Economics
The Open-Access, Open-Assessment E-Journal

Discussion Paper
No. 2017-79 | September 28, 2017 | http://www.economics-ejournal.org/economics/discussionpapers/2017-79

Should you choose to do so… A replication paradigm

Richard G. Anderson

Abstract
This note introduces the concept of the *replication paradigm*, a framework that can (and should) be followed in every replication attempt. The paradigm expands, in part, on Bruce McCullough's well-known paraphrase of Berkeley computer scientist Jon Claerbout's insight – "An applied economics article is only the advertising for the data and code that produced the results" – and on the view that the primary social and scientific value of replication is to measure the scientific contribution of the inferences in an empirical study. The paradigm has four steps. First, in the "candidate study," identify and state clearly the hypotheses advanced by the study's authors. Second, provide a clear statement of the authors' econometric methods. Third, discuss the data. Fourth, discuss the authors' statistical inference. The author´s purpose in this ordering is to reverse the too-frequent focus in the replication literature on "data." The correct data, of course, are critical to the replication. But "replication" *as a scientific endeavor* will never achieve respectability unless and until it abandons a narrow focus on data and expands its focus to the underlying scientific inferences.

(Published in Special Issue The practice of replication)

JEL B41
Keywords Replication; paradigm

Authors
Richard G. Anderson, ✉ Robert W. Plaster School of Business and Entrepreneurship, Lindenwood University, St Charles, Missouri, USA, rganderson@alum.mit.edu

Citation Richard G. Anderson (2017). Should you choose to do so… A replication paradigm. Economics Discussion Papers, No 2017-79, Kiel Institute for the World Economy. http://www.economics-ejournal.org/economics/discussionpapers/2017-79

Received August 25, 2017 **Accepted as Economics Discussion Paper** September 25, 2017
Published September 28, 2017

1. The Scientific Role of Replication[1]

A number of authors have discussed why a researcher might wish to replicate the work of others. Fame and fortune seldom are the motivation—even those who have uncovered meaningful errors by prominent economists seldom gain lasting fame. The primary social and scientific value of replication is to measure the scientific contribution of the inferences in a published empirical article. That is, I take a replicator's primary motivation as a desire to build accurately on previous work, not to embarrass or frustrate another researcher.

The inherent conflicts in replication as a scholarly endeavor are discussed by Dewald et al (1986) and Anderson et al (2008).[2] First, replications, if shared with others (and how many researchers tackle a replication intending to keep it secret?), are public goods:

> all who read the journal benefit from the knowledge that the research reported in its articles has been more care-fully monitored by the researcher; the quan-tity of benefits available to any single reader is not reduced by others reading the journal; and it is difficult to induce the reader to reveal his or her true value (price) for better quality articles. A single researcher faces high costs in time and money from undertaking replication of a study and finds no ready marketplace which correctly prices the social and individual value of the good.

A specific study might be selected for replication because of its closeness to the hypothesis being tested, or perhaps because it is the most recent study available. In any case, I assume that the researcher has selected a single "candidate" empirical study, to be replicated, on top of which she will pursue new work.

It is important to remember that "replication" in economics, except for experimental economics, is a *deterministic* process: "Beginning with an author's dataset and applying the mathematical operations specified by the author, a researcher should obtain the author's numerical results." (Anderson and Kichkha, 2017, p.56). Kane (1984) labeled this "econometric auditing" and Hamermesh (2007) labeled it "pure replication." Unfortunately, analysts continue to confound and confuse the concept of "replication" as used in the physical sciences, where it refers to repeating an experiment, and the concept in economics, where it is solely a matter of calculation. [3]

The first temptation to be avoided is being distracted by the mechanics of the candidate study, that is, racing to locate the data, write/run the estimation, and compare the empirical results. Rather, begin by "reverse engineering" the paper, emphasizing the hypotheses being tested.

[1] This section builds on Anderson and Dewald (1994), which has a similar title.
[2] In a recent paper, these same arguments have been re-discovered by Galiani, Gertler and Romero (2017).
[3] The same argument is made by Anderson and Dewald (1994). Christensen and Miguel (2017, forthcoming) repeat the assertion of Ioannidis (2005) that "most published research findings are false." Unfortunately, they intertwine the concept of "replication" in *experimental* sciences (e.g., medicine) with the concept in non-experimental sciences (e.g., non-experimental empirical economics). See also, for example, Duvendack, Palmer-Jones and Reed (2017) and Hamermesh (2017). An additional aspect is "true" vs "false" in terms of rejection of a null hypothesis when issues of size-adjusted power are ignored, that is, when *p* values alone are used to judge the strength of assertions. For an excellent discussion, see Harvey (2017).

Doing so guards against misinterpreting the empirical tests due to imprecise descriptions by the candidate study's author(s).

The second temptation to be avoided is ending the replication when the new empirical results do not match those in the candidate study. If a replication attempt is being conducted as a type of "econometric audit" (Kane, 1984) then failing to obtain the author(s) results might be an adequate ending. But, if the replication is being conducted as an early step in a scientific quest to understand the extant literature before moving forward, the researcher should explore, to every possible extent, the set of potential causes for the failure to reproduce. Unfortunately, the set of such causes might be large, requiring both persistence and creativity. Here is the true payoff (and pain) from replication: *Given a hypothesis to be tested, what is the universe of possible empirical results that might be obtained when the search space is constrained by the experimental design of the candidate study?*

Building the "replication space" for a candidate article is, itself, a formidable task. Dewald et al (1986), p. 589, argue that a replicator might anticipate a substantive difference between candidate studies where the journal has required submission of data and programs and where it has not:

> Our findings suggest that the existence of a requirement that authors submit to the journal their programs and data along with each manuscript would significantly reduce the frequency and magnitude of errors. We found that the very process of authors compiling their programs and data for submission reveals to them ambiguities, errors, and oversights which otherwise would be un-detected.

2. The Concept of the "Replication Paradigm"

The "replication space" for a candidate study includes three elements: the hypotheses examined, the statistical/econometric methods used, and the data. The "replication paradigm" is a flexible tool that formalizes the concept of a replication space. A researcher undertaking a replication has accepted a high moral throne: the researcher, implicitly or explicitly, has set himself up as a judge of the quality of the original study. The replicator must be as certain as possible that he/she understands fully the analysis in the original paper before declaring that the work cannot be replicated (thereby seeking to avoid the widespread bias/belief that most published empirical studies in economics contain errors). Hence, "reverse engineering" the original study, while tedious, is essential.

(1) Collecting the hypotheses examined is a straightforward, if sometimes tedious, endeavor, requiring sifting through the study's "motivation" and other boilerplate materials plus careful examination of the econometrics. When the specific hypotheses and empirical tests cannot be identified, an article is of uncertain scientific value.

(2) Econometric estimation is a second hurdle. Today, menu-driven econometrics programs that report large batteries of test statistics ease publishing articles with a formidable appearance even while possessing only a passing familiarity with the underlying econometric methods. But the

replicator has a higher burden: both to produce the reported point estimates and, often, to explore/explain why the replicated results do not match the published results. So doing requires exploring the edges and frailties of estimators and tests, when they work and when they fail, when they are robust and when they are not. Worse for the replicator, it may be necessary to pursue such tests in more than one computer package, especially if the authors do not report the package used.

(3) Finally, data collection is troublesome. In general, a replicator without access to the author(s)'s data should collect all vintages within the scope of the research. Fortunately, interest rates (generally) do not revise. For other variables, the replicator faces a higher hurdle than the original author, who might have successfully published the article after collecting but one set of data. It is always wise to graph the data. If the original article contains charts, it should go without further saying that either the charts must match or the replicator must explore the cause of the differences.

3. A Classic Study

Dewald, Thursby and Anderson (1986), section V, describe the replication of a published study of the US banking industry. Here, I revisit that replication within the umbrella of the replication paradigm suggested above.

The specific study we examined is Lawrence Goldberg and Anthony Saunders, "The Growth of Organizational Forms of Foreign Banks in the U.S.," *Journal of Money, Credit and Banking*, August 1981, hereafter referred to as GS (1981). The choice was based, in part, on tractability: the article was relatively brief with a straightforward three-equation linear regression model of the growth of agencies, branches and subsidiaries of foreign banks in the United States, and the authors responded promptly to our queries.

(1) Hypothesis

- Foreign banks operate in the U.S. through three organizational forms: branches, agencies, and subsidiaries. It is asserted that the growth of all is affected by four variables (current profits, future domestic business, future international business, and regulatory change) and that the effects perhaps differ among the three organizational types.

(2) Estimation

- One equation was estimated for each organizational form by "generalized least squares," although the specific estimator is not specified. The equations are independent, with no cross-equation restrictions (that is, no formal tests are reported for the equality of coefficients across equations). A "$\hat{\rho}$ " is printed for two equations, without any description or discussion.

(3) Data

- Banking data, quarterly, for 1972 Q4 to 1980 Q1. The specific source/publication is not stated in the article; the data likely are from the G.11 release, which was published

monthly November 1972 to July 1980, containing figures for the final business day of the month. It appears that the data are no longer available from the source.[4] As with most quantity data, revisions might have occurred. We did not check for revisions in our replication.

- Macroeconomic data on "domestic investment," "imports," and GNP, according to the authors, were obtained from the *Survey of Current Business*. The published article provides neither the complete names of the series (e.g., real gross private domestic investment) nor the dates/issues from which the data were collected.

(4) Inference

- Inference is based on the signs and t-statistics of the estimated coefficients in this regression, estimated separately for each business form

$$a_{i,t} = \beta_{i,0} + \beta_{i,1} S_{t-1} + \beta_{i,2} I_t + \beta_{i,3} \left(M/Y \right)_t + \beta_{i,4} D_t + u_{i,t} \ ,$$

where a_{it} is total U.S. assets (i = agency, branch, or subsidiary), S_t is an interest rate spread, I_t is gross private domestic investment[5], M_t is imports[6], Y_t is GNP,[7] and D_t is a dummy reflecting passage of the International Banking Act. There are no specific null or alternative hypotheses; rather, the authors discuss the signs and t-statistics of the estimated coefficients. They write in summary (p. 372): "This paper has sought to

[4] A researcher seeking to replicate the Goldberg and Sanders analysis as a basis for further research must be cautious. Currently, the Board publishes a release entitled "Assets and Liabilities of U.S. Branches and Agencies of Foreign Banks." A footnote on that release suggests that Goldberg and Saunders' data is from the discontinued G.11 release:

> "Data [shown] are aggregates of categories reported on the quarterly form FFIEC 002, `Report of Assets and Liabilities of U.S. Branches and Agencies of Foreign Banks.` The form was first used for reporting data as of June 30, 1980, and was revised as of December 31, 1985. From November 1972 through May 1980, U.S. branches and agencies of foreign banks had filed a monthly FR 886a report. Aggregate data from that report were available through the Federal Reserve monthly statistical release G.11, last issued on July 10, 1980. Data in this table and in the G.11 tables are not strictly comparable because of differences in reporting panels and in definitions of balance sheet items."

The Federal Reserve Board's web site no longer provides data prior to 1980 (worse, and perhaps more disturbing, is that even *currently published* foreign banking data are not available for download in machine-readable form, only as HTML files). Further, the Federal Reserve Bank of St. Louis FRASER archive (which roughly seeks to hold all historical Federal Reserve data releases) omits issues for all of 1972-1975, provides the 1976 issues for only February and December, and provides all issues May 1977 through July 1980. The final issue displays figures for the last business day of May 1980. In brief: A wise researcher likely should avoid seeking to replicate GS (2008) as a base for future research. We did not explore the extent to which data from the Board's Z.1 release, *Financial Accounts of the United States* (formerly, the Flow of Funds Accounts) might be an adequate substitute for the lost G.11 data..

[5] The article refers to this variable as "the current rate of domestic investment." I assume the complete variable name, if provided, would be "nominal gross private domestic investment, seasonally adjusted annual rate, billions of current dollars."

[6] The article refers to this variable as "U.S. imports." I assume the complete variable name would be "nominal imports, seasonally adjusted annual rate, billions of current dollars."

[7] The article refers to this variable as "U.S. GNP." I assume the complete variable name would be "nominal GNP, seasonally adjusted annual rate, billions of current dollars."

determine the variables impacting on the growth of different types of foreign banks in the U.S. ... The empirical results suggested that these factors tended to impact on agencies in a different manner from branches and subsidiaries."

- The replicator then has three dimensions to discuss: (1) Is the pattern of signs the same in the published and replicated regressions? (2) Is the pattern of t-statistics the same among coefficients in the published and replicated regressions? (3) Are the *numerical values* of parameters (coefficients and t-statistics) the same in the published and replicated regressions?

Responding to our queries, GS furnished their banking data but were unable to furnish the macroeconomic data, asserting that it was unimportant to have retained that data since any interested researcher could readily obtain the data. GS stated that they collected their macroeconomic data manually from the *Survey of Current Business*; it seems likely that this was during the summer of 1980.[8] But how, exactly, were the data collected? Facing a stack of printed *Survey* issues, a person might perhaps begin with the first-published issue and end with the last-published (collecting the "originally-published" observations) or, alternatively, work backward, beginning with the last-published issue and ending with the first-published (collecting the "most-recently published" observations). [9] In our experiment, we collected *all* figures for imports, investment and GNP that were published in the 118 monthly issues of the *Survey* published during 1972 Q4 to 1982 Q3.

In our replication, we re-estimated GS's model 500 times, each time using an independent draw from the *Survey* dataset. In addition, we estimated the model using the "most-recently published" data and the "originally published" data. All models were estimated using both Cochrane-Orcutt and Prais-Winsten estimators to adjust for putative AR(1) error processes.[10]

[8] Seeking to bracket the period during which they collected the data, we note that the last observation of the banking data used by GS was published by the Federal Reserve Board on May 12, 1980, and their article was published in the August 1981 issue of the *JMCB*.

.

[9] We do not necessarily assert that GS collected data in this manner; indeed, they could not recall precisely when or how they had collected their data from the *Survey*. Younger readers of this article might wince at a discussion of data collection. When both GS's study and many classic articles were written, electronic databases generally were unavailable to academic researchers. One of the first machine-readable databases was Citibase, distributed by Citibank Database Services. Although I have been unable to locate a definitive date for the first release of this database, OCLC's WorldCat shows 1982 as the cataloging date for the print book "CITIBASE: Citibank economic database." WorldCat, in a related entry, also, under "Details," includes: "Received on tape reels: 9 tracks, EBCDIC, 1600 bpi, 80 characters per record, block size 8,640; System requirements: IBM 4361." The IBM 4300 series were System 370-compatible minicomputers sold from 1979 to 1992; the 4361 was introduced September 1983 (Computerworld, 1983). It is puzzling why the Model 4361 was required since earlier models also would have handled such 9-track IBM-format tapes which were introduced in 1964 with the IBM System 360. Later, under the guidance of research director Robert Rasche, concerns with regard to data vintage led to the creation of ALFRED at the Federal Reserve Bank of St. Louis, a system in which every data point for every variable in the FRED database is tagged with a "vintage" date. For background, see Anderson (2006).

[10] GS write that the regressions are adjusted for first-order autocorrelation but do not specify the estimator used. On Cochrane-Orcutt and Prais-Winsten estimators, see Judge et al. (1985), chapter 8.

Our results for the coefficient on investment, $\hat{\beta}_2$ and $t_{\hat{\beta}_2} = \left(\hat{\beta}_2 / \hat{\sigma}_{\hat{\beta}_2}\right)$, are summarized in Figures 1-12 of Dewald et al. (1986); the figures are reproduced below.[11] Estimates using the most-recently published data are labeled "A" and those using originally-published data are labeled "B." The parameters in the published GS article are labeled "C".

Although none of our datasets produced estimated parameters matching exactly those reported by GS, those based on originally-published data (with the Cochrane-Orcutt estimator) were quite close to GS's published results—in other words, conditional on using the first-published data and the Cochrane-Orcutt estimator, we "approximately" replicated the GS results.[12]

Speaking generally, although our estimates always differed numerically from published parameter values, the estimates using *first-published Survey* data and Cochrane-Orcutt were quite close to GS's published results while the estimates using Prais-Winsten were not.

We concluded our replication by emphasizing the role of the vintage of the data (the estimates labeled "A" are quite far from those labeled "C"):

> Our experiment demonstrates that Goldberg and Saunders' results are not easily replicated using data from published sources. A researcher using either the most-recently published data or a mixture of vintages of data from the Survey of Current Business would be unlikely to reproduce the Gold-berg-Saunders findings and, in turn, may be misled regarding the value of Goldberg and Saunders' results as a foundation for future research."

In summary:

- When "replication" (reproduction) of the results in a prior study is successful and the study's inferences confirmed, be pleased and move on with your new follow-on research.
- When the replication is unsuccessful, pursue likely avenues to illuminate *why* the replication failed. This is the essential importance of replication: How does a scientist judge the contribution to knowledge made by an empirical study?

[11] So far as I am aware, unfortunately, the underlying data, programs and regression results from our experiment have been lost.

[12] Because the large-sample properties of the one-step and iterated estimators are the same, it is difficult to prefer one over the other.

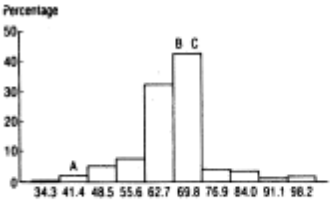


FIGURE 1. AGENCIES EQUATION

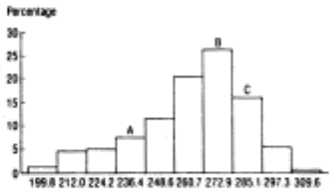


FIGURE 3. BRANCHES EQUATION

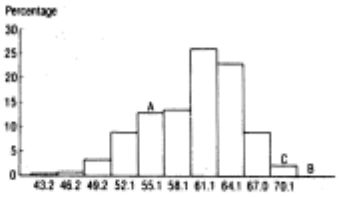


FIGURE 5. SUBSIDIARIES EQUATION

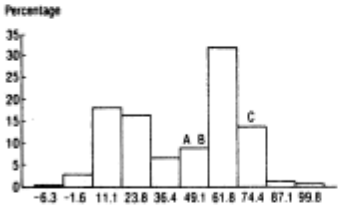


FIGURE 7. AGENCIES EQUATION

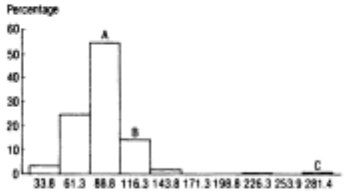


FIGURE 9. BRANCHES EQUATION

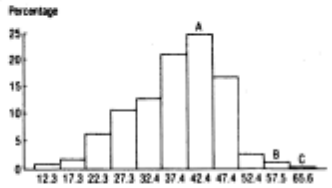


FIGURE 11. SUBSIDIARIES EQUATION

Note: Frequency Distribution of Investment Coefficient (Coefficient value, Midpoint of interval). Figure 1: Ordinary least squares estimator; Figures 3 and 5: Single-iteration Cochrane-Orcutt estimator; Figures 7, 9, and 11: Iterative Prais-Winsten estimator. A = estimate with most recently published data; B = estimate with originally published data; C = estimate in published article.

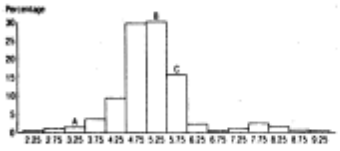


FIGURE 2. AGENCIES EQUATION INVESTMENT COEFFICIENT

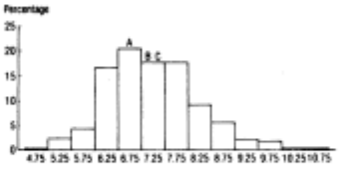


FIGURE 4. BRANCHES EQUATION INVESTMENT COEFFICIENT

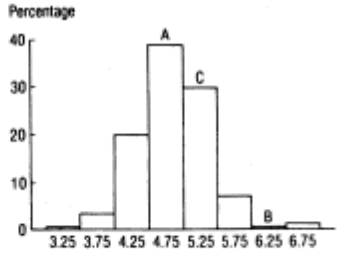


FIGURE 6. AGENCIES EQUATION INVESTMENT COEFFICIENT

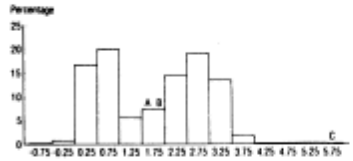


FIGURE 8. AGENCIES EQUATION INVESTMENT COEFFICIENT

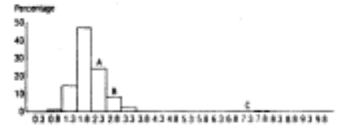


FIGURE 10. BRANCHES EQUATION INVESTMENT COEFFICIENT

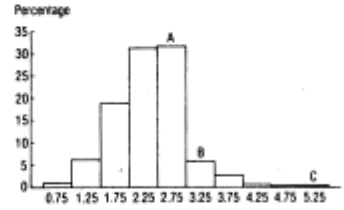


FIGURE 12. SUBSIDIARIES EQUATION INVESTMENT COEFFICIENT

Note: Frequency Distribution of t-Ratio (Value of t-ratio, midpoint of interval). Figure 2: Ordinary least squares estimator; Figures 4 and 6: Single-iteration Cochrane-Orcutt estimator; Figures 8, 10, and 12: Iterative Prais-Winsten estimator. A, B, C are as defined above.

4. A Recent Study

In this section, I apply the replication paradigm to a more recently published paper. Results of the replication are omitted.

The study is James Payne and George Waters, "Interest Rate Pass-Through and Asymmetric Adjustment: Evidence from the Federal Funds Rate Operating Target." *Applied Economics*, 2008, hereafter referred to as PW (2008). The theme of the study is an important one: that American financial markets have both market-determined and "administered" interest rates, the latter being distinguished by having an organization or firm that seeks to limit their day-to-day movement and/or to sustain the rates near a target/desired level. The United States government, for example, participates widely in financial markets, its actions affecting a number of interest rates, including the federal funds rate, FHA/VA mortgage rates, student loan rates, farm commodity program (CCC) loan rates, and small business SBA 7(a) interest rates. The private sector also, to the extent permitted by anti-trust laws, administers interest rates, including the bank prime rate and LIBOR (although the latter has been largely nationalized following the rate-fixing scandal). Relationships among various market-determined and administered interest rates are an important empirical topic.

The replication paradigm suggests that we begin by "reverse engineering" the empirical mechanics of the study, in four steps: (i) the null and alternative hypotheses, (ii) the econometric methods, (iii) the data, and (iv) the statistical inferences. These four steps should be written as "neutral" (or scientific) as possible. Their purpose is to impose discipline and clarity of communication on the replicator so as to ease the burden of those who seek to read the replication report. Replications that are poorly written (or that force the reader herself to reverse engineer the original article) are of limited value and are unlikely to advance the goal of improving scientific standards via replication. Critique and criticism should (usually) be delayed until after the replication effort.

(1) The principal hypotheses of the study:

- That changes in the monthly average level of the federal funds rate during the period February 1987 to October 2005 were followed by changes in the monthly average value of the prime rate of approximately the same size.
- That the size of the change in the prime rate (immediate and cumulative) following a change in the prime rate differed depending on whether the federal funds rate increased or decreased.

(2) The econometric methods:

- The analysis begins with the equation $P_t = \alpha_P + \beta_P FFR_t + \varepsilon_t$, where P_t denotes the prime rate and FFR_t denotes the federal funds. The ADF, PP, and KPSS tests are used to explore the stationarity of P_t and FFR_t.

- Next, the possible simultaneous existence of cointegration and a "structural" break is explored (Gregory-Hansen, 1996) via the equation $P_t = \alpha_P + \alpha_P^D D_t + \beta_P FFR_t + \varepsilon_t$, where $D_t = \begin{cases} 0 & \text{if } t \le \theta \\ 1 & \text{if } t > \theta \end{cases}$, and α_P^D measures the size of the shift in the intercept.
- Conditional on the $\hat{\theta}$ chosen via the Gregory-Hansen test, estimate the model $P_t = \alpha_P + \alpha_P^D D_t + \beta_P FFR_t + \varepsilon_t$ using Stock-Watson's DOLS, augmenting with $\Delta FFR_t, \Delta FFR_{t-1}, \Delta FFR_{t-2}, \Delta FFR_{t+1}, \Delta FFR_{t+2}$
- Next, the authors estimate a momentum threshold autoregressive (MTAR) model

$$\Delta\hat{\varepsilon}_t = I_t \rho_1 \hat{\varepsilon}_{t-1} + (1 - I_t)\rho_2 \hat{\varepsilon}_{t-1} + \sum_{i=1}^{p} \alpha_i \Delta\hat{\varepsilon}_{t-p} + u_t \ ,$$

  where $\hat{\varepsilon}_t = P_t - \left(\hat{\alpha}_P + \hat{\alpha}_P^D \tilde{D}_t + \hat{\beta}_P FFR_t\right)$ and $I_t = \begin{cases} 1 & \text{if } \Delta\hat{\varepsilon}_{t-1} \ge \tau \\ 0 & \text{if } \Delta\hat{\varepsilon}_{t-1} < \tau \end{cases}$.

  Unfortunately, the authors do not state what estimator was used to obtain $\hat{\varepsilon}_t$ (OLS or DOLS), only that "the MTAR model uses the residuals generated from Equation 2" which, in the article, is $P_t = \alpha_P + \alpha_P^D D_t + \beta_P FFR_t + \varepsilon_t$. The authors conclude that the estimates support both cointegration and asymmetric adjustment.

- Finally, motivated by the MTAR results, the authors fit the asymmetric error-correction model

$$\Delta P_t = \alpha_0 + \sum_{i=1}^{n} \alpha_i \Delta P_{t-i} + \sum_{i=1}^{q} \gamma_i \Delta FFR_{t-i} + I_t \rho_1 \hat{\varepsilon}_{t-1} + (1 - I_t)\rho_2 \hat{\varepsilon}_{t-1} + u_{1t}$$

$$\Delta FFR_t = \tilde{\alpha}_0 + \sum_{i=1}^{n} \tilde{\alpha}_i \Delta P_{t-i} + \sum_{i=1}^{q} \tilde{\gamma}_i \Delta FFR_{t-i} + I_t \tilde{\rho}_1 \hat{\varepsilon}_{t-1} + (1 - I_t)\tilde{\rho}_2 \hat{\varepsilon}_{t-1} + u_{2t}$$

  where $\hat{\varepsilon}_t = P_t - \left(\hat{\alpha}_P + \hat{\alpha}_P^D \tilde{D}_t + \hat{\beta}_P FFR_t\right)$. The $\hat{\varepsilon}_t$ in this model likely are the same as in the MTAR model but, again, the authors do not state what estimator was used to obtain $\hat{\varepsilon}_t$ (OLS or DOLS).

- The authors do not specify what software was used to do the calculations.

(3) The data:

- Two variables, both at a monthly frequency: the federal funds rate and the prime rate.
- The dataset was not available from the journal. (I did not request the data from the authors.)
- The authors state that the data are from the Federal Reserve Bank of St. Louis FRED database, although they do not provide the variables' names. My assumption is that the federal funds rate is the monthly average of the daily "effective federal funds rate" as published by the Federal Reserve [FRED name: fedfunds] and that the prime rate is the monthly average of the daily prime rate [FRED name: mprime], both as printed on the

Federal Reserve Board's H.15 statistical release. If so, "vintage" issues are unimportant because these data do not revise after publication.

(4) Inference:

- The authors' inference begins by asking if P_t and FFR_t are well-modeled as I(1). PW conclude that both are well-modeled as I(1), that is, stationary after first differencing.
- Next, the authors fit the Engle-Granger regression $P_t = \hat{\alpha}_P + \hat{\beta}_P FFR_t$, familiar as a test for cointegration. But, seeking to move beyond cointegration, they note that "pass through," as measured by $\hat{\beta}_P = 0.841$, is less than complete. The coefficient's t-statistic, however, is 36.5, almost certainly indicating that both rates are simultaneously determined. Inference on the size of the estimated coefficient is misplaced.
- The authors suspect that cointegration perhaps is rejected due to a structural shift. Using the Gregory-Hansen test (1996), they reject $H_0 : \alpha_P^D = 0$ and set θ = April 1996, that is, a "structural break" occurred in April 1996. A DOLS regression $P_t = \hat{\alpha}_P + \hat{\alpha}_P^D D_t + \hat{\beta}_P FFR_t$ returns $\hat{\beta}_P = 1.06$ with a t-statistic of 7.6; again, inference regarding the degree of pass through is clouded by likely feedback between the two rates.
- The authors, unfortunately, do not present all estimated coefficients for the MTAR and EC models but, rather, focus on hypothesis test statistics. As a replicator, I would prefer to see all estimated coefficients because the sensitivity of individual coefficients to changes/errors in the data and the estimators differs. Some coefficients might be judged "close" while others might be judged "far away." I continue this discussion below.

Comments

Some replicators ask if there is a "make or break" coefficient in an analysis, that is, does the contribution of the analysis to scientific knowledge depend more on some coefficients than others and, if so, should a replicator devote more time to a specific coefficient (or hypothesis)? My general answer is no—failure to replicate in any part of the analysis must call into question all inferences in the analysis (how can we know where the data and/or programming errors, if any, are buried?)

In the article from Dewald et al. discussed in section 3, above, for example, the hypotheses were no more than whether some putative explanatory variables had significant t-statistics; generally, each coefficient could be judged alone. In the article of this section, all estimation and inference beyond the initial stationarity tests is conditional on the Gregory-Hansen test's choice of θ= April 1996 and rejection of $H_0 : \alpha_P^D = 0$. Suppose, for example, that the replicator finds herself unable to replicate the Gregory-Hansen test results, that is, the test suggests an alternative date. Although the replicator will be tempted to halt at that point, t*he replication paradigm suggested in this article recommends that the replicator should explore alternative estimates of the DOLS, MTAR and EC models conditional on a variety of values of* θ, that is, the replicator (ideally) should explore the shapes of the vector of functions

$$\frac{\partial \hat{\alpha}_P(\theta)}{\partial \theta}, \frac{\partial \hat{\alpha}_P^D(\theta)}{\partial \theta}, \frac{\partial \hat{\beta}_P(\theta)}{\partial \theta}.$$

To summarize her findings, a response surface regression analysis might be reasonable. If subsequent model estimates vary sharply with θ, the scientific contribution might be labeled "fragile" (not matter if similar results are obtained); if the model estimates are largely invariant to the choice of θ, then the contribution might be labeled "robust." By so doing, the replicator disavows the "binary" choice of whether the article is "replicated" or not, and draws forward the scientific contribution of the article.

My final comment, specifically addressed to this study, cautions not to underestimate the power of simple exploratory data analysis (that is, graphical analysis of data). Replications should begin with charts of the data because sometimes charts suggest interesting avenues of exploration. Figure 1 (below) displays levels of the prime and federal funds rate retrieved from FRED; ocular econometrics (holding both charts together in front of my desk lamp) suggests it reproduces figure 1 in Payne and Waters (2008). Figure 2 displays the difference between the series. The chart hints of a possible break in late 1990 or early 1991, rather than 1996 (subsequent hypothesis testing does not reject a break circa 1990/1991, while rejecting a break in 1996).

Fig 1: Prime Rate and Federal Funds Rate

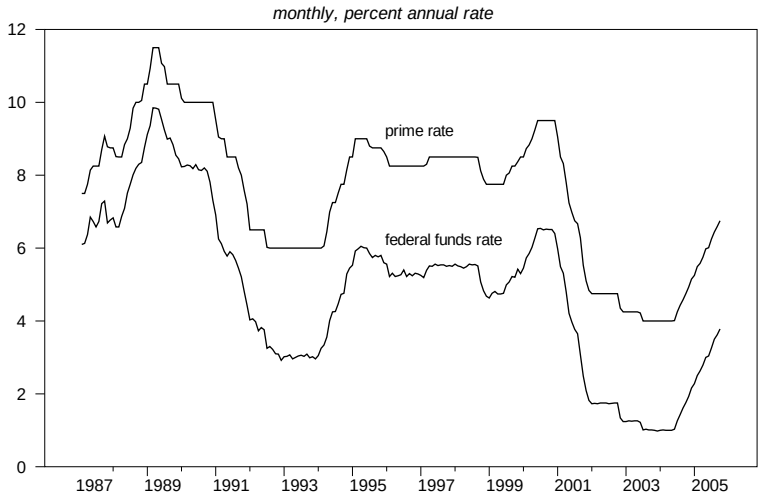
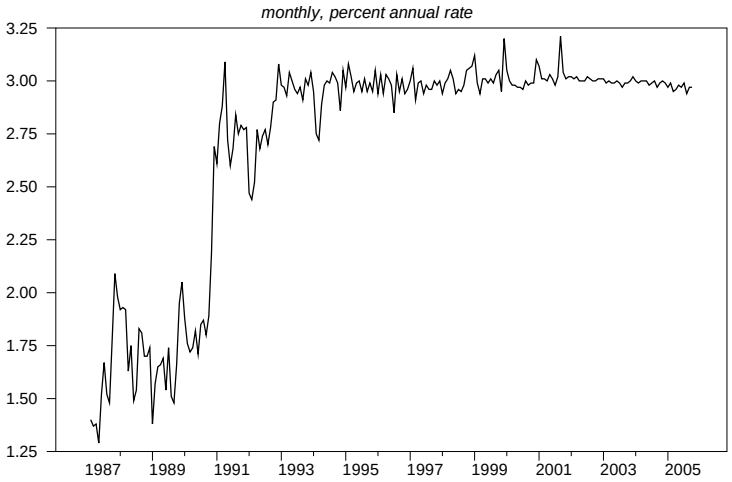


Fig 2: Spread, Prime Rate - Federal Funds Rate



5. How Do We Judge When a Replication Is Successful?

"Replication" can either be studied as a epistemological subject or a statistical one. The former, in my experience, leads researchers/replicators to discuss binary outcomes: either the study is "replicated" or it is not, that is, all parameters are reproduced precisely or they are not. In analyses where the author(s) carefully store the data and programs, exact reproduction is possible (e.g., Dewald et al, 1986).

In empirical economics (beyond experimental economics and stochastic computer simulations), as noted above, replication—in theory—is a deterministic process: applying a well-defined set of computer/algebraic manipulations to the data should produce, each time, the same numerical results. [13] It has been recognized for some time, however, that the truth is more complex.[14]

Replication, better, should be regarded as a statistical decision problem. Lehmann (1959), p. 2, notes that, quite obviously, "the methods required for the solution of a specific statistical problem depend quite strongly on the three elements that define it: the class to which the distribution of X is assumed to belong, the structure of the space D of possible decisions *d*, and the form of the loss function *L*." The difficulty of doing so in replication has been formidable but, absent the statistical formulation, how do we decide if a study is adequately "replicated" or not? In Payne and Waters (2008), how do we decide which parameters are critical and which are not? Do we focus more attention on some parameters than others? And, if so, what criteria form the weighting? Which parameters should cause us to reject the conclusions of the study?

From a statistical viewpoint, failure to replicate is a multivariate problem. To be more definite, let $\Gamma^P = \left\{\beta^P, \Omega^P, Z^P\right\}$ and $\Gamma^R = \left\{\beta^R, \Omega^R, Z^R\right\}$, respectively, denote the published and replicated parameter vectors (β), covariance matrices (Ω), and datasets (Z) of a generic study, such that the object of inference is $\Gamma^R - \Gamma^P$. [15]

Can a distance measure and statistical distribution be defined over this object so as to test $H_0 : \Gamma^R - \Gamma^P = 0$? So far as I am aware, this is an open question. It seems likely that additional, a priori, information must be included: Does a replicator have reason to believe that some variables in the dataset are measured less precisely (say, to fewer decimal digits of precision)

[13] In experimental fields, including medicine, the results are necessarily statistical because the focus shifts to the data generating process, which necessarily is stochastic.

[14] A number of somewhat older studies examined the distinction between computational accuracy in regression and inferences regarding statistical accuracy via, for example, t-statistics or p-ratios. These authors emphasized that computational accuracy depends, in part, on interactions between the model and the data: within the data, multicollinearity matters but the model may place larger or smaller information demands on the data. The coefficients of a demanding model fit to highly collinear data may suggest seemingly precise statistical inferences (t-statistics) while being very inaccurately/imprecisely estimated. It is unknown, pending future research, to what extent this distinction might cause a replication attempt to be judged a failure—particularly when the original software and hardware are not available. See for example Moulet (1992), Beaton, Rubin and Barone (1976), and Longley (1967).

[15] Of course, if the author(s) furnish their data, $Z^R \equiv Z^P$.

than others? And, do these specific variables have a stronger impact on parameters regarded as more important, relative to other parameters?

While an extended discussion of this topic is beyond the scope of this paper, I urge all interested to read Campbell Harvey's 2017 American Finance Association presidential address (Harvey, 2017) in which he recommends the minimum Bayes factor (MBF) as a testing regimen. It seems (to me) that Lehmann's requirements for a defensible testing regime for $H_0 : \Gamma^R - \Gamma^P = 0$ are infeasible outside a Bayesian environment. Harvey (and Lehmann before him) discouraged blind use of p-values (or t-statistics) but, rather, encouraged careful consideration of the tradeoff between size and power; as Harvey notes, too often the process of publication is blind to this tradeoff, focused (almost) entirely on p-values.[16] Further, due to "p-hacking" and the filter of publication bias, some "statistically significant" parameters might be quite computationally fragile, that is, the parameters may be inaccurately estimated, making replication difficult; see Longley (1959) and Beaton et al. (1976). But, when the distribution of the null is difficult to specify and interpretation of the testing environment does not lie within the classical paradigm of "repeated sampling," Bayesian methods are an alternative. Harvey (2017), section VII, notes that the MBF explicitly combines Bayesian-style *a priori* information and Lehmann's tradeoff between size and power, providing a flexible robust testing environment.

6. Conclusions

Replication efforts are important if they contribute to building scientific knowledge in a field. Studies should seek to be not only (in Ed Kane's well-known phrase) "econometric audits" but also to assess the power (robustness) of a study's scientific contribution. To better define that process and make replication contributions more accessible, this essay proposes the framework of the *replication paradigm.*

Eventually, the researcher replicating an empirical study reaches a decision: (1) the published paper is fully correct and can be reproduced/replicated; (2) the paper can be replicated but the replication has revealed errors; or (3) assert that the published paper cannot be replicated. Conclusion (1) is the easiest: essentially, say all is fine and move on. Conclusions (2) and (3) are grave because the scientific contribution of the replication is unclear. The replication paradigm presses on the replicator the task of seeking to "parametrize" the cause(s) and consequence(s) of the replication failure. The unfortunate result is that, by the conclusion of the replication effort, the replicator very well may have done more work on the model than the original researcher. Yet, without so doing, how can we know if the researcher was only painting dots on mice?

Finally, the "success" or "failure" of a replication effort should be regarded as a decision problem. Developing the required statistical framework is a challenging topic for future research.

[16] Harvey, in discussing the Neyman-Pearson framework, unfortunately does not mention that much of his discussion was foreshadowed by Leymann (1959), ch 3.

References

Anderson, Richard G. and Areerat Kichkha (2017). "Replication, Meta-Analysis, and Research Synthesis in Economics." *American Economic Review*. 107(5), pp. 56-59. May.

Anderson, Richard G. and William G. Dewald (1994). "Replication and Scientific Standards in Applied Economics a Decade After the *Journal of Money, Credit and Banking* Project," Federal Reserve Bank of St Louis *Review*, 76(6), November/December, pp. 79-83.

Anderson, Richard G., William H. Greene, Bruce D. McCullough, and H. D. Vinod (2008). "The Role of Data & Program Code Archives in the Future of Economic Research" *Journal of Economic Methodology*, March 2008, *15*(1).

Anderson, Richard G. (2006). "Replicability, Real-Time Data, and the Science of Economic Research: FRED, ALFRED and VDC." Federal Reserve Bank of St. Louis *Review*, January/February 2006.

Beaton, Albert E., Donald B. Rubin, and John L. Barone (1976). "The Acceptability of Regression Solutions: Another Look at Computational Accuracy." *Journal of the American Statistical Association*. 71(353) 158-168. March.

Christensen, Garret S and Miguel, Edward (2017). "Transparency, Reproducibility, and the Credibility of Economics Research." *Journal of Economic Literature* (forthcoming).

Computerworld (1983). "IBM's 4300 Entries Seen Rocking Supermini Mart," September 19, 1983, page 4.

Dewald, William G., Jerry G. Thursby, and Richard G. Anderson (1986). "Replication in Empirical Economics: The *Journal of Money, Credit and Banking* Project." *The American Economic Review*, Vol. 76, No. 4, (Sep., 1986), pp. 587-603.

Duvendack, Maren, Richard Palmer-Jones and W. Robert Reed (2017). "What Is Meant by 'Replication' and Why Does it Encounter Resistance in Economics?" *American Economic Review*. 107(5), 46-51.

Galiani, Sebastian, Paul Gertler, and Mauricio Romero (2017). "Incentives for Replication in Economics." NBER working paper 23576, July.

Goldberg, Lawrence and Anthony Saunders (1981), "The Growth of Organizational Forms of Foreign Banks in the U.S." *Journal of Money, Credit and Banking*, vol. 13, pp. 365-374, August.

Gregory, Allan W. and Bruce E. Hansen (1996). "Residual-based tests of cointegration in models with regime shifts." *Journal of Econometrics*, 70, pp. 99–126.

Hamermesh, Daniel S. (2007). "Viewpoint: Replication in Economics." *Canadian Journal of Economics.* 40(3), pp. 715-733.

Hamermesh, Daniel S. (2017). "Replication in Labor Economics: Evidence from Data, and What It Suggests." *American Economic Review*. 107(5), pp. 37-40.

Harvey, Campbell R. (2017). "Presidential Address: The Scientific Outlook in Financial Economics." *The Journal of Finance.* 72 (4) 1399-1440. August.

Ioannidis, John P.A. (2005). "Why Most Published Research Findings Are False," PLoS Med. 2(8) (August)

Judge, George G., W.E. Griffiths, R. Carter Hill, Helmut Lütkepohl, and Tsoung-Chao Lee (1985). *The Theory and Practice of Econometrics.* John Wiley and Sons.

Kane, Edwarrd J. (1984). "Why Journal Editors Should Encourage the Replication of Applied Econometric Research." *Quarterly Journal of Business and Economics.* 23(1), pp. 3-8.

Lehmann, Edward L. (1959). *Testing Statistical Hypotheses.* Wiley.

Longley, James W. (1967). "An Appraisal of Least Squares Programs for the Electronic Computer from the Point of View of the User." *Journal of the American Statistical Association*. 62 (319) 819-841 (September).

Moulet, Marjorie (1992). "A symbolic algorithm for computing coefficients' accuracy in regression." In *Machine Learning*, Derek Sleeman and Peter Edwards, eds., pp. 332-337. Proceedings of the Ninth International Workshop, Aberdeen, Scotland. Morgan Kaufman, San Francisco.

Payne, James E. and George A. Waters (2008). "Interest rate pass through and asymmetric adjustment: evidence from the federal funds rate operating target period." *Applied Economics*. Vol. 40, 1355-1362.