicator that measures to what extent results reported as statistically significant in original studies are statistically significant with an effect in the same direction also when they are reproduced or replicated. The second indicator measures the relative effect sizes of replication studies

compared to original studies. These two indicators have been commonly used in systematic replication studies and can be used also for reproducibility studies. These indicators are for original results reported as statistically significant and below we comment also on indicators of original results not reported as statistically significant. For robustness reproducibility we also propose two additional indicators: one indicator based on the variation in results across alternative analyses and one indicator that aggregates the results of different robustness tests into a pooled hypothesis test.

In Table 2 we define our proposed reproducibility indicators. In column (1) we describe the indicator for evaluating one result (one tested hypothesis) in a paper, and in column (2) and (3) we describe how the indicator can be pooled for evaluating several results in a paper and how it can be pooled across papers to evaluate a group of studies. For computational reproducibility the perhaps most natural indicator is to measure (yes/no) if all the results can be exactly reproduced in the paper or not, as the goal of for instance a journal Data Editor is to ensure this. Most work on computational reproducibility also uses some version of that indicator (that may also allow for minor deviations in the original results), but we do not include this indicator in the table as it is less applicable to other forms of reproducibility and replicability. That indicator will also not show to what extent a lack of computational reproducibility leads to systematic bias in reported results or or if this is a form of random measurement error. We describe the proposed indicators further below.

Table 2. Recommended reproducibility indicators (for results reported as statistically significant in original studies).

Reproducibility indicator	**(1). Description for one original result in one paper**	**(2). Pooling across separate original results within one paper**	**(3). Pooling across papers**
Statistical significance indicator	The fraction of statistically significant effect sizes in the same direction as the original study among all the reproducibility tests* of that result.	Average of (1) across separate original results.	Average of (2) across papers.
Relative effect sizes	A: The average effect size of all the reproducibility tests* of that original result divided by the effect size of the main specification in the published paper. B: If effect sizes cannot be compared across robustness	Average of (1) across separate original results.	Average of (2) across papers (can be tested if it differs statistically significantly from 1).

	tests: Estimate the mean t/z-value of all the reproducibility tests of that original result divided by the t/z-value of the main specification in the published paper.		
Robustness ratio	A: Estimate first the mean squared deviation of each robustness test effect size from the original effect size; and take the square root of this mean. Divide this "standard deviation" by the standard error of the original estimate.§ B: If effect sizes cannot be compared across robustness tests: Estimate the mean squared deviation of each robustness test t/z-value from the original t/z-value; and take the square root of this mean (note that this "standard deviation" measure is already scaled in standard error units).	Average of (1) across separate results.	Average of (2) across papers.
Pooled hypothesis test	A: z-test statistic estimated as the average effect size among all the robustness tests and the original effect size divided by the square root of the average variance of these tests.# B. If effect sizes cannot be compared across robustness tests: z-test statistic estimated as the average t/z-value of all the robustness tests and the original t/z-value.#	Fraction of original results that are statistically significant in the pooled test (with an effect in the same direction as the original study).	Average of (2) across papers.

* For computational and recrate reproducibility this will be one test; for robustness reproducibility it will typically be several tests.
§ The absolute "standard deviation" measure should also be reported here and can be viewed as an absolute robustness measure.
# This can also exclude the original result if it is not considered a plausible analysis path.

3.1. The statistical significance indicator

This indicator defines reproducibility as finding a statistically significant effect size in the same direction as the original study (typically evaluated at the 5% level based on two-sided p-values). For robustness reproducibility several robustness tests of the same original result will typically be carried out and this indicator is then measured as the fraction of robustness tests that fulfil

this indicator (whereas for computational and recreate reproducibility this would typically be one reproducibility test per original result in the original study).

3.2. Relative effect sizes

The relative effect size of each original result is estimated as the average effect size of all the reproducibility tests of that original result divided by the effect size of the main specification in the published paper. This measure will show to what extent results in original studies systematically overestimate effect sizes or not.

In some cases, the effect sizes of all the reproducibility tests of the same original result cannot be measured in the same effect size units as the original study, for instance using a log transformation in one robustness test. In those cases, the relative effect size measure can be defined in terms of t/z-values instead. Even if the effect sizes are measured in comparable units across reproducibility tests so that the first relative effect size measure can be constructed and reported, it can be good to add the relative t/z-values as an additional indicator as this indicator will also reflect variation in standard errors across robustness tests.

3.3. Robustness ratio

This indicator is only proposed for robustness reproducibility and is a measure of how much the results of the robustness tests varies compared to the original result. This measure was proposed by Athey & Imbens (2015) as a measure of the robustness to alternative regression analysis specifications. The measure is based on the standard deviation of the effect sizes of the robustness tests compared to the original effect size. This standard deviation is then divided by the standard error of the original effect size so that the variation is measured relative to the sampling variation. Note that this measure will be affected both by variation among the robustness tests, and by how much the robustness tests differ from the original effect size.[6] It is also useful to report the standard deviation measure as such as it can be viewed as an absolute measure of robustness (especially if a standardized effect size measure such as Cohen's d is used).

In some cases, the effect size may not be comparable for all the robustness tests of the same original result, but in such cases the robustness ratio can be estimated based on only the t/z-values of all the robustness tests and the original result. As this standard deviation measure is

[6] If one wants to isolate only the variation among the robustness tests, this indicator can be defined based on the standard deviation of the robustness tests, instead of basing the standard deviation measure on the deviation from the original effect size.

already measured in standard error units this measure should not be divided by the average standard errors of the original estimate (an increase in a t/z-value by 1 implies that the effect size increases by the magnitude of one standard error). This is an alternative measure of how robust the result is, which will also incorporate variation in standard errors across the robustness tests.

3.3. Pooled hypothesis test

This final indicator is also only proposed for robustness reproducibility and can be viewed as a modification of the statistical significance indicator. A pooled hypothesis test of all the robustness tests can be carried out separately for each original result in a paper, based on the following formula:

(1) Mean effect size/√(mean variance)=z-test statistic
(2) Mean t/z-value=z test statistic

If the effect sizes of all robustness tests of an original result is estimated in the same units, we recommend using equation (1) and otherwise we recommend using equation (2). It is not obvious whether to include the original result in this pooled test and this depends on whether the original test is viewed as one of the possible plausible analyses or not, and one possibility is to report the pooled tests results with and without the original result. It would also be possible to construct more complex pooled tests based on for instance bootstrap methods, but these simple pooled measures can be a natural starting point for a pooled test.

3.4. Indicators for original null results

For reproducibility tests of non-significant results (null results) in the original paper we recommend reporting an adjusted version of the statistical significance indicator. The adjusted version of the statistical significance indicator estimates the fraction of significant reproducibility tests irrespective of direction of the effect size (and this indicator can be pooled across non-significant original results within a paper and across papers for a group of papers in the same way as for our other proposed indicators). Note that the interpretation of this measure will be in the other direction compared to using the statistical significance indicator for original statistically significant results. A low fraction of statistically significant findings now suggest that the original null result have high reproducibility. We do not recommend estimating relative effect sizes for original null results as this is complicated for results where the original finding may be close to zero (the ratios may “blow up”). If the original result is argued to be a null result, the relative effect size measure is also difficult to interpret in a meaningful way. For

robustness reproducibility, the robustness ratio and the pooled hypothesis test can also be used as reproducibility indicators for original null results.

4. Replicability indicators

For replicability we also propose using the statistical significance and the relative effect size indicators. These indicators have been used in systematic replication studies (Open Science Collaboration, 2015; Camerer et al. 2016, 2018). For replications, where new data is collected, we make a distinction between statistical tests of replicability and descriptive indicators of replicability. We furthermore make a distinction between replications of one original result reported as statistically significant per original paper, and replications of several original results per paper reported as statistically significant. In addition, we make a distinction between replication indicators of a single original study and replication indicators of a group of studies (which opens up possibilities for statistical tests of the degree of replicability for the group of studies). Our recommendations for replications of results reported as statistically significant are summarized in Table 3. Below we also comment on replication indicators for results not reported as statistically significant (null results) in original studies, and we also briefly comment on some additional replication indicators proposed in the literature.

Table 3. Recommended replication indicators (for results reported as statistically significant in original studies).

	Statistical test indicator	**Descriptive indicator**
Individual replications: one original result replicated per paper	Replication effect size statistically significantly>0 (where >0 implies an effect in the same direction as the original study).	Relative effect size: Replication effect size/original effect size.
Individual replications: several separate original results replicated per paper	Replication effect size statistically significantly>0 tested for each result in a paper (where >0 implies an effect in the same direction as the original study).	A: Fraction of results that replicated according to the statistical significance indicator. B: The relative effect size of the replication for each result. C: The mean relative effect size where the mean replication effect size across the results is divided by the mean original effect size across results (requires that effect sizes are measured in

		the same units across results). D: The mean relative effect size of the replications for all tests (where the relative effect size is first estimated for each result and then the mean is taken of this variable).
A group of replication studies: one original result replicated per paper	A: If effect sizes are in comparable units across papers (e.g. Cohen's d units): A paired test (t-test or Wilxocon) of if the replication effect size differs statistically significantly from the original effect size (each original study and replication forms one pair). B: If effect sizes are not comparable across papers. Either a paired sign test of if the replication effect size differs statistically significantly from the original effect size (each original study and replication forms one pair). Or test if the relative effect size of the replications differs statistically significantly from 1 (100%) (one observation per replication study).	A: The fraction of results/papers that replicated according to the statistical significance indicator. B: The mean relative effect size estimated as the mean replication effect size across the studies divided by the mean original effect size across the original studies. C: The mean relative effect of the replications where the mean relative effect size is first estimated for each replication and then the mean of this variable is estimated (one observation per paper).
A group of replication studies: several separate original results replicated per paper	A: If effect sizes of all results in a paper have the same units and effect sizes are in comparable units across papers (e.g. Cohen's d units): A paired test of if the average replication effect size per paper differs statistically significantly from the average original effect size per paper (each original study and replication forms one pair). B: If effect sizes of all results in a paper have the same	A: The mean fraction of results that replicated per paper according to the statistical significance indicator (where this mean is first estimated per paper so that there is one observation per paper and the mean is then estimated for this variable). B: The mean relative effect size per paper (where the relative effect size per paper is estimated as the mean replication effect size of the

	units but effect sizes are not comparable across papers: A test of if the average relative effect size per paper differs statistically significantly from 1 (100%) (one observation per paper where the relative effect size per paper is estimated as the mean replication effect size of the results of that paper divided by the mean original effect size of the results of that paper). C: If effect sizes of all results in a paper do not have the same units and effect sizes are not comparable across papers: A test of if the average relative effect size per paper differs statistically significantly from 1 (100%) (one observation per paper estimated as the mean relative effect size of all results replicated in that paper).	results of that paper divided by the mean original effect size of the results of that paper). C: The mean relative effect size per paper (where the mean relative effect size per paper is first estimated as the mean relative effect size of the results replicated for that paper and then the mean is estimated for this variable with one observation per paper).

4.1. The statistical significance indicator

This indicator is the standard null hypothesis test in the literature, with the addition that the significant effect also has to be in the same direction as the original study.[7] This replication indicator focuses on to what extent the replication data support the hypothesis claimed to be statistically significant in the original study. It has the same pros and cons as null hypothesis testing in other settings. The statistical power of the replication is important and if the replication has low power the risk of false negatives is high; if for instance a replication is based on the same sample size as an original study that reported a statistically significant effect with a p-value just below 0.05, the probability that the replication will find a significant effect in the same direction is only about 50% even if the original effect size corresponds to the true

[7] We can think of this as testing a one-sided hypothesis, although using a two-sided hypothesis test that is more conservative and a test at the 5% level implies a false positive risk of 2.5%.

effect size. It is therefore very important with high powered replications, which also need to consider that the effect sizes of true positive original results are likely to be overestimated. We recommend having at least 90% power to detect 50% of the original effect size.

4.2. The relative effect size indicator

This indicator is defined as the effect size of the replication divided by the effect size of the original study. This is a continuous measure of the "degree of replication". The drawback of this indicator is that it is difficult to apply as a statistical test of replication for individual replications, but we recommend that it is reported as a descriptive indicator also for individual replications.

For a group of replication studies the relative effect size indicator can also be used as a statistical test. If a common standardized effect size unit (such as Cohen's d) is used in all original and replication studies, the mean relative effect size for a group of replications can be estimated as the mean effect size of all the replication studies and the mean effect size of all the original studies and then estimating the ratio between these two means. It can be tested if the mean replication effect size differs from the mean original effect size in a paired t-test or a Wilcoxon non-parametric test. We think this is the most useful statistical test of the replication rate for a group of replication studies (although the number of studies needs to be sufficiently large for the test to be high-powered). The mean relative effect size for a group of studies can also be estimated in a second way where the relative effect size is first estimated for each replication and then the mean is estimated of this variable. It can be tested if this mean relative effect size is significantly different from 1 in a one-sample t-test. This second mean relative effect size measure only requires comparability of effect sizes between the original study and the replication study for each original-replication study pair.

These two aggregated measures of the relative effect size will typically not be identical as the weighting of the difference between the replication and original study will differ for each original-replication study pair (the first aggregated measure aggregates absolute differences before the relative effect size measure is formed and the second aggregate measure aggregates relative differences). The second measure is more sensitive to outliers in relative effect sizes. We recommend reporting both relative effect size measures descriptively for studies using standardized effect sizes but using the first more robust aggregated measure for the main statistical test.

One advantage of the relative effect size indicators is that the mean relative effect size is not affected by the power of the replications, and it is therefore suitable for comparing the replicability across large scale systematic replication studies (that may differ in terms of the statistical power of the replications).

4.3 Replication indicators of original null results

For replication tests of non-significant results (null results) in the original paper we recommend reporting an adjusted version of the statistical significance indicator. The adjusted version of the statistical significance indicator defines replication as finding a non-significant effect size also in the replication and defines failed replication as finding a statistically significant effect size (irrespective of direction) in the replication. This measure can also be aggregated for replication of multiple original null results per paper and for multiple papers. Note that using this indicator for null results leads to the opposite relationship to statistical power than for replicating statistically significant original results; low statistical power now increases the likelihood of replication for this indicator (if the original result is a true positive). For the same reasons as for reproducibility indicators of original null results above, we do not recommend using the relative effect size indicator for replication of original null results.

4.4. Additional replication indicators proposed in the literature

Several additional replication indicators have been proposed in the literature. One is the "prediction interval approach" (Patil et al. 2016), which entails testing for a statistically significant difference between the replication effect size and the original effect size in a z-test. This replication indicator has important disadvantages for individual replication studies as it has a low likelihood to detect a lower replication effect size for original studies with a p-value close to 0.05 (the replication effect size needs to be in the opposite direction of the original effect size to have a chance of being statistically significantly lower than the original effect size). The original studies that are the most likely to be false positives are thus more than 50% likely to be classified as replicating with this criteria even if there is a true null effect. But note that for a group of studies it is useful to test if the replication effect sizes are smaller on average than the original effect sizes, and this corresponds to our recommended tests/indicators based on relative effect sizes in Table 3.

The "small telescopes" indicator involves testing if the replication effect size is significantly smaller (at the 5% level in a one-sided test) than a "small effect size" defined as the effect size the original study had 33% power to detect (Simonsohn 2015). If the replication effect size is

significantly smaller than the small effect size it counts as a failed replication and otherwise it counts as a successful replication. An important limitation of this indicator is that the "small effect size" is arbitrarily determined by the original sample size leading to substantially larger "small effect sizes" for small underpowered original studies than large high-powered original studies.

Another proposed replication indicator is the Bayes factor (BF) of the likelihood of the original hypothesis versus the null hypothesis of no effect based on the replication data; often referred to as the one-sided default Bayes factor (Wagenmakers et al. 2018). The so-called replication Bayes Factor has also been proposed, that estimates the likelihood of the original effect size versus the null hypothesis of no effect based on the replication data (Verhagen & Wagenmakers 2014). Reporting these Bayes Factors can be a useful complement or substitute to the statistical significance indicator for individual replication studies (the default Bayes Factor can be expected to be highly correlated with the p-value of the test of if the replication effect size is statistically significant).

5. Discussion

Threre exist some previous definitions of various types of replications in economics such as those proposed by Hammermesh (2007) and Clemens (2017); see the recent paper by Ankel-Peters et al. (2023) for a comprehensive comparison and discussion of the various proposed classifications.[8] An advantage of our proposed typology over previous proposals is that it aligns the use of the terms reproducibility and replicability with what is becoming the standard use of these terms in the social sciences. It also retains the typical use of direct and conceptual replications, as well as the growing use of computational reproducibility. The two newer categories, recreate reproducibility and robustness reproducibility, reflect the growing interest in these type of reproducibility studies.

With the above definition of replication (direct and conceptual replication), a replication is any study that tests the same hypothesis as a previous study but with new data, and where the result of the replication will affect beliefs about the likelihood of the tested hypothesis being true. This is in line with the recent definition of replication by Nosek and Errington (2020), although they define the beliefs part as that the following two conditions has to hold: (i) the beliefs in the original claim increase if the replication finds a result consistent with the original study and (ii) the beliefs in the original claim decrease if the replication finds a result that is inconsistent

[8] See also the proposed classification for sociology by Freese and Peterson (2017).

with the original hypothesis. This is a kind of symmetry condition that has to hold in their definition, but we find it hard to see how one of these conditions can be fulfilled without the other also being fulfilled; i.e. if a result consistent with the original study increases the beliefs in the hypothesis tested in the original study a result that is inconsistent with the original study must presumably decrease the beliefs in the hypothesis tested in the original study (although the effect can be very small).

Nosek and Errington (2020) also argue for abandoning the distinction between direct and conceptual replications. We are not convinced of this as a large fraction of papers in the scientific literature that would not classify themselves as replications will be replications with their definition (their definition would mean combining direct and conceptual replications in our definition). We do not think this is in line with the common understanding of the term. We therefore still prefer to make a distinction between direct and conceptual replications, where we think direct replications is what most researchers have in mind when using the term replication. There is, however, a degree of arbitrariness in drawing the line between a direct and a conceptual replication. A study testing the same hypothesis as a previous study can differ in (at least) three dimensions: the population included in the study, the research design used to test the hypothesis, and the analysis used to test the hypothesis. In our definition of a direct replication we argue that the research design and analysis should be the same, while we divide direct replications in different sub-groups depending on the population included. Even though these definitions are in some sense precise, there is a degree of arbitrariness in defining what constitutes the same research design and analysis (and the same, similar or different populations). The population, research design and analysis can differ along a continuous scale between two studies testing the same hypothesis and it becomes a more or less arbitrary decision where to draw the line between direct and conceptual replications along these continuous scales. In our classification it may also be controversial to define a study using the same research design and analysis implemented in a different population as a direct replication and this could potentially also have been defined as a conceptual replication.

As more systematic reproducibility and replication projects take place in economics, we believe that the usage of the proposed typology and indicators will facilitate the discussion and dissemination of reproducibility and replication results. This will probably also lead to refinements of the typology and the proposed indicators and to the development of new indicators, but we believe that our proposed ones provide a solid starting point.

References

Ankel-Peters J, Fiala N, Neubauer F. Do economists replicate?, I4R Discussion Paper Series, No. 13, Institute for Replication (I4R), 2023.

Athey S, Imbens G. A measure of robustness to misspecification. American Economic Review: Papers & Proceedings 2015;105:476-480.

Bergh DD, Sharp BM, Aguinis H, Li M. Is there a credibility crisis in strategic management research? Evidence on the reproducibility of study findings. Strategic Organization 2017;15:423-436.

Berkeley Initiative for Transparency in the Social Sciences. Guide for Advancing Computational Reproducibility in the Social Sciences. 2020. https://bitss.github.io/ACRE/.

Black B, Desai H, Litvak K, Yoo W, Yu JJ. The SEC's short-sale experiment: Evidence on causal channels and on the importance of specification choice in randomized and natural experiments. ECGI Working Paper Series in Finance, Working Paper No 813/2022.

Brodeur A, Lé M, Sangnier M, Zylberberg Y. Star wars: the empirics strikes back. American Economic Journal: Applied 2016:8:1-32.

Brodeur A, Cook N, Heyes A. Methods matter: p-hacking and publication bias in causal analysis in economics. American Economic Review 2020;110:3634-3660.

Brown AN, Wood BDK. Replication studies of development impact evaluations. Journal of Development Studies 2019:55:917-925.

Button KS, Ioannidis JPA, Mokrysz C, Nosek BA, Flint J, Robinson ESJ, Munafo MR. Power failure: Why small sample size undermines the reliability of neuroscience. Nature Reviews Neuroscience 2013;14:365-376.

Camerer CF, et al. Evaluating replicability of laboratory experiments in economics. Science 2016;351:1433-1436.

Camerer CF, et al. Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. Nature Human Behaviour 2018;2:637-644.

Chang AC, Li P. A preanalysis plan to replicate sixty economics research papers that worked half of the time. American Economic Review Papers and Proceedings 2017;107:60-64.

Christensen G, Miguel E. Transparency, reproducibility, and the credibility of economics research. Journal of Economic Literature 2018;56:920-980.

Clemens MA. The meaning of failed replications: A review and proposal. Journal of Economic Surveys 2017;31:326–342.

Delios A, Clemente EG, Wu T, Tan H, Wang Y, Gordon M, Viganola D, Chen Z, Dreber A, Johannesson M, Pfeiffer T, Generalizability Tests Forecasting Collaboration, Uhlmann EL. Examining the generalizability of research findings from archival data. PNAS 2022;119:e2120377119.

Dewald WG, Thursby J, Anderson R. Replication in empirical economics: The Journal of Money, Credit and Banking project. American Economic Review 1986;76:587-603.

Dreber A, Pfeiffer T, Almenberg J, Isaksson S, Wilson B, Chen Y, Nosek BA, Johannesson M. Using prediction markets to estimate the reproducibility of scientific research. PNAS 2015;112:15343-15347.

Ebersole CR, et al. Many Labs 3: Evaluating participant pool quality across the academic semester via replication. Journal of Experimental Social Psychology 2016;67:68-82.

Franco A., Malhotra N, Simonovits G. Publication bias in the social sciences: Unlocking the file drawer. *Science* 2014;345:1502-1505.

Franco A, Malhotra N & Simonovits G. Underreporting in political science survey experiments: Comparing questionnaires to published results. Political Analysis *2015;*23:306-312.

Freese J, Peterson D. Replication in social science. Annual Review of Sociology 2017;43:147–165.

Gelman A & Loken E. The statistical crisis in science. American Scientist 2014;102:460-465.

Gertler P, Galiani S, Romero M. How to make replication the norm. Nature 2018;554:417-419.

Glandon PJ. Appendix to the Report of the Editor: Report on the American Economic Review data availability compliance project. American Economic Review Papers & Proceedings 2011;101:695-699.

Hamermesh DS. Viewpoint: Replication in economics. Canadian Journal of Economics 2007;40:715–733.

Herbert S, Kingi H, Stanchi F, Vilhuber L. The reproducibility of economics research: A case study. Banque de France Working Paper Series, WP 85#3, 2021.

Hou K, Xue C, Zhang L. Replicating anomalies. Review of Financial Studies 2020;33:2019–2133.

Huntington-Klein N, et al. The influence of hidden researcher decisions in applied microeconomics. Economic Inquiry 2021;59:944-960.

Ioannidis JPA. Why most published research findings are false. PLoS Medicine 2005;2:e124.

Ioannidis JPA, Stanley TD, Doucouliagos H. The power of bias in economics research. Economic Journal 2017;127:F236-F265.

Jensen TI, Kelly BT, Pedersen LH. Is there a replication crises in finance? Journal of Finance 2022, forthcoming.

John LK, Loewenstein G & Prelec D. Measuring the prevalence of questionable research practices with incentives for truth telling. Psychological Science 2012;23:524-532.

Johnson VE, Payne RD, Wang T, Asher A, Mandal S. On the reproducibility of psycological science. Journal of the American Statistical Association 2017:112:1-10.

Klein RA, et al. Investigating variation in replicability: A “many labs” replication project. Social Psychology 2014:45:142-152.

Klein RA, et al. Many Labs 2: Investigating variation in replicability across samples and settings. Advances in Methods and Practices in Psychological Science 2018:1:443-490.

Leamer EE. Let’s take the con out of econometrics. American Economic Review 1983;73:31-43.

Maniadis Z, Tufano F, List JA. One swallow doesn’t make a summer: New evidence of anchoring effects. American Economic Review 2014:104(1):277-290.

McCullough BD, McGeary KA, Harrison TD. Lessons from the JMCB archive. Journal of Money, Credit and Banking 2006;38:1093-1107.

McCullough BD, McGeary KA, Harrison TD. Do economics journal archives promote replicable research? Canadian Journal of Economics 2008;41:1406-1420.

Mueller-Langer F, Fecher B, Harhoff D, Wagner GG. Replication studies in economics: How many and which papers are chosen for replication, and why? Research Policy 2019;48:62-83.

Nelson LD, Simmons J & Simonsohn U. Psychology's renaissance. Annual Review of Psychology 2018;69:511-534.

Nosek BN, Errington TM. What is replication? PLoS Biology 2020;18:e3000691.
Open Science Collaboration. Estimating the reproducibility of psychological science. Science 2015;349:aac4716.

Patil P, Peng RD, Leek JT. What should we expect when we replicate? A statistical view of replicability in psychological science. Perspectives of Psychological Science 2016;11:539-544.

Perignon C, et al. Reproducibility of empirical results: Evidence from 1,000 tests in finance. SSRN Working paper, 2022.

Ryan JC, Tipu SAA. Business and management research: Low instances of replication studies and a lack of author independence in replications. Research Policy 2022;51:104408.

Simmons JP, Nelson LD, Simonsohn U. False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. Psychological Science 2011:22:1359-1366.

Simonsohn U, Simmons JP, Nelson LD. Specification curve analysis. Nature Human Behaviour 2020;4:1208-1214.

Simonsohn U. Small telescopes: Detectability and the evaluation of replication results. Psychological Science 2015;26:559-569.

Steegen S, Tuerlinckx F, Gelman A, Vanpaemel W. Increasing transparency through a multiverse analysis. Perspectives on Psychological Science 2016;11:702-712.

Stern JM, Simes RJ. Publication bias: evidence of delayed publication in a cohort of clinical research projects. *BMJ* 1997;315:640-645.

Verhagen J, Wagenmakers E.-J. Bayesian tests to quantify the result of a replication attempt. Journal of Experimental Psychology: General 2014:143:1457-1475.

Wagenmakers E.-J., et al. Bayesian inference for psychology. Part II: Example applications with JASP. Psychonomic Bulletin & Review 2018;25:58-76.