Dreber, Anna; Johannesson, Magnus

Working Paper

A framework for evaluating reproducibility and replicability in economics

Ruhr Economic Papers, No. 1055

Provided in Cooperation with:
RWI – Leibniz-Institut für Wirtschaftsforschung, Essen

Suggested Citation: Dreber, Anna; Johannesson, Magnus (2023) : A framework for evaluating reproducibility and replicability in economics, Ruhr Economic Papers, No. 1055, ISBN 978-3-96973-225-0, RWI - Leibniz-Institut für Wirtschaftsforschung, Essen, https://doi.org/10.4419/96973225

This Version is available at:
https://hdl.handle.net/10419/281196

RUHR
ECONOMIC PAPERS

Anna Dreber
Magnus Johannesson

A Framework for Evaluating Reproducibility and Replicability in Economics

#1055

Imprint

Ruhr Economic Papers

Published by

RWI – Leibniz-Institut für Wirtschaftsforschung
Hohenzollernstr. 1-3, 45128 Essen, Germany

Ruhr-Universität Bochum (RUB), Department of Economics
Universitätsstr. 150, 44801 Bochum, Germany

Technische Universität Dortmund, Department of Economic and Social Sciences
Vogelpothsweg 87, 44227 Dortmund, Germany

Universität Duisburg-Essen, Department of Economics
Universitätsstr. 12, 45117 Essen, Germany

Editors

Prof. Dr. Thomas K. Bauer
RUB, Department of Economics, Empirical Economics
Phone: +49 (0) 234/3 22 83 41, e-mail: thomas.bauer@rub.de

Prof. Dr. Wolfgang Leininger
Technische Universität Dortmund, Department of Economic and Social Sciences
Economics – Microeconomics
Phone: +49 (0) 231/7 55-3297, e-mail: W.Leininger@tu-dortmund.de

Prof. Dr. Volker Clausen
University of Duisburg-Essen, Department of Economics
International Economics
Phone: +49 (0) 201/1 83-3655, e-mail: vclausen@vwl.uni-due.de

Prof. Dr. Ronald Bachmann, Prof. Dr. Manuel Frondel, Prof. Dr. Torsten Schmidt, Prof. Dr. Ansgar Wübker
RWI, Phone: +49 (0) 201/81 49-213, e-mail: presse@rwi-essen.de

Editorial Office

Sabine Weiler
RWI, Phone: +49 (0) 201/81 49-213, e-mail: sabine.weiler@rwi-essen.de

Ruhr Economic Papers #1055

Responsible Editor: Manuel Frondel

ISSN 1864-4872 (online) – ISBN 978-3-96973-225-0

Ruhr Economic Papers #1055

Anna Dreber and Magnus Johannesson

A Framework for Evaluating Reproducibility and Replicability in Economics

Bibliografische Informationen
der Deutschen Nationalbibliothek

The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available on the Internet at **http://dnb.dnb.de**

RWI is funded by the Federal Government and the federal state of North Rhine-Westphalia.

http://dx.doi.org/10.4419/96973225
ISSN 1864-4872 (online)
ISBN 978-3-96973-225-0

Anna Dreber and Magnus Johannesson[1]

A Framework for Evaluating Reproducibility and Replicability in Economics

Abstract

We propose a framework for evaluating reproducibility and replicability in economics. 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 further divide reproducibility and replicability studies into five types: computational reproducibility, recreate reproducibility, robustness reproducibility, direct replicability and conceptual replicability. In addition to this typology we propose indicators to measure the degree of reproducibility and replicability in both individual studies and for a group of studies.

JEL-Code: B41, C81, C90

Keywords: Reproducibility; replicability; metascience

December 2023

1 Anna Dreber, Stockholm School of Economics, University of Innsbruck and RWI Research Network; Magnus Johannesson, Stockholm School of Economics and RWI Research Network. *- 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.), Riksbankens Jubileumsfond (grant P21-0168 to M.J.), and the German Research Foundation (DFG) [Grant No. 3473/1–1; METAREP (SPP 2317)]. We thank Jörg Ankel-Peters, Abel Brodeur, Fernando Hoces de la Guardia and Séverine Toussaert for helpful comments.* *– All correspondence to: Anna Dreber, Stockholm School of Economics, Sveavägen 65, 113 83 Stockholm, Sweden, e-mail: anna.dreber@hhs.se*

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, 2023; Nelson et al. 2018; Ferraro & Shukla 2020), 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) (Bollen et al. 2015).

In this article we propose a framework for evaluating reproducibility and replicability in economics.[2] The article is divided into two parts. In the first part we propose a typology of reproducibility and replicability building on the existing literature. We divide reproducibility and replicability studies into five types: computational reproducibility, recreate reproducibility, robustness reproducibility, direct replicability and conceptual replicability. In the second part, we propose indicators to measure the degree of reproducibility and replicability of each type, and we show how these indicators can be used for both individual reproducibility and replicability studies and aggregated for a group of studies.

2. Typology of reproducibility and replicability studies

Hamermesh (2007) and Clemens (2017) have previously provided definitions of different types of replications in economics. Their definitions are however not aligned with what is becoming the standard use of the terms reproducibility and replicability in the social sciences.[3] An updated typology is therefore needed. In the Discussion section we further compare our typology to the proposals of Hamermesh (2007) and Clemens (2007).

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) (Bollen et al. 2015). 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, but are introduced due to the increased

[2] We believe this can be applied also to other quantitative fields in the social sciences.
[3] See for instance the definitions of reproducibility and replicability by a U.S. National Science Foundation (NSF) subcommittee on replicability in science (Bollen et al. 2015).

interest in those type of studies in recent years.[4] 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 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.[5] One would expect computational reproducibility to be high as it essentially implies testing for errors in running the original code on the original

[4] The term "robustness" in robustness reproducibility is also in line with how the term "robustness analysis" is used in economics.

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

data, 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 conducted 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 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), with the most challenging case being 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. 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.[6]

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 or specification curve analysis that explores all combinations of plausible analytical decisions (Steegen et al. 2016; Simonsohn et al. 2020). The ideal test of robustness reproducibility may be moving towards conducting multiverse analysis.

There are many studies testing the robustness of individual papers in economics, often published as comments (Ankel-Peters et al. 2023b). It is however 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. It is interesting to note that several systematic studies are currently being conducted (see, e.g., the work from the Institute for Replication) and we thus expect more systematic work to be published on robustness reproducibility in the coming years.[7]

2.4. Direct replicability

For experimental studies, a direct replication as closely as possible uses the same experimental design and the same analysis as the original study to test the same hypothesis (ideally using the experimental instructions, software and analysis code used in the original study, which should preferably 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 to the population in the original study. However, completely ruling out any systematic difference between the sample of the original study and the replication study would involve randomly drawing the sample from the same population. In

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

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 (as otherwise it could be the case that the population has 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 we 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 replications 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.

The systematic replication projects on experimental economics and experimental social sciences by Camerer et al. (2016, 2018) are examples of systematic direct replication projects 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, we can distinguish between conceptual replications based on new samples from the same population, a similar population, or a different population.

We 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, which would imply a sizeable literature on conceptual replications in economics. The recent literature on replicating anomalies in finance can also be viewed as conceptual replications; see for instance the studies by Hou et al. (2020) and Jensen et al. (2023).

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 and the second is the relative effect size indicator. 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 one additional indicator, the variation indicator, based on the variation in results across alternative analyses.

In Table 2 we define our proposed reproducibility indicators.[8] In column (1) we describe the indicator for evaluating one original result 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, 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 if this is a form of random measurement error. We describe the proposed indicators further below.

[8] In estimating the reproducibility indicators in Table 2 it is important to correctly incorporate the signs of the effect sizes and t/z-values so that an effect in the same direction as the original study has the same sign as the original effect size (t/z-value) and an effect in the opposite direction of the original study has the opposite sign as the original effect size (t/z-value). This can be achieved by always assigning the effect size and t/z-value of the original study a positive sign, and assigning effect sizes and t/z-values of reproducibility tests a positive sign if the effect goes in the same direction as the original study and assigning effect sizes and t/z-values of reproducibility tests a negative sign if the effect goes in the opposite direction of the original study. This also applies to the replicability indicators in Table 3.

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

Reproducibility indicator	(1). One original result reproduced in one paper	(2). Pooling across separate original results reproduced within one paper	(3). Pooling across papers
Statistical significance indicator	**X/N** X=number of statistically significant reproducibility tests of the original result with an effect in the same direction as the original result. N=number of reproducibility tests of the original result.*	Average of (1) across separate original results.	Average of (2) across papers.
Relative effect size indicator	**A: Mean ES_r/ES_o** Mean ES_r=mean effect size of all the reproducibility tests of the original result.* ES_o=effect size of the original result. **B: Mean t_r/t_o** Mean t_r=mean t/z-value of all the reproducibility tests of the original result.* t_o=t/z-value of the original result.	Average of (1) across separate original results.	Average of (2) across papers.&
Variation indicator§	**A: SD_r/SE_o** SD_r=standard deviation of the effect size of all the robustness tests of the original result.# SE_o=standard error of the original effect size. B: **SDt_r** SDt_r=standard deviation of the t/z-value of all the robustness tests of the original result.#	Average of (1) across separate results.	Average of (2) across papers.

***** For computational and recreate reproducibility this will be one test/effect size/t-value (based on one reproducibility test); for robustness reproducibility it will typically be several tests/effect sizes/t-values (based on several robustness tests).

§ The numerator (standard deviation measure) can also be reported here and viewed as an indicator of absolute variation (but can only be compared across studies using the same effect size units).

The original result should also be included among the robustness tests in this estimation if it is considered a reasonable analysis.

& It can be tested if this average differs statistically significantly from 1 to test for systematic bias in original effect sizes; where a value below 1 implies systematically lower reproducibility effect sizes than original effect sizes.

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).[9] 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.[10] This replication indicator focuses on to what extent the reproduction 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. If a robustness reproducibility study is carried out including 10 robustness tests and 7 of these tests are statistically significant with an effect in the same direction as the original study, the value of the statistical significance indicator for this robustness reproducibility study will be 0.7 (7/10).

The statistical significance indicator is binary (0/1) if only one reproducibility test is carried out of one original result, but will be a continuous indicator between 0-1 if more than one reproducibility test is carried out or results are pooled across results within a paper or across papers; a lower value implies lower reproducibility. It can be interpreted as the strength of support of the hypotheses tested in the original studies included in a reproducibility study.

3.2. The relative effect size indicator

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 indicator 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 if a log transformation is used 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, the relative t/z-values can be reported as an

[9] For robustness reproducibility another potential related indicator would be to include the average t/z-values of the robustness tests of an original result (and the original analysis if it is considered a reasonable analysis). The average t/z-value could be used to conduct a simple pooled hypothesis test, and can be viewed as a modification of the statistical significance indicator. The average t/z-value could also be considered a continuous measure of the strength of support in the hypothesis tested in the original study.

[10] 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%.

additional indicator as this indicator will also reflect variation in standard errors across robustness tests.

If a robustness reproducibility study is carried out including 10 robustness tests and the average effect size in these robustness tests is 0.2 and the original effect size is 0.5, the value of the relative effect size indicator for this robustness reproducibility study will be 0.4 (0.2/0.5). The relative effect size indicator is a continuous reproducibility indicator and if it is less than 1 it suggests that the original effect sizes are systematically overestimated. This indicator thus provides important information in addition to the statistical significance indicator as it is an indicator of systematic bias in original studies. The statistical significance indicator in a robustness reproducibility study can for instance be relatively low even if the average effect size in the robustness tests are not lower than the original effect size, if the results vary widely across the robustness tests. In our view the relative effect size indicator is the most important indicator of reproducibility as it directly measures systematic bias in original studies.

3.3. The variation indicator

This indicator is only proposed for robustness reproducibility.[11] The variation indicator is a measure of the variation across robustness tests, including also the original result if that is deemed a reasonable analysis. It is defined as the standard deviation in effect sizes among all the robustness tests divided by the standard error of the original effect size. It is divided by the standard error of the original study to get a measure of the variation across robustness tests relative to the reported sampling uncertainty in the original estimate, and to be able to compare the indicator across studies. If a robustness reproducibility study finds a standard deviation in effect sizes across the robustness test of 0.4 and the standard error of the effect size of the original study is 0.4, the variation indicator is 1, implying that the variation between robustness tests is as large as the sampling variation (standard error) in the original study. The higher is the variation indicator, the lower is the robustness reproducibility.

[11] An additional related potential robustness reproducibility indicator would be to use an indicator proposed by Athey & Imbens (2015) as a measure of the robustness to alternative regression analysis specifications. This measure is based on the square root of the mean absolute deviation between each robustness test and the original result. This "standard deviation" is then divided by the standard error of the original effect size (unless the measure is based on t/z-values rather than effect sizes, in which case it should not be divided by the standard error). This indicator will capture both the variation captured by the variation indicator, and the systematic deviation between the robustness tests and the original effect size and we referred to this indicator as the "Robustness ratio" in a previous version of this paper. One limitation of this indicator is that it will not depend on the direction of the systematic bias, making it complicated to pool results if the systematic deviation goes in different directions for different results.

The variation indicator is related to measures of heterogeneity reported in multi-analyst studies (Huntington-Klein et al. 2021; Menkveld et al. 2023). It provides additional information compared to the statistical significance indicator and the relative effect size indicator as it is a pure measure of the variation across robustness tests. The statistical significance indicator and the relative effect size indicator can for instance be close to 1 for an original study, even though there is substantial variation across robustness tests and this variation will be picked up by the variation indicator (this can happen if an original study has a very high original t-value and the effect sizes in the robustness tests are on average similar to the original effect size; but there is still large variation between the robustness tests).

Also this indicator can be defined in terms of t/z-values instead of effect sizes. As the standard deviation of t/z-values is 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). When the indicator is defined based on t/z-values it will also incorporate variation in standard errors across the robustness tests.

3.4. Indicators for original null results

For reproducibility tests of non-significant results in the original paper we recommend reporting an adjusted version of the statistical significance indicator. The adjusted version 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 suggests that the original null result has 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. The variation indicator can also be used as robustness reproducibility indicator for original null results.

4. Replicability indicators

For replicability we also propose using the statistical significance indicator and the relative effect size indicators. These indicators have been used in systematic replication studies (Open

Science Collaboration 2015; Camerer et al. 2016, 2018). Our recommendations for replicability indicators of original results reported as statistically significant are summarized in Table 3. The Table also provides information for how results can be pooled within a paper and across papers. For the systematic replication projects the pooled indicators across papers have been the primary results of these studies; i.e. the replication rate across the included papers and the average relative effect size of the replications.

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

Replicability indicator	**(1). One original result replicated in one paper.**	**Pooling across separate original results replicated within one paper (2)**	**(3). Pooling across papers**
Statistical significance indicator	**1=replication effect in the original direction and statistically significant.** **0=otherwise.**	Average of (1) across separate original results.	Average of (2) across papers.
Relative effect size indicator	**ES_r/ES_o** ES_r=the effect size of the replication. ES_o=the effect size of the original result.	Average of (1) across separate original results.	Average of (2) across papers.*

* It can be tested if it differs statistically significantly from 1; where a value below 1 implies systematically lower replication effect sizes than original effect sizes.

4.1. The statistical significance indicator

An important difference in applying this indicator to replicability rather than reproducibility is that for replications the sample size will often differ between the original study and the replication study. The replication sample size and thereby statistical power of the replication is important and if the replication has low power the risk of false negatives is high for this replicability indicator. It is therefore crucial 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 ideally having at least 90% power to detect 50% of the original effect size. An example of the use of this replication indicator is in the Experimental Economics Replication Project (Camerer et al. 2016), in which 11/18 studies replicated with this indicator and the

replicability was thus 0.61 (11/18). The indicator can vary between 0 and 1 for a group of studies (although it is a binary indicator in each replication test).

4.2. The relative effect size indicator

The relative effect size indicator is a continuous indicator of the degree of replicability.[12] For individual repliction studies a drawback of this indicator is that it is difficult to apply as a statistical test of replication. For a group of replication studies it can be used as a statistical test of systematically lower effect sizes in the replication studies, and this test was used to test for systematically lower effect sizes in the RPP (Open Science Collaboration 2016), the Experimental Economics Replication Project (EERP) (Camerer et al. 2016), and the Social Science Replication Project (SSRP) (Camerer et al. 2018).[13] 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 average relative effect size was 0.44 in RPP (Open Science Collaboration 2015), 0.66 in EERP (camerer et al. 2016), and 0.46 in SSRP (Camerer et al. 2018), suggesting a replicability of about 50% in these studies.[14]

One important advantage of the relative effect size indicator 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 systematic replication studies (that may differ in terms of the statistical power of the replications). For reporting the replication results of a group of studies we thus consider it the most important indicator of replicability as it measures the degree of replication, and is not affected by statistical power or more or less arbitrary binary categorizations into successful or failed replications.

For reproducibility, a relative effect size measure could also be constructed based on t/z-values, but this is not possible for replicability as the sample size typically varies between original

[12] For a group of replication studies the relative effect size indicator can be defined in two different ways. For simplicity we only show the first version in Table 3, which can be used even if effect sizes are not measured in the same units across the included studies. 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 also be estimated in a second way, by dividing the mean effect size of all the replication studies by the mean effect size of all the original studies. We recommend reporting both relative effect size measures descriptively for studies using standardized effect sizes across replications.

[13] If the effect sizes are measured in standardized effect sizes across the replications a paired t-test or a Wilcoxon non-parametric test can be used (Open Science Collaboration 2015, Camerer et al. 2016, 2018).

[14] For the second way of estimating the mean relative effect size, the mean relative effect size was 0.49, in RPP (Open Science Collaboration 2015), 0.59 in EERP (Camerer et al 2016), and 0.54 in SSRP (Camerer et al. 2018).

studies and replication studies and the sample size affects the standard error of the effect size and thereby the t/z-value.

4.3 Replicability indicators of original null results

For replication tests of non-significant results in the original paper the recommendations in the previous section on reproducibility indicators applies here as well, with the exception that the variation indicator cannot be used for replicability. Note that using the statistical significance indicator for null results leads to the opposite relationship to statistical power than for replicating statistically significant original results; for original null results low statistical power increases the likelihood of replication.

4.4. Additional replicability 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 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 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 indicators based on relative effect sizes.

The "small telescopes" indicator involves testing if the replication effect size is significantly smaller 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 based on the replication data (Wagenmakers et al. 2018). Reporting 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 testing if the replication effect size is statistically significant).

5. Discussion

There exist some previous classifications of replications in economics (Hamermesh 2007; Clemens 2017); see the recent paper by Ankel-Peters et al. (2023) for a comprehensive comparison and discussion of the various proposed classifications.[15] Hamermesh (2007) proposed two types of replication studies: pure replications using the same data as the original study and scientific replications using new data. Pure replication corresponds to reproducibility and scientific replication to replicability in our typology. Clemens (2017) proposed four types of studies: replication: verification; replication: reproduction; robustness: reanalysis; and robustness: extension. In our typology, these four types approximately correspond to: computational reproducibility, direct replicability, robustness reproducibility, and conceptual replicability. Goodman et al. (2016) also proposed the following typology for biomedical sciences: methods reproducibility, results reproducibility, and inferential reproducibility. Methods reproducibility approximately corresponds to computational reproducibility and results reproducibility to direct replicability in our typology. Inferential reproducibility is about different researchers drawing the same conclusion from the same study and data and is not matched by any of our categories. Goodman et al. (2016) also separately discuss robustness but without integrating it into their reproducibility categories. As the Hamermesh (2007), Clemens (2017), or Goodman et al. (2016) terminologies have not been widely adopted in economics or the social sciences we believe there is room for our recommended typology.

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 terms, recreate reproducibility and robustness reproducibility, reflect the growing interest in these type of reproducibility studies. There are currently several large ongoing projects about robustness reproducibility such as the replication games organized by the Institute for Replication, and several studies that would be classified as recreate reproducibility have recently been published or conducted (Bergh et al. 2017; Delios et al. 2022; Black et al. 2022). Although there is substantial overlap between our typology and the Clemens (2017) typology as such, we would

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

argue that our terminology is semantically more straightforward and more in line with how these terms are currently used in the literature.

Having a clear terminology is a necessary condition for more clarity in the profession, but as emphasized by Clemens (2017), increased clarity will only be achieved if researchers actually use the terminology and refer to the specific type of reproducibility and replicability study in their papers. Ankel-Peters (2023a) relatedly introduced the term ”policing replication”, implying that replications should embrace their role policing and challenging a previously published result. This is also close to the diagnostic motives for replication discussed by Peterson and Panofsky (2021). To fulfill this “policing function” researchers need to use very clear language and definitions to communicate the type of reproducibility or replicability study that is being conducted. Without the use of clear definitions it may become more difficult for replications and reproductions to achieve its disciplinary and scientific self-correction effect. This is also an argument for conducting ”deep reproducibility” studies, digging deeply into the original study including the raw data and coding decisions to detect potential coding errors or qustionable coding decisions. Computational reproducibility tests conducted by data editors at for instance the American Economic Association journals will detect mismatches between code and reported results, but may not detect important coding errors that are incorporated into the reported results; see the discussion on this by Ankel-Peters et al (2023b).

With the above definition of replication (direct and conceptual), 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 & Errington (2020), although they define the beliefs part as the following two conditions: (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 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.

This definition of replication implies that many studies could be defined as conceptual replications. All the studies typically pooled in a meta-analysis to estimate a pooled effect size of some hypothesis will be conceptual replications with this definition. That the category of

conceptual replications becomes too broad could be argued to be a weakness of our proposed typology, but it is consistent with the definition of replication of Nosek & Errington (2020). One possibility to narrow the conceptual replication category would be to require a study to label itself as a repliction and clearly refer to the replicated paper to be considered a conceptual replication. This may also be necessary for conceptual replications to have a disciplinary effect on the incentives of researchers.

Nosek & Errington (2020) 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.

Our proposed indicators provides important information about the degree of reproducibility and replicability. However, there are also important aspects that are not covered by our framework. One such issue is the sharing of data, code and research materials which will facilitate reproducibility and replicability studies. Another important issue is the choice of robustness tests to implement in robustness reproducibility studies. Ideally all reasonable analysis paths should be included in a multiverse type of analysis, but it is not clear how to determine these analysis paths and different scholars can disagree about which specifications are reasonable (Steegen et al. 2016; Simonsohn et al. 2020). This is also evidenced in the study

by Ankel-Peters et al. (2023b) about reproducibility debates in comments and replies in the American Economic Review. Multi-analyst studies where many analysts independently test the same hypotheses using the same data also show wide variation in the analytical decisions considered reasonable by the analysts (Silberzahn et al. 2018; Botvinik-Nezer et al. 2020; Huntington-Klein et al. 2021; Breznau et al. 2022; Menkveld et al. 2023). An important topic in future work is how to identify reasonable analysis paths for tests of robustness reproducibility, possibly by the use of crowd-sourcing methods.

As more systematic reproducibility and replication projects take place in economics, we believe that the usage of our 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. (2023a). Do economists replicate? Journal of Economic Behavior & Organization 212: 219-232.

Ankel-Peters, J., Fiala, N. &Neubauer F. (2023b). Is economics self-correcting? Replications in the American Economic Review. Ruhr Economic Papers 105.

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

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

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

Black, B., Desai, H., Litvak, K., Yoo, W. & Yu, J.J. (2022). 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.

Botvinik-Nezer R., Holzmeister, F., Camerer, C.F. et al. (2020). Variability in the analysis of a single neuroimaging dataset by many teams. Nature 582: 84–88.

Bollen, K., Cacioppo, J.T., Kaplan, R., Krosnick, J. & Olds J.L. (2015). Social, behavioral, and economic sciences perspectives on robust and reliable science. National Science Foundation, Arlington, VA.

Breznau, N., Rinke E.M., Wuttke, A. et al. (2022). Observing many researchers using the same data and hypothesis reveals a hidden universe of uncertainty. Proceedings of the National Academy of Sciences 119: e2203150119.

Brodeur, A., Carrell, S.E., Figlio, D.N. & Lusher, L.R. (2023). Unpacking p-hacking and publication bias. American Economic Review, forthcoming.

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

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

Brown, A.N. & Wood, B.D.K. (2019). Replication studies of development impact evaluations. Journal of Development Studies 55: 917-925.

Button, K.S., Ioannidis, J.P.A., Mokrysz, C., Nosek, B.A., Flint, J., Robinson, E.S.J. & Munafo, M.R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. Nature Reviews Neuroscience 14: 365-376.

Camerer, C.F., Dreber, A, Forsell, E. et al. (2016). Evaluating replicability of laboratory experiments in economics. Science 351: 1433-1436.

Camerer, C.F., Dreber, A., Holzmeister, F., et al. (2018). Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. Nature Human Behaviour 2: 637-644.

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

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

Clemens, M.A. (2017). The meaning of failed replications: A review and proposal. Journal of Economic Surveys 31: 326–342.

Delios, A., Clemente, E.G., Wu, T. et al. (2022). Examining the generalizability of research findings from archival data. Proceedings of the National Academy of Sciences 119: e2120377119.

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

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

Ebersole, C.R., Atherton, O.E., Belanger, A.L. et al. (2016). Many Labs 3: Evaluating participant pool quality across the academic semester via replication. Journal of Experimental Social Psychology 67: 68-82.

Ferraro, P.J. & Shukla, P. (2020). Feature—is a replicability crisis on the horizon for environmental and resource economics? Review of Environmental Economics and Policy 14: 339-351.

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

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

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

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

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

Glandon, P.J. (2011). Appendix to the Report of the Editor: Report on the American Economic Review data availability compliance project. American Economic Review Papers & Proceedings 101: 695-699.

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

Hamermesh, D.S. (2007). Viewpoint: Replication in economics. Canadian Journal of Economics 40: 715–733.

Hedges, L.V. (1992). Modeling publication selection effects in meta-analysis. Statistical Science 7: 246-255.

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

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

Huntington-Klein, N., Areans, A., Beam, E. et al. (2021). The influence of hidden researcher decisions in applied microeconomics. Economic Inquiry 59: 944-960.

Ioannidis, J.P.A. (2005). Why most published research findings are false. PLoS Medicine 2: e124.

Ioannidis, J.P.A., Stanley, T.D. & Doucouliagos, H. (2017). The power of bias in economics research. Economic Journal 127: F236-F265.

Jensen, T.I., Kelly, B.T., Pedersen, L.H. (2023). Is there a replication crises in finance? Journal of Finance 78: 2465-2518.

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

Johnson, V.E., Payne, R.D., Wang, T., Asher, A. & Mandal, S. (2017). On the reproducibility of psychological science. Journal of the American Statistical Association 112: 1-10.

Klein, R.A., Ratliff, K.A., Vianello, M. et al. (2014). Investigating variation in replicability: A "many labs" replication project. Social Psychology 45: 142-152.

Klein, R.A., Vianello, M, Hasselman, F. et al. (2018). Many Labs 2: Investigating variation in replicability across samples and settings. Advances in Methods and Practices in Psychological Science 1: 443-490.

Leamer, E.E. (1983). Let's take the con out of econometrics. American Economic Review 73: 31-43.

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

Menkveld, A., Dreber, A., Holzmeister, F. et al. (2023). Non-standard errors. Journal of Finance, forthcoming.

McCullough, B.D., McGeary, K.A. & Harrison, T.D. (2006). Lessons from the JMCB archive. Journal of Money, Credit and Banking 38: 1093-1107.

McCullough, B.D., McGeary, K.A. & Harrison, T.D. (2008). Do economics journal archives promote replicable research? Canadian Journal of Economics 41: 1406-1420.

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

Nelson, L.D, Simmons, J. & Simonsohn, U. (2018). Psychology's renaissance. Annual Review of Psychology 69: 511-534.

Nosek, B.N. & Errington, T.M. (2020). What is replication? PLoS Biology 18: e3000691.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. Science 349: aac4716.

Patil, P., Peng, R.D. & Leek, J.T. (2016). What should we expect when we replicate? A statistical view of replicability in psychological science. Perspectives of Psychological Science 11: 539-544.

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

Peterson, D. & Panofsky, A. (2021). Self-correction in science: The diagnostic and integrative motives for replication. Social Studies of Science 51: 583-605.

Ryan, J.C. & Tipu, S.A.A. (2022). Business and management research: Low instances of replication studies and a lack of author independence in replications. Research Policy 51: 104408.

Silberzahn, R., Uhlmann, E.L., Martin, D.P. et al. (2018). Many analysts, one dataset: Making transparent how variation in analytical choices affect results. Advances in Methods and Practices in Psychological Science 1: 337-356.

Simmons, J.P., Nelson, L.D. & Simonsohn, U. (2011). False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. Psychological Science 22: 1359-1366.

Simonsohn, U., Simmons, J.P. & Nelson, L.D. (2020). Specification curve analysis. Nature Human Behaviour 4: 1208-1214.

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

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

Stern, J.M. & Simes, R.J. (1997). Publication bias: evidence of delayed publication in a cohort of clinical research projects. *BMJ* 315: 640-645.

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