The Meaning of Failed Replications: A Review and Proposal

Michael Clemens

Abstract

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 should not be described as such. Adopting this definition can improve incentives for researchers, encouraging more and better replication tests.

JEL Codes: B40, C18, C80.
The Meaning of Failed Replications: A Review and Proposal

Michael Clemens
Center for Global Development
IZA

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.

CGD is grateful for contributions from its funders, including Good Ventures and the William and Flora Hewlett Foundation, in support of this work.


The Center for Global Development is an independent, nonprofit policy research organization dedicated to reducing global poverty and inequality and to making globalization work for the poor. Use and dissemination of this Working Paper is encouraged; however, reproduced copies may not be used for commercial purposes. Further usage is permitted under the terms of the Creative Commons License.

The views expressed in CGD Working Papers are those of the authors and should not be attributed to the board of directors or funders of the Center for Global Development.
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. 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.

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

<table>
<thead>
<tr>
<th></th>
<th>Sampling distribution for parameter estimates</th>
<th>Sufficient conditions for discrepancy</th>
<th>Types</th>
<th>Methods in follow-up study versus methods reported in original:</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Replication</strong></td>
<td>Same</td>
<td>Random chance, error, or fraud</td>
<td>Verification</td>
<td>Same specification</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Reproduction</td>
<td>Same population</td>
<td>Yes</td>
</tr>
<tr>
<td><strong>Robustness</strong></td>
<td>Different</td>
<td>Sampling distribution has changed</td>
<td>Reanalysis</td>
<td>Same sample</td>
<td>No</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Extension</td>
<td></td>
<td>Yes</td>
</tr>
</tbody>
</table>

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).
ing 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. 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).


---

1Authors 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.
visit 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, “Reproductions 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 include 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”.

7
Numerous researchers also work with a different definition than that proposed in Table 1. For Hamermesh (1997), 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 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
Table 2: Correspondence between Table 1 Definitions and Previous Definitions

<table>
<thead>
<tr>
<th>Source</th>
<th>Replication test</th>
<th>Robustness test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mittelstaedt and Zorn 1984</td>
<td>Type I replication, Econometric audit, ‘Reproduction’ replication</td>
<td>Type II, III, IV replication, Improvisational replication</td>
</tr>
<tr>
<td>Fuess 1996</td>
<td>‘Narrow sense’ replication, Pure replication</td>
<td>‘Wide sense’ replication, Statistical/Scientific replication, Stress test</td>
</tr>
<tr>
<td>Kniesner 1997</td>
<td>Replication, Replication/reproduction, Repeatability/Strict replication</td>
<td>—</td>
</tr>
<tr>
<td>Arulampalam et al. 1997</td>
<td>—</td>
<td>Conceptual replication, Replication</td>
</tr>
<tr>
<td>Pesaran 2003</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Vinod 2009</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Koenker and Zeileis 2009</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Ioannidis and Doucouliagos 2013</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Data policy of AER, JPE, etc.</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Lindsay and Ehrenberg 1993</td>
<td>Differentiated replication, Computational reproduction</td>
<td>Extension</td>
</tr>
<tr>
<td>Donoho 2010</td>
<td>—</td>
<td>‘Replication’ reanalysis, Extension, improvement</td>
</tr>
<tr>
<td>Stodd 2010</td>
<td>—</td>
<td>Broad replication</td>
</tr>
<tr>
<td>Peng 2011</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>King 1995</td>
<td>Extension</td>
<td>Extension</td>
</tr>
<tr>
<td>Herrnson 1995</td>
<td>‘Verification’ reanalysis, Replication</td>
<td>Extension</td>
</tr>
<tr>
<td>King 2006</td>
<td>Replication</td>
<td>Broad replication</td>
</tr>
<tr>
<td>Dafoe 2014</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>La Sorte 1972</td>
<td>Independent/ theoretical replication, Identical replication, Replication type a</td>
<td>——</td>
</tr>
<tr>
<td>Collins 1991</td>
<td>Repetition/checking</td>
<td>Replication, robustness, Virtual/systematic replication</td>
</tr>
<tr>
<td>Hendrick 1990</td>
<td>—</td>
<td>Replication, type b…p</td>
</tr>
<tr>
<td>Thompson 1994</td>
<td>Internal replication, Direct replication</td>
<td>Replication</td>
</tr>
<tr>
<td>Schmidt 2009</td>
<td>Conceptual replication, Exact replication</td>
<td>Reproduction, internal replication</td>
</tr>
<tr>
<td>Lykken 1968</td>
<td>Constructive replication, Literal/operational replication, Replication, checking</td>
<td>Conceptual replication</td>
</tr>
<tr>
<td>Cronbach 1975</td>
<td>Quasi-replication, Operational replication, Exact replication</td>
<td>Operational replication, Conceptual replication, Checking, Replication, etc.</td>
</tr>
<tr>
<td>Hendrick 1990</td>
<td>Partial/conceptual replication, Significant sameness</td>
<td>Operation, reanalysis, extension, etc.</td>
</tr>
<tr>
<td>Thompson 1994</td>
<td>Extended replication</td>
<td>Partial/conceptual replication, Extension</td>
</tr>
<tr>
<td>Schmidt 2009</td>
<td>Conceptual replication, Exact replication</td>
<td>Replication, reanalysis, extension, etc.</td>
</tr>
<tr>
<td>Leone and Schultz 1980</td>
<td>Non-experimental/ corroboration replication, Perfect replication, Representative replication, Strict replication, Duplication, Checking, Strict replication, Replication, Type II, I replication, Statistical replication, Replication</td>
<td>Replication, reanalysis, extension, etc.</td>
</tr>
<tr>
<td>Farley et al. 1981</td>
<td>Imperfect replication, Extension, Significant sameness</td>
<td>Partial/conceptual replication, Extension</td>
</tr>
<tr>
<td>Hubbard and Armstrong 1994</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Barwise 1995</td>
<td>Operational replication, Significant sameness, Operational replication, Replication</td>
<td>—</td>
</tr>
<tr>
<td>Madden et al. 1995</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Tsang and Kwan 1999</td>
<td>Conceptual replication</td>
<td>—</td>
</tr>
<tr>
<td>Darley 2000</td>
<td>Conceptual replication, Extension</td>
<td>—</td>
</tr>
<tr>
<td>Easley et al. 2000</td>
<td>Type II, III replication, Scientific replication, Replication with extension</td>
<td>—</td>
</tr>
<tr>
<td>Hunter 2001</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Evanschitzky et al. 2007</td>
<td>—</td>
<td>—</td>
</tr>
</tbody>
</table>

*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.
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.”

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-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”.

2Worse, ‘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.”
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 origi-
nal, 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 literature. 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 documenting 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 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”
**Table 3: An Application of Proposed Table 1 Definitions to Selected Literature**

<table>
<thead>
<tr>
<th>Replication tests</th>
<th>Original papers and replies</th>
<th>Note</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baker (2013)*</td>
<td>Feldstein (1996)</td>
<td>✓ Gov. dataset may have changed with revisions</td>
</tr>
<tr>
<td>Herndon et al. (2014)</td>
<td>Reinhart and Rogoff (2010, 2013)</td>
<td>✓ Coding errors (also contains reanalysis)</td>
</tr>
<tr>
<td>Aiken et al. (2014a)</td>
<td>Miguel and Kremer (2004), Hicks et al. (2014a)</td>
<td>⇝ Coding error; relevance to findings disputed</td>
</tr>
</tbody>
</table>

| Robustness tests | |
|------------------|-----------------------------|------|
| Johnson et al. (2013) | Mankiw et al. (1992), Jones and Olken (2005), etc. | ● New, updated dataset |
| Iversen et al. (2013) | Banerjee and Iyer (2005) | ◊ Different dataset by recoding key regressor |
| Aiken et al. (2014b) | Miguel and Kremer (2004), Hicks et al. (2014b) | ◊ Different subsets of data, new specifications |

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.
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


Chambers, Chris, “You can’t replicate a concept,” Neurochambers weblog; School of Psychology, Cardiff University; March 26 2012.


Cronbach, Lee J, “Beyond the two disciplines of scientific psychology,” American Psychologist, 1975, 30 (2), 116–


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.

