Do economics journal archives promote replicable research?

B.D. McCullough, Kerry Anne McGeary, and Teresa D. Harrison
Department of Economics, Drexel University

Abstract. All the long-standing archives at economics journals do not facilitate the reproduction of published results. The data-only archives at *Journal of Business and Economic Statistics* and *Economic Journal* fail in part because most authors do not contribute data. Results published in the FRB St. Louis *Review* can rarely be reproduced using the data+code in the journal archive. Recently created archives at top journals should avoid the mistakes of their predecessors. We categorize reasons for archives' failures and identify successful policies. JEL classification: B40, C80

Est-ce que les archives des revues économiques promeuvent les travaux de réplication? Toutes les archives de longue date des revues économiques ne facilitent pas la réplication des résultats publiés. Les archives de données du *Journal of Business* and *Economic Statistics*, et du *Economic Journal* achoppent en partie parce que la plupart des auteurs ne leur confient pas leurs données. Les résultats publiés dans la FRB St. Louis *Review* peuvent rarement être répliqués en utilisant les données et codes des archives de la revue. Les archives créées récemment par les meilleures revues devraient pouvoir éviter les erreurs de leurs prédécesseurs. On catégorise les raisons pour lesquelles les archives n'ont pas bien joué leur rôle et on identifie les politiques plus prometteuses.

1. Introduction

> The ability to replicate a study is typically the gold standard by which the reliability of scientific claims are judged.
>
> National Research Council (2002, 7)

Twenty-two years ago in the pages of the *American Economic Review* (*AER*), Dewald, Thursby, and Anderson (1986; DTA) attempted to replicate articles

A previous version of this paper was presented at the 2005 American Economics Association annual meeting in Philadelphia and the 2006 Canadian Economics Association annual meeting in Montreal. Thanks for comments to Dick Anderson, Paul Jensen, Ross McKitrick, Julie Murphy, and Charles Renfro. Email: bdmccullough@drexel.edu

Canadian Journal of Economics / Revue canadienne d'Economique, Vol. 41, No. 4
November / novembre 2008. Printed in Canada / Imprimé au Canada

0008-4085 / 08 / 1406–1420 / © Canadian Economics Association

published in the *Journal of Money, Credit and Banking* (*JMCB*). They were able to replicate only 2 of 54 articles; that is, published research, they found, was not replicable. This finding, of course, was of no minor consequence. Of particular note, DTA (589) observed that unless authors organized their data and code *before* submitting their articles, there was little hope of later reproducing the published results. If the data and code were not organized prior to submission, then they found that they were unable to reproduce the published results even with the original author's help. Consequently, DTA did not recommend that journals require authors to share data and code upon request (a 'replication policy'), but instead recommended that journals create archives, where the data and code would be archived prior to publication. The disincentives for authors to comply with replication policies (or archives, for that matter) that rely on the honour system rather than enforcement mechanisms have been analyzed theoretically by Mirowski and Sklivas (1991) and Feigenbaum and Levy (1993), among others. Both in theory and as shown in this paper and elsewhere, the honour system does not work.

In response to DTA, the *JMCB* adopted such a mandatory data+code archive. In sharp contrast, and despite the obvious incentive problems, the *AER* adopted a replication policy (Ashenfelter et al. 1986), whereby authors pledged to provide data and code to would-be replicators. Several other journals soon followed the lead of the *AER*. Only two journals, the Federal Reserve Bank of St. Louis *Review* (*Review*) and later *Macroeconomic Dynamics* (*MD*) followed the lead of *JMCB*. The *MD* archive, however, is no longer functioning. Three other journals, *Journal of Business and Economic Statistics* (*JBES*), *Economic Journal* (*EJ*), and *Journal of Applied Econometrics* (*JAE*) have archives, but their policy is 'data only'; that is, authors need not supply code, despite the fact that data alone are insufficient.[1]

When McCullough and Vinod (2003) attempted to replicate all the empirical articles in a single issue of the *AER*, fully half of the authors failed to honour the replication policy. The problem was fixed quickly; the *AER* implemented a mandatory data+code archive (Bernanke 2004). McCullough and Vinod (2004) wondered whether other journals would follow suit. The answer, we are happy to report, is a resounding 'Yes.' Recently, three other top journals have joined the mandatory data+code club: *Econometrica* (in 2005), *Review of Economic Studies* (in 2005), and *Journal of Political Economy* (in 2006). More recently, *Spanish Economic Review* (in 2007) and this journal (2008) have joined the club.

The importance of journals adopting strict mandatory data+code archives cannot be understated, because it is not unknown for editors to defend published work that is not demonstrably replicable, regardless of whether the journal has some sort of commitment to replicability of research. Two recent episodes illustrate this point. First, the *JPE* recently published the Oberholzer and Strumpf (2007) paper, which claims online file sharing does not hurt music sales. The

[1] For all but the simplest empirical articles, the article cannot possibly describe *every* detail of what was done with the data. See section 4 for examples.

working paper version circulated for a couple years, during which time Liebowitz found he was unable to replicate much of the paper and that most of its claims were wrong (see Liebowitz 2007 for full details). Liebowitz made his critique known to the JPE prior to publication. The *JPE* defended the publication of Oberholzer and Strumpf paper by saying that it was submitted prior to the *JPE*'s adoption of its replication policy (Handelsblatt 2008). Second is the the Hoxby/Rothstein case, in which Rothstein (2005) failed to replicate Hoxby's (2000) *AER* article on school choice. *AER* editor Robert A. Moffitt (who was not editor when the Hoxby article was published) said that the *AER*'s replication was not strictly enforced at that time; specifically, Moffitt's comment was that Hoxby's paper was published 'prior to the strengthening of our data availability policy' (Hernandez 2005). The corrosive effect on the scientific method of such a casual editorial approach needs no elaboration.

The replication policies of the aforementioned journals can be found at their websites; they usually run two to three single-spaced pages, with one set of requirements for traditional empirical work and another set of requirements for experimental work. Despite these extensive requirements, we found it easy to discover violations of the policies. At *AER* and *RES* we easily found papers that had data but no code, so these journals are not enforcing their own rules. We could not even examine the *JPE* archive.[2] At *Econometrica* there were precious few articles with data and code, despite so many empirical articles. We conjecture that the empirical articles without data and code entered the refereeing process before the policy was implemented, and so are not required to be replicable. Yet it would be nice if *Econometrica* (and other journals) made a note to this effect in the archive, so that readers would know whether data and code should be available for an article.

But do archives work? If these new archives follow in the footsteps of the older archives, will they be making a collective mistake? It seems obvious that a data+code archive should be more conducive to replicable research than a mere replication policy, but how much more? To answer this question, McCullough, McGeary and Harrison (2006; MMH) attempted to replicate every empirical article published in the *JMCB* since 1996. Of 186 empirical articles, only 69 had archive entries. Of these, replication could not be attempted for 7, owing to lack of software or the use of proprietary data. Of the remaining 62, the results of 14 articles could be replicated. This is better than the 2 of 54 that Dewald, Thursby, and Anderson (1986) could replicate, but hardly cause for enthusiasm. The primary reason so few empirical articles had archive entries is that nobody was checking – it was entirely up to the author to submit his data and code. For those articles that had archive entries, there are many reasons that submitted

[2] Apparently *JPE* archive access is available only to *JPE* subscribers. Our library gets *JPE* via electronic journal aggregators such as Proquest, Gale, and EBSCO and therefore we can access the pdf for an article but not the data and code. Our librarians were unable to secure archive access for us. The *JPE* is the only journal of which we are aware that limits data and code access to subscribers.

data and code failed to reproduce published results, and MMH made specific recommendations for both researchers and journals – these recommendations are presented in an appendix. In response to MMH, the *JMCB* revised its procedures (The Editors 2006). Yet three years after MMH was accepted, and a year after it was published, the *JMCB* archive still failed to contain data and code for many published empirical articles (McCullough 2007). Again, we see a casual editorial approach from a journal that has professed a commitment to replication.

The *JMCB* archive had/has two primary failings: (1) failure of authors to contribute data and code; and (2) failure of contributed data and code to replicate published results. It is unclear whether these failings were specific to the *JMCB* or endemic to archives in general. To ascertain this, it is necessary to examine other archives. In the present article, therefore, we examine the archives at *Review*, *JBES*, *JAE*, and *EJ*, focusing on the two primary failings. Many economists think that most articles are replicable, and lack of replicability of an article is not a matter for concern. To disabuse persons of these errant notions, in section 2 we discuss some important articles that turned out not to be replicable. Section 3 compares the four archives on the basis of authors actually contributing code to the archive; we find that only one journal does this well. Section 4 assesses the extent to which the data and/or code in the various archives can be used to reproduce the published results. We find that none of them supports replication. Section 5 presents the conclusions.

2. The need for replication: examples and benefits

The primary purpose of an archive is *not* to ensure replicability (King 1995, 494) but to enhance extensibility (which presumes replicability). Thus, an archive should make it easier for one researcher to build on the work of another, and part of this 'building' is, of course, being able to reconstruct (replicate) what the first researcher did. In this regard, any data-only archive fails miserably. The would-be replicator must invest resources in recreating code that someone else already has written; this unnecessary waste of time cannot encourage researchers to build on another's work. Yet mere replication does have an important purpose: it assures the reader – be he researcher or policymaker – that there are data and code that produce the published results.

There is a not uncommon sentiment among economists that replication is unnecessary (see, e.g., Hamermesh 1997). If replication is unnecessary, why would any journal have an archive? Why would economists such as Bronwyn Hall, Mark Watson, Bruce Hansen, and James Hamilton, to name a few, maintain their own archives? Nonetheless, we briefly present a few case studies of failed replication attempts. To demonstrate the pernicious effects of unreplicated, erroneous research, we first present the case of the Phillips Curve. We then present more recent episodes that have had at least the potential to impact public policy in significant ways. It is important to note that a properly functioning data+code archive would have significantly meliorated, if not completely eliminated, all the problems presented.

At very young and impressionable ages, economists of all stripes are indoctrinated with the Phillips Curve, where it is portrayed as an empirically derived law showing a tradeoff between wage inflation and unemployment that, most regrettably, shifts outward over the long run. For example, in his intermediate macro text, Mankiw (2003, 361) writes, 'Phillips observed a negative relationship between the unemployment rate and the rate of wage inflation in data for the United Kingdom.' This is not the truth. Macroeconomics textbooks never reveal that Phillips (1958) did not stumble upon his curve and then conclude that there is a tradeoff between wage inflation and unemployment. Rather, he hypothesized such a relationship and then found data that would support his belief (Wulwick 1989). Examining the historical record, it is hard to escape the conclusion that the Phillips Curve is the result of data mining.[3]

In a 1997 article in the *AER*, Levitt had an article in which he claimed to demonstrate that increases in police lead to marked decreases in crime. McCrary (2002) found a coding error: Levitt had intended to give more weight to crimes that are not so variable, but he actually gave more weight to the more variable crimes. Correcting for this error reversed the conclusion. Further, according to Levitt, the major contribution of his article was an instrumental variable relating to the timing of mayoral and gubernatorial elections. McCrary was unable to reproduce this data series from the references in Levitt's paper. In his reply to McCrary, Levitt (2002) also was unable to reconstruct this variable! Levitt's work was not replicable. How might policymakers and subsequent researchers have treated Levitt's result but for McCrary?

In Australia, analyses of the relationship between a woman's lifetime earnings and her number of children (Chapman et al. 2001; Gray and Chapman 2001) generated many headlines and were frequently cited in policy debates. Breusch and Gray (2004) reanalyzed the data and found that the original paper contained many errors and the policy implications of the original work were changed dramatically: the extent of forgone earnings is greater, the effect of second and third children is much larger, and the relationship of forgone earnings to educational background is stronger than originally reported. The effort expended in the replication effort

[3] Phillips did not run a regression on his 53 observations for unemployment and inflation data. He binned the data and then fit a curve to the resulting averages within each bin. This is evident when one considers the 'points' that Phillips used to describe his 'curve'; with U (unemployment) on the abscissa, the bins are defined by 0-2, 2-3, 3-4, 4-5, 5-7, and 7-11: the bins containing 6, 12, 5, 11, and 9 observations, respectively. Why such unequal class widths (with such differing numbers of observations in the classes)? Other choices for intervals did not support his theory. For example, the intervals 0-2, 2-3, 3-4, 4-6, 6-8, and 8-11 show a *positive* relation between wage inflation and unemployment for high unemployment. All this is well described in Wulwick (1996). What is also little known is that Phillips never intended this work as serious scholarship; it was just some casual work he did over a weekend at the suggestion of a colleague, and he was embarrassed by it. See the papers by Phillips in Leesom (2000) for further discussion of this point.

was so great that it required a separate article to describe it all (Breusch and Gray 2006).

Many researchers around the world study the underground economy, one key question being 'How large is the underground economy?' Giles and Tedds (2002) wrote a book on the Canadian underground economy in which they estimated that the underground economy grew from 3.6% of official GDP in 1976 to 15.6% in 1995. They used the econometric method known as MIMIC (multiple indicator, multiple cause), and this method for analyzing the underground economy has been adopted by other researchers. In the course of replicating the work by Giles and Tedds (2002), Breusch (2005) concluded, 'there is almost no difference between the measure of the underground economy calculated by Giles and Tedds and just *one* of their causal variables.' In other words, the MIMIC model, as applied, was useless. Breusch also showed that the model Giles and Tedds used for benchmarking is unidentified, both locally and globally. In sum, Breusch showed that policymakers ought not to rely on results from MIMIC models when seeking advice on the underground economy.

In the *Quarterly Journal of Economics*, Donohue and Levitt (2001) published a controversial paper arguing that more abortions results in less crime. The *QJE* had no replication policy at the time (and still doesn't), but in the best scientific tradition, Levitt and Donohue made their data and code available for inspection. Foote and Goetz (2005) discovered a programming error in the code, which was much trumpeted in the popular press. What was not much discussed was the fact that programming error made little difference to the result of the paper. The important point here is that despite his exchange with McCrary, Levitt continued to make his data and code available even when the journals didn't require him to do so.

Caroline Hoxby (2000) published an influential paper on school choice in the *American Economic Review*. As usual for a paper of such complexity, Hoxby's paper did not fully describe either the data set used or the precise calculations used to produce the published results.[4] Yet she was unable to produce the details upon request when Rothstein sought to replicate her results. After Rothstein (2004) attempted to recreate Hoxby's data and wrote his paper,[5] Hoxby then produced a 'redistribution CD' containing data and code. Using the redistribution CD, Rothstein (2005) was unable to replicate Hoxby's results. This is really not surprising, since, by Hoxby's own admission, these data and code do not reproduce the results published in the original paper: data errors were corrected (Hoxby 2005, 5) and the code was simplified (Hoxby 2005, 8). Thus, the redistribution

[4] For this reason, 'data only' archives are not conducive to replication; only data+code archives can facilitate replication. Even when data are proprietary, the code should still be made available, so that other researchers can apply the same method to new data or to check the accuracy of the code. See our recommendation 8 in the appendix.

[5] Rothstein (2004, 19–20) wrote, 'The researcher must make many decisions, many arbitrary, about the creation of a sample and the definition of control variables . . . Evidently Hoxby made different decisions – surely equally defensible – than did I, with important consequences for the results.'

CD produces results different from the published results. Whether these different results support or undermine the original results is the subject of a debate between Hoxby and Rothstein; all parties at least tacitly agree that the data and code necessary to replicate the original published results have yet to be produced.

The half-life of a false result is proportional to the effort with which the result can be checked by others. Most economics journals provide no mechanism whereby false results can be discovered. The above instances of the scientific method purging bad results from the cumulative body of knowledge occurred almost accidentally; certainly not because of any institutional efforts on the part of the journals. Do authors who refuse to honour replication policies incur any penalty? No. Do authors who refuse to deposit data and code in an archive incur any penalties? No. The lack of penalties is not consistent with a properly functioning incentive scheme. Such a scheme can simultaneously achieve the following goals: (1) ensure that published articles do have data and code to back up the published results; (2) make it easy for one researcher to build on the work of another; and (3) provide a means for purging the journals of incorrect results.

The benefits of an archive are many and are described in detail in Anderson et al. (2008); here we briefly review a few of them. The primary benefit is that because a data+code archive preserves a record of how the published research was produced, one researcher wishing to build on another researcher's work no longer has to reinvent the wheel. An additional benefit of a mandatory archive is that it compels compliance from authors who wouldn't otherwise permit their work to be subjected to replication or easily used as a basis for further research. Further, with data+code archives, checking robustness is a simple matter, as is extending results. Consequently, not only do well-functioning archives increase the accuracy of published results, but they will be better examined for robustness and more easily extended, thus increasing the quality of research. But all these benefits accrue only if (1) authors actually contribute to the archive, and (2) the contribution actually reproduces the published results. We investigate each of these conditions in the next two sections.

3. Do authors contribute to the archive?

We began our analysis of archives in the summer of 2004, so we examined journals only up to the end of 2003 (we did not extend the analysis as time progressed, because we believe that some journals already have altered their habits as a result of our previous work, MMH). We proceeded in the following way. First, we examined each article in the journal, classifying each as requiring an archive or not. To determine which articles required an archive entry we made the following rule: *any article that displayed or otherwise represented numbers required an archive*. This included not just classically 'empirical' articles, but computational economics articles as well. Even articles that display only graphs of numbers need

TABLE 1
By year, percentage of empirical articles with archive entries

	Fed. St. Louis *Review*		
Year	Empirical articles	Archive entries	%
1993	22	22	100
1994	25	13	52
1995	18	11	61
1996	19	15	79
1997	23	10	43
1998	24	7	29
1999	22	11	50
2000	22	14	64
2001	27	13	48
2002	25	15	60
2003	24	12	50
Total	251	143	57

an archive, so that the accuracy of the figures can be assessed,[6] though there were only a few of these. We then visited the archive to see which articles actually had an entry. Results for the 'data+code' archive are presented in table 1. For each issue, we give the number of articles that used data and should have an archive entry, the number of articles that actually have an archive entry, and the 'compliance' percentage of articles that should have an entry that actually have an entry. We first consider the data+code archive, and then the data-only archives.

The archive for the *Review* has been in existence since 1993. The inside cover of the journal states: 'All nonproprietary and nonconfidential data and programs for the articles written by Federal Reserve Bank of St. Louis staff and published in the *Review* are also available to our readers on [the *Review*'s] website.' As shown in table 1, it started off well, with 100% compliance (22/22) for the first year. Apparently the powers-that-be had a method for ensuring compliance but, for some unknown reason, abandoned that method.[7] Overall, of 406 total articles published during the period 1993–2003, 251 of them should have had an archive entry, but only 143 did; 143/251 = 57%. This compares favourably with what we found at the *JMCB* (69/193 = 36%).

[6] There are (or used to be) many programs that perform calculations in double precision and hand the results off to a graphics routine that is written in single precision. Such software packages can drop points from a graph without warning. See McCullough (2004, 263) for an example. Further, the information that can be extracted from a graph depends on the type of graph (Cleveland 1993). For an example of how simply altering the aspect ratio can dramatically alter a graph, see Spanos (1999, 190–1).

[7] It is worth noting that each year there is a special issue that contains an atypically large number of articles. For the first two years, these issues complied with the policy. In later years, they did not. There is nothing in the archive to indicate that these issues were exempted from the policy, nor any reason that such prominent results should be presented outside the scientific method (i.e., their reliability as a basis for policy or further research cannot be established).

TABLE 2
By year, percentage of empirical articles with archive entries

	JBES			*JAE*			*EJ*		
Year	Empirical articles	Entries	%	Empirical articles	Entries	%	Empirical articles	Entries	%
1997	42	30	71	30	30	100	53	3	06
1998	60	16	26	30	30	100	36	5	14
1999	46	9	20	29	29	100	36	1	03
2000	44	11	25	31	31	100	37	3	08
2001	43	18	42	32	32	100	40	3	08
2002	33	6	18	29	27	93	33	2	06
2003	44	22	50	32	32	100	44	18	41
Total	312	112	36	213	211	99	279	35	13

We next turn to the three ‘data only’ archives. The *JBES* has been published since 1983 by the American Statistical Association and has had an archive since the July 1993 issue. The official journal policy states: ‘For papers using datasets built up from publicly available data, providing the data will be a requirement for final acceptance. (The policy can be waived on a case-by-case basis when the data are proprietary or prohibitively expensive.)’ The *JAE* has been published since 1986, and all papers accepted after January 1994 have been required to archive data. The official policy is ‘Authors of accepted papers are expected to deposit in electronic form a complete set of data used onto the Journal’s Data Archive, unless they are confidential.’ The *EJ* archive dates back to 1995. While the journal page at the publisher, Blackwell, makes no mention of an archive, the journal’s owner, the Royal Economic Society, maintains the data archive and its website says: ‘The Economic Journal includes datasets associated with articles published in the Journal. The datasets are supplied by authors where copyright and technical issues permit, and they are made available online here.’ Further, the inside cover of the journal says: ‘The Royal Economic Society and Blackwell Publishers are pleased to announce the availability of data sets from the papers published in THE ECONOMIC JOURNAL.’

The first thing to note from table 2 is that *JAE* has near-perfect compliance, $211/213 = 99\%$. Meanwhile, $JBES = 112/312 = 36\%$ and $EJ = 35/279 = 13\%$. What is the reason for the *JAE*’s high compliance rate? In contrast to the other journals, only the *JAE* already follows two of the recommendations that we made in MMH: (1) the archive manager should be an editorial position; and (2) it should be the policy of the journal not to publish an article until the author has deposited his contribution to the archive.[8]

[8] This last requirement is harder to implement for special issues of the journal; note that the only two missing entries from the *JAE* archive were from a special issue. Nonetheless, this is a problem that could be fixed easily.

4. Do archive contents reproduce the published results?

While we have already noted that 'data only' archives do not support replication, it is instructive to take an example directly from *JBES* (Racine 2001). One *JBES* article claimed that neural nets outperformed linear regression for a specific type of problem. Racine had published in this area, and wanted to verify the result. The original author could not find his code, so Racine tried to implement the original author's work based solely on the article. He was unable to do so. Racine even solicited and received the original author's help in reconstructing the code and could not reproduce the published result. Despite this, the journal did not subsequently change its policy to require code, too.

For an example from *JAE*, Bai and Perron (2003) published an article about structural change models, and implemented their work in the package GAUSS. Kleiber and Zeileis (2005) wished to port this code to the package 'R.' In the course of so doing, they were unable to replicate some confidence intervals for breakpoints. For example, for quarterly data both packages estimated a breakpoint at 1972:3, though the GAUSS interval was [1970:3, 1972:4], while 'R' calculated [1969:1, 1972:4]. The discrepancies could not be attributed to rounding error. In a fine example of numerical detective work, they determined that the normal CDF in GAUSS was weak in the tails. The results published in Bai and Perron (2003) were wrong. But for the efforts of Kleiber and Zeileis, policymakers and researchers who relied on them would be misled, and other researchers who used the Bai/Perron code would obtain wrong answers and so forth. This is the scientific method at work: one researcher building on the work of another and purging the cumulated body of knowledge of incorrect results.

Kleiber and Zeileis could not have done this with just the Bai and Perron data: they needed the code, too. Indeed, they wrote:

> At present, the *JAE* only requires data, more often than not the corresponding code is not available. We are therefore grateful to Professors Bai and Perron for making their code publicly available without being obliged to. Without their code, this replication project would have had to stop halfway: GAUSS returns this, 'R' returns that, and it would have been much harder to determine the source of the observed differences. With their code, we can confidently say that numerical problems in GAUSS are responsible for most of the differences.

Given the above admission published in its own pages – in its own 'Replication Section' – it is surprising that the *JAE* still does not require code, too. The ostensible purpose of the 'data only' archive is for purposes of replication, yet the journal has all but admitted that data-only is insufficient for replication.

We now consider whether the data+code archive at the *Review* was capable of reproducing the published results. A very curious aspect of the process whereby data and code were archived at the *Review* needs to be explained. The author did not post the data and code. Rather, he gave it to a Reserve Bank research assistant. This research assistant, apparently under orders to post only two files, would concatenate all the code files into one file, and similarly for data. If the author had only one program and one dataset, this procedure presented no problem. If the author had several programs, and each program did not have a clearly identifiable beginning (e.g., a 'header') and/or ending, then it was impossible to separate the large file into its constituent components. The concatenation of data files also is pernicious (especially if the variable names are given in the code, not in the data file). Hence, for many articles we found it impossible to use the archived data and code to produce data and code that would run on a computer program. Even when data and code files were available (or when we could produce them after some cutting and pasting), we usually failed to replicate, for the same reasons we discussed at length in MMH (1101–4).

Of the 143 articles with archived entries, 18 employed software that we did not have and 8 used proprietary data – we excluded these from further consideration. For the remaining 117 articles, we attempted to reproduce all of the results, which is defined as successful replication in MMH (1094). When all was said and done, we were able to replicate only 9 of the 117 articles. It is our considered opinion that if the authors had been allowed to prepare the archived files, then we would have had much greater success in replicating articles, perhaps doubling or even tripling the proportion of replicable articles.

5. Conclusions

New technologies have decreased the costs of creating and maintaining journal archives, so presumbably the supply of articles for which data and code are available has increased. As our work shows, this potential supply is reduced when editors fail to enforce and authors do not comply with journal archiving policies. Moreover, the demand for the data and code remains negligible: not much replication is being done. Indeed, in a recent viewpoint in this journal, Hamermesh (2006) addresses this problem at length. Among other things, he notes: 'few researchers are availing themselves of the opportunities offered [to engage in replication]' because the economics profession considers replication 'as an ideal to be professed but not to be practised.' Hamermesh explores several reasons and suggests several possible solutions. One major reason for this lack of replication activity is that replication is still viewed as a parasitic activity by many researchers (Hamermesh 1997). To combat this perception, Hamermesh proposes that a few top journals commission prominent researchers to perform some replications. Obviously, these top researchers would be performing a professional service rather than pursuing their own research agendas. The obvious benefit is that if these researchers were visibly performing replications, it would be difficult for other researchers to stigmatize the activity. With the stigma removed, armies of econometrics students could easily replicate many articles. Successful replications could be mentioned in the journal's archive, unsuccessful replications could be handled by the journal's 'replication section.'

MMH showed that the archive at the *Journal of Money, Credit and Banking* failed on two counts: a large percentage of empirical articles had nothing in the archive, and most of the archived material would not replicate the published articles. The present article shows that the archive at the Federal Reserve Bank of St. Louis *Review* also fails on both counts. In short, we have attempted to replicate all the articles in the only two long-standing data+code archives in the economics profession and found that they do not support replication.

Our tables 1 and 2 show the journals' compliance with their own policies: *JAE* excepted, it is hard to square these numbers with the ostensible purpose of the archives. As far as advancing replicable research is concerned, one might suggest that these journals' archive policies are similar to the 'replication policy' of the *American Economic Review*: treating replication 'as an ideal to be professed but not to be practiced.' If archives are to be successful, they must enable successful replication of published results. So far, they have not.

We have shown that editors can, with low cost, implement and enforce the appropriate incentive mechanisms. Whether the new archives will learn from the mistakes of their predecessors and whether the old archives will mend their ways remain to be seen. It will be most interesting to wait a couple years, and see whether any of the archives is producing replicable research.

Appendix: Recommendations for an effective archive

1. The readme file lists all the replication files with a brief description of each. It clearly indicates which programs correspond to which results in the paper.
2. A data dictionary is included in the replication files. Additionally, the first program prints summary statistics on all the variables, so that subsequent researchers can be sure that they have loaded the data correctly.
3. The author provides code such that the data and code, when placed in the same subdirectory, will execute; and that the output from doing this also will be provided. The author checks to make sure that this runs correctly and produces the results in his paper.
4. Programs are commented so that users of other packages can port the code and structured to make clear which parts of the output constitute the results in the paper.
5. All data are provided in ASCII format, and the version of the code submitted to the archive calls these same ASCII files.
6. Authors provide the primary data from which the final data set is derived – the code includes all the transformations that are used to produce the final dataset.

7. The author must 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.
8. The archive, which lists each paper regardless of whether or not the paper has an archive entry, specifically states that a paper has been exempted from the requirement. If data are exempted, the code should still be required.
9. The journal issues conditional acceptance letters, with a formal acceptance letter being sent only after the data/code have been archived.
10. The archive must be mandatory.
11. Managing the archive should be an editorial function.
12. The journal should institute a replications section that stands ready to publish a single-page summary of failed replication attempts, with supporting materials placed in the archive. Successful replications can be reported in the archive.

References

Anderson, Richard, William H. Greene, B.D. McCullough, and H.D. Vinod (2008) 'The role of data & program code archives in the future of economic research,' *Journal of Economic Methodology* 15, 99–119

Ashenfelter, Orley, Robert H. Haveman, John G. Riley, and John T. Taylor (1986) 'Editorial statement,' *American Economic Review* 76, v

Bai, Jushan, and Pierre Perron (2003) 'Computation and analysis of multiple structural change models,' *Journal of Applied Econometrics* 18, 1–22

Bernanke, Ben S. (2004) 'Editorial statement,' *American Economic Review* 94, 404

Breusch, Trevor (2005) 'The Canadian underground economy: an examination of Giles and Tedds,' *Canadian Tax Journal* 53, 367–91

Breusch, Trevor, and Edith Gray (2004) 'New estimates of mothers' forgone earnings using HILDA data,' *Australian Journal of Labour Economics* 7, 125–50

— (2006) 'Replicating a study of mothers' forgone earnings in Australia,' *Journal of Economic and Social Measurement* 31, 107–25

Chapman, B., Y. Dunlop, M. Gray, A. Liu, and D. Mitchell (2001) 'The impact of children on the lifetime earnings of Australian women,' *Australian Economic Review* 7, 373–89

Cleveland, Willliam S. (1993) *Visualizing Data* (Summit, NJ: Hobart Press)

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

Donohue, John J. III, and Steven D. Levitt (2001) 'The impact of legalized abortion on crime,' *Quarterly Journal of Economics* 116, 379–420

Feigenbaum, S., and D. Levy (1993) 'The market for (Ir)reproducible econometrics,' *Social Epistemology* 7, 215–32

Foote, Christopher L., and Christopher F. Goetz (2005) 'Testing economic hypotheses with state-level data: a comment on Donohue and Levitt (2001),' Federal Reserve Bank of Boston Working Paper 05-15

Giles, D.E.A., and Lindsay M. Tedds (2002) *Taxes and the Canadian Underground Economy* (Toronto: Canadian Tax Foundation)

Gray, M., and B. Chapman (2001) 'Foregone earnings from child rearing: changes between 1986 and 1997,' *Family Matters* 58, 4–9

Hamermesh, Daniel S. (1997) 'Some thoughts on replications and reviews,' *Labour Economics* 4, 107–9

— (2006) 'Viewpoint: Replication in economics,' *Canadian Journal of Economics* 40, 715–33

Handelsblatt (2008) 'Der Download-Krieg der konomen' ('The economists' war about downloading'), 4 March

Hernandez, Javier C. (2005) 'Star ec prof caught in academic feud,' *Harvard Crimson*, 8 July 2005

Hoxby, Caroline (2000) 'Does competiton among public schools benefit students and taxpayers?' *American Economic Review* 90, 1209–38

— (2005) 'Competition among public schools: a reply to Rothstein (2004),' NBER Working Paper No. 11216

King, Gary (1995) 'A revised proposal proposal,' *Political Science & Politics* Sept, 494–9

Kleiber, Christian, and Achim Zeileis (2005) 'Validating multiple structural change models-a case study,' *Journal of Applied Econometrics* 20, 685–90

Leesom, Robert, ed. (2000) *A. W.H. Phillips: Collected Works in Contemporary Perspective* (New York: Cambridge University Press)

Levitt, Steven D. (1997) 'Using electoral cycles in police hiring to estimate the effect of police on crime,' *American Economic Review* 87, 270–90

— (2002) 'Using electoral cycles in police hiring to estimate the effect of police on crime: reply,' *American Economic Review* 92, 1244–50

Liebowitz, Stan J. (2007) 'How reliable is the Oberholzer-Gee and Strumpf paper on file-sharing?' http://ssrn.com/abstract=1014399

Mankiw, N. Gregory (2003) *Macroeconomics*. 5th ed. (New York: Worth)

McCrary, Justin (2002) 'Using electoral cycles in police hiring to estimate the effect of police on crime: comment,' *American Economic Review* 92, 1236–43

McCullough, B.D. (2004) 'Wilkinson's tests and econometric software,' *Journal of Economic and Social Measurement* 29, 261–70

— (2007) 'Got replicability? The Journal of Money, Credit and Banking Archive,' *Econ Journal Watch* 4, 326–37

McCullough, B.D., and H.D. Vinod (2003) 'Verifying the solution from a nonlinear solver: a case study,' *American Economic Review*, 93, 873–92

— (2004) 'Verifying the solution from a nonlinear solver: reply,' *American Economic Review* 94, 391–6

McCullough, B.D., Kerry Anne McGeary, and Teresa D. Harrison (2006) 'Lessons from the JMCB Archive,' *Journal of Money, Credit and Banking* 38, 1093–108

Mirowski, Philip, and Steven Sklivas (1991) 'Why econometricians don't replicate (although they do reproduce),' *Review of Political Economy* 3, 146–63

National Research Council (2002) *Access to Research Data in the 21st Century: An Ongoing Dialogue among Interested Parties. Report of a Workshop* (Washington, DC: National Academy Press)

Oberholzer-Gee, Felix, and Koleman Strumpf (2007) 'The effect of file sharing on record sales: an empirical analysis,' *Journal of Political Economy* 115, 1–42

Phillips, A.W.H. (1958) 'The relation between unemployment and the rate of change of money wage rates in the United Kingdom, 1861–1957,' *Economica* 25, 283–99

Racine, J.S. (2001) 'On the nonlinear predictability of stock returns using financial and economic variables,' *Journal of Business and Economic Statistics* 19, 380–2

Rothstein, Jesse (2004) 'Does competition among public schools benefit students and taxpayers? A comment on Hoxby (2000),' Princeton University, Education Research Section Working Paper No. 10, December

— (2005) 'Does competition among public schools benefit students and taxpayers? A comment on Hoxby (2000),' NBER Working Paper No. 11215

Spanos, Aris (1999) *Probability Theory and Statistical Inference: Econometric Modeling with Observational Data* (Cambridge: Cambridge University Press)

The Editors (2006) 'An editors' comment on "Lessons from the JMCB Archive," by B.D. McCullough, Kerry Anne McGreary and Teresa D. Harrison,' *Journal of Money, Credit and Banking* 38, 1109–10

Wulwick, N.J. (1989) 'Phillips's approximate regression,' *Oxford Economic Papers* 41, 170–88

— (1996) 'Two econometric replications,' *History of Political Economy* 28, 391–439