B.D. MCCULLOUGH
KERRY ANNE MCGEARY
TERESA D. HARRISON

Lessons from the *JMCB* Archive

We examine the online archive of the *Journal of Money, Credit, and Banking*, in which an author is required to deposit the data and code that replicate the results of his paper. We find that most authors do not fulfill this requirement. Of more than 150 empirical articles, fewer than 15 could be replicated. Despite all this, there is no doubt that a data/code archive is more conducive to replicable research than the alternatives. We make recommendations to improve the functioning of the archive.

JEL codes: B4, C8
Keyword: replication.

IN MARCH 1989 at the University of Utah, Stanley Pons and Martin Fleischman claimed to have produced “cold fusion” which, if true, would have led to an almost limitless supply of cheap energy. Upon the announcement, researchers around the world began attempting to replicate this already-famous result. Very shortly, “confirmations” of the experiment were issued by researchers at Georgia Tech and Texas A&M. Before too much longer, MIT, Cal Tech, Harwell, and others reported failures to confirm. Tech and A&M, upon re-examining their results, found errors and retracted their confirmations. The scientific method had revealed the claim of cold fusion to be false.

If Pons and Fleischman had published their cold fusion result in an economics journal, the world would still be awaiting lower utility bills. Distinct from most sciences, economics has not fully embraced the scientific method; in particular, there is no tradition of replication in economics. Results published in economics

For comments, we thank Charles Renfro, James MacKinnon, Phil Mirowski, Roger McCain, H.D. Vinod, Kishore Kulkarni, Paul Jensen, Brett Katzman, and Tansel Yilmazer, as well as seminar participants at Drexel University, the University of Pennsylvania, and the Federal Communications Commission. We are especially grateful to Paul Evans, the outgoing editor, for support and encouragement.

B.D. MCCULLOUGH *is an associate professor in the Department of Decision Sciences at Drexel University* (*E-mail:* bdmccullough@drexel.edu). KERRY ANNE MCGEARY *is an assistant professor in the Department of Economics at Drexel University* (*E-mail:* kmcgeary@drexel.edu). TERESA D. HARRISON *is an assistant professor in the Department of Economics at Drexel University* (*E-mail:* tharrison@drexel.edu).

Received August 22, 2003; and accepted in revised form December 12, 2003.

Journal of Money, Credit, and Banking, Vol. 38, No. 4 (June 2006)

journals are accepted at face value and rarely subjected to the independent verification that is the cornerstone of the scientific method.[1] Most results published in economics journals cannot be subjected to verification, even in principle, because authors typically are not required to make their data and code available for verification. Moreover, when such a requirement does exist at a particular journal, it can be (and is) ignored with impunity. This lack of emphasis on replication in the economics profession is regrettable because the importance of replication in the scientific process cannot be understated: "It is attempts at replication that check whether a genuine advance in knowledge has been made or a puzzle encountered, or whether either mistake or fraud lies behind the results" (O'Brien 1992, p. 263).

The literature on replication offers many different definitions for different types of replication (see, e.g., Fuess 1996, pp. 4–5). Here, we employ the following commonly used definitions. *Replication* refers to the duplication of published results. The sole purpose of a replication study is to attempt to reproduce the results of some paper. *Successful replication*, *partially successful replication*, and *unsuccessful replication* refer to duplicating all the results, enough of the results that the conclusions of the article remain intact, and not enough of the results to sustain the article's conclusions, respectively. A successful replication can still lead to reversal of an article's conclusions, for example, by finding a programming error that reverses the results. By contrast, *reproduction*, also called replication-with-extension, has as its main goal the use of newer methods on an existing dataset, or the use of new data with an existing methodology, and replication is only an incidental (and not always necessary) part of the task. It has long been known, informally, that an author is much more likely to make his data and code available to someone seeking reproduction rather than replication, because the reproducer is more likely to cite the author positively, while the replicator can only get published by overturning the author's already published results.

Mirowski and Sklivas (1991) formally modelled the reproduce/replicate decision, and concluded not only that reproductions would be much more common than replications but that the equilibrium number of replications would be near zero, since there are many disincentives to replication. The market for replications is very limited. Perhaps only the journal that published the original article might be interested and sometimes even that is not the case. Given such a limited market for publication, very few persons would undertake the effort of writing a replication study. Additionally, journals tend to favor the reproduction to the exclusion of the replication. For example, both the *Quarterly Journal of Business and Economics* (Fuess 1996, p. 6) and *Labour Economics* (Arulampalam et al. 1997, p. 100) formally encourage reproduction and discourage replication. As reported by Mirowski and Sklivas (1991, pp. 159–161), when the *Journal of Political Economy* had a replications section (1976–87), of the 36 notes that appeared in this section, only 14% attempted replications (and only one actually was successful). Hubbard and Vetter (1997) studied

[1] "Neither originality, logical rigor or any other criterion is ranked as 'essential' by so many natural scientists as replicability." (Mayer 1980)

replication in the finance literature for the period 1975–94. Of 1423 empirical articles, 144 were reproductions, and not one was a replication.

Given the profession's historical and institutional lack of interest in replication, it was nonetheless quite a shock to the profession when Dewald, Thursby, and Anderson (1986, "DTA") showed that much economic research is not replicable: they were able to replicate only 2 of 54 articles, 3.7%. In response to this disturbing finding, most journals did nothing, but a few journals required authors to make their data and code available. Consequently, each economics journal can be said to operate in one of two distinct markets: the market for reproducible results, and the market for irreproducible results.

The market for irreproducible econometric results has been analyzed by, among others, Feigenbaum and Levy (1993), Wible (1991), and Mirowski and Sklivas (1991). The argument is very simple. Making sure that one's results are replicable is an enormous amount of work and, since no one is checking, the rational economist will not invest the amount of time necessary to ensure that his results are replicable. The journals do not check because they do not want to run the risk of admitting that they published irreproducible results. Moreover, the vast majority of journal editors simply is not willing to do what it takes to ensure that they are publishing replicable research in the first place.

The market for reproducible econometric results can be broken down into journals that have "policies" requiring authors to make their data and code available on request (e.g., *Labour Economics*, *International Journal of Industrial Organization*, *Journal of Human Resources*) and journals that have data and code archives (e.g., this journal, *Macroeconomic Dynamics,* and *Fed. Res. St. Louis Review*). McCullough and Vinod (1999) argued that "policies" are honored more often in the breach, and that only a mandatory data/code archive might bring replicability of research to the economic science. McCullough and Vinod (2003, section 4) showed that such journal replication policies are ineffectual because authors generally ignore them, i.e., authors simply refuse to supply data and code upon request. Whether this is evidence of widespread moral failure on the part of economists who refuse to honor a previously agreed-to policy is debatable. It is certain that the editorial board easily could have provided sufficient incentive to induce the researchers to honor their commitments. Will the editor publish an embarrassing note listing authors who fail to comply, and informing the profession that their results cannot be verified? No. Will the editor bar noncomplying authors from future submissions? No. The goals of the replication policies were incompatible with the incentive mechanisms implemented (or not) by the journals.

The question remains whether data/code archives are effective. This journal has had one since 1996:[2] authors are asked to provide "the data and programs used in generating the reported results." This experiment has been running long enough

[2] Dewald, who edited this journal from 1975 to 1983, started and maintained a floppy-disk based archive. Under a subsequent editor, not only was the policy of archiving data and code abandoned, but the files that Dewald accumulated were discarded.

that it is time to analyze the data and draw some conclusions, which we do in this article.

Section 1 presents some benefits of replication. Section 2 discusses the importance of data and code for purposes of replication. Section 3 gives a recent example from the replication literature that illustrates many of our points. In Section 4 we describe how we surveyed the archive, and present summary results. In Section 5 we describe in detail what we found in the archive. Section 6 argues that more replication is necessary. Section 7 presents the conclusions.

1. THE BENEFITS OF REPLICATION

Replication is a rarity despite the fact that its value can be stated unequivocally: "Whether it is successful or unsuccessful, replication changes the evidence for (and therefore the probability of) some scientific hypothesis" (Kane 1984, p. 4). Negative replications obviously have value in that they purge the profession's body of knowledge of errors. Positive replications have value, too.[3] Though they may not expand the frontier of knowledge, they do demonstrate that there exist data and code behind the numbers. Of course, that a result is replicable does not mean that it is correct: data can be incorrectly copied from book to computer, equations can be incorrectly coded, etc. The value of a replication can be increased if, in addition to merely replicating the published results, the replicator also verifies that there are no programming errors, etc.

Given the frequency with which different software packages give different answers to the same problem, an additional value for positive replication has been demonstrated by a pair of econometricians at Stata Corp., who used a canned routine in Stata to replicate a previously-published and independently programmed panel data estimation. This increases our confidence not only that the original article is correct, but that also the panel data routine in Stata is correct, too. See Drukker and Guan (2003) for details.

As King (1995b, p. 494) points out, simply checking to make sure that published results are correct is the least important reason for having an archive. The primary reason is to save other researchers the trouble of reinventing the wheel (Kniesner 1997, p. 115) so that "the second researcher will receive all the benefits of the first researcher's hard work" (King 1995a, p. 445). Gleditsch and Metelits (2003) show that an article that makes data available has twice the impact of an article that does not; this impact can only increase when reliable code is made available, too. Building on existing research is problematic enough, without having to rewrite code that someone else already has written, or collect data that someone else already has collected, or puzzle through code that is not intelligible even to its author.[4]

[3] See Tomek (1993) for a discussion of the benefits of confirming published results.

[4] We are all familiar with looking at code we wrote only weeks ago and wondering, "Why did I do this?" or "What does this do?"

For example, in this journal at least two articles used the Hansen-Heaton-Ogaki (1988) code (Hansen, Heaton, and Ogaki 1988). Ogaki wrote his code so that the others could understand what the code was doing and adapt it to their own purposes. As another example, of Quandt disequilibrium models and Hamilton Markov switching Renfro (2003, p. 42) writes, "These two models are actually very closely related but the Quandt models never caught on because essentially Dick Quandt never gave out the software to implement them. Hamilton on the other hand gave out GAUSS code for everyone to do it and so created a whole industry."

2. DATA AND CODE

DTA (p. 591) noted that "Data are useless to another researcher unless accurately recorded and properly documented." Yet, they further observed, "Many authors cited only general sources such as *Survey of Current Business*, *Federal Reserve Bulletin*, or *International Financial Statistics*, but did not identify the specific issues, tables, and pages from which the data had been extracted." Since data often are revised many times after initial publication, different issues of the same periodical may yield different values for the same observation. We observed that many articles in this journal are inadequately documented, and the same is true of many other economics journals. This type of research is far beneath professional standards, since introductory textbooks, when instructing the novice on how to write a research paper, stress the importance of clearly identifying the data sources. For example, in his introductory textbook Wooldridge (2003, p. 661) advises the novice, "Enough information should be presented so that a reader could, in principle, obtain the data and redo your analysis."

It is true that print is a limiting medium, and many authors believe there is no space to specify each issue of the *Survey of Current Business* that was consulted for each observation recorded. There can be no such sentiment when electronic archives are available to hold the files that are sufficient to replicate the published results, i.e., the replication files. Therefore, *we recommend that a data dictionary be included in the replication files, as part of a readme file. The data dictionary describes the variables and also gives the provenance of all the data.* Even data that is subject to an embargo[5] should have their provenance documented.

The objective of an embargo is to permit the author to have sole use of the dataset that he collected. Even without an embargo, some authors will be hesitant to provide data, arguing that it infringes upon the author's competitive advantage.[6] McCullough and Vinod (2003) and Gill and Meier (2000) have noted that such articles cannot be relied upon at least until the embargo period ends, since the accuracy of such

[5] Some journals permit authors to "embargo" originally collected data, i.e., the author does not have to make his data available to other researchers for some period of time, e.g., two years.

[6] By the time the article appears in print, the author should already have a second article submitted and be working on his third. If an author really wants to have the dataset all to himself, he can simply write all the articles he wants, and then submit them to journals simultaneously rather than seriatim—the cost to the author is that he has to wait to submit his articles.

articles cannot be independently verified by other researchers.[7] Not providing data or code, even if due to an embargo, is merely a method by which some journals permit the author to shift the cost of keeping the dataset to himself (delayed publication) onto the journal-reading public (in the form of articles whose accuracy cannot be assessed) with the added expense of retarding scientific progress. Additionally, if the data do have value, then the author will gain citations as others use his data.[8] We pointedly note that neither this journal nor the *American Economic Review* permits embargoes on data solely to permit the author to have exclusive use of the data he collected.

Not only should the data be properly documented but the code should be documented, as well. Putting code in an archive is not simply a matter of depositing uncommented, unclear code. In fact, the code is a better record of what was actually done to the data than is the article. Moreover, the myriad minor decisions for which there simply is insufficient space in print are revealed in the code, and hopefully made clear via extensive commenting of the code. *We recommend that the readme file list all the replication files with a brief description of each.*

Writing code for replicable research is not as easy as it sounds (see Altman and McDonald, 2003, for a cautionary tale). It takes an enormous amount of effort but, then again, this code does represent a contribution to the cumulative body of knowledge: in terms of actually building the body of knowledge and enabling others to make use of an author's research, the code is in many ways no less important than the article itself. For example, the code should be written and commented so that someone with a different package can build on the existing research. Consequently, if an author writes an article using RATS and a second researcher uses TSP, as a general rule the second researcher should not need to obtain the RATS reference manual and learn RATS just to extend the author's results. As a simple example, consider the following RATS code:

```
linreg(spread=x4) y
# constant x1 x2 x3
restrict 1
# 2
# 1 −1
```

It is obvious enough that this is some sort of regression of *y* on a constant and three independent variables, but the rest of the code is not intuitive. It may reproduce the results in the paper, but it hardly helps a TSP user understand what was done. Even RATS users may have to consult the reference manual. A few comment lines would remedy the situation:

[7] This argument applies not just to data collected by a researcher, but also to code that is written, too.

[8] Of course, the author does not have to supply *all* the data that he collected, just those data from which the final dataset is produced.

```
* weighted least squares with the variance proportional
to x4
linreg(spread=x4) y
# constant x1 x2 x3
* test a single restriction on the regression equation
restrict 1
* the restriction is on the second coefficient
# 2
* test that ``1'' times the second coefficient equals ``−1''
# 1 −1
```

Additionally, we note that names of the variables/parameters in the code should correspond, at least up to a mnemonic transformation, to the variables/parameters in the paper.

3. AN EXAMPLE

McCrary's (2002) recent partially successful replication attempt of Levitt's (1997) paper highlights many of the issues we have raised. Levitt published a paper in which he found that increases in police substantially reduce crime. Levitt supplied McCrary with data/code that would not replicate even the OLS results in Levitt's paper. Coefficients that Levitt reported as −0.06, −0.31, and 0.11, in his paper were estimated as 0.00, −0.28, and 0.17, respectively, by his own data and code.[9] Mere (partially) successful replication does not ensure that published results are correct, only that they can be (partially) duplicated. McCrary also found a serious coding error that reversed Levitt's conclusion. Levitt had intended to give more weight to crimes with lower year-to-year variability, but actually gave highly variable crimes the most weight. When McCrary corrected this error, he found that police do not have an effect on crime.

Of his instrumental variable Levitt wrote (1997, p. 271), "The primary innovation in the paper is the approach used to break the simultaneity between police and crime...The instrument employed in this paper is the timing of mayoral and gubernatorial elections." Yet Levitt's article, in violation of Wooldridge's admonition, provided no sources for this variable. McCrary attempted to recreate this variable, but his version differed substantially from Levitt's version. When McCrary pointed this out, Levitt's response (Levitt 2002) was not to provide his source for this variable, but to point out that he called the mayor's office of 13 of the disputed cities and found that in seven of the cities the date provided by the mayor's office disagreed with both McCrary and Levitt. This falls short of the appropriate response, which is the provision of the source or an admission that such had been lost in the intervening years: Levitt could not document the "primary innovation" of his paper. If the journal

[9] The issue is not whether the coefficients are qualitatively similar, but whether there exist data and code that can duplicate the published results.

in question had required a data dictionary, the source would have been recorded at the time of publication. If the journal permitted an embargo, then McCrary might never have uncovered these many errors, leaving subsequent researchers of the police/crime connection to explain to referees why their results differed from those of Levitt (1997).

The larger point here is that if one of the top researchers in the profession can make these kinds of mistakes, then anyone can. DTA showed that economic research is rife with these mistakes, and we have found many similar mistakes in the *JMCB* archive. Journals should institute hard-and-fast rules to make these kinds of mistakes much more rare. A mandatory archive, together with the policies that we recommend, will go a long way toward making economic research more free of these errors.

4. METHODOLOGY

As a first step in assessing the archive, each of us took a specific year and categorized all the articles, comments, and replies in regular issues as either requiring an archive entry or not.[10] We decided that if the author of a paper uses data and code to obtain numerical results, then these data and code should be archived. Another of us then repeated the exercise for the same year and the two compared results. This was repeated for each year, 1996–2002, inclusive. Having decided whether each article should or should not have an archive entry, we then checked the archive to see whether the articles that should have an entry did, in fact, have an entry. Results are presented in Table 1.

In only two cases (the first and third issues of 1999) did we observe 100% compliance.[11] Certainly, some of the articles were exempted from the archive requirement due to confidentiality of data, but we have no idea which ones. *We recommend that the archive, which lists each paper regardless of whether or not the paper has an archive entry, should specifically state that a paper has been exempted from the requirement.*[12] It is certainly the case that some authors should have archive entries, but for some unknown reason do not. In order to increase compliance with the policy, *we recommend that the journal issue conditional acceptance letters, with a formal acceptance letter being sent only after the data/code have been archived.* Arguably, the replication files should be submitted when the paper is initially submitted so that they will be available for inspection by the referees, if the referees so desire.

[10] Special issues seem to be exempt from the requirement, though there is no obvious reason. We did not include them.

[11] The reason for the many zeroes in the last column of the final years is that the person responsible for archiving the data and code stopped doing this part of his job. The new holder of this job has commenced contacting authors from these years to obtain their data and code for inclusion in the archive.

[12] We agree that such papers should have their data exempted from the requirement, but we see no reason that the code should be exempt, too.

TABLE 1

Replication Policy Compliance: Do Authors Submit Data and Code?

Volume (Issue)	Articles	Should be Archived	Is Archived	Percent Compliant	Volume (Issue)	Articles	Should be Archived	Is Archived	Percent Compliant
		1996					2000		
28(1)	10	8	4	50	32(1)	11	5	2	40
28(2)	8	7	3	43	32(2)	9	7	3	43
28(3)	9	9	5	56	32(3)	9	7	4	57
28(4)	7	5	2	40	32(4)	7	6	2	33
		1997					2001		
29(1)	9	5	2	40	33(1)	7	6	1	17
29(2)	7	5	4	80	33(2)	10	7	2	29
29(3)	10	8	2	25	33(3)	8	7	0	0
29(4)	8	5	2	40	33(4)	8	6	0	0
		1998					2002		
30(1)	7	6	5	83	34(1)	13	10	0	0
30(2)	8	7	3	43	34(2)	14	10	1	10
30(3)	8	6	3	50	34(3)	7	4	0	0
30(4)	11	5	3	60	34(4)	7	7	0	0
		1999					2003		
31(1)	9	5	5	100	35(1)	7	4	0	0
31(2)	7	5	3	60	35(2)	7	4	0	0
31(3)	8	5	5	100	35(3)	8	8	0	0
31(4)	8	4	3	75					

Notes: "should be archived" is the number of articles that used a computer program and data, while "is archived" is the number of archive entries for that issue. N.B. This table does not include special issues, none of which had an archive entry.

5. EXAMINATION OF THE ARCHIVE

It is not our purpose to embarrass any researchers whose adherence to the policy was less than acceptable. Therefore, when we take an example of less than exemplary research, rather than cite the author, we will simply cite the volume and issue in which the article was published, e.g., 30(4), if the article appeared in the fourth issue of the thirtieth volume. The interested reader will then have to search only amongst a very few archive entries to find the author of the article in question.

Of the 266 articles (remember that this excludes special issues), 193 should have had archive entries, but only 69 did. Clearly, authors can refuse to comply with an archive the same way they refuse to comply with a replication policy, by ignoring it with impunity. *We recommend that the archive be mandatory.* Of the 69 archive entries, for which we expected to find data and code, we found only data for 11 of them (28(2), 28(4), 29(2), 29(4), 31(1), 31(3), 32(1), 32(3), 33(1), 33(2), 34(2)). Thus, 16% of the archive entries have no code. The secretary managing the archive was not checking to see that both data and code were being submitted. *We recommend that managing the archive be an editorial function.*

To try and understand the reasons for failing to submit code, we sent e-mails to all the above authors and asked them three things: why is there no code in the archive?

what software package did they use? and would they please send us the code that, with their already archived data, would replicate their published results?

Of the eleven, three reported that their econometric package only required them to click a mouse, and they never had any code. One wrote that he submitted data, but the journal never asked for code. Two replied by saying that they just ran "simple regressions" so there was no need for code.[13] Four never responded, and one said that he did submit data and code, had no idea why it was not in the archive, and sent us the data and code.

The remaining 58 archive entries had at least some data and some code. For each, we attempted to use the supplied data and code to replicate the published results. We made minor alterations to data and code to try to get the code to run with the data, but we did not attempt major alterations. Consequently, for some articles that we could not replicate, it is possible that there exists some data and code combination that will reproduce the reported results, but we are certain that it is not the combination that is in the *JMCB* archive. Since journal policy requires authors to deposit in the archive the data and code that will replicate their results, if we cannot use the data and code in the archive to replicate an author's results, it is fair to say that the author did not honor the policy.

Many authors took what can only be considered a desultory approach to fulfilling the requirement, not even caring whether the data would run with the code. In many cases, the author had specifically not provided the data that ran with the code, and instead had provided the data in some alternate format. The obvious implication of such an action is that it makes replication difficult, sometimes requiring much effort to put the data into a format that would run with the code. For example, one author's code reads an ".xls" file, but the code provided is in ".prn" format (28(3)). Another author's code (30(2)) calls a ".rat" file and his provided data are just the output from a "print" command that would take great effort to turn into a machine readable data file. Occasionally, authors provided data in a program-specific format, so that researchers who did not have that program could not access the data (30(1), 30(3)). One author (35(4)) provides no readme file and two data files with no column headers: we are supposed to guess the names of the variables! *We recommend that all data be provided in ASCII format, and that the version of the code submitted to the archive call these same ASCII files. Additionally, the first program should print summary statistics on all the variables, so that subsequent researchers can be sure that they have loaded the data correctly.*

Other authors seem to think that the entire world shares the exact same hard drive layout, with "C:\MYDATA\MYPROJECT\" sprinkled liberally throughout their

[13] In one case, the author used the Cochrane–Orcutt method and so indicated in a footnote. Failure to replicate resulted in correspondence, in which the author identified the software package used. This package, by default, implemented the Prais–Winsten transformation in the course of performing a Cochrane–Orcutt regression. In another case, the author's article merely referred to use of the Hodrick–Prescott (HP) filter and presented results for many subsamples. It was unclear whether the HP residuals were obtained from a regression over the entire sample which was then broken into subsamples or whether the HP filter was run on each subsample. We tried both methods and failed to replicate. Perhaps, the reason for failure to replicate is that different software packages give different results for the same HP filtering problem.

code. Of course, a would-be replicator has to find and change all these. Moreover, the author might not realize all the data/subroutine files that his code utilizes, and forget to include said data/subroutine in his replication files. For example, some authors forgot to include code for a subroutine that existed in yet another subdirectory (32(4), 30(1)), and similarly other researchers forgot data files (29(4), 31(1)). *We recommend that the author provide code such that the data and code, when placed in the same subdirectory, will execute; and that the output from doing this also be provided. The author should check to make sure that this runs correctly and produces the results in his paper.* An exception to this would be when substantial preprocessing must be done to produce the final dataset. In such a case, the preprocessing should all be done in one subdirectory, and the actual data analysis should be done in another. In more than one case, the author only provided the final dataset (30(4), 33(2)), so it was impossible to check to see whether he had preprocessed the data correctly, or to determine that the original data were correct. *We recommend that authors provide the primary data from which the final dataset is derived.*

Simply having data and code and possessing the same software package does not guarantee that a researcher will be able to replicate the author's results. We know of a package that in two successive releases produced different results for the same "calculate the correlation matrix" problem. We know also of a few packages that, in successive releases, produced different answers for the same nonlinear estimation problem. It is important, therefore, for the author to identify the version of the software he used. Currently, the only journal of which we are aware that requires this information is the *Journal of Economic and Social Measurement*. Moreover, the operating system can have an important effect on the results obtained (or not, as the case may be) as shown by McCullough and Vinod (2003, footnote 12). *We recommend that the author identify the version of the software he uses (by version number and/or release date) and similarly for the operating system on which the software is run.*

Besides the authors who provided only data, other authors as well made no pretense of providing code that would replicate the results in the paper, providing just code fragments that would not run (e.g., 31(2)). Some researchers provided only partial code, perhaps indicating that the reader could "easily" adapt the programs provided to produce the rest of the results (28(1)). Notwithstanding the fact that this will be trivial only for someone who knows the programming language, it provides no record of how the other results were obtained.[14]

Sometimes the code was poorly organized, and it was difficult to determine which of several programs produced which results in the paper (28(3), 32(2)). One author (30(1)) had four programs, one of which produced over 22,000 lines of output. We could not figure out which parts of this output were supposed to be results in the paper. *We recommend that the readme file should clearly indicate which*

[14] One particularly egregious example was a researcher who wrote in his readme file that due to "considerable experimentation, [the program in the archive] is not necessarily the one that produced the results reported in the paper" (31(2)). Indeed, he was correct: his program does not replicate the results in his paper.

programs correspond to what results in the paper, and further that the program(s) should be commented and perhaps structured to make clear which parts of the output constitute the results in the paper.

Most of the code we saw was poorly commented (if at all), not indented, and was not written so that someone familiar with another language could understand it. Indeed, much of the code was written so that someone familiar with the language would have a hard time figuring out what the code was doing! This is not the way to make it easy for other researchers to build on what already (ostensibly) has been done. The basic premise is that another researcher should not need the user guide and reference manual for the package used by the author: options invoked and special commands need to be explained. We recommend the excellent article by Nagler (1995) on how to write code so that others can understand it.

6. THE NEED FOR MORE REPLICATION

In a theory journal, the referees are expected to check the proofs of theorems to ensure the article's correctness. There is no corresponding effort for empirical journals, nor do all believe that there should be. An oft-made argument against replication holds that eventually, enough correct articles will be published that the incorrect article will not be believed. A typical presentation of this argument is given by Hamermesh (1997), who argues that "Mindlessly taking the exact same data and checking to see if the author 'made a mistake' is not a useful activity in the social sciences" because "other similar studies will show the result to be such an outlier that its importance in our appraisals of how the world works will quickly approach zero." This approach to correctness in scientific research never determines whether the article in question is an anomaly or an error. We think it a more effective use of resources to have one note discredit an incorrect article, rather than have several articles outweigh the incorrect result. Further, the first few "outweighing" papers must account for their discrepancy with the incorrect article; this may be very difficult to do unless one assumes (without proof!) that the original article was in error. Moreover, we wonder how many lower-tier articles it would take to drive the importance of an *AER* article to zero, assuming that those who submit such articles could overcome the logical referee's objection: "Your results contradict the results in an *AER* article, how do you account for this?" What is the appropriate response to the referee, that the *AER* author must have made a mistake?

Moreover, as Kane (1984, p. 4) put it, "Incentives to cheat and to rest content with results that are less than thoroughly checked out cannot be wished away." This journal's formal replication policy is correct when it states that the threat of replication provides an incentive for careful empirical research, but for researchers to respond to the incentive, the threat must be credible. The sloppy approach to providing data/code taken by many authors will continue until authors realize that another researcher might attempt to use the data/code to replicate and fail in the attempt, *and that this embarrassing result will be published.* However, we could not find a

Comment in this journal that discussed an attempt to replicate an article from this journal, successful or otherwise. *We recommend that this journal (and all journals) institute a replications section that stands ready to publish a single page summary of replication attempts, with supporting materials placed in the archive*; the *Journal of Applied Econometrics* has recently instituted such a section. In equilibrium there will be the need for few such pages, but until equilibrium is reached there will be the need for many such pages, and many authors will be embarrassed.

It is imperative that prospective replicators believe that they have a reasonable chance of getting a publication out of their efforts, if there is to be a sufficient amount of replication. As already noted, the market for any particular replication is extremely limited, and an editor and a replicator might disagree on whether the errors uncovered merited journal space. To provide would-be replicators with an option, the *Indian Journal of Economics and Business* (*IJEB*) and the *Journal of Economic and Social Measurement* (*JESM*) both have recently initiated replication sections, and will consider for publication attempts to replicate articles from any economics journal. *IJEB* will publish refereed, one page summaries with supporting documentation placed at the journal's website (www.ijeb.com), while *JESM* will publish short articles. Thus, a standard literature search will turn up reports of articles whose results are not replicable, or in some cases, whose results *are* replicable. We chose an issue of this journal and attempted to replicate all the articles for which there was an entry in the archive, and the *IJEB* has published our results (two successful, one partially successful): Harrison (2003), McCullough (2003), and McGeary (2003).

To foster the spirit of replication in the profession, in addition to teaching good programming style, professors of econometrics can send their students to archives as advocated by Feigenbaum and Levy (1993, pp. 230–231) and Vinod (2001, p. 87), perhaps with instructions to replicate published results using a different software package.

7. CONCLUSIONS

The *JMCB* policy requires that authors submit data and code that will reproduce the published results. While we did contact authors for various reasons, for purposes of replication we used only the information in the archive.[15] Of the 69 archive entries we examined, we did not have the resources to attempt replication of seven. Of the remaining 62, we could replicate all the results from 14.[16] While 14/62 = 22% is substantially better than the 2/54 = 3.7% found by DTA, it is not nearly good enough.[17] Whether the $62-14$ = 48 noncomplying authors were incapable of

[15] King (1995a, p. 444) emphasizes the importance of being able to use replication files to reproduce results *without any additional information from the author.*

[16] The titles and authors of the 14 will be provided on request.

[17] The 22% ignores all the empirical articles for which there was no archive entry, in which case the appropriate percentage is 14/186 = 7.5%.

producing functional replication files or just did not care is immaterial: sufficient incentives (e.g., a credible threat of irreproducibility being exposed in the journal) combined with minor checks (e.g., making sure that authors actually put data and code into the archive prior to publication) should correct the problem. It is time to change the rules of the game and run the experiment again.[18]

LITERATURE CITED

Altman, M., and M.P. McDonald (2003). “Replication with Attention to Numerical Accuracy.” *Political Analysis* 11, 302–307.

Arulampalam, Wiji, Joop Hartog, Tom MacCurdy, and Jules Theeuwes (1997). “Replication and Re-analysis.” *Labour Economics* 4, 99–105.

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

Drukker, David, and Weihua Guan (2003). “Replicating the Results in ‘On Efficient Estimation with Panel Data: An Empirical Comparison of Instrumental Variables Estimators’.” *Journal of Applied Econometrics* 18, 119.

Feigenbaum, S., and D. Levy (1993). “The Market for (Ir)Reproducible Econometrics.” *Social Epistemology* 7, 215–232.

Fuess, Scott M. (1996). “On Replication in Business and Economics Research: The QJBE Case.” *Quarterly Journal of Business and Economics* 35, 3–13.

Gill, Jeff, and Kenneth Meier (2000). “Public Administration Research and Practice: A Methodological Manifesto.” *Journal of Public Administration Research and Theory* 10, 570–599.

Gleditsch, N.P., and C. Metelits (2003). “Posting Your Data: Will You Be Scooped or Will You Be Famous?” *International Studies Perspectives* 4, 89–97.

Hamermesh, Daniel S. (1997). “Some Thoughts on Replications and Reviews.” *Labour Economics* 4, 107–109.

Hansen, Lars Peter, John C. Heaton, and Masao Ogaki (1988). “Efficiency Bounds Implied by Multiperiod Conditional Moment Restrictions” *Journal of the American Statistical Association* 83, 863–871.

Harrison, Teresa D. (2003). “Successful Replication of Thornton’s (2000) *JMCB* Article.” *Indian Journal of Economics and Business* 2, 285.

Hubbard, Raymond, and Daniel E. Vetter (1997). “Journal Prestige and the Publication Frequency of Replication Research in the Finance Literature.” *Quarterly Journal of Business and Economics* 36, 3–14.

Kane, E.J. (1984). “Why Journal Editors Should Encourage the Replication of Applied Econometric Research.” *Quarterly Journal of Business and Economics* 1, 3–8.

King, Gary (1995a). “Replication, Replication.” *Political Science & Politics* (September 1995), 444–452.

[18] It appears that the *American Economic Review* may be repeating the experiment already conducted by the *JMCB*. In response to McCullough and Vinod (2003, section 4), the *AER* switched from a “replication policy” to a “mandatory archive.” A casual perusal of the first couple of months of the *AER* archive shows some authors submitting data only, etc. It will be interesting to see whether repeating the experiment produces a different outcome. Time will tell.

King, Gary (1995b). "A Revised Proposal, Proposal." *Political Science & Politics* (September 1995), 494–499.

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

Levitt, Steven D. (1997). "Using Electoral Cycles in Police Hiring to Estimate the Effect of Police on Crime." *American Economic Review* 87, 1230–1250.

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

Mayer, T. (1980). "Economics as a Hard Science: Realistic Goal or Wishful Thinking?" *Economic Inquiry* 18, 165–178.

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

McCullough, B.D. (2003). "Partially Successful Replication of Brunner's (2000) *JMCB* Article." *Indian Journal of Economics and Business* 2, 289–290.

McCullough, B.D., and H.D. Vinod (1999). "The Numerical Reliability of Econometric Software." *Journal of Economic Literature* 37, 633–665.

McCullough, B.D., and H.D. Vinod (2003). "Verifying the Solution from a Nonlinear Solver: A Case Study." *American Economic Review* 93, 873–892.

McGeary, Kerry A. (2003). "Successful Replication of Wong's (2000) *JMCB* Article." *Indian Journal of Economics and Business* 2, 287–288.

Mirowski, Philip, and Steven Sklivas (1991). "Why Econometricians Don't Replicate (Although They Do Reproduce)." *Review of Political Economy* 3, 146–163.

Nagler, J. (1995). "Coding Style and Good Computing Practices." *Political Science & Politics* (September 1995), 488–492.

O'Brien, D.P. (1992). "Economists and Data." *British Journal of Industrial Relations* 30, 253–285.

Renfro, Charles (2003). "Econometric Software: The First Fifty Years in Perspective." *Journal of Economic and Social Measurement* 29, 9–107.

Tomek, William G. (1993). "Confirmation and Replication in Empirical Econometrics: A Step Toward Improved Scholarship." *American Journal of Agricultural Economics* 75, 6–14.

Vinod, H.D. (2001). "Care and Feeding of Reproducible Econometrics." *Journal of Econometrics* 100, 87–88.

Wible, James R. (1991). "Maximization, Replication, and the Economics Rationality of Positive Economic Science." *Review of Political Economy* 3, 164–186.

Wooldridge, J.M. (2003). *Introductory Econometrics,* 3rd edition. Mason, OH: Southwestern Publishing.