Clemens, Michael A.

Working Paper

The Meaning of Failed Replications: A Review and Proposal

IZA Discussion Papers, No. 9000

Provided in Cooperation with:
IZA – Institute of Labor Economics

Suggested Citation: Clemens, Michael A. (2015) : The Meaning of Failed Replications: A Review and Proposal, IZA Discussion Papers, No. 9000, Institute for the Study of Labor (IZA), Bonn

This Version is available at:
https://hdl.handle.net/10419/110735

I Z A

DISCUSSION PAPER SERIES

IZA DP No. 9000

The Meaning of Failed Replications: A Review and Proposal

Michael A. Clemens

April 2015

Forschungsinstitut
zur Zukunft der Arbeit
Institute for the Study
of Labor

The Meaning of Failed Replications: A Review and Proposal

Michael A. Clemens
Center for Global Development
and IZA

Discussion Paper No. 9000
April 2015

IZA

P.O. Box 7240
53072 Bonn
Germany

Phone: +49-228-3894-0
Fax: +49-228-3894-180
E-mail: iza@iza.org

IZA Discussion Paper No. 9000
April 2015

ABSTRACT

The Meaning of Failed Replications: A Review and Proposal*

The welcome rise of replication tests in economics has not been accompanied by a single, clear definition of replication. A discrepant replication, in current usage of the term, can signal anything from an unremarkable disagreement over methods to scientific incompetence or misconduct. This paper proposes an unambiguous definition of replication, one that reflects currently common but unstandardized use. It contrasts this definition with decades of unsuccessful attempts to standardize terminology, and argues that many prominent results described as replication tests – in labor, development, and other fields of economics – should not be described as such. Adopting this definition can improve incentives for researchers, encouraging more and better replication tests.

NON-TECHNICAL SUMMARY

Economists are increasingly using publicly shared data and code to check each other's work, an exercise often called 'replication' testing. But this much-needed trend has not been accompanied by a consensus about what 'replication' means. Many follow-up studies on influential papers in labor economics and other fields have been unable to 'replicate' the original result. But according to current usage of the term, this can mean anything from an unremarkable disagreement over methods to scientific incompetence or misconduct.

This paper proposes an unambiguous definition of replication. Many social scientists already use the term in the way suggested here, but many more do not. The paper contrasts this definition with decades of unsuccessful attempts to standardize terminology, and argues that many prominent results described as replication tests should not be described as such. It argues that professional associations should formally adopt this definition, thereby improving incentives for researchers to conduct more and better replication tests.

JEL Classification: B40, C18, C80

Keywords: replication, robustness, transparency, open data, ethics, reproducible, replicate, misconduct, fraud, error, code, registry

Corresponding author:

Michael A. Clemens
Center for Global Development
2055 L Street NW, 5th floor
Washington, DC 20036
USA
E-mail: mclemens@cgdev.org

* This research was generously supported by Good Ventures. Chris Blattman, Annette Brown, Angus Deaton, Gabriel Demombynes, Stefan Dercon, John Hoddinott, Macartan Humphreys, Stephan Klasen, Ted Miguel, Emily Oster, Justin Sandefur, Bill Savedoff, and Ben Wood provided helpful comments. All viewpoints and any errors are the sole responsibility of the author and do not represent CGD, its Board of Directors, or its funders.

1 The problem

Social science is benefiting from a surge of interest in subjecting published research to replication tests. But economics and other social sciences have yet to clearly define what a replication is. Thus if a replication test gives discrepant results, under current usage of the term, this could mean a wide spectrum of things—from signaling a legitimate disagreement over the best methods (science), to signaling incompetence and fraud (pseudoscience). Terminology that lumps together fundamentally different things impedes scientific progress, hobbling researchers with fruitless debates and poor incentives.

This paper argues that the movement for replication in social science will become stronger with clear terminology. It begins by proposing an unambiguous definition of replication. It shows that usage compatible with the proposed definition is already widespread in the literature, but so is usage that is incompatible. It then reviews decades of attempts to resolve this conceptual confusion across the social sciences, and shows how the terminology proposed here addresses the problem. The remaining sections argue that the proposed definition creates better incentives for researchers, and applies this definition to classify many recent and prominent critique papers in labor economics, development economics, and other subfields. It concludes by arguing that the need for this terminology arises from a generational shift in how empirical social science is conducted.

2 A proposal to define replication and robustness

Consider the proposed definitions of replication and robustness tests in Table 1. They are distinguished by whether or not the follow-up test should give, in expectation, exactly the same quantitative result.

Table 1: A Proposed Definition to Distinguish Replication and Robustness Tests

	Sampling distribution for parameter estimates	Sufficient conditions for discrepancy	Types	Methods in follow-up study versus methods *reported* in original: Same specification	Same population	Same sample	Examples
Replication	*Same*	*Random chance, error, or fraud*	Verification	*Yes*	*Yes*	*Yes*	*Fix faulty measurement, code, dataset*
			Reproduction	*Yes*	*Yes*	*No*	*Remedy sampling error, low power*
Robustness	*Different*	*Sampling distribution has changed*	Reanalysis	*No*	*Yes*	*Yes/No*	*Alter specification, recode variables*
			Extension	*Yes*	*No*	*No*	*Alter place or time; drop outliers*

The "same" specification, population, or sample means the same as *reported* in the original paper, not necessarily what was contained in the code and data used by the original paper. Thus for example if code used in the original paper contains an error such that it does not run exactly the regressions that the original paper said it does, new code that fixes the error is nevertheless using the "same" specifications (as described in the paper).

2.1 What sets a replication apart

A *replication* test estimates parameters drawn from the same sampling distribution as those in the original study. A replication test can take two forms: A *verification* test means ensuring that the exact statistical analysis reported in the original paper gives materially the same results reported in the paper, either using the original dataset or remeasuring with identical methods the same traits of the same sample of subjects. This form of replication can remedy measurement error, coding errors, or errors in dataset construction. It can also uncover scientific misconduct. A *reproduction* test means resampling precisely the same population but otherwise using identical methods to the original study. This form of replication can remedy sampling error or low power, in addition to the errors addressed by a verification test.

A *robustness* test estimates parameters drawn from a different sampling distribution from those in the original study. A robustness test can take two forms: A *reanalysis* test means altering the computer code from the original study. It is exclusively a reanalysis if it uses exactly the same dataset or a new sample representative of the same population. This includes new regression specifications and variable coding. An *extension* test means using new data—gathered on a sample representative of a different population, or gathered on the same sample at a substantially different time, or both. This includes dropping influential observations, since a truncated sample cannot represent the same population. It is exclusively an extension test if it runs identical computer code on the new data. Both forms of robustness test estimate population parameters that are different from those in the original study, thus they need not give identical results in expectation. Many robustness tests are a mix of reanalysis and extension.

2.2 Examples

Restricting the term replication to this meaning fits the intuitive meaning that social science borrows from natural science. Lewis et al. (1989), for example, failed to *replicate* the "cold fusion" that Fleischmann and Pons (1989) notoriously claimed to generate with palladium and heavy water. This type of replication test was a reproduction test in the terms of Table 1, using new samples of palladium and heavy water (not the original dataset of Fleischmann and

Pons).

Lewis et al.'s failure-to-replicate meant that they got different results when they took identical actions with different samples of identical materials. It did not mean that they got different results when rhodium was substituted for palladium. In much of modern empirical social science, the actions of inquiry are contained in computer code and the materials are datasets that represent some population. If new code does different things than the original paper describes, or a new dataset represents a different population than the original paper describes, we are no longer speaking of replication in this intuitive sense.

Replication tests include: fixing coding errors so that the code does exactly what the original paper *describes* (verification), having the same sample of students take the same exam again to remedy measurement error (verification), and re-sampling the same population to remedy sampling error or low power with otherwise identical methods (reproduction). Discrepant results in any of these settings are properly described as negative replications of an original inquiry. In expectation, these tests are supposed to yield estimates identical to the original study. If they do not, then either the original or the replication contains a fluke, a mistake, or fraud.

Robustness tests include: re-coding or re-periodizing the same underlying dataset (reanalysis), changing the set of covariates or method of calculating standard errors (reanalysis), updating the dataset with new observations (extension), doing the same analysis on a sample that is representative of a different village or country (extension), and testing subsets of the original data (extension). Results discrepant from the original should not be described as a replication issue; there is no reason these tests should yield identical results in expectation. They are all testing quantitatively different hypotheses than the original study tested, because they all change the sampling distribution for the parameter of interest. Discrepant results in these settings are properly described, under the proposed definitions, by saying for example that the original study is not robust to reanalysis with new covariates, or not robust to extending the data to a different country. They are not properly described as negative replications.

The critical distinction between replication and robustness is whether or not the follow-up test

should give, in expectation, exactly the same quantitative result as the original test. The word quantitative is key. Suppose an original study gathered data on city *A*. You might replicate that study by gathering data from a new sample of people representing city *A*, analyzed with identical regressions. If you did such replications a large number of times, the estimates would converge towards a single number—in the absence of measurement error, sampling error, fraud, and so on. But if you add a new interaction term to the regressions on city *A* data, or run the original regressions on data from city *B*, the new estimates need not converge toward the original estimates, no matter how many times this is repeated. A discrepant result in these new tests does not mean that the original study failed to properly measure the population parameter of city *A* that it sought to measure, so there is no failure to replicate.

This remains true if the new results are "qualitatively" different, such as rejecting the null in city *A* but failing to reject in city *B*, or getting a different sign with and without including an interaction term. It is trivial to find some dataset, or some subset of the original dataset, or some regression specification where coefficients lose significance or change sign. If this is the criterion of failure-to-replicate, then any result can easily fail replication, and the value of the term is gone.

Thus declaring a failure to replicate requires demonstrating that the new estimate should be quantitatively identical to the old result in expectation. An investigator claiming a "replication" bears this burden of proof, due to the special status of replication in science. Other discrepant follow-up estimates should be described by stating that the original result was not robust to certain alterations.

For example, suppose a follow-up study of the same African district 15 years after the original study tries the same experimental intervention on a new sample at the same location. The original study found a large effect, the follow-up finds none. If potential confounders change slowly over time, this would constitute re-sampling a population that is materially the same—a replication test attempting to *reproduce* the original result. If confounders can change more quickly, the population in the follow-up is materially different; the sampling distribution for the new estimates is not the same, and the follow-up study is a robustness test (extension to new data). Thus a test of providing free mobile phones 15 years after an original study might

be a test of *extension* to a new time period, given that the mobile communications landscape is rapidly changing in Africa. But a test of providing the same vitamin supplement in the same place 15 years later might be a *replication*, if the factors that shape the impact of the vitamin change very slowly. This must be shown, and the standard should be high. If the location of the follow-up study differs at all, this is certainly an extension.

3 How the term replication is used now

There is no settled definition of the term replication in the tradition of economics or social science in general (Wulwick 1996; Hamermesh 2007). "The term nearly defies precise definition" (Mittelstaedt and Zorn 1984).

3.1 Usage compatible with this proposal is widespread

Many economics journals already endorse the key goal of Table 1: restricting the meaning of the word replication to a sense that does not include robustness tests. Authors in the *American Economic Review* "are expected to send their data, programs, and sufficient details to permit replication." Here, the term replication unambiguously means using the original data and code to get exactly the same results as appear in the paper.[1] The same policy has been adopted by the *Journal of Political Economy* and other leading journals. Dewald et al.'s (1986) famous scrutiny of articles in the *Journal of Money, Credit, and Banking* "collected programs and data from authors and attempted to replicate their published results"—that is, duplicate them precisely. The editorial policy of *Labour Economics* distinguishes replication—"repeat their estimated model with their method on their data"—from reanalysis—"changes in empirical specifications and estimation methods" or "a different data set" (Arulampalam et al. 1997).

[1] Authors of experimental papers must provide "sufficient explanation to make it possible to use the submitted computer programs to replicate the data analysis". This usage of the term replication is strictly incompatible with a meaning that includes new regression specifications or new data, since the authors of the original paper could not logically be required to "explain" to other authors in what ways they should modify the original work. The *AER* "Data Availability Policy" is available at `https://www.aeaweb.org/aer/data.php`. The *JPE* "Data Policy" is available at `http://www.press.uchicago.edu/journals/jpe/datapolicy.html`.

Many researchers, too, already use terminology consistent with the restricted meaning of replication in Table 1. Summers (1991) sharply distinguishes Dewald et al.'s (1986) "attempts at replication" from "the evaluation of robustness". Hubbard and Vetter (1996) separate "replications" ("substantial duplication") from "extensions". When Houtenville and Burkhauser (2004) revisit the findings of Acemoğlu and Angrist (2001) with new definitions of key variables, they distinguish this "robustness" test from a "replication". Easterly et al. (2004) revisit the results of Burnside and Dollar (2000) with new data, and describe their inquiry as a "robustness" test, not a replication test. Rothstein (2007a) describes his attempts to exactly reproduce Hoxby's (2000) coefficients as "replication" but distinguishes his "plausible alterations to Hoxby's specification" as a "reanalysis". To Anderson et al. (2008), replication means "another researcher using the same data and the same computer software should achieve the same results." McCullough (2009b) defines empirical economic research as replicable if "there exist data and code that can reproduce the published results." Vinod (2009) writes, "Replications merely check whether the results reported by authors are independently verifiable, not whether they are reliable, robust and stable." Miguel and Satyanath (2011) write that "Ciccone (2011) is not a replication critique, but rather a critique of the regression functional form that we use". Albouy (2012) describes his critique of Acemoğlu et al. (2001) with a different dataset as a test of "reliability", "robustness", and "sensitivity", but never suggests that the original study could not be "replicated".

3.2 Incompatible usage is also widespread

But there is no consensus meaning of the term replication. Many journals and organizations work with a definition that is irreconcilable with Table 1 and the usage in subsection 3.1. That is, they define replication so that follow-up studies can fail to replicate an original paper's findings even when the original study's code and data are correct and reproduce that study's results precisely.

Pesaran's (2003) editorial policy for the *Journal of Applied Econometrics* considers "replication" to include testing "if the substantive empirical finding of the paper can be replicated using data from other periods, countries, regions, or other entities as appropriate." Burman et al.'s (2010) editorial policy for *Public Finance Review* defines "negative replication" to in-

clude results that "are not robust to substantial extensions over time, data sets, explanatory variables, functional forms, software, and/or alternative estimation procedures." Selection criteria for the Replication Program of the International Initiative for Impact Evaluation (3ie 2012) include studies that reach "results that contradict previous findings in the literature" through "innovative methodology or estimation techniques".

Numerous researchers also work with a different definition than that proposed in Table 1. For Hamermesh (1997), in a *Labour Economics* symposium on replication, economics "can never be a field where mere duplication could be of any interest. ... The best replication studies... will attempt duplication as their starting point, but go far beyond that" to "try alternative methods and other specifications" or "time-series data... outside the original sample period." To Kniesner (1997), "the best replication study is a broad parameter robustness check" whose "prototypical" example alters the original study's "definition of the wage variable, pay scheme, inclusion of income taxation, instrument set, curvature, and parameterization of latent heterogeneity." Dority and Fuess (2007) alter both the specifications and dataset of Layard et al. (1994) and describe this exercise as "replication". Johnson et al. (2013) run other studies' regressions on a new, extended dataset unavailable to the previous authors and describe this inquiry as "replication". Camfield and Palmer-Jones (2013) call Albouy's aforementioned work a "replication", though Albouy does not. And Brown et al. (2014) define "internal replication" to include "redefining and recalculating the variables of interest, introducing additional control or interaction variables, and using alternative estimation methodologies."

3.3 Past attempts at a definition have not worked

The social science literature recognizes this confusion but has not resolved it. The literature is chronically afflicted with attempts to define replication. Those efforts have yielded a disappointing mess, summarized in Table 2. It shows different terms previously proposed in the literature to describe the concepts of "replication" and "robustness" distinguished in Table 1. This confusion suggests three lessons.

First, Table 2 reveals an enduring need for the conceptual distinction drawn by Table 1. There are decades of attempts, across the social sciences, to distinguish two things: studies that

Table 2: CORRESPONDENCE BETWEEN TABLE 1 DEFINITIONS AND PREVIOUS DEFINITIONS

	Replication test*	**Robustness test***	Source
Economics	Type I replication	Type II, III, IV replication	Mittelstaedt and Zorn 1984
	Econometric audit	Improvisational replication	Kane 1984
	'Reproduction' replication	'Reexamination' replication	Fuess 1996
	'Reproduction' replication	'Robustness' replication	Kniesner 1997
	Replication of the first degree	Higher-order replication/ reanalysis	Arulampalam et al. 1997
	'Narrow sense' replication	'Wide sense' replication	Pesaran 2003
	Pure replication	Statistical/Scientific replication	Hamermesh 2007
	Replication	—	McCullough et al. 2008
	Replication	Stress test	Vinod 2009
	Replication/reproduction	—	Koenker and Zeileis 2009
	Repeatability/Strict replication	Conceptual replication	Ioannidis and Doucouliagos 2013
	Replication	—	Data policy of *AER*, *JPE*, etc.
Statistics	Close replication	Differentiated replication	Lindsay and Ehrenberg 1993
	Computational reproduction	—	Donoho 2010
	Replication/reproduction	—	Stodden 2010
	Reproduction	Replication	Peng 2011
Political science	Replication	Extension	King 1995
	'Verification' reanalysis	'Replication' reanalysis	Herrnson 1995
	Replication	Extension, improvement	King 2006
	Narrow replication	Broad replication	Dafoe 2014
Sociology	Retest/internal replication	Independent/ theoretical replication	La Sorte 1972
	Identical replication	Virtual/systematic replication	Finifter 1972
	Replication type a	Replication type $b \ldots p$	Bahr et al. 1983
	Repetition/checking	Replication	Collins 1991
	Replication	Reproduction, robustness	Cartwright 1991
Psychology	Literal/operational replication	Constructive replication	Lykken 1968
	Replication	Quasi-replication	Cronbach 1975
	Exact replication	Partial/conceptual replication	Hendrick 1990
	Internal replication	External replication	Thompson 1994
	Direct replication	Conceptual replication	Schmidt 2009
Business	Experimental replication	Non-experimental/ corroboration replication	Leone and Schultz 1980
	Perfect replication	Imperfect replication	Farley et al. 1981
	Replication	Extension	Hubbard and Armstrong 1994
	Strict replication	Significant sameness	Barwise 1995
	Duplication	Operational replication	Madden et al. 1995
	Checking	Replication, reanalysis, extension, etc.	Tsang and Kwan 1999
	Strict replication	Partial/conceptual replication	Darley 2000
	Replication	Extension	Easley and Madden 2000
	Type 0, I replication	Type II, III replication	Easley et al. 2000
	Statistical replication	Scientific replication	Hunter 2001
	Replication	Replication with extension	Evanschitzky et al. 2007

*Column headings used as defined in Table 1. Each row contains terms used in the paper cited with meanings corresponding to the concept in each column. The word replication is appears in red to reveal how often it is applied to both concepts.
Note: Cartwright (1991) appears under 'sociology' because it comments on a work of sociology.

revisit an earlier paper by strictly reproducing its findings with the same data and methods it describes, and studies that revisit those findings by changing the data and/or methods.

Second, the word replication is routinely used to describe both kinds of studies. This occurs in every field. The attempted solution has been to use qualifiers to distinguish flavors of replication, but none of these have become standard. Thus if a 'replication' study finds a different result, that could mean that the study used identical data and methods or completely different data and methods.

Third, Table 2 shows not just a range of blurry meanings, but strictly incompatible meanings. In economics, McCullough (2009b) and Vinod (2009) use the term replication to exclude altering regression specifications and changing datasets, while Pesaran (2003) and Hamermesh (2007) explicitly include them. In political science, some researchers endorse the distinction in Table 1: King (1995) writes that replication has occurred when the same data and code "reproduce the numerical results in the article." Herrnson (1995) sides with the Committee on National Statistics (1985) and insists that replication only occurs when "different, independently collected data are used to study the same problem."[2]

4 Why this unambiguous distinction is needed

It is imperative for social science to distinguish between the concepts of replication and robustness distinguished in Table 1. This is because the the two concepts carry sharply different normative messages about the original research, with consequences for the field as a whole.

If a paper fails a *replication* test it is because there was something indisputably *wrong* in the original work or in the replication. At best this can mean measurement error or a minor, good-

[2]Worse, 'replication' is sometimes used with incompatible meanings within a single paper. Peng (2011) sharply distinguishes 'reproducibility' from 'replication': "The standard of reproducibility calls for the data and the computer code used to analyze the data be made available to others. This standard falls short of full replication because the same data are analyzed again, rather than analyzing independently collected data" (pp. 1–2). But then in Peng's Figure 1, "full replication" is described as a form of "reproducibility". Dafoe (2014, p. 66) uses the terms with incompatible meanings in one pair of sentences: "I use the term 'reproducible' to refer specifically to research for which the analysis is replicable. Sharing of replication files foremost promotes reproducible research (replications of analysis), though it might also promote replications of studies if the greater transparency facilitates the execution of the study on a new sample."

faith oversight, and even that best case—without any suggestion of scientific misconduct—is traumatic to authors: For Levitt (2002), a failed replication arising from an oversight in his original work was "unacceptable" and a source of "tremendous personal embarrassment". At worst, failed replications are linked to "fraud" (Trikalinos et al. 2008) and "doubts about scientific ethics" (Furman et al. 2012). "Replication speaks to ethical professional practice," write Camfield and Palmer-Jones (2013), and its motive is often to "uncover error or fraud". The American Political Science Association's policy on replication is expressed in its *Guide to Professional Ethics* (APSA 2012).

But if a paper fails a *robustness* test, it is because the original paper exhibits a choice that is legitimately debatable. It is not beyond question what the right choice was, and divergence of opinions has nothing at all to do with "ethics" and "fraud". Robustness tests often speak of "plausible" alterations to regression specifications, but the original specifications can seem just as "plausible" to another competent researcher. Robustness tests descriptively establish what would have happened if the original researchers had not done X; only replication tests normatively claim that the original researcher *indisputably should not* have done X. As Collins (1991, p. 136) puts it, "Replication is a matter of establishing ought, not is." And ought must be established. A replication critique bears the burden of proving that the original authors indisputably should not have made a certain choice.

These are two fundamentally different situations. We harm scientific progress when we confuse them, as we must, by referring to both with the same word (Table 2). Harm can arise in two ways, by shaping the incentives of authors on both sides of debates.

First, confused use of the term replication harms research by reducing original authors' incentives to collaborate across bona fide disagreements in method. All papers have legitimately debatable shortcomings, and science proceeds by collaborative discussions of better approaches. But many authors informed of a different result upon 'replication' of their work feel compelled to an adversarial, defensive stance (Camfield and Palmer-Jones 2013). It is understandable that they perceive failed replication as a threat; common usage of the word includes cases where it signifies incompetence or fraud. So muddled terminology makes authors fret as much about insubstantive misunderstandings as about the substance of research.

Thus Miguel and Satyanath (2011) feel obliged to begin by clarifying that Ciccone (2011) "is not a replication critique." Dercon et al. (2015), learning that their paper had been classified as "unable to replicate" based on Bowser's (2015) results, did not find the new results materially discrepant and protested the claim as "a fairly serious factual misstatement of the findings with reputational costs for all concerned." When Ash and Robinson (2009) claimed that they failed to "replicate" Deaton and Lubotsky's (2003) results due to a coding error in the original, Deaton and Lubotsky (2009) were obliged to counter, "This is not correct. Our regressions were run exactly as we claim, though it is certainly possible to challenge our choice." What unites these episodes is a strong concern by the original authors that readers could confuse replication tests (signifying mistakes or worse) with robustness tests (signifying legitimately arguable choices). This may be part of why the responding authors clearly feel targeted for attack rather more than they feel engaged in collaboration to advance science.

To be clear: I do not criticize these reactions, but consider them inevitable sequelae of the field's confused terminology. And confusion harms science. Misunderstood claims of failed "replications", in which the original researcher in fact did nothing indisputably wrong, "will make it much more difficult for serious policy-relevant researchers to do their job," continues Deaton (2013). "Scholars will also be much less willing to share data than is currently the case; doing so allows anyone who is unscrupulous enough to turn your cooperation against you." That is bad news for social science, which has a "desperate need for replications" (Hunter 2001).

Second, confusion in the meaning of replication harms research by creating perverse incentives for those conducting replication and robustness checks. Anyone can find 'plausible' ways to change someone else's regressions so that coefficient estimates change. Casey et al. (2012) show that they could have gotten nearly any result they might have wanted, with 'plausible' alterations to their *own* regressions, had their hands not been tied by a pre-analysis plan. Likewise, "if you want to debunk a paper, working through it equation by equation. . . you will eventually find something that changes" (Deaton 2013). A well-known form of this problem is that it is a simple matter to change any result into a null result by running modified versions of the same test that are underpowered by construction (Ottenbacher 1996; Hicks et al. 2014b; Bazzi and Bhavnani 2015). Thus if the meaning of a failed replication includes cases with

different specifications and data, then any empirical study can be made to "fail to replicate" by a person with a computer and sufficient determination. Failed replications attract more and faster attention than successful ones—the "Proteus phenomenon" documented by Ioannidis and Trikalinos (2005)—with obvious perverse incentives for those seeking academic or public notice. This problem, too, can be limited by clear terminology that distinguishes works altering the original research design from works of replication.

Things can go better with crisp distinctions by clear terminology. A successful example of clear language is the critiques of Oster's (2005) result on the potential for Hepatitis B infection to explain gender ratios in Asia. None of the major critiques claim any failed replication, or even contain the word replication (e.g. Das Gupta 2006; Lin and Luoh 2008; Klasen 2008). This accurately described their critiques: They disagreed with choices made in the original paper without suggesting that those choices were indisputably illegitimate, or that they reflected incompetence or fraud. They thus encouraged further analysis, without backing Oster into a corner with unintended innuendo about basic competence or scientific ethics. In the end, Oster herself refuted the original result with an extension test (Oster et al. 2010). Scientific debate worked, and Oster was widely praised (e.g. Cowen 2008).

Ultimately, clear terminology may encourage more replication and robustness testing. Clear terms make these exercises more of an opportunity for research and less of a perceived threat to researchers.

5 Most prominent critiques are not replications

The definitions proposed in Table 1 would clarify the nature of scientific critiques in the liteature. To take a famous example, Herndon et al. (2014) present a *replication* test of Reinhart and Rogoff (2010), in that they discuss what is indisputably an error (specifically, it is a verification test). Égert (2013), in contrast, presents a *robustness* test of the same paper, with *reanalysis* and *extension*: he uses alternative estimators and data to challenge choices in Reinhart and Rogoff (2010) that are legitimately disputable. In another well-known example, Foote and Goetz (2008) present a *replication* test of Donohue and Levitt (2001) by document-

ing a mistake beyond dispute; Joyce's (2004) critique of the same work regards robustness, not replication. These two types of critiques have sharply different implications for the original research and researchers; they need different names.

Table 3 carries out this classification for many of the best-known replication and robustness tests in the recent economics and political science literature. Of the 34 critiques, 35% are replication tests, by the definition in Table 1. The rest, a large majority, are robustness tests. And of those, 59% are reanalyses with new methods, 27% are extensions to new data, and 14% carry out both reanalysis and extension.

In short, definition matters. Most of these noted critiques do *not* fit the definition of a replication test proposed here and used by many researchers already (section 3.1). But all 34 of them *do* fit the definition of a replication test used by so many others (sections 3.2–3.3), which accommodates substantial changes in method and data. Most of these papers do not deserve the vague associations with incompetence or ethics that accompany failed "replications", but most of them could receive that label in the absence of terminology that clearly distinguishes the two types of critiques.

6 Replication yesterday and tomorrow

"[T]he replication standard is extremely important to the further development of the discipline," writes King (1995). For important things we need clear terms. The meaning of replication needs to be standardized just as the meaning of "statistically significant" once was.

The root of confusion about replication's meaning may lie in the changing nature of empirical investigation. Morgenstern (1951), in his time, saw little role for repetition in economics to mimic that found in the natural sciences. This was because, for example, public statistics on industrial production in a given year are only gathered once, and cannot be infinitely re-measured like the velocity of light. Thus Finifter (1972) finds that for social science, "on reflection, the notion of identical replication in a strict one-to-one duplication is eventually abandoned as an unattainable goal", and Madden et al. (1995) agree that "literal replication

Table 3: An Application of Proposed Table 1 Definitions to Selected Literature

Comment papers and rejoinders	Original papers and replies	*Note*
Replication tests		
Leimer and Lesnoy (1982)	Feldstein (1974, 1982)	✓ Fix programming error (also contains reanalysis)
Day and Liebowitz (1998)	Munnell et al. (1996)	✓ Dataset not as presented due to error in gov. dataset
McCrary (2002)	Levitt (1997, 2002)	✓ Fix programming & classification errors
Breusch and Gray (2006)	Chapman et al. (2001, 2006)	✓ Fix programming & classification errors
Rothstein (2007a, 2007b)	Hoxby (2000, 2007)	⇆ Alleged "errors in data and computer programs"
Foote and Goetz (2008)	Donohue and Levitt (2001, 2008)	✓ Paper describes specifications different from code
Ash and Robinson (2009)	Deaton and Lubotsky (2003, 2009)	⇆ Alleged coding error
Bailey (2009)	Bailey (2006)	✓ Rebuilds code with somewhat discrepant results
Bump et al. (2012)	Pronyk et al. (2012), Pronyk (2012)	✓ Errors in calculating changes in main outcome
Baker (2013)*	Feldstein (1996)	✓ Gov. dataset may have changed with revisions
Herndon et al. (2014)	Reinhart and Rogoff (2010, 2013)	✓ Coding errors (also contains reanalysis)
Aiken et al. (2014a)	Miguel and Kremer (2004), Hicks et al. (2014a)	⇆ Coding error; relevance to findings disputed
Robustness tests		
Boyce and Ravallion (1991)	Khan (1984)	⬗ New specifications, full data
Mack and Wulwick (1991)	Phillips (1958)	◇ Nonparametric estimator (Sleeman 2011)
Harrison (1998)	Munnell et al. (1996)	◇ New specification, additional variables
Dai (2002, 2006)	Mansfield et al. (2000, 2002)	◇ Alters assumptions of original model
Joyce (2004, 2006, 2009)	Donohue and Levitt (2001, 2004, 2008)	◇ New ident. strategy, serial corr. adjustment
Easterly et al. (2004)	Burnside and Dollar (2000)	◆ Updated data
Houtenville and Burkhauser (2004)	Acemoğlu and Angrist (2001)	◇ New definitions of disability & employment
Das Gupta (2006), Lin and Luoh (2008)	Oster (2005), Oster et al. (2010)	◆ Large new individual-level dataset
Dority and Fuess (2007)	Layard et al. (1994)	⬗ New specifications, updated data
Lott and Whitley (2007)	Donohue and Levitt (2001)	◆ Alternative dataset
Özer Ballı and Sørensen (2010, 2013)	Rajan and Zingales (1998), Easterly et al. (2004), etc.	◇ De-meaning interaction term
Ciccone (2011)	Miguel et al. (2004), Miguel and Satyanath (2011)	◇ New specifications
Albouy (2012)	Acemoğlu et al. (2001, 2012)	◆ Different dataset by recoding key regressor
Clemens et al. (2012)	Boone (1996), Burnside and Dollar (2000), etc.	⬗ New specifications and data
Johnson et al. (2013)	Mankiw et al. (1992), Jones and Olken (2005), etc.	◆ New, updated dataset
Davis (2013)	Sachs and Warner (1997)	◇ Different indep. variable, specification
Égert (2013)	Reinhart and Rogoff (2010)	◇ Alternative debt measures, specifications
Iversen et al. (2013)	Banerjee and Iyer (2005)	◇ Different dataset by recoding key regressor
Martel García (2013)	Ross (2006)	◇ Recoding regressor of interest
Iversen and Palmer-Jones (2014)	Jensen and Oster (2009, 2012, 2014)	◇ Rebuild index (immaterial corrections)
Aiken et al. (2014b)	Miguel and Kremer (2004), Hicks et al. (2014b)	◇ Different subsets of data, new specifications
Bowser (2015)	Dercon et al. (2009, 2015)	◆ Data from new survey round

Replication tests: ✓ = Little dispute that some substantial replication failure occurred. ⇆ = Responding authors dispute substantial replication failure.
Robustness tests: ◇ = Reanalysis. ◆ = Extension. ⬗ = Reanalysis and extension. Terms defined in Table 1. *Does not report full results.

is probably not possible in the social sciences." This reflects a view in which social scientists are passive and infallible observers of unique, one-off phenomena in the world outside, not fallible executors of repeatable inquiries within their own departments and offices. From this point of view, mere repetition of the same analysis on the same data is seen as "uninventive" (Kane 1984). "Mindlessly taking the exact same data and checking to see if an author 'made a mistake' is not a useful activity in the social sciences," writes (Hamermesh 1997). This helps explain why "[t]he credo of experimental repetition never has taken hold in economics" (Wulwick 1996).

Modern social science has evolved into something quite different. In particular, empirical economics today consists largely of the application of computer code to computerized datasets. Empirical economics is acquiring important traits of computational science. In computational science, "[m]ost of the work in a modern research project is hidden in computational scripts that go to produce the reported results," writes Donoho (2010). "An article about a computational result is advertising, not scholarship. The actual scholarship is the full software environment, code and data, that produced the result." In modern empirical social science, problems within that environment are common, as Dewald et al. (1986) revealed and Table 3 here confirms. These go far beyond researcher error. Two statistical software packages ostensibly performing the same calculation can yield very different results (McCullough 2009a; Bazzi and Clemens 2013, footnote 30), and the underlying data used by the original authors can be incorrectly constructed, through no fault of theirs (Day and Liebowitz 1998). As datasets, code, and underlying software become more complex, replication "is increasingly important because our intuition fails in high dimensions" (Baggerly and Berry 2011). In this view, repeating a set of empirical calculations exactly *is* replicating scientific inquiry, because the code and data *are* the scholarship. Replication thus conceived, far from being "not possible" (Madden et al. 1995), is a necessary condition for science.

This is a profound evolution in methods and concepts. Our terminology has not caught up with it, and many researchers have noticed. Psychologist Chris Chambers (2012) asks for an end to describing robustness tests, sometimes called "conceptual replications" (Table 2) as any form of replication: "[W]e must jettison the flawed notion of conceptual replication. It is vital to seek converging evidence for particular phenomena using different methodologies. But *this*

isn't replication, and it should never be regarded as a substitute for replication."

Social science needs more replication work. It will get more and better replication work by standardizing the meaning of replication, ending confusion and doubt about the meaning of past and future failures-to-replicate. Standardization requires leadership—in this case by professional associations and by institutions championing the noble work of replication. They should act to enshrine a single definition of the term. The American Economic Association, for example, could definitively solve this problem by creating separate *Journal of Economic Literature* codes for replication tests and robustness tests, with nonintersecting definitions. This would transform the word replication from a blurry, fraught locution into an exact technical term, a tool for scientific progress.

References

3ie, "3ie Replication Programme: Programme Description," International Initiative for Impact Evaluation 2012.

Acemoğlu, Daron and Joshua D Angrist, "Consequences of Employment Protection? The Case of the Americans with Disabilities Act," *Journal of Political Economy*, 2001, *109* (5), 915–957.

__ , Simon Johnson, and James A Robinson, "The Colonial Origins of Comparative Development: An Empirical Investigation," *American Economic Review*, 2001, *91* (5), 1369–1401.

__ , __ , and __ , "The colonial origins of comparative development: an empirical investigation: reply," *American Economic Review*, 2012, *102* (6), 3077–3110.

Aiken, Alexander M, Calum Davey, James R Hargreaves, and Richard J Hayes, "Reanalysis of health and educational impacts of a school-based deworming program in western Kenya, Part 1: pure replication," 3ie Replication Series Paper 3, Part 1, New Delhi: International Initiative for Impact Evaluation (3ie) 2014.

__ , __ , __ , and __ , "Reanalysis of health and educational impacts of a school-based deworming program in western Kenya, Part 2: Alternative analyses," 3ie Replication Series Paper 3, Part 2, New Delhi: International Initiative for Impact Evaluation (3ie) 2014.

Albouy, David Y, "The colonial origins of comparative development: an empirical investigation: comment," *American Economic Review*, 2012, *102* (6), 3059–3076.

Anderson, Richard G, William H Greene, Bruce D McCullough, and Hrishikesh D Vinod, "The role of data/code archives in the future of economic research," *Journal of Economic Methodology*, 2008, *15* (1), 99–119.

APSA, *A Guide to Professional Ethics in Political Science*, 2nd ed., Washington, DC: American Political Science Association, 2012.

Arulampalam, Wiji, Joop Hartog, Tom MaCurdy, and Jules Theeuwes, "Replication and re-analysis," *Labour Economics*, 1997, *4* (2), 99–105.

Ash, Michael and Dean E Robinson, "Inequality, race, and mortality in US cities: a political and econometric review of Deaton and Lubotsky (56: 6, 1139–1153, 2003)," *Social Science & Medicine*, 2009, *68* (11), 1909–1913.

Baggerly, Keith A and Donald A Berry, "Reproducible research," *Amstat News*, 2011, *January*.

Bahr, Howard M, Theodore Caplow, and Bruce A Chadwick, "Middletown III: Problems of Replication, Longitudinal Measurement, and Triangulation," *Annual Review of Sociology*, 1983, *9* (1), 243–264.

Bailey, Martha J, "More Power to the Pill: The Impact of Contraceptive Freedom on Women's Life Cycle Labor Supply," *Quarterly Journal of Economics*, 2006, *121* (1), 289–320.

Bailey, Martha J., "Erratum and Addendum: 'More Power to the Pill,' *Quarterly Journal of Economics*, February 2006," Working Paper, University of Michigan 2009.

Baker, Dean, "In History of Economic Errors, Martin Feldstein Deserves Mention," *Beat the Press* blog, Center for Economic Policy and Research, April 17 2013.

Banerjee, Abhijit and Lakshmi Iyer, "History, Institutions, and Economic Performance: The Legacy of Colonial Land Tenure Systems in India," *American Economic Review*, 2005, *95* (4), 1190–1213.

Barwise, Patrick, "Good empirical generalizations," *Marketing Science*, 1995, *14* (3, Part 2), G29–G35.

Bazzi, Samuel and Michael A Clemens, "Blunt instruments: Avoiding common pitfalls in identifying the causes of economic growth," *American Economic Journal: Macroeconomics*, 2013, *5* (2), 152–186.

__ and Rikhil Bhavnani, "A Reply to 'A Replication of "Counting Chickens When They Hatch" (*Economic Journal* 2012)'," *Public Finance Review*, 2015, *43* (2), 282–286.

Boone, Peter, "Politics and the effectiveness of foreign aid," *European Economic Review*, 1996, *40* (2), 289–329.

Bowser, William H., "The long and short of returns to public investments in fifteen Ethiopian villages," 3ie Replication Series Paper 4, New Delhi: International Initiative for Impact Evaluation (3ie) 2015.

Boyce, James K and Martin Ravallion, "A dynamic econometric model of agricultural wage determination in Bangladesh," *Oxford Bulletin of Economics and Statistics*, 1991, *53* (4), 361–376.

Breusch, Trevor and Edith Gray, "Replicating a study of mothers' forgone earnings in Australia," *Journal of Economic and Social Measurement*, 2006, *31* (1), 107–125.

Brown, Annette N, Drew B Cameron, and Benjamin DK Wood, "Quality evidence for policymaking: I'll believe it when I see the replication," *Journal of Development Effectiveness*, 2014, *6* (3), 215–235.

Bump, Jesse B, Michael A Clemens, Gabriel Demombynes, and Lawrence Haddad, "Concerns about the Millennium Villages project report," *The Lancet*, 2012, *379* (9830), 1945.

Burman, Leonard E, W Robert Reed, and James Alm, "A call for replication studies," *Public Finance Review*, 2010, *38* (6), 787–793.

Burnside, Craig and David Dollar, "Aid, Policies, and Growth," *American Economic Review*, 2000, *90* (4), 847–868.

Camfield, Laura and Richard Palmer-Jones, "Three 'Rs' of Econometrics: Repetition, Reproduction and Replication," *Journal of Development Studies*, 2013, *49* (12), 1607–1614.

Cartwright, Nancy, "Replicability, reproducibility, and robustness: Comments on Harry Collins," *History of Political Economy*, 1991, *23* (1), 143–155.

Casey, Katherine, Rachel Glennerster, and Edward Miguel, "Reshaping Institutions: Evidence on Aid Impacts Using a Preanalysis Plan," *Quarterly Journal of Economics*, 2012, *127* (4), 1755–1812.

Chambers, Chris, "You can't replicate a concept," *Neurochambers* weblog; School of Psychology, Cardiff University; March 26 2012.

Chapman, Bruce, "Response to Breusch and Gray," *Journal of Economic and Social Measurement*, 2006, *31* (1), 127–138.

__ , Yvonne Dunlop, Matthew Gray, Amy Liu, and Deborah Mitchell, "The impact of children on the lifetime earnings of Australian women: Evidence from the 1990s," *Australian Economic Review*, 2001, *34* (4), 373–389.

Ciccone, Antonio, "Economic shocks and civil conflict: A comment," *American Economic Journal: Applied Economics*, 2011, *3* (4), 215–227.

Clemens, Michael A, Steven Radelet, Rikhil R Bhavnani, and Samuel Bazzi, "Counting chickens when they hatch: Timing and the effects of aid on growth*," *Economic Journal*, 2012, *122* (561), 590–617.

Collins, Harry M, "The meaning of replication and the science of economics," *History of Political Economy*, 1991, *23* (1), 123–142.

Committee on National Statistics, *Sharing Research Data*, Washington, DC: National Academy Press, 1985.

Cowen, Tyler, "Hail Emily Oster!," *Marginal Revolution* weblog, May 12 2008.

Cronbach, Lee J, "Beyond the two disciplines of scientific psychology.," *American Psychologist*, 1975, *30* (2), 116–127.

Dafoe, Allan, "Science deserves better: the imperative to share complete replication files," *PS: Political Science & Politics*, 2014, *47* (01), 60–66.

Dai, Xinyuan, "Political regimes and international trade: The democratic difference revisited," *American Political Science Review*, 2002, *96* (1), 159–165.

__ , "Dyadic Myth and Monadic Advantage Conceptualizing the Effect of Democratic Constraints on Trade," *Journal of Theoretical Politics*, 2006, *18* (3), 267–297.

Darley, William K, "Status of replication studies in marketing: A validation and extension," *Marketing Management Journal*, 2000, *10* (2), 121–132.

Das Gupta, Monica, "Cultural versus biological factors in explaining Asia's 'missing women': Response to Oster," *Population and Development Review*, 2006, *32* (2), 328–332.

Davis, Graham A, "Replicating Sachs and Warner's working papers on the resource curse," *Journal of Development Studies*, 2013, *49* (12), 1615–1630.

Day, Theodore E and Stan J Liebowitz, "Mortgage lending to minorities: Where's the bias?," *Economic Inquiry*, 1998, *36* (1), 3–28.

Deaton, Angus, "On weights and coding errors: odd coincidence or dress rehearsal?," On James D. Hamilton and Menzie Chinn, eds., *Econbrowser: Analysis of Current Economic Conditions and Policy* weblog, October 9 2013.

__ **and Darren Lubotsky**, "Mortality, inequality and race in American cities and states," *Social Science & Medicine*, 2003, *56* (6), 1139–1153.

__ **and** __ , "Income inequality and mortality in US cities: Weighing the evidence. A response to Ash," *Social Science & Medicine*, 2009, *68* (11), 1914–1917.

Dercon, Stefan, Daniel O Gilligan, John Hoddinott, and Tassew Woldehanna, "The impact of agricultural extension and roads on poverty and consumption growth in fifteen Ethiopian villages," *American Journal of Agricultural Economics*, 2009, *91* (4), 1007–1021.

__ , __ , __ , **and** __ , "The Impact of Agricultural Extension and Roads on Poverty and Consumption Growth in Fifteen Ethiopian Villages: Response to William Bowser," *Response* section of '3ie Replication Paper 4', New Delhi: International Initiative for Impact Evaluation (3ie) 2015.

Dewald, William G, Jerry Thursby, and Richard G Anderson, "Replication in Empirical Economics: The Journal of Money, Credit and Banking Project," *American Economic Review*, 1986, *76* (4), 587–603.

Donoho, David L, "An invitation to reproducible computational research," *Biostatistics*, 2010, *11* (3), 385–388.

Donohue, John J and Steven D Levitt, "The Impact of Legalized Abortion on Crime," *Quarterly Journal of Economics*, 2001, *116* (2), 379–420.

__ **and** __ , "Further Evidence that Legalized Abortion Lowered Crime A Reply to Joyce," *Journal of Human Resources*, 2004, *39* (1), 29–49.

__ **and** __ , "Measurement Error, Legalized Abortion, and the Decline in Crime: A Response to Foote and Goetz," *Quarterly Journal of Economics*, 2008, *123* (1), 425–440.

Dority, Bree and Scott M Fuess, "Labor Market Institutions and Unemployment: Can Earlier Findings be Replicated?," *Quarterly Journal of Business & Economics*, 2007, *46* (4), 23–44.

Easley, Richard W and Charles S Madden, "Replications and extensions in marketing and management research," *Journal of Business Research*, 2000, *48* (1), 1–3.

__ , __ , **and Mark G Dunn**, "Conducting marketing science: The role of replication in the research process," *Journal of Business Research*, 2000, *48* (1), 83–92.

Easterly, William, Ross Levine, and David Roodman, "Aid, Policies, and Growth: Comment," *American Economic Review*, 2004, *94* (3), 774–780.

Égert, Balázs, "The 90% Public Debt Threshold: The Rise and Fall of a Stylised Fact," OECD Economics Department Working Paper 1055. Paris: OECD 2013.

Evanschitzky, Heiner, Carsten Baumgarth, Raymond Hubbard, and J Scott Armstrong, "Replication research's disturbing trend," *Journal of Business Research*, 2007, *60* (4), 411–415.

Farley, John U, Donald R Lehmann, and Michael J Ryan, "Generalizing from" Imperfect" Replication," *Journal of Business*, 1981, *54* (4), 597–610.

Feldstein, Martin S, "Social Security, Induced Retirement, and Aggregate Capital Accumulation," *Journal of Political Economy*, 1974, *82* (5), 905–926.

__ , "Social Security and Private Saving: Reply," *Journal of Political Economy*, 1982, *90* (3), 630–42.

__ , "Social Security and Saving: New Time Series Evidence," *National Tax Journal*, 1996, *49* (2), 151–64.

Finifter, Bernard M, "The generation of confidence: Evaluating research findings by random subsample replication," *Sociological Methodology*, 1972, *4*, 112–175.

Fleischmann, Martin and Stanley Pons, "Electrochemically induced nuclear fusion of deuterium," *Journal of Electroanalytical Chemistry and Interfacial Electrochemistry*, 1989, *261* (2), 301–308.

Foote, Christopher L and Christopher F Goetz, "The Impact of Legalized Abortion on Crime: Comment," *Quarterly Journal of Economics*, 2008, *123* (1), 407–423.

Fuess, Scott M, "On replication in business and economics research: The QJBE case," *Quarterly Journal of Business and Economics*, 1996, *35* (2), 3–13.

Furman, Jeffrey L, Kyle Jensen, and Fiona Murray, "Governing knowledge in the scientific community: Exploring the role of retractions in biomedicine," *Research Policy*, 2012, *41* (2), 276–290.

García, Fernando Martel, "Scientific Progress in the Absence of New Data: A Procedural Replication of Ross (2006)," Working Paper, New York University 2013.

Hamermesh, Daniel S, "Some thoughts on replications and reviews," *Labour Economics*, 1997, *4* (2), 107–109.

__ , "Viewpoint: Replication in economics," *Canadian Journal of Economics/Revue canadienne d'économique*, 2007, *40* (3), 715–733.

Harrison, Glenn W, "Mortgage lending in Boston: A reconsideration of the evidence," *Economic Inquiry*, 1998, *36* (1), 29–38.

Hendrick, Clyde, "Replications, strict replications, and conceptual replications: are they important?," *Journal of Social Behavior & Personality*, 1990, *5* (4), 41–49.

Herndon, Thomas, Michael Ash, and Robert Pollin, "Does high public debt consistently stifle economic growth? A critique of Reinhart and Rogoff," *Cambridge Journal of Economics*, 2014, *38* (2), 257–279.

Herrnson, Paul S, "Replication, verification, secondary analysis, and data collection in political science," *PS: Political Science & Politics*, 1995, *28* (03), 452–455.

Hicks, Joan Hamory, Michael Kremer, and Edward Miguel, "Estimating deworming school participation impacts and externalities in Kenya: A Comment on Aiken et al. (2014)," *Response* section of '3ie Replication Paper 3, Part 1', New Delhi: International Initiative for Impact Evaluation (3ie) 2014.

__ , __ , and __ , "Estimating deworming school participation impacts in Kenya: A Comment on Aiken et al. (2014b)," *Response* section of '3ie Replication Paper 3, Part 2', New Delhi: International Initiative for Impact Evaluation (3ie) 2014.

Houtenville, Andrew J and Richard V Burkhauser, "Did the Employment of People with Disabilities Decline in the 1990s, and was the ADA Responsible? A Replication and Robustness Check of Acemoğlu and Angrist (2001)," Ithaca, NY: Research and Rehabilitation Training Center, Cornell University 2004.

Hoxby, Caroline, "Does Competition Among Public Schools Benefit Students and Taxpayers? Reply," *American Economic Review*, 2007, *97* (5), 2038–2055.

Hoxby, Caroline M, "Does Competition among Public Schools Benefit Students and Taxpayers?," *American Economic Review*, 2000, *90* (5), 1209–1238.

Hubbard, Raymond and Daniel E Vetter, "An empirical comparison of published replication research in accounting, economics, finance, management, and marketing," *Journal of Business Research*, 1996, *35* (2), 153–164.

__ and J Scott Armstrong, "Replications and extensions in marketing: Rarely published but quite contrary," *International Journal of Research in Marketing*, 1994, *11* (3), 233–248.

Hunter, John E, "The desperate need for replications," *Journal of Consumer Research*, 2001, *28* (1), 149–158.

Ioannidis, John and Chris Doucouliagos, "What's To Know about the Credibility of Empirical Economics?," *Journal of Economic Surveys*, 2013, *27* (5), 997–1004.

__ and Thomas A Trikalinos, "Early extreme contradictory estimates may appear in published research: The Proteus phenomenon in molecular genetics research and randomized trials," *Journal of Clinical Epidemiology*, 2005, *58* (6), 543–549.

Iversen, Vegard and Richard Palmer-Jones, "TV, Female Empowerment and Demographic Change in Rural India," 3ie Replication Series Paper 2, New Delhi: International Initiative for Impact Evaluation (3ie) 2014.

__ , __ , and Kunal Sen, "On the colonial origins of agricultural development in India: A re-examination of Banerjee and Iyer, 'History, institutions and economic performance'," *Journal of Development Studies*, 2013, *49* (12), 1631–1646.

Jensen, Robert and Emily Oster, "The power of TV: Cable television and women's status in India," *Quarterly Journal of Economics*, 2009, *124* (3), 1057–1094.

__ and __ , "Corrigendum, 'The Power of TV', July 2012," Correction note, University of Chicago 2012.

__ and __ , "TV, Female Empowerment and Fertility Decline in Rural India: Response to Iversen and Palmer-Jones," *Response* section of '3ie Replication Paper 2', New Delhi: International Initiative for Impact Evaluation (3ie) 2014.

Johnson, Simon, William Larson, Chris Papageorgiou, and Arvind Subramanian, "Is newer better? Penn World Table revisions and their impact on growth estimates," *Journal of Monetary Economics*, 2013, *60* (2), 255–274.

Jones, Benjamin F and Benjamin A Olken, "Do Leaders Matter? National Leadership and Growth Since World War II," *Quarterly Journal of Economics*, 2005, *120* (3), 835–864.

Joyce, Ted, "Did legalized abortion lower crime?," *Journal of Human Resources*, 2004, *39* (1), 1–28.

__ , "Further Tests of Abortion and Crime: A Response to Donohue and Levitt (2001, 2004, 2006)," NBER Working Paper 12607. Cambridge, MA: National Bureau of Economic Research 2006.

__ , "A simple test of abortion and crime," *Review of Economics and Statistics*, 2009, *91* (1), 112–123.

Kane, Edward J, "Why journal editors should encourage the replication of applied econometric research," *Quar-*

terly Journal of Business and Economics, 1984, *23* (1), 3–8.

Khan, A Rahman, "Real wages of agricultural workers in Bangladesh," *Economic and Political Weekly*, 1984, *19* (4), PE40–PE48.

King, Gary, "Replication, replication," *PS: Political Science & Politics*, 1995, *28* (3), 444–452.

__ , "Publication, publication," *PS: Political Science & Politics*, 2006, *39* (01), 119–125.

Klasen, Stephan, "Some Recent Controversies on Levels and Trends in Gender Bias in Mortality," in Kaushik Basu and Ravi Kanbur, eds., *Arguments for a Better World: Essays in Honor of Amartya Sen, Volume 2: Society, Institutions, and Development*, Oxford: Oxford University Press, 2008, pp. 280–302.

Kniesner, Thomas J, "Replication? Yes. But how?," *Labour Economics*, 1997, *4* (2), 115–119.

Koenker, Roger and Achim Zeileis, "On reproducible econometric research," *Journal of Applied Econometrics*, 2009, *24* (5), 833–847.

Layard, Richard, Stephen Nickell, and Richard Jackman, *The Unemployment Crisis*, Oxford: Oxford University Press, 1994.

Leimer, Dean R and Selig D Lesnoy, "Social Security and Private Saving: New Time-Series Evidence," *Journal of Political Economy*, 1982, *90* (3), 606–29.

Leone, Robert P and Randall L Schultz, "A study of marketing generalizations," *Journal of Marketing*, 1980, *44* (1), 10–18.

Levitt, Steven D, "Using Electoral Cycles in Police Hiring to Estimate the Effect of Police on Crime," *American Economic Review*, 1997, *87* (3), 270–90.

__ , "Using Electoral Cycles in Police Hiring to Estimate the Effects of Police on Crime: Reply," *American Economic Review*, 2002, *92* (4), 1244–1250.

Lewis, NS, CA Barnes, MJ Heben, A Kumar, SR Lunt, GE McManis, GM Miskelly, RM Penner, MJ Sailor, PG Santangelo et al., "Searches for low-temperature nuclear fusion of deuterium in palladium," *Nature*, 1989, *340* (6234), 525–530.

Lin, Ming-Jen and Ming-Ching Luoh, "Can Hepatitis B Mothers Account for the Number of Missing Women? Evidence from Three Million Newborns in Taiwan," *American Economic Review*, 2008, *98* (5), 2259–2273.

Lindsay, R Murray and Andrew SC Ehrenberg, "The design of replicated studies," *The American Statistician*, 1993, *47* (3), 217–228.

Lott, John R and John Whitley, "Abortion and Crime: Unwanted Children and Out-of-Wedlock Births," *Economic Inquiry*, 2007, *45* (2), 304–324.

Lykken, David T, "Statistical significance in psychological research.," *Psychological Bulletin*, 1968, *70* (3), 151–159.

Mack, Y P and Nancy J Wulwick, "Nonparametric Regression Analysis of Some Economic Data," in George Roussas, ed., *Nonparametric Functional Estimation and Related Topics*, Vol. 335 of *NATO ASI Series*, Springer Netherlands, 1991, pp. 361–374.

Madden, Charles S, Richard W Easley, and Mark G Dunn, "How journal editors view replication research," *Journal of Advertising*, 1995, *24* (4), 77–87.

Mankiw, N Gregory, David Romer, and David N Weil, "A Contribution to the Empirics of Economic Growth," *Quarterly Journal of Economics*, 1992, *107* (2), 407–437.

Mansfield, Edward D, Helen V Milner, and B Peter Rosendorff, "Free to trade: Democracies, autocracies, and international trade," *American Political Science Review*, 2000, *94* (2), 305–321.

__ , __ , **and** __ , "Replication, realism, and robustness: Analyzing political regimes and international trade," *American Political Science Review*, 2002, *96* (1), 167–169.

McCrary, Justin, "Using Electoral Cycles in Police Hiring to Estimate the Effect of Police on Crime: Comment," *American Economic Review*, 2002, *92* (4), 1236–1243.

McCullough, Bruce D, "The accuracy of econometric software," in David A. Belsley and Erricos John Kontoghiorghes, eds., *Handbook of Computational Econometrics*, West Sussex, UK: John Wiley & Sons, 2009, chapter 2, pp. 55–80.

__ , "Open access economics journals and the market for reproducible economic research," *Economic Analysis and Policy*, 2009, *39* (1), 117–126.

__ **, Kerry Anne McGeary, and Teresa D Harrison**, "Do economics journal archives promote replicable research?," *Canadian Journal of Economics/Revue canadienne d'économique*, 2008, *41* (4), 1406–1420.

Miguel, Edward and Michael Kremer, "Worms: identifying impacts on education and health in the presence of treatment externalities," *Econometrica*, 2004, *72* (1), 159–217.

__ **and Shanker Satyanath**, "Re-examining economic shocks and civil conflict," *American Economic Journal: Applied Economics*, 2011, *3* (4), 228–232.

__ , __ **, and Ernest Sergenti**, "Economic shocks and civil conflict: An instrumental variables approach," *Journal of Political Economy*, 2004, *112* (4), 725–753.

Mittelstaedt, Robert A and Thomas S Zorn, "Econometric replication: Lessons from the experimental sciences," *Quarterly Journal of Business and Economics*, 1984, *23* (1), 9–15.

Morgenstern, Oskar, "The Accuracy of Economic Observations," in Tjalling C Koopmans, Armen A Alchian, George B Dantzig, Nicholas Georgescu-Roegen, Paul A Samuelson, and Albert W Tucker, eds., *Activity Snalysis of Production and Allocation: Proceedings of a Conference*, New York, NY: Wiley & Sons, 1951, pp. 282–284.

Munnell, Alicia H, Geoffrey MB Tootell, Lynn E Browne, and James McEneaney, "Mortgage lending in Boston: Interpreting HMDA data," *American Economic Review*, 1996, *86* (1), 25–53.

Oster, Emily, "Hepatitis B and the Case of the Missing Women," *Journal of Political Economy*, 2005, *113* (6), 1163–1216.

__ **, Gang Chen, Xinsen Yu, and Wenyao Lin**, "Hepatitis B does not explain male-biased sex ratios in China," *Economics Letters*, 2010, *107* (2), 142–144.

Ottenbacher, Kenneth J, "The power of replications and replications of power," *The American Statistician*, 1996, *50* (3), 271–275.

Özer Ballı, Hatice and Bent E Sørensen, "Interaction effects in econometrics," CEPR Discussion Paper DP7929. London: Centre for Economic Policy Research 2010.

__ **and** __ , "Interaction effects in econometrics," *Empirical Economics*, 2013, *45* (1), 583–603.

Peng, Roger D, "Reproducible research in computational science," *Science*, 2011, *334* (6060), 1226–1227.

Pesaran, Hashem, "Introducing a replication section," *Journal of Applied Econometrics*, 2003, *18* (1), 111–111.

Phillips, Alban W, "The Relation Between Unemployment and the Rate of Change of Money Wage Rates in the United Kingdom, 1861–1957," *Economica*, 1958, *25* (100), 283–299.

Pronyk, Paul, "Errors in a paper on the Millennium Villages project," *Lancet*, 2012, *379* (9830), 1946.

Pronyk, Paul M, Maria Muniz, Ben Nemser, Marie-Andrée Somers, Lucy McClellan, Cheryl A Palm, Uyen Kim Huynh, Yanis Ben Amor, Belay Begashaw, John W McArthur, Jeffrey D Sachs et al., "The effect of an integrated multisector model for achieving the Millennium Development Goals and improving child survival in rural sub-Saharan Africa: a non-randomised controlled assessment," *The Lancet*, 2012, *379* (9832), 2179–2188.

Rajan, Raghuram G and Luigi Zingales, "Financial Dependence and Growth," *American Economic Review*, 1998, *88* (3), 559–86.

Reinhart, Carmen M and Kenneth S Rogoff, "Growth in a Time of Debt," *American Economic Review Papers & Proceedings*, 2010, *100* (2), 573–8.

__ **and** __ , "Errata: 'Growth in a Time of Debt'," Working Paper. Harvard University Dept. of Economics 2013.

Ross, Michael, "Is democracy good for the poor?," *American Journal of Political Science*, 2006, *50* (4), 860–874.

Rothstein, Jesse M, "Does Competition Among Public Schools Benefit Students and Taxpayers? Comment," *American Economic Review*, 2007, *97* (5), 2026–2037.

__ , "Rejoinder to Hoxby," Working Paper, Princeton University 2007.

Sachs, Jeffrey D and Andrew M Warner, "Natural Resource Abundance and Economic Growth," Working Paper, Center for International Development. Cambridge, MA: Harvard Kennedy School 1997.

Schmidt, Stefan, "Shall we really do it again? The powerful concept of replication is neglected in the social sciences.," *Review of General Psychology*, 2009, *13* (2), 90.

Sleeman, AG, "Retrospectives: The Phillips Curve: A Rushed Job?," *Journal of Economic Perspectives*, 2011, *25* (1), 223–38.

Sorte, Michael A La, "Replication as a verification technique in survey research: A paradigm," *The Sociological Quarterly*, 1972, *13* (2), 218–227.

Stodden, Victoria C, "Reproducible research: Addressing the need for data and code sharing in computational science (Yale Law School Roundtable on Data and Code Sharing)," *Computing in Science & Engineering*, 2010, *12* (5), 8–12.

Summers, Lawrence H, "The Scientific Illusion in Empirical Macroeconomics," *Scandinavian Journal of Economics*, 1991, *93* (2), 129–48.

Thompson, Bruce, "The pivotal role of replication in psychological research: Empirically evaluating the replicability of sample results," *Journal of Personality*, 1994, *62* (2), 157–176.

Trikalinos, Nikolaos A, Evangelos Evangelou, and John PA Ioannidis, "Falsified papers in high-impact journals were slow to retract and indistinguishable from nonfraudulent papers," *Journal of clinical epidemiology*, 2008, *61* (5), 464–470.

Tsang, Eric W K and Kai-Man Kwan, "Replication and theory development in organizational science: A critical realist perspective," *Academy of Management review*, 1999, *24* (4), 759–780.

Vinod, Hrishikesh D, "Stress testing of econometric results using archived code for replication," *Journal of Economic and Social Measurement*, 2009, *34* (2), 205–217.

Wulwick, Nancy J, "Two econometric replications: The historic Phillips and Lipsey-Phillips curves," *History of Political Economy*, 1996, *28* (3), 391–439.