Studies in
Risk & Regulation

35 YEARS 1974-2009 FRASER INSTITUTE

February 2009

Check the Numbers: The Case for Due Diligence in Policy Formation

by B. D. McCullough and Ross McKitrick

Check the Numbers

The Case for Due Diligence in Policy Formation

Table of Contents

Executive summary

Empirical research in academic journals is often cited as the basis for public policy decisions, in part because people think that the journals have checked the accuracy of the research. Yet such work is rarely subjected to independent checks for accuracy during the peer review process, and the data and computational methods are so seldom disclosed that post-publication verification is equally rare. This study argues that researchers and journals have allowed habits of secrecy to persist that severely inhibit independent replication. Non-disclosure of essential research materials may have deleterious scientific consequences, but our concern herein is something different: the possible negative effects on public policy formation. When a piece of academic research takes on a public role, such as becoming the basis for public policy decisions, practices that obstruct independent replication, such as refusal to disclose data, or the concealment of details about computational methods, prevent the proper functioning of the scientific process and can lead to poor public decision making. This study shows that such practices are surprisingly common, and that researchers, users of research, and the public need to consider ways to address the situation. We offer suggestions that journals, funding agencies, and policy makers can implement to improve the transparency of the publication process and enhance the replicability of the research that is published.

Introduction

> [I]f the sums do not add up, the science is wrong. If there are no sums to be added up, no one can tell whether the science is right or wrong.
> – Donald Laming

In recent years a considerable amount of attention has been paid to mechanisms for ensuring transparency and veracity in financial reports from publicly-traded corporations. Penalties for failure to meet these requirements are based on the recognition that there is a fiduciary trust at stake when investments are solicited. The public policymaking process also involves large amounts of spending, but the documents and research reports that motivate such spending may provide no comparable guarantees of transparency and veracity. Specifically, empirical research in academic journal articles is often cited as the basis for decisions, yet such work is rarely subject to independent checks for accuracy during the peer review process, and the data and computational methods are so seldom disclosed that post-publication verification is equally rare. The relevant mechanisms for ensuring transparency in private-sector documents fall under the heading of "due diligence." This study questions whether due diligence is adequately undertaken within academia, and whether researchers themselves have allowed habits to persist that prevent it from occurring. Of particular interest to us is whether due diligence is delayed or prevented for the subset of academic studies that form a basis for public policy decisions.

This study arose out of our experiences in attempting to replicate published empirical research, as well as our observations of the way empirical research is used in public policy formation. Much of this paper documents examples of lack of transparency in academic research. Non-disclosure of essential research materials may have deleterious scientific consequences, but our concern herein is something different: the possible negative effects on public policy formation. Nobody would recommend basing policy on flawed research. Precisely to avoid such situations, the research must be demonstrably sound. If the researcher has obtained his results in accordance with the principles of the scientific method (i.e., kept a log book, clearly identified all data used, documented the computer code, etc.) then the burden on the researcher to disclose data and methods will be negligible. On the other hand, if the researcher has not followed the principles of the scientific method and has not kept an audit trail of his research, then the "research" should not be used as a basis for policy.

Scholars must have the unhindered right to publish their research and make their points of view known without fear of reprisal. But when a piece of academic research takes on a public role, such as becoming the basis for public policy decisions, then practices that obstruct independent replication, such as refusal to disclose data or the concealment of details about computational methods, prevent the proper func-

tioning of the scientific process and can lead to poor public decision making. In this study we will show that such practices are surprisingly common, and that researchers, users of research, and the public need to consider ways to address the situation.

The lack of replication work in the academic community suggests that research replication is an under-provided public good. Academic journals have not resolved the problem of non-disclosure of data. As we will show, few journals require that data be archived, and those that have such requirements do not reliably enforce them. Even fewer require authors to disclose the software they used for statistical computations. As a result, the time cost of attempting to replicate published studies is extremely high, and replication efforts are rare. The few systematic attempts of which we are aware give surprisingly strong grounds for pessimism regarding the veracity and reproducibility of much empirical research.

The term "peer review" is often invoked as a guarantor of quality. But conventional journal peer review does not typically provide a check of underlying data and findings. Public misunderstanding of this point occasionally comes to public attention when research frauds are uncovered, such as the South Korean stem cell experiments of Woo Suk Hwang. Some of the Hwang et al. papers were published in the prestigious journal *Science*. Dr. Donald Kennedy, editor of *Science*, was asked in an interview why the fraudulent results weren't detected prior to publication. He emphasized that journal peer review does not involve actual scrutiny of the underlying data and analysis:

> What we can't do is ask our peer reviewers to go into the laboratories of the submitting authors and demand their lab notebooks. Were we to do that, we would create a huge administrative cost, and we would in some sense dishonor and rob the entire scientific enterprise of the integrity that 99.9 percent of it has ... it all depends on trust at the end, and the journal has to trust its reviewers; it has to trust the source. It can't go in and demand the data books. (PBS Online Newshour, December 27, 2005)

Commenting on the same scandal, Dr. David Scadden of the Harvard Stem Cell Institute pointed out that the real review process happens after an article has been published, when other scientists try to reproduce the findings:

> [A study] is disseminated through the journal to the public. That allows the—both the lay public to then review it in terms of the interpretation by the press but also then scientists can repeat the work. And that's really the critical step that validates it. The scientific method assures that it be repeated and replicated before it is regarded as fact. (PBS Online Newshour, December 27, 2005)

This is a crucial point. Journal peer review does not generally provide any guarantee that the research results are correct, or even that they have been checked. It only signals

that a journal has decided to put results out to the scientific community for debate and examination. That is only the start of formal scientific review, which involves, inter alia, other researchers checking the data, repeating the analysis and verifying the conclusions.

For this scrutiny to take place, the data and methodology (in the form of executable computer code) must be accessible, and qualified researchers must be willing to undertake the work. In practice it is rare for scientists to make their data and code accessible, and it is rare for scientists to replicate one another's work, in part because it can be so difficult to get the data.

As a preliminary example, in a study from the Institute of Health Policy at Massachusetts General Hospital, more than a thousand graduate students and postdoctoral fellows from several scientific disciplines were asked about their experiences in obtaining data from other researchers.

> Respondents from the 50 US universities that grant the most degrees in the fields surveyed were asked about their own experiences with data withholding, the consequences of withholding, the competitiveness of their lab or research group, and whether their research received industry support.
>
> One quarter of the trainee respondents reported that their own requests for data, information, materials, or programming had been denied. Withholding was more likely to have been experienced by life scientists, by postdoctoral fellows rather than graduate students, and in settings described as highly competitive (Massachusetts General Hospital, 2006).

The same authors had earlier surveyed over 1,800 life scientists, and in a 1998 paper reported even worse findings.

> The survey showed that 47 percent of geneticists who asked other faculty for additional information, data, or materials relating to published scientific findings had been denied at least once in the past three years.
>
> Overall, 10 percent of all post-publication requests for additional information in genetics were denied; 28 percent of geneticists said they had been unable to replicate published research results because of a lack of access, and a quarter had to delay their own publications because of data withholding by their peers. Despite some speculation in earlier reports that data withholding was more common in genetics, the geneticists were no more likely to report denial of their requests than were the non-geneticists. (Massachusetts General Hospital, 2002)

Numerous examples of data secrecy and replication problems will be explored below, from across scientific disciplines. Section 1 focuses on the evidence that eco-

nomics journals do not facilitate replication of empirical research, and that successful replication is the exception, not the rule. Section 2 reviews some cases from other disciplines. We remind the reader that *most* work is *never* checked for accuracy by outside parties, so the examples point to the potential of an even more pervasive problem. Section 3 discusses possible remedies.

1. Do economics journals publish replicable research?

> Replication is the cornerstone of science. Research that cannot be replicated is not science, and cannot be trusted either as part of the profession's accumulated body of knowledge or as a basis for policy. Authors may think they have written perfect code for their bug-free software and correctly transcribed each data point, but readers cannot safely assume that these error-prone activities have been executed flawlessly until the authors' efforts have been independently verified.
> —McCullough and Vinod, 2003: 888

Current economic issues are often the subject of academic research, creating an obvious connection between economics journals and the policy process. If interesting or influential results are published, how easy would it be for an independent researcher to verify them? Not easy at all, as has been established by a considerable body of evidence.

In the *Journal of Money, Credit and Banking* (JMCB) Project, researchers Dewald, Thursby and Anderson (hereafter, "DTA") obtained a National Science Foundation grant to investigate whether published results in economics could be replicated. The focus of the investigation was the *Journal of Money, Credit and Banking*. The journal's articles were divided into two groups: a control group of 62 articles that had been published prior to July 1982, and an experimental group consisting of 95 subsequent articles that had either been accepted but not yet published, or were being refereed.

In the control group, the authors of the articles were sent a letter requesting the data and code for their articles. One third never even responded, despite repeated requests. Twenty refused to supply their data and code, only two of whom cited confidentiality of the data. Twenty-two submitted either data or code, thus of the 62 articles in this group, only 24 responded favorably to a request for data and code (we include in this the two who cited confidentiality).

Of the 92 in the experimental group, 75 responded to the letter, of whom only 68 supplied data *or* code, despite the fact that these authors were made aware, at the time of submission, that their data and code would be requested. DTA carefully inspected

the first 54 data sets, finding only eight were sufficiently complete to facilitate replication. There were many reasons for this incompleteness, but DTA noted, "... data sets were often so inadequately documented that we could not identify the variables which had been used in calculating the published empirical results" (Dewald, Thursby, and Anderson, 1986: 592). DTA also noted that failure to replicate was commonplace, even with the active assistance of the original author. All told, DTA were able to replicate only two of the 54 articles. Even when the data were complete, the usual result was that the article could not be replicated because the text did not describe every thing that had been done to the data. Only the actual code can provide such necessary detail. Hence, data alone (without code) are not sufficient to permit replication.

As a result of their investigation, DTA recommended that each journal establish an archive of data and code (Dewald, Thursby, and Anderson, 1986: 601). Researchers' disincentives to supply data and code are described in detail in Mirowski and Sklivas, 1991, and Feigenbaum and Levy, 1993. In response to DTA, the *Journal of Money, Credit and Banking* established an archive, whereby authors of accepted papers had to supply their data and code. The failure of this initiative is discussed in Anderson and Dewald, 1994 (see "Another JMCB project" below). The flagship journal of the economics profession, the *American Economic Review*, decided to ignore the above-mentioned incentive problems and adopted a "policy" that authors must supply their data and code upon request (Ashenfelter, Haveman, Riley, and Taylor, 1986). The policy notably lacked any enforcement provision: authors who refused to supply data and code did not face any sanction. It was little more than window dressing.

The American Economic Review

In 2002, B.D. McCullough and H.D. Vinod decided to test the efficacy of the *AER* replication policy. They tried to replicate the eight empirical papers published in the (then just-released) June 2002 edition of the *American Economic Review* (McCullough and Vinod, 2003). McCullough and Vinod sent letters to the eight lead authors requesting their data and code. Four of the eight refused: two provided unusable files, one replied that the data were lost, and one claimed to have the files but was unwilling to take the time to send them. McCullough and Vinod also tested two other journals with similar "replication policies" (*International Journal of Industrial Organization* and *Journal of International Economics*) and found even less compliance. These results are summarized in table 1. In light of the refusal of so many authors to provide their data and code, the *American Economic Review* adopted a new policy as of March 2004 requiring authors, as a precondition for publication, to archive on the journal's web site their data and code (*American Economic Review*, 2004: 404). According to this policy, the data and code should be sufficient to reproduce the published results.

Table 1: Do authors honor replication policies? Results from one issue of each journal

Journal	Asked to supply data and code	Actually supplied data and code
International Journal of Industrial Organization	3	1
Journal of International Economics	4	1
American Economic Review	8	4
Total	15	6

Another JMCB project

As mentioned, in response to DTA, the *Journal of Money, Credit and Banking* adopted a mandatory data and code archive. Yet only a few years later, in 1993, a subsequent editor abandoned the archiving policy as well as all the data and code that had been collected. The archive was reborn in 1996, with a policy stating that for empirical articles, the author "must provide the *JMCB* with the data and programs used in generating the reported results or persuade the editor that doing so is infeasible." McCullough, McGeary, and Harrison (2006) attempted to use the archive to determine whether a mandatory data/code archive produces replicable research.

The first thing they found was that the journal did not collect data and code for every article. Of the 266 articles published, 186 were empirical and should have had data and code archived. Only 69 articles actually had an archive entry, including 11 with only data and no code. Thus, the journal failed to collect data and code for nearly two-thirds of the empirical articles, despite its policy. Obviously, no one on the editorial board was ensuring that the policy was being followed. Of the 69 archive entries, McCullough et al. were unable to attempt replication for seven because they did not have the necessary software. Of the 62 articles they could assess, they encountered many of the experiences enumerated by DTA: shoddy programming, incomplete data, incomplete code, etc. One author who supplied incomplete data and incomplete code casually pointed out that "since there was considerable experimentation, the version of the [program in the archive] is not necessarily one that produced the results in the paper." In all, only 14 papers could be replicated. In response, the editors of the *JMCB* adopted new policies concerning the archive, and presumably monitor it to ensure better compliance.

Table 2: Archive policy compliance for various journals, 1997-2003

Journal	Number of articles that should have archive entries	Actual number of entries	Percent of compliant articles
Journal of Applied Econometrics	213	211	99
Federal Reserve Bank of St. Louis *Review*	167	82	49
Journal of Money, Credit and Banking	164	44	33
Journal of Business and Economic Statistics	312	112	36
Macroeconomic Dynamics	143	20	14

The Federal Reserve Bank of St. Louis Review

Was the archive at the *JMCB* unique or was it characteristic of archives in general? Continuing with their investigation into the role that archives have in producing replicable research, McCullough, McGeary, and Harrison (2008) turned their attention to the data/code archive for the *Review* published by the Federal Reserve Bank of St. Louis. This archive has been in existence since 1993. Of the 384 articles published during the period, 219 of them were empirical and should have had an archive entry. There were only 162 archive entries. This is a much better compliance rate than the *JMCB*, but not nearly as good as *Journal of Applied Econometrics*. Of the 162 *Review* archive entries, 29 could not be examined because they did not have the necessary software or the articles employed proprietary data. Of the remaining 133, only nine could be replicated.

Other archives

The *Journal of Business and Economic Statistics* and the *Journal of Applied Econometrics* have long had data-only archives. "Data only" is insufficient to support replication, as shown by DTA and others. One has to wonder why these journals do not simply require the authors to supply code, too. *Macroeconomic Dynamics* had a data/code archive, but it was discontinued in about 2004. These journal archives were surveyed to determine whether authors at least submitted something to the archive, i.e., to determine whether the editors were doing anything more than paying lip service to the idea of an archive.

Table 2 makes clear that most journals with archives are not seriously committed to their archives; the notable exception is the *JAE*. The minor blemishes in the *JAE*'s otherwise perfect record were in special issues, over which the archive manager had no control.

We draw two conclusions. First, economics journals do not ensure replicability, even when their own policies require it. Second, to the limited extent the question has been researched, the evidence strongly suggests that most results published in economics journals are not independently reproducible, or verifiable. Where verification is feasible in principle, the time cost to do so is typically very high, requiring multiple requests to reluctant or hostile authors, resulting in underprovision of replication studies.[1]

Should this concern policymakers? We believe it should. Where policy decisions are based on empirical results, it does not fall on a critic to prove that the results are non-reproducible, since in most cases the data and code are unavailable to do so. Instead the policymaker faces an obligation to check that the results *are* reproducible.

2. Other cases of prominent or policy-relevant research

> Especially when massive amounts of public monies and human lives are at stake, academic work should have a more intense level of scrutiny and review. It is especially the case that authors of policy-related documents like the IPCC report, *Climate Change 2001: The Scientific Basis*, should not be the same people as those that constructed the academic papers.
> —Edward Wegman et al., 2006

Above we have described a general failure to replicate samples of published economics research. We now turn to specific cases, from economics as well as other disciplines, in which policy-relevant empirical results were difficult or impossible to reproduce due to data problems or secrecy. In some cases the data were eventually obtained, but only after such a long delay as to make the results largely pointless. And in some cases the disclosure required direct government intervention. In general, the examples illustrate the obstacles facing researchers trying to replicate high-profile studies.

The Harvard Six Cities study

In 1993, a team of researchers led by D.W. Dockery and C.A. Pope published a study in the *New England Journal of Medicine* supposedly showing a statistically significant correlation between atmospheric fine particulate levels and premature mortality in six US cities (Dockery, Pope, et al., 1993). The "Harvard Six Cities" (HSC) study, as it came to be called, attracted considerable attention and has since been repeatedly cited in assessment reports, including those prepared for the Ontario government, the Toronto board of public health and the Ontario medical association. In each case the reports have used the HSC study to recommend tighter air quality standards or other costly pollution control measures.

Shortly after HSC was published, the US Environmental Protection Agency (EPA) announced plans to tighten the existing fine particle standards, based largely on the HSC findings as well as a follow-up study by the same authors for the American Cancer Society. However, other researchers who wanted to critique the findings found that the data were not available for independent inspection. There were doubts about whether the results were robust with respect to controls for smoking and educational status, but the original authors would not make their data available. In early 1994, the Clean Air Scientific Advisory Committee of the US EPA wrote to the EPA administrator asking her to obtain the data behind the study. At the same time, several groups filed Freedom of Information requests to obtain the data from the EPA. In its response, the EPA admitted it did not have the data for the HSC study since the authors had not released it (Fumento, 1997). Meanwhile, the regulatory process continued: new rules for fine particles were announced by the EPA in early 1997 (US Environmental Protection Agency, 2008).

The US House Commerce Committee asked the EPA about the availability of HSC data, but an EPA official responded that since the study was published in a peer-reviewed journal there was no need for them to obtain it. Finally, after continuing pressure, Dockery and Pope gave their data to a third party research group called the Health Effects Institute (HEI), which agreed to conduct an audit of the findings. In 2000, fully six years after the CASAC request, and three years after the new air quality regulations had been introduced, the HEI completed its reanalysis. The audit of the HSC data reported no material problems in replicating the original results, though there were a few coding errors (Health Effects Institute, 2000). However, their sensitivity analysis showed the risk originally attributed to particles became insignificant when sulphur dioxide was included in the model, and the estimated health effects differed by educational attainment and region, weakening the plausibility of the original findings (Heuss and Wolff, 2006). The HEI also found that there were simultaneous effects of different pollutants that needed to be included in the analysis to obtain more accurate results.

In this case, timely replication was not possible since the data were not available, and direct intervention by Congress was necessary to facilitate replication and critical analysis. The reassessment of the original results did not arrive in time to affect the policy decisions. The example shows that the academic community is not necessarily able to deal with a refusal to disclose research materials, and even if the research community had eventually obtained the data, the timeliness issue must be addressed.

The Boston Fed Study

The United States is experiencing an unprecedented financial crisis that has its origins in the accumulation of trillions of dollars of bad mortgage debt and related financial derivatives. The buildup of bad debt was guided by federal rule changes since the 1970s that intentionally expanded the number of low-income borrowers who could obtain a mortgage. In 1974 the US passed the Equal Credit Opportunity Act (ECOA), prescribing fines for lenders found to be discriminating against minority mortgage applicants. In 1975 the US Congress passed the Home Mortgage Disclosure Act (HMDA), requiring mortgage lenders to disclose information (including race) about their applicants. This disclosure led to concerns that minority applicants were being unfairly denied mortgages. In 1977 Congress passed the Community Reinvestment Act (CRA), requiring banks to show that they attempted to loan money to minorities, even if such loans were bad business risks. US banks typically re-sell mortgages to two large government-sponsored entities, Fannie Mae and Freddie Mac, who then repackage them and sell them to US and foreign investors. After 1977, banks began to complain that Fannie Mae and Freddie Mac wouldn't buy many of the low-standard, or so-called sub-prime mortgages they were required to provide. Although there had been political pressure on banks to increase lending to minorities, there was no legitimate justification for doing so until the Federal Reserve Bank of Boston released a now-famous working paper in 1992 entitled *Mortgage Lending in Boston: Interpreting HMDA Data*, which purported to show widespread discrimination against minorities in the Boston mortgage market. This led to a series of rapid rule changes affecting bank lending practices. These coincided with passage of the 1992 Federal Housing Enterprises Financial Safety and Soundness Act, which forced Fannie Mae and Freddie Mac to accept sub-prime loans, thus removing from the banks the risks associated with making bad loans.

It soon became possible to obtain a mortgage without verifiable income and with no down payment. Under the post-1992 rules, hundreds of billions of dollars of sub-prime loans were issued, then repackaged and sold as derivative products, leveraged many times over by investment banks. Disaster has struck in the form of collapsing housing prices, soaring default rates, and the seizing up of global credit markets.

The story, which is told in many places (see, for example, Liebowitz, 2009 and Husock, 2003), has numerous linked components. For our purposes, we focus on the role of the so-called Boston Fed Study.

The study was written by four Federal Reserve Bank of Boston economists: Alicia H. Munnell, Lynn E. Browne, James McEneaney, and Geoffrey M.B. Tootell (1992). They took loan application information from the Home Mortgage Disclosure Act data for conventional mortgage applications in the Boston area in 1990, including all 1,200 applications from blacks and Hispanics, and a random sample of 3,300 applications from whites. These data were augmented with economic information (such as credit worthiness) from the actual loan applications, as well as some economic data from the census tracts in which the house was located. The final data set was used, in the words of the report, "to test whether race was a significant factor in the lending decision once financial, employment, and neighborhood characteristics were taken into account." The study concluded, "[I]n the end, a statistically significant gap remains, which is associated with race" (Munnell et al., 1992).

Liebowitz described well what happened next:

> Most politicians jumped to support the study. "This study is definitive," and "it changes the landscape" said a spokeswoman for the Office of the Comptroller of the Currency. "This comports completely with common sense" and "I don't think you need a lot more studies like this," said Richard F. Syron, president of the Boston Fed (and now head of Freddie Mac). One of the study's authors, Alicia Munnell said, without any apparent concern for academic modesty, "the study eliminates all the other possible factors that could be influencing [mortgage] decisions." When quotes like these are made by important functionaries, you know that the fix is in and that scientific enquiry is out. (Liebowitz, 2009)

Due to the media attention it received and its almost immediate impact on government policy, the study quickly became known simply as "The Boston Fed Study." Shortly after its release, the Federal Reserve Bank of Boston issued new guidelines reminding mortgage lenders that failure to comply with the Equal Credit Opportunity Act could result in fines of $10,000 per instance, increasing the pressure for mortgage lenders to ignore credit history, down payments, and sources of income when evaluating a loan they knew would ultimately be transferred to Fannie Mae or Freddie Mac.

In 1995 Congress toughened the Community Reinvestment Act. No longer was it sufficient for a bank to show that it tried to make loans to minorities, they were forced to do so, even if the loans were bad business risks. Regulators also required banks to respond to complaints, in particular from organizations that got involved to promote mortgages for minorities, such as the Association of Community Organizations for Reform Now (ACORN) and Neighborhood Assistance Corporation of Amer-

ica (NACA). These groups soon became embedded in the financial intermediation process. In 2000, the role of NACA was described as follows:

> With "delegated underwriting authority" from the banks, NACA itself, not the bank, determines whether a mortgage applicant is qualified, and it closes sales right in its own offices. It expects to close 5,000 mortgages next year, earning a $2,000 origination fee on each. Its annual budget exceeds $10 million. (Husock, 2000)

A new industry of activist groups emerged under the toughened CRA to pressure banks into making more sub-prime loans. Meanwhile, with Fannie Mae and Freddie Mac obliged to buy the loans, not only did banks no longer bear the financial risk from making them, under the revised CRA they faced serious legal consequences for not making them.

Standards got predictably looser and looser. Entire banks dedicated themselves to issuing these loans. For example, Countrywide, once a small financial institution, rode the sub-prime wave to become the country's largest mortgage provider, at its peak providing 17 percent of the country's mortgages before its collapse in 2008. Total originations of sub-prime loans grew from $35 billion in 1994 to $160 billion in 1999, to $600 billion in 2006, and to $1 trillion as of March 2007.

It is in the light of the post-1992 sub-prime mortgage bubble that we turn to the question of the reliability of the Boston Fed Study. The results were based on a data set containing 3,062 observations. The authors released a data set that only contained 2,932 observations. A later version of the paper based on 2,925 observations was published in *American Economic Review* (Munnell et al., 1996), which had its "replication policy" in effect at that time. The authors did not provide the data necessary to replicate their published results, though later replication efforts (Day, and Liebowitz, 1998 and Harrison, 1998), yielded qualitatively similar results. Even though the study was very controversial, the *AER* editor refused to run comments (Liebowitz, 2009, footnote 1).

More than a year after the Boston Fed study was released, the first suggestion that there might be problems was published in the editorial pages of the *Wall Street Journal* (Liebowitz, 1993), where serious questions about the quality of the data were raised. In 1994, David Horne, an employee of the Federal Deposit Insurance Corporation (FDIC), examined the FDIC files of persons in the Boston Fed sample and found 26 cases in which the Boston Fed authors had classified an applicant as a rejection when the applicant had actually been accepted, or the loan had been rejected by government program administrators after bank approval had been given, or the borrower had been offered a loan but decided not to take it. Day and Liebowitz (1998) filed a Freedom of Information Act request to obtain identifiers for these observations so they could re-run the analysis without them. They also noted that the Boston Fed

authors (Munnell et al., 1992) did not use the applicant's credit score as generated by the bank, but had replaced it with three alternate indicators they themselves constructed, which Day and Liebowitz found had omitted many standard indicators of creditworthiness. Day and Liebowitz showed that simply reverting to the bank's own credit score and correcting the 26 misclassified observations caused the discrimination coefficient to drop to zero.

Harrison (1998) noted that the Boston Fed data set included many more variables than the authors had actually used. These included measures such as marital status, age, and whether the application contained information the bank was unable to verify. These variables were significant when added back in, and their inclusion caused the discrimination effects to drop to zero even without correcting the data errors noted by Day and Liebowitz.

Thus, the original Boston Fed conclusions were eventually shown to be wholly insupportable. But due to various delays these studies were not published until 1998 in *Economic Inquiry*, six years after the original study's release and at least three years after the policy changes had been made that ultimately led to today's credit crisis.

The "hockey stick" graph

The Mann, Bradley, and Hughes (1998; 1999) "hockey stick" graph, shown in figure 1, was a key piece of evidence used by the Intergovernmental Panel on Climate Change in its 2001 *Third Assessment Report* to conclude that humans are causing climate change (Working Group I, IPCC, 2001, ch. 2, fig. 2.7c and ch. 2, fig. 2.20). The graph has a striking visual effect, suggesting the Earth's climate (represented by the average northern hemisphere temperature) was stable for nine centuries prior to industrialization, then underwent a rapid warming in the 20th century. The hockey stick graph appeared five times in the *Third Assessment Report*, each time in an unusually large and colorful format compared to other data series. It was widely reproduced on government web sites around the world and played an influential role in the debates that took place in many countries between 2001 and 2004 over whether to ratify the Kyoto Protocol.

One of the key selling points of the hockey stick graph (as emphasized in the 2001 IPCC report) was its supposedly robust computational methodology and the high level of statistical significance it supposedly attained in multiple tests. However, in the underlying paper itself (Mann, Bradley, and Hughes, 1998), while mention was made of two tests, called the R^2 and RE statistics, only the RE scores were reported. The RE (Reduction of Error) score is a somewhat obscure test of the fit between a model and its data. Since there are no standard tables for it, significance benchmarks must be computed for each study using a simulation procedure called Monte Carlo analysis.

The data behind the hockey stick was not available on line as of 2003, nor was the methodology clearly described in the underlying articles. In April 2003, a Toronto businessman named Stephen McIntyre decided to try to replicate the graph, and contacted the lead author (Michael Mann) to ask for the data. Mann initially said that they were not all in one place and it would take a while to gather them up, but eventually he provided a text file containing the data set.

Over the next six months McIntyre and coauthor McKitrick studied the data and implemented the methods described in the original paper, concluding that the original results could not be reproduced. After several attempts to clarify methodological points Mann cut off all further inquiries. After McIntyre and McKitrick published a paper in late 2003 detailing numerous problems in the data and methods (McIntyre and McKitrick, 2003), Mann released a new version of his data set and some hitherto undisclosed details of his methodology. However, the new version of the data set conflicted with the description in the original paper. McIntyre and McKitrick filed a materials complaint with *Nature*, which was reviewed and upheld, leading to an order for Mann to publish a correct data listing. This appeared in July 2004, six years after the original publication. Mann et al. published a *corrigendum* and a new data archive (Mann, Bradley, and Hughes, 2004) but were not required, and indeed refused, to supply either their code or a complete mathematical description of their methods, despite

Figure 1: The hockey stick graph from the IPCC Third Assessment Report

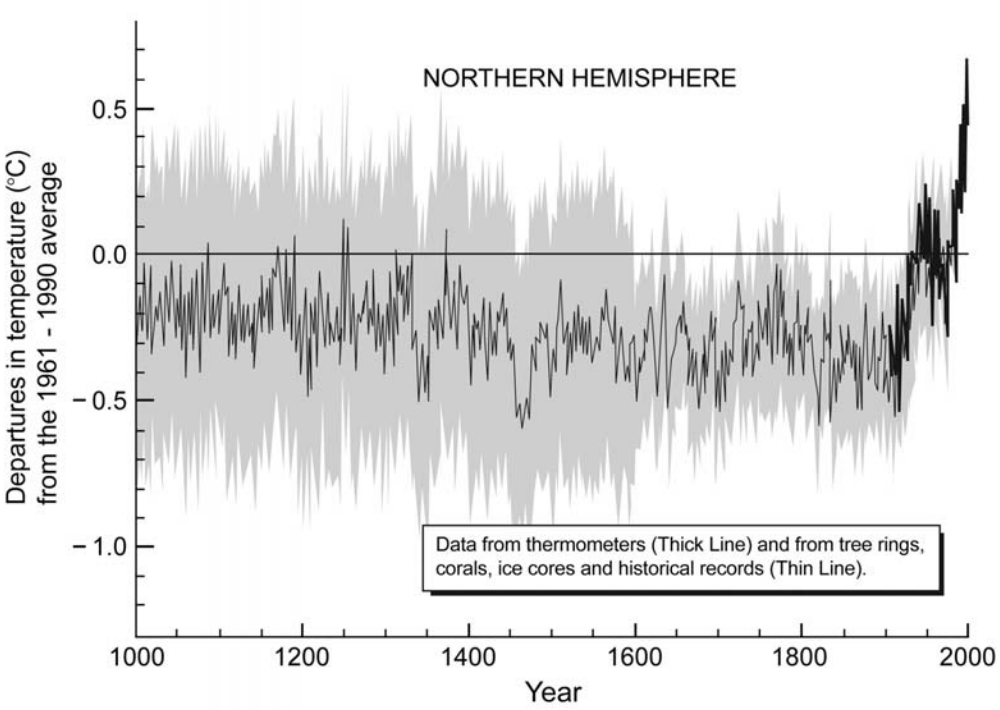
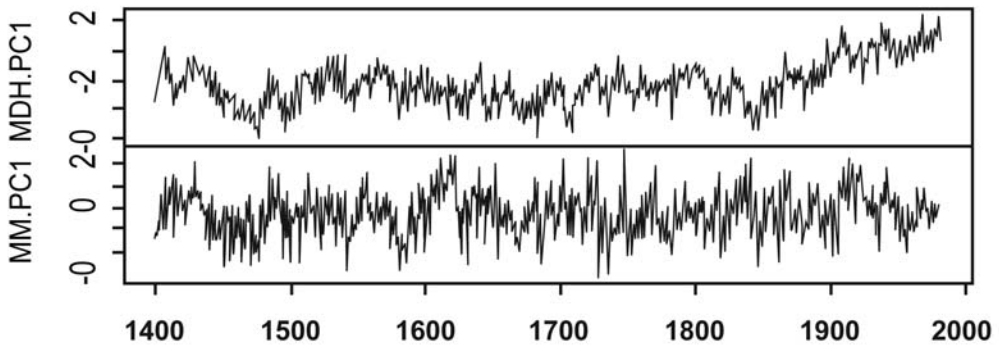


Figure 2
Top: First principal component of North American tree ring network as computed by Mann et al. (1998)
Bottom: same, but computed using standard algorithm



the claim (Mann, Bradley, and Hughes, 1998: 779) that the principal contribution of their study was their "new statistical approach" to analyze global climatic patterns.

The "new statistical approach" turned out to involve a standard method popular among applied statisticians and available in every statistical software package, called "principal components analysis" (PCA). Mann, however, wrote his own PCA code in Fortran, despite the fact that well-debugged and polished versions of the algorithm were readily available in existing statistical packages. Mann's PCA code did not give the same answer as SAS, SPSS, or other statistical packages would have.

McIntyre and McKitrick subsequently published two more studies (McIntyre and McKitrick, 2005a; 2005b) diagnosing the main error in Mann's method. Prior to the PCA step, the proxy data were transformed in such a way as to inflate the weight assigned to proxies with upward slopes in the 20th century.

Figure 2, top panel, shows the first principal component of the largest data network (North American tree rings) as computed using Mann's method. The strong hockey stick shape drives the final result (figure 1), and its overall influence on the result was said to be justified because, in PCA, the first principal component represents the dominant pattern in the data. The bottom panel of figure 2 shows the first principal component of the same tree ring data set when computed using a correct method, on the centered covariance matrix. Using correct methods, the hockey stick shape falls to the fourth principal component and is only associated with a small but controversial series of tree ring records from the western United States (the "bristlecone pines," 16 of over 400 proxy series) which prior researchers (as well as the 1995 IPCC report) had warned were invalid for the study of past climate changes.

The influence of the faulty PC method was not accounted for in the Monte Carlo algorithm, yielding an incorrect critical value for the RE score. The corrected critical

value (McIntyre and McKitrick, 2005a) showed that the reconstruction was invalid for projecting temperatures back 600 years. Also, McIntyre and McKitrick (2005a) reported that the R^2 value confirmed the revised RE score but had not been reported. They also showed that the characteristic hockey stick shape disappeared from the result with the removal of the bristlecone pines from the data set. And they reported that the exact form of the hockey stick graph could not be replicated, nor could portions of the data set be identified without the code, since they were constructed by splicing segments of PCs together at truncation points that were only disclosed in the code itself. Since Mann refused to release most of his computer code there was no way to sort out the remaining discrepancies.

The US National Science Foundation, which had funded the research, refused a request in 2003 from McIntyre to compel Mann to release his code. Likewise, *Nature* would not compel release of the code, accepting instead a verbal description of the algorithm for the *corrigendum*. In June 2005, the US House Committee on Energy and Commerce intervened to demand Mann release his code. This prompted letters of protest from, among others, the American Meteorological Society, the American Geophysical Union, and the American Association for the Advancement of Science, none of whom had ever objected to Mann's refusal to disclose his data and code in the first place. Mann released an incomplete portion of his code in July 2005, seven years after his paper had been published. Among other things, the code revealed that Mann et al. had calculated the insignificant R^2 statistics, but had failed to report it. Also, files found on Mann's FTP site showed that he had re-run his analysis specifically excluding the bristlecone pine data, which yielded the result that the hockey sick shape disappeared, but this result was also not reported.

In the context of private sector due diligence, a failure to disclose adverse performance is a misrepresentation. It is no less serious in scientific contexts.

In 2005, the House Science Committee asked the National Research Council (NRC) to investigate the controversy over the hockey stick. Prior to beginning its work, the NRC revised its terms of reference to exclude any specific assessment of Mann's work. The Energy and Commerce Committee then asked Edward Wegman, Professor of Statistics at George Mason University and Chairman of the National Academy of Sciences Committee on Theoretical and Applied Statistics, to assemble a separate panel to assess Mann's methods and results. The NRC report ended up critiquing the hockey stick anyway, noting that it failed key statistical significance tests (National Research Council, 2006: 91), relied on invalid bristlecone data for its shape (pp. 50, 106-7), used a PC technique that biased the shape (p. 106), and, like other proxy reconstructions that followed it, systematically underestimated the associated uncertainties (p. 107). The Wegman panel report was published in July 2006 (Wegman et al., 2006). It upheld the findings of McIntyre and McKitrick (p. 4). Among other

things, the panel reported that, despite downloading the materials from Mann's web site, they were unable to replicate the hockey stick results (p. 29).

The hockey stick episode illustrates, among other things, the inability or unwillingness of granting agencies, academic societies, and journals to enforce disclosure to a degree sufficient for the purposes of replication. Government intervention in this case resulted in release of essential code. Unless granting agencies and journals deal with this issue forcefully, policy makers should be prepared to accept a responsibility to act if their decisions are going to be based on the findings of unreplicated academic research.

The US obesity epidemic

In March 2004, the *Journal of the American Medical Association* published a paper by Dr. Julie Gerberding, Director of the Centers for Disease Control and Prevention (CDC), and three other staff scientists, claiming that being overweight caused the deaths of 400,000 Americans annually, up from 300,000 in 1990 (Mokdad, Marks, Stroup, and Gerberding, 2004). This study, and the 400,000 deaths figure, was the subject of considerable media attention and was immediately cited by then-US Health and Human Services Secretary Tommy Thompson in a March 9, 2004 press release announcing a major new public policy initiative on obesity, a $20 million increase in funding for obesity-related programs and a further $40 million increase the following year (US Department of Health and Human Services, 2004).

However, questions were raised almost immediately about the data and methodology used in the study, and whether it had undergone appropriate review procedures at the CDC. Less than a year later (January 2005) the same authors published a downward revision of the estimate to 365,000 (Mokdad, Marks, Stroup and Gerberding, 2005). However, other CDC staff vocally objected to this estimate as well. A group of CDC scientists and experts at the National Institutes of Health published a study three months later estimating fewer than 26,000 annual deaths were attributable to being overweight or obese; indeed, being moderately overweight was found less risky than having "normal" weight (Flegal, Graubard, Williamson, and Gail, 2005). No revision to the public policy initiative on obesity was announced in the aftermath.

The CDC soon found itself under intense criticism over the chaotic statistics and the issue of whether internal dissent was suppressed. In response, it appointed an internal review panel to investigate, but the resulting report has never been made public. Some portions were released after Freedom of Information requests were made. The report makes scathing comments about the poor quality of the Gerberding study, the lack of expertise of the authors, the use of outdated data, and the political overtones to the paper (Couzin, 2005). The report also found that the authors knew their

work was flawed prior to publication but that since all the authors were attached to the Office of the Director, internal reviewers did not press for revisions.

In light of the episode, the CDC has revised some of its pronouncements on obesity, downplaying any specific numerical estimate. However, it still links to the April 2004 *JAMA* paper from its web site and has not published the internal review. A lesson here is that agencies should not be left to audit their own research, and that audit reports should be available to the public.

The Arctic Climate Impact Assessment

In late 2004, a summary report entitled the *Arctic Climate Impact Assessment* (ACIA) was released by the Arctic Council, an intergovernmental organization formed to discuss policy issues related to the Arctic region. The council had convened a team of scientists to survey available scientific information related to climate change and the Arctic. *Impacts of a Warming Arctic: Highlights* (Arctic Council, 2004) was released to considerable international media fanfare, and prompted hearings before a US Senate committee on November 16, 2004 (the full report did not appear until August 2005). Among other things, the *Highlights* document stated that the Arctic region was warming faster than the rest of the world, that the Arctic was now warmer than at any time since the late 19th century, that sea-ice extent had declined 15 to 20 percent over the past 30 years and that the area of Greenland susceptible to melting had increased by 16 percent in the past 30 years.

Shortly after its publication, critics started noting on web sites that the main summary graph (Arctic Council, 2004, *Highlights*: 4) showing unprecedented warmth in the Arctic had never appeared in a peer-reviewed journal (Taylor, 2004; Soon, Baliunas, Legates, and Taylor, 2004), and the claims of unprecedented warming were at odds with numerous published Arctic climate histories in the peer-reviewed literature (Michaels, 2004). Neither the data used nor an explanation of the graph's methodology were made available (Taylor, 2004; Soon, Baliunas, Legates, and Taylor, 2004). When the final report was released eight months later, it explained that they had used only land-based weather stations, even though the region is two-thirds ocean, and had re-defined the boundaries of the Arctic southwards to 60N, thereby including some regions of Siberia with poor quality data and anomalously strong warming trends. Other recently published climatology papers that used land- and ocean-based data had concluded that the Arctic was, on average, cooler than it had been in the late 1930s (Polyakov et al., 2002). But while these studies were cited in the full report, their findings were not mentioned as caveats against the dramatic conclusions of the ACIA summary, nor were their data sets presented graphically.

This example indicates that empirical claims in assessment reports may need to be audited if they present new data or calculations; or to ensure that the findings are based on published, peer-reviewed journal articles (which themselves can be audited) if the mandate of the panel doing the report is confined to citing only published research. It also highlights the importance of timeliness. If a summary document is released to great fanfare, and contradictory information is quietly disclosed eight months later, the later information may not affect the way the issue was framed by the summary.

The Donato study of post-fire logging and forest regeneration

On January 5, 2006, an article entitled "Post-wildfire logging hinders regeneration and increases fire risk" appeared in *Science Express*, the pre-publication venue for accepted articles in *Science* (Donato, Fontaine, Campbell, Robinson, Kauffman, and Law, 2006a). The paper examined logging activity in Oregon's Biscuit Forest following a 2002 fire. It argued that logging reduced by 71 percent the density of viable seedlings during the recovery period, and led to an accumulation of slash on the ground, increasing potential fuel levels for future fires. The article drew attention to legislation pending before the US Congress, H.R. 4200, which mandated rapid salvage logging on federal lands following a fire. The authors concluded that post-fire logging "can be counterproductive to stated goals of post-fire forest regeneration." The article was quickly cited by opponents of H.R. 4200 as authoritative scientific evidence against it (eg., Earth Justice, 2006).

While the print edition of the paper (Donato, Fontaine, Campbell, Robinson, Kauffman, and Law, 2006b) removed the reference to H.R. 4200, the study nevertheless prompted considerable controversy. Democratic Congressman and bill co-sponsor Brian Baird published a rebuttal questioning the study's sampling methodology and relevance to the legislation (Baird, 2006). A second critique followed arguing that Donato et al. (2006b) lacked sufficient contextual and supporting information to justify their conclusions or their sweeping title (Newton et al., 2006). That logging can damage seedlings is well-known. In the case of the Biscuit Forest, since it was intended to be a rapid post-fire salvage operation during which conifer seedlings would not yet have sprouted, helicopters and elevated skylines were permitted. But because of protracted environmental litigation, the salvage operation was delayed for two years, during which time some seedlings sprouted and were thus vulnerable to damage from the logging equipment (Skinner, 2006). Hence the findings, if anything, pointed to the advantage of *rapid* post-fire salvage, the intent of the legislation.

In their response, Donato, Fontaine, Campbell, Robinson, Kauffman, and Law (2006c) acknowledged that their findings were less general than their title suggested,

but they defended their sampling methodology and conclusions. At this point their critics asked to inspect the data and the sites where the data were gathered. The authors refused to disclose this information. Following publication of the exchange in *Science*, Newton and coauthors have repeatedly requested the underlying data collected at the measurement sites, as well as the locations of the specific sample sites, so they can examine how the seedling density measurements were done. These requests have been refused by Donato and coauthors (J. Sessions, pers. comm.), as have been similar data requests from Congressman Baird (Skinner, 2006).

Abortion and crime

In 2001, Donohue and Levitt published an article in which they argued that abortion reduces the crime rate. Their theory was that "unwanted children" are more likely to grow up to be criminals than "wanted children" and that, when the former are aborted, they do not grow up to commit crimes (Donohue and Levitt, 2001). The paper attracted considerable attention at the time, as it had obvious implications in the debate about the social implications of abortion laws. In this case, Donohue and Levitt made all their data and code available. This had an important impact: when a pair of economists at the Federal Reserve Bank of Boston examined it they found a serious programming error that, when corrected, cut the effect in half (Foote and Goetz, 2005). Further details can be found in the *Wall Street Journal* (Hilsenrath, November 28, 2005) and the magazine *The Economist* (December 1, 2005). We cite the case as an example of the importance of availability of data and code. Had Donohue and Levitt not followed the scientific method and made their data and code available, the error would not have been found.

Detecting cancer

In 2002, *The Lancet* published a study by Petricoin et al. (2002), who claimed to have found a test that would almost perfectly detect the presence or absence of ovarian cancer with far greater accuracy than anything else available. The US Congress passed a resolution applauding the result and calling for more money to be channeled into the area. The test employed "mass spectrometry" to analyze blood proteins via something called the *m/z* value (Check, 2004). Petricoin et al. released part of their data set, and after examining it, two biostatisticians at the University of Maryland argued that the test might not be observing biologic differences in the blood proteins, but instead differences in the way the specimens were collected or processed (Sorace and Zhan, 2003). This conclusion was seconded and extended by biostatisticians at the M.D. An-

derson Cancer Center (Baggerly, Morris, and Coombes, 2004). The Society of Gynecologic Oncologists (SGO) then weighed in on the subject, saying, "In the opinion of SGO, more research is needed to validate the test's effectiveness before offering it to the public" (SGO, February 7, 2004).

However, the test (now called "OvaCheck") has already been commercialized. The test is owned by Correlogic Systems and has been licensed to two commercial testing laboratories, Quest Diagnostics and Laboratory Corporation of America. According to Check (2004), Correlogic claims to have refined the test. However, Correlogic refuses to release its data and its claim cannot be independently verified. A mini-symposium on the issue comprising three articles (Baggerly et al., 2005; Liotta et al., 2005; Ransohoff, 2005) was published in the *Journal of the National Cancer Institute*. A primary point of contention addressed in the symposium is whether the published results are reproducible, with Petricoin (Liotta et al., 2005) and Baggerly (Baggerly et al., 2005) taking opposing sides. Summarizing the exchange, Ransohoff concluded, "On the basis of the concerns raised by Baggerly et al. and the response of Liotta et al., it would seem that such reproducibility has not been clearly demonstrated for the pattern-recognition serum proteomics discussed herein" (Ransohoff, 2005).

OvaCheck may indeed work, but we do not know, because the data are unavailable. However, we know enough to question whether it works because Petricoin et al., to their credit, put their initial data out for inspection. Had they not done so, no debate would have been possible.

The Bellesiles affair

In 2000, to great fanfare, Knopf Publishing released *Arming America: The Origins of a National Gun Culture*. Written by Michael A. Bellesiles, then a professor of history at Emory University, the book purported to show that prior to the Civil War, guns were rare in America and Americans had little interest in owning guns. Other history professors wrote glowing reviews of the book: Garry Wills in the *New York Times Review of Books*, Edmund Morgan in the *New York Review of Books*, and Fred Anderson in the *Los Angeles Times*. The *Washington Post* did publish a critical review (Chambers, October 29, 2000), but it was a rarity. The book was promptly awarded Columbia University's prestigious "Bancroft Prize" for its contribution to American history.

Arming America was immediately recognized as having considerable social and political importance. Professional historians supplied promotional blurbs to Knopf such as: "NRA zealots beware! This splendidly subversive book will convince any sane reader that America's 'gun culture' owes little to personal self defense in its pioneer past—or even to putting meat on the table... but instead a relentlessly insistent federal government" and "Michael Bellesiles' work shifts the terms of the debate about the

gun's place in the modern United States... His research raises fundamental issues that go to the heart of widely-held but apparently erroneous assumptions about American gun culture" (*Arming America*, dust jacket).

Many of Bellesiles' claims were based on his alleged examination of over 11,000 probate inventories from the period between 1765 and 1859, which led to his much-quoted assertion that only 14 percent of colonial Americans owned firearms. Despite the political importance of the topic, professional historians did not actively scrutinize Bellesiles' thesis. Instead it was non-historians who began the process of due diligence. Stephen Halbrook, a lawyer, checked the probate records for Thomas Jefferson's three estates (Halbrook, 2000). He found no record of any firearm, despite the fact that Jefferson is known to have been a lifelong owner of firearms, putting into question the usefulness of probate records for the purpose. Soon after, a software engineer named Clayton Cramer began checking Bellesiles' sources. Cramer, who has a master's degree in history, found dates changed and quotations substantively altered. However, Cramer was unable to get academic journals to publish his findings. Instead he began sending articles to magazines such as the *National Review Online* and *Shotgun News*. He compiled an extensive list of errors, numbering in the hundreds, and went so far as to scan original documents and post them on his website so historians would check the original documents against the text of Bellesiles' book (Cramer, 2006).

At this point, Northwestern University law professor and probate expert James Lindgren published an article in the *Yale Law Review* detailing several false claims made by Bellesiles (Lindgren, 2002). In January 2002, *William and Mary Quarterly* journal published three articles in which four history professors critiqued *Arming America*, as well as Bellesiles' response. These articles, while noting some errors, stopped short of levying any serious charges against the book. However, other reviewers joined Cramer and Lindgren in their attacks on the veracity of the book, for example Joyce Lee Malcom (2001) in *Reason Magazine*, and Melissa Seckora (2001a, 2001b, 2002) in a three-part series in *National Review Online*, among others.

Bellesiles claimed to have examined hundreds of San Francisco probate records from the 1850s. When confronted with the fact that all the San Francisco probate records had been destroyed in the 1906 earthquake, Bellesiles claimed that he obtained them from the Contra Costa County Historical Society. But the Society stated that it did not possess the requisite records. Bellesiles soon resorted to *ad hominem,* claiming that the amateur critics could not be trusted because they lack credentials. Referring to Clayton Cramer, Bellesiles said, "It is not my intention to give an introductory history lesson, but as a non-historian, Mr. Cramer may not appreciate that historians do not just chronicle the past, but attempt to analyze events and ideas while providing contexts for documents" (Bellesiles, 2001). Note that Bellesiles could have, at any time, ended the controversy by simply supplying his data to his critics, something he refused to do.

At one point Bellesiles claimed he could not provide his data because all his notes had been destroyed when a fire-sprinkler system malfunctioned and flooded his university office. In April 2000, a sprinkler did flood several offices in the Emory History Department building. While many other faculty availed themselves of the library's specialized services to restore the water-damaged papers, Bellesiles did not. Instead, according to him, he took the waterlogged documents home and stored them in his attic for several months, after which he took them to his garage and spread them on the floor to dry them out in an unsuccessful attempt to salvage the documents. Belleiles' changing stories for why he was unable to produce his data were described in detail by Sternstein (2002a; 2002b).

In November 2001, Emory University demanded that Bellesiles respond to his critics. In February 2002 the University appointed an internal panel to review the case. This was followed by an external panel that focused on a narrow aspect of *Arming America*: probate records and militia counts. The panel decided that "[t]he best that can be said of his work with the probate and militia records is that he is guilty of unprofessional and misleading work. Every aspect of his work in the probate records is deeply flawed" (Katz, Gray, and Ulrich, 2002: 18). They also noted that trying to reconstruct the table that described the probate results was "an exercise in frustration because it is almost impossible to tell where Bellesiles got his information" (Katz, Gray, and Ulrich, 2002: 6). After the release of the report, on October 25, 2002, Emory University issued a press release announcing that Bellesiles had resigned his tenured position effective at the end of that year (Paul, 2002). On December13, 2002, the Columbia University Board of Trustees rescinded Bellesiles' Bancroft Prize.

Droughts in Australia

In July 2008, the Australian Bureau of Meteorology and the Commonwealth Science and Industrial Research Organization (CSIRO) released a report entitled *An Assessment of the Impact of Climate Change on the Nature and Frequency of Exceptional Climatic Events*. It received considerable media attention for what appeared to be predictions of a dramatic increase in drought. News coverage by the Australian Broadcasting Corporation began, "A new report is predicting a dramatic loss of soil moisture, increased evaporation and reduced ground water levels across much of Australia's farming regions, as temperatures begin to rise exponentially" (ABC Rural, July 7, 2008).

Shortly after its release, David Stockwell, an ecological systems modeler and Australian expatriate living in San Diego, became doubtful about whether the models had any demonstrated ability to predict known past events and whether the forecast changes were statistically significant–i.e., distinguishable from random guesses. How-

ever, neither the data nor the methodology were sufficiently well described in the report to allow him to investigate. Stockwell emailed CSIRO to request the data used for the claims in the report. The request was promptly refused. He was told on July 15, 2008, that the data would not be sent to him "due to restrictions on Intellectual Property" (Niche Modeling, July 15, 2008). About a month after Stockwell's requests began to get media and Internet attention, CSIRO changed course and released their data. Stockwell quickly found that the models were unable to replicate observed historical trends, typically generating patterns that were opposite to those in the data. Hence their predictions of future trends did not have the credibility CSIRO had claimed (Niche Modeling, August 28th, 2008). By this time, however, media interest in the report had substantially died away so the initial impression was not corrected.

File sharing

In February 2007, the *Journal of Political Economy* published a paper by Felix Oberholzer-Gee and Koleman Strumpf called "The effect of file sharing on record sales." It argued that the rise of Internet file sharing had not cut into music sales, which went strongly against conventional views and attracted considerable attention, including media coverage, due to its commercial and policy implications. Prior to its publication, the findings had been criticized by another researcher, Stan Liebowitz of the University of Texas at Dallas. He had tried to persuade the authors to correct what he argued were some errors in their analysis, and as far back as 2004 had asked to see their data to check some of their results. His initial data requests were refused, though the authors gave conflicting reasons why they would not share the data (Liebowitz, 2008).

Following publication, Liebowitz again asked for their data, but the authors did not reply. He then obtained comparable market data and found he could not replicate the findings of Oberholzer-Gee and Strumpf; instead repetition of their analysis on similar data yielded completely different conclusions. Some of the Oberholzer-Gee and Strumpf results were based on publicly available data, which Liebowitz could not replicate. In the fall of 2007 he again requested their data to find out why his results were so different. The correct response, at least in the case of publicly available data, is to provide the data and the code that produced the published results. This request was ignored again. Liebowitz then submitted a comment to the *Journal*, whose editor is Steven Levitt, mentioned above. Levitt rejected the comment, citing a report by an anonymous referee, in part because Liebowitz was unable to prove conclusively that the original results could not be replicated.

However, as Liebowitz has pointed out, he could not prove conclusively that the original results could not be replicated because the authors had refused to give him their data (Liebowitz, 2008). Even more disturbingly, the referee turned out to be Strumpf, one of the authors of the original paper! In other words, the editor had asked

an author who had concealed his data and was obviously biased against his critic to decide if the criticism should be published. While it is not without precedent for an editor to ask an original author to be a referee, it should be noted that there was a second, impartial referee, who did not have a vested interest, who had recommended publication of Liebowitz's comment. Among other things, this episode illustrates the dismal record of academic journals in ensuring proper disclosure of data, even in high profile studies.

Conclusions from examples

A key lesson of the Bellesiles affair, as well as the hockey stick debate, is that, within an academic milieu, there can be strong peer pressure not to question politically popular results, even when they are *prima facie* doubtful. The initial work of debunking Bellesiles's book fell to amateurs, while the professionals who ought to have done so instead formed a protective cheering squad. Wegman noted the same wagon-circling effect regarding the hockey stick affair. Rather than scrutinizing the result for errors, or calling out Mann for not disclosing his data and methods, scientists working in the area formed a "self-reinforcing feedback mechanism" (Wegman, 2006: 65) that made it effectively impossible for them to critically assess his work, while dismissing the efforts of outsiders who were trying to do so. For that reason, it should not be assumed that the scientific process will reliably correct erroneous research: the sociological process within science is just as likely to protect false results from scrutiny.

Nor should the fact that we present only a few examples in which famous, policy-relevant research has been refuted by outsiders to the field be taken as evidence that the problem is isolated and rare. Instead, it is rare that amateurs appear who have the time, skill, and thick skin necessary to investigate high-profile academic research, and it is rare that they can get access to the data needed to undertake such checking.

In the preceding sections we reviewed some cases in which empirical research could not be replicated by peers. Sometimes government intervention forced data disclosure. In other cases the data were never shared. The obstruction of replication efforts either thwarted independent scrutiny or delayed it so long as to make it much less relevant for the policy process. Many more examples could be cited along these lines, but what we have presented should suffice to establish our basic point. Journal articles and expert assessment reports cannot be assumed to be correct just because they went through some sort of "peer-review" process. If (as the quotation at the start of this section notes) massive amounts of public monies and human lives are at stake, academic research must be subject to *timely*, *objective,* and *independent* audit. The next section discusses what needs to be verified in an audit and delineates the disclosure required to do such checking.

3. The required disclosure for replication

Journal articles

Supposing that an article is selected for replication, the following items are the basic "checklist" that should be verified.

a. The data have been published in a form that permits other researchers to check it;
b. The data described in the article were actually used for the analysis;
c. Computer code used for numerical calculations has been published or is made available for other researchers to examine;
d. The calculations described in the paper correspond to the code actually used;
e. The results listed in the paper can be independently reproduced using the published data and methods;
f. If empirical findings arise during the analysis that are materially adverse to the stated conclusions of the paper, this has been acknowledged in the article and an explanation is offered to reconcile them to the conclusions.

One hopes that all these things are true of a scientific article. But we emphasize, once again, that they are not verified during the typical peer review process. In discussing this point with academics and government staff we have noticed two typical responses: some government staff are surprised to find out that peer review does not involve checking data and calculations, while some academics are surprised that anyone thought it did. At present, readers of a study have no way of knowing which, if any, of the above conditions hold, without doing a great deal of work checking into such things themselves. And if the data and methods are not published, such checking is effectively stymied.

The implicit replication standard of science journals is to assume that another researcher could, if desired, replicate the articles' results. But modern empirical research is too complex for such a simple assertion—and has been for a half century or more. Few journals would even attempt to publish a description of all of an article's data sources and every programming step. However, without knowledge of these details, results frequently cannot be replicated or, at times, even fully understood. Recognizing this fact, it is apparent that much of the discussion on replication has been misguided because it treats the article itself as if it were the sole contribution to scholarship. It is not. We assert that Jon Claerbout's insight for computer science, slightly modified, also applies more broadly: an applied science article is only the advertising

for the data and code that produced the published results. We would like to see a mechanism for enforcing "truth in advertising."

Checking item (d) would require some verification that appropriate software was used. As indicated in note 1, different software packages sometimes give different answers to the same problem. And in some cases, authors may have used a package that is poorly suited for the estimations being done.

In most cases, a particular scientific article is of limited interest. It may not be of any practical consequence that the conditions listed above do not all hold. The paper may only be read by a few people in a specific area and have little influence; or the result may be so unremarkable that there is no need to check it in detail. But there are times when a published study becomes very influential on public understanding and public policy. The results may be unexpected and/or momentous. In such cases it is essential, if confidence is to be placed on such studies for the purpose of setting public policy, that a process exist to verify conditions (a) to (f). It is not sufficient to depend on the scientific community eventually checking the analysis. Replication studies are quite rare at the best of times, and if conditions (b), (c), and (d) are not met, then the academic debate cannot even begin, since other researchers will not have access to the research materials.

Checking these conditions in no way intrudes upon the proper, independent functioning of the research community. Instead, it speeds up a process that needs to happen anyway, and ensures that users of research results can have confidence in the basic findings. Nor is there any presumption of guilt in calling for these things to be checked. Public corporations are routinely audited without presuming guilt. Checks and balances are entirely proper where a major public trust is exercised.

Scientific assessment reports

When policymakers need to survey research literature for the purpose of providing guidance on the state of knowledge, it is a common practice to appoint an individual or a panel to provide a scientific assessment report. Such reports (for example those from the Intergovernmental Panel on Climate Change, the US National Research Council, the Ontario Medical Association, the US Environmental Protection Agency, Health Canada, and so forth), tend to obtain a privileged standing in subsequent public debates because of the perception that they represent authoritative and unbiased surveys of information.

Therefore, it is advisable to ensure that such assessment reports are in fact authoritative and unbiased. This is usually addressed by including two "peer review" requirements: research cited in the report must be from peer-reviewed journal articles, and the assessment report itself must go through some form of peer review. However, as noted above, journal peer review is not sufficient as a quality control rule,

especially for recently published results. Moreover, peer review for assessment reports is even more problematic, since the authors of the report sometimes choose their own reviewers, and disputes with reviewers are not necessarily resolved. Considering the influence such reports have these days on domestic and foreign policy, a further audit process is needed to verify:

> g. The key conclusions of an assessment report are based on an identifiable group of published studies,
> h. For each of these studies, conditions (a) to (f) are met, and
> i. If conflicting evidence is available in the expert literature and one side is given enhanced prominence over another, the range of published evidence is nonetheless acknowledged and a credible explanation is provided as to why the one side is emphasized.

If we knew that an assessment report failed to meet these conditions, it would be imprudent to base major policies on it, just as it would be imprudent for a firm to release financial statements that failed an internal audit, or for an investor to put money into a company or project whose financial statements could not pass an audit. Since there is no other mechanism to check that journal articles and science assessment reports satisfy these conditions, if policymakers want to avoid basing decisions on research that is erroneous, fabricated, cherry-picked, or unreproducible, it will be necessary for them to institute a process that specifically checks if the conditions are met.

The responsibility of researchers

The first step in promoting transparency in empirical research is to call for researchers to adopt better disclosure practices. With the availability of computers and the Internet, publishing data and code is trivially inexpensive, though a bit time-consuming. Authors who want their work to influence the thinking of others, and even to guide public policy, should simply be prepared to publish all their data and code. Authors who want to keep their data or code to themselves should keep their results to themselves too.

The responsibility of journals

Some journals have tried to encourage replication work by actively soliciting such studies. *Labour Economics* called for replication papers in 1993, but dropped the section after 1997 because they had received no submissions. Likewise the *Journal of Political Economy* ran a section for replication work starting in 1977, but pure replications were very rare at the time the section was eliminated in 1999 (Hamermesh, 2007).

The main contribution that journals could make towards ensuring research transparency would be to adopt and enforce policies requiring archiving of both data and code at the time of publication. Data alone is not enough. It is the code that produces the results, and only the code itself can clear up ambiguities in written descriptions of methodology (see Anderson, Greene, McCullough, and Vinod, 2008).

The responsibility of granting agencies

Federal and provincial agencies that provide taxpayer monies to support research can rightly claim that the results of the research should be published. Such agencies are in a strong position to compel disclosure of data and code. In Canada, groups like the Natural Sciences and Engineering Research Council and the Social Sciences and Humanities Research Council do not currently require data and code arising from funded research to be disclosed as part of the publication process. We suggest that some attention be given by these councils to developing policies on this matter.

The caveat emptor principle

The above steps call for better practices on the supply side, but there is also the demand, or user's side. Since the government uses research in the policy making process, it needs to adopt a "buyer beware" principle, and exercise the appropriate amount of due diligence. Critical assessment of research ought to be, and indeed is, part of the policy making process. Staff who undertake replication work exist within the federal government. Recently, one of us (McKitrick) was contacted by an economist at the department of finance, who asked for the data and code used for an empirical paper published in 1999 so it could be replicated as part of the department's examination of the topic.

However, sitting members of Parliament do not have access to ministry staff for the purpose of handling detailed technical work, such as replication studies. And not all ministries necessarily have staff with the quantitative training required to handle such requests. It might therefore be advisable to identify a qualified group either within the civil service or under contract to Parliament who could establish a standard research audit procedure, to ensure that replication checks as listed in items (a) to (i) above are handled in a neutral and competent manner. But the first step is for users of research to become aware that such checks are necessary.

4. Conclusions

In recent years, there has been considerable attention paid to the question of whether financial statements and other data from corporations are adequately reviewed prior to release. An analogous question concerns the data and findings in academic papers which sometimes influence public sector decisions. Disclosure of data and code for the purpose of permitting independent replication in no way intrudes on or imperils academic freedom; instead, it should be seen as essential to good scientific practice, as well as a contribution to better public decisionmaking.

Notes

1 Note that simply being able to replicate published results does not imply that the results are technically correct, since the literature is filled with examples of different software packages giving different answers to the same problems. (See Brooks, Burke, Persand, 2001; Newbold, Agiakloglou, and Miller, 1994; McCullough and Renfro, 1999; McCullough, 1999a; McCullough, 1999b; Stokes, 2004; McCullough and Wilson, 1999; McCullough and Wilson, 2002; McCullough and Wilson, 2005).

References

American Economic Review (2004, March). Editorial Statement 94(1): 404.

Anderson, R.G., and W.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* (Nov): 79-83.

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

Arctic Council (2004). *Arctic Climate Impact Assessment. Impacts of a Warming Arctic: Highlights*. Cambridge University Press. <http://amap.no/acia/Highlights.pdf>, as of November 19, 2008. Final Report, Chapter 2 <http://www.acia.uaf.edu/>, as of November 19, 2008.

Ashenfelter, Orley, Robert H. Haveman, John G. Riley, and John T. Taylor (1986). Editorial Statement. *American Economic Review* 76(4): v.

Australian Broadcasting Corporation (2008, July 7). Drier soil, less water predicted in weather report. ABC Rural web site <http://www.abc.net.au/rural/news/content/200807/s2296263.htm>, as of November 20, 2008.

Baggerly, Keith A., Jeffrey S. Morris and Kevin R. Coombes (2004). Reproducibility of SELDI-TOF protein patterns in serum: Comparing data sets from different experiments. *Bioinformatics* 20(5): 777-785.

Baggerly, Keith A., Jeffrey S. Morris, Sarah R. Edmonson and Kevin R. Coombes (2005). Signal in noise: Evaluating reproducibility of serum proteomic tests for ovarian cancer. *Journal of the National Cancer Institute* 97(4): 307-309.

B.N. Baird (2006). Comment on "Post-wildfire logging hinders regeneration and increases fire risk." *Science* 4 August, 313 (5787), 615b.

Bellesiles, Michael A. (2000). *Arming America: The Origins of a National Gun Culture.* Knopf.

Bellesiles, Michael A. (2001, October 27). Letters to the editor. *Chronicle of Higher Education*.

Brooks C., S.P. Burke, and G. Persand (2001). Benchmarks and the accuracy of GARCH model estimation. *International Journal of Forecasting* 17(1): 45-56.

Chambers, John Whiteclay (2000, October 29). *Washington Post*.

Check, Erika (2004). Running before we can walk. *Nature* 429 (3 June): 496-497.

Couzin, J. (2005). A heavyweight battle over CDC's obesity forecasts. *Science* 308(5723), 6 May 2005: 770-771.

Cramer, C.E. (2006). Why footnotes matter: checking *Arming America's* claims. *Plagiary: Cross-Disciplinary Studies in Plagiarism, Fabrication, and Falsification* 1(11): 1-31.

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

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(4): 587-603.

Dockery D.W., C. Arden Pope, et. al (1993). An association between air pollution and mortality in six US cities. *New England Journal of Medicine* 329(24): 1753-1759.

D.C. Donato, J.B. Fontaine, J.L. Campbell, W.D. Robinson, J.B. Kauffman, B.E. Law (2006a). Post-wildfire logging hinders regeneration and increases fire risk. *Science Express*, 05 January.

D.C. Donato, J.B. Fontaine, J.L. Campbell, W.D. Robinson, J.B. Kauffman, B.E. Law (2006b). Post-wildfire logging hinders regeneration and increases fire risk. *Science* 20 January 311(5759): 352.

D.C. Donato, J.B. Fontaine, J.L. Campbell, W.D. Robinson, J.B. Kauffman, B.E. Law (2006c). Response to comments on "Post-wildfire logging hinders regeneration and increases fire risk." *Science* 4 August, 313 (5787), 615c.

Donohue, John J. III, and Steven D. Levitt (2001). The impact of legalized abortion on crime. *Quarterly Journal of Economics* 116(2): 379-420.

Earth Justice (2006).OPPOSE H.R. 4200: The Walden logging bill: An assault on public involvement, public forests, and restoration science (May 9). Earth Justice web site. <http://www.earthjustice.org/library/policy_factsheets/oppose-the-walden-logging-bill.pdf>, as of November 19, 2008.

The Economist (2005, December 1). Oops-onomics. *The Economist.*

Feigenbaum S., and D. Levy (1993). The market for (ir)reproducible econometrics. *Social Epistemology* 7(3): 215-32.

Flegal, K.M., B.I. Graubard, D.F.Williamson, and M.H. Gail (2005). Excess deaths associated with underweight, overweight and obesity. *Journal of the American Medical Association* 293(15): 1861-1867.

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.

Fumento, Michael (1997). Polluted Science. *Reason Magazine* (August-September).

Halbrook, Stephen (2000, November 5). Were the founding fathers in favor of gun ownership? *The Washington Times:* 31.

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

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

Health Effects Institute (2000). Reanalysis of the Harvard Six Cities study and the American Cancer Society study of particulate air pollution and mortality: a special re-

port of the institute's particle epidemiology reanalysis project. Health Effects Institute. http://pubs.healtheffects.org/getfile.php?u=273, as of November 17, 2008.

Heuss, Jon M., George T. Wolff (2006). Comments on "Health effects of fine particulate air pollution: lines that connect." *Journal of the Air and Waste Management Association* 56:1375-1378.

Hilsenrath, Jon E. (2005, November 28). "Freakonomics" abortion research is faulted by a pair of economists. *Wall Street Journal*: A2.

Husock, Howard (2000). The trillion-dollar bank shakedown that bodes ill for cities. *City Journal* (Winter).

Husock, Howard (2003). *America's Trillion-Dollar Housing Mistake*. Ivan R. Dee.

Intergovernmental Panel on Climate Change (2001). *Climate Change 2001: The Scientific Basis.* Cambridge University Press.

Katz, Stanley, Hanna H. Gray, and Laurel Thatcher Ulrich (2002). *Report of the Investigative Committee in the Matter of Professor Michael Bellesiles.* Emory University. <http://www.emory.edu/central/NEWS/Releases/Final_Report.pdf>, as of December 11, 2008.

Liebowitz, Stan J. (1993). A study that deserves no credit. *Wall Street Journal* (September 1): A14.

Liebowitz, Stan J. (2008). Sequel to Liebowitz's Comment on the Oberholzer-Gee and Strumpf Paper on Filesharing (July 5). <http://ssrn.com/abstract=1155764>, as of November 20, 2008.

Liebowitz, Stan J. (2009). Anatomy of a train wreck: Causes of the mortgage meltdown. In Benjamin Powell and Randall Holcomb, eds. *Housing America: Building out of a Crisis.* Transaction Publishers. <http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1211822>, as of November 28, 2008.

Lindgren, James (2002). Fall from grace: Arming America and the Bellesiles scandal. *Yale Law Journal* 111(8): 2195-2249.

Liotta, Lance A., Mark Lowenthal, Arpita Mehta, Thomas P. Conrads, Timothy D. Veenstra, David A. Fishman, Emmanual F. Petricoin III (2005). Importance of communication between producers and consumers of publicly available experimental data. *Journal of the National Cancer Institute* 97(4): 310-314.

Malcolm, Joyce Lee (2001). Concealed weapons: The controversial book *Arming America* has the facts all wrong. *Reason Magazine* (Jan): 46-49.

Mann, Michael, Raymond Bradley, and Malcolm Hughes (1998). Globalscale Temperature Patterns and Climate Forcings over the Past Six Centuries. *Nature* 392: 779-787.

Mann, Michael, Raymond Bradley, and Malcolm Hughes (1999). Northern Hemisphere Temperatures During the Past Millennium: Inferences, Uncertainties and Limitations. *Geophysical Research Letters* 26: 759-762.

Mann, Michael, Raymond Bradley, and Malcolm Hughes (2004). Corrigendum. *Nature* 430 (July 1) (105): 391-396.

Massachusetts General Hospital (2002). Study examines data withholding in academic genetics: Many genetic researchers denied access to resources related to published studies. News release (January 22). Massachusetts General Hospital. <http://www.massgeneral.org/news/releases/012202data.htm>, as of November 17, 2008.

Massachusetts General Hospital (2006). Studies examine withholding of scientific data among researchers, trainees: Relationships with industry, competitive environments associated with research secrecy. News release (January 25). Massachusetts General Hospital. http://www.massgeneral.org/news/releases/012506campbell.html, as of November 17, 2008.

McCullough, B.D. (1999a). Econometric software reliability: EViews, LIMDEP, SHAZAM and TSP. *Journal of Applied Econometrics* 14(2): 191-202; with Comment and Reply at 15(1): 107-111.

McCullough, B.D. (1999b). Assessing the reliability of statistical software: part II. *The American Statistician*, 53(2): 149-159.

McCullough, B.D., Kerry Anne McGeary, and Teresa D. Harrison (2006). Lessons from the JMCB archive. *Journal of Money, Credit and Banking* 38(4): 1093-1107.

McCullough, B.D., Kerry Anne McGeary, and Teresa D. Harrison (2008). Do economics journal archives promote replicable research? *Canadian Journal of Economics* (to appear).

McCullough, B.D., and C.G. Renfro (1999). Benchmarks and software standards: a case study of GARCH procedures. *Journal of Economic and Social Measurement* 25(2): 59-71.

McCullough, B.D., and H.D. Vinod (2003). Verifying the solution from a nonlinear solver: a case study. *American Economic Review* 93(3): 873-892.

McCullough, B.D., and Berry Wilson (1999). On the accuracy of statistical procedures in Microsoft Excel 97. *Computational Statistics and Data Analysis* 31(1): 27-37.

McCullough, B.D., and Berry Wilson (2002). On the accuracy of statistical procedures in Microsoft Excel 2000 and Excel XP. *Computational Statistics and Data Analysis* 40(4): 713-721.

McCullough, B.D., and Berry Wilson (2005). On the accuracy of statistical procedures in Microsoft Excel 2003. *Computational Statistics and Data Analysis* 49(4): 1244-1252.

McIntyre, Stephen, and Ross McKitrick (2003). Corrections to the Mann et. al. (1998) proxy data base and northern hemispheric average temperature series. *Energy and Environment* 14(6): 751-771.

McIntyre, Stephen, and Ross McKitrick (2005a). Hockey sticks, principal components and spurious significance. *Geophysical Research Letters* 32(3) L03710 10.1029/2004GL021750 12 (February 2005).

McIntyre, Stephen and Ross McKitrick (2005b). The MM critique of the MBH98 northern hemisphere climate index: update and implications. *Energy and Environment* 16: 69-99.

Michaels, P.J. (2004). A long-term perspective. *World Climate Report* (November 30). <http://www.worldclimatereport.com/index.php/2004/11/30/a-long-term-perspective/>, as of November 19, 2008.

Mirowski, Philip, and Steven Sklivas (1991). Why econometricians don't replicate (although they do reproduce). *Review of Political Economy*, 3(2): 146-163.

Mokdad, A.H., J.F. Marks, D.F. Stroup, and J.L. Gerberding (2004). Actual Causes of Death in the United States, 2000. *Journal of the American Medical Association* 291(10): 1238-1245.

Mokdad, A.H., J.F. Marks, D.F. Stroup and J.L. Gerberding (2005). Correction: Actual causes of death in the United States, 2000. *Journal of the American Medical Association* 293(3): 293-294.

Munnell, Alicia H., Lynn E. Browne, James McEneaney, and Geoffrey M.B. Tootell (1992). *Mortgage Lending in Boston: Interpreting the HMDA Data*. Federal Reserve Bank of Boston Working Paper #92-1 (October).

Munnell, Alicia H., Lynn E. Browne, James McEneaney, and Geoffrey M.B. Tootell (1996). Mortgage lending in Boston: Interpreting the HMDA data. *American Economic Review* 82(1): 25-54.

Newbold, P., C. Agiakloglou, and J. Miller (1994). Adventures with ARIMA software. *International Journal of Forecasting* 10(4): 573-581.

Newton, M., S. Fitzgerald, R. R. Rose, P.W. Adams, S. D. Tesch, J. Sessions, T. Atzet, R. F. Powers, C. Skinner (2006). Comment on "Post-wildfire logging hinders regeneration and increases fire risk." *Science* 4 August, 313 (5787), 615a.

National Research Council (2006). Surface temperature reconstructions for the past 1200 years. National Academies Press.

Niche Modeling web site (2008, July 15). CSIRO data policy: Go pound sand. <http://landshape.org/enm/csiro-data-policy-go-pound-sand/>, as of Nov. 20, 2008).

Niche Modeling web site (2008, August 28th). Comparison of models and observations in CSIRO/BoM DECR. <http://landshape.org/enm/comparison-of-models-and-observations-in-csiro-decr/>, as of November 20, 2008.

Oberholzer-Gee, Felix, and Koleman S. Strumpf (2007). The effect of file sharing on record sales. *The Journal of Political Economy* 115(February): 1-42.

Paul, Robert A. (2002). Michael Bellesiles resigns from Emory faculty October 25, 2002. News Release. <http://www.news.emory.edu/Releases/bellesiles1035563546.html>, as of December 11, 2008.

PBS Online Newshour (2005, December 27). Stem cell scandal. <http://www.pbs.org/newshour/bb/science/july-dec05/scandal_12-27.html>, as of Nov. 17, 2008.

Petricoin, Emanuel F. III, Ali M. Ardekani, Ben A. Hitt, Peter J. Levine, Vincent A. Fusaro, Seth M. Steinberg, Gordon B. Mills, Charles Simone, David A. Fishman, Elise C. Kohn, and Lance A. Liotta (2002). Use of proteomic patterns in serum to identify ovarian cancer. *The Lancet* 359(9306): 572-577.

Polyakov, I., et al. (2002). Trends and Variations in Arctic Climate Systems. *Eos, Transactions* American Geophysical Union, 83, 547548.293(15): 1861-1867.

Ransohoff, David F. (2005). Lessons from controversy: Ovarian cancer screening and serum proteomics. *Journal of the National Cancer Institute* 97(4): 315-319.

Seckora, Melissa (2001a). Disarming America. *National Review Online* (October 15). <http://www.nationalreview.com/15oct01/seckora101501.shtml>, as of Dec. 11, 2008.

Seckora, Melissa (2001b). Disarming America, Part II. *National Review Online* (Nov. 26). <http://www.nationalreview.com/nr_comment/nr_comment112601.shtml>, as of December 11, 2008.

Seckora, Melissa (2002). Disarming America, Part III. *National Review Online* (January 29). <http://www.nationalreview.com/nr_comment/nr_comment012902.shtml>, as of December 11, 2008.

D. Skinner (2006). The Donato-Law fiasco mixing politics and science: Alchemy at OSU. *Evergreen* (Winter 2006-07): 4-31.

Society of Gynecological Oncologists (SGO) (2004, February 7). Statement regarding OvaCheck. Position statement. SGO.

Soon, W, S. Baliunas, D. Legates, and G. Taylor (2004). What defines the arctic? A discussion of the Arctic Climate Impact Assessment. Tech Central Station. TCS Daily web site (December 20). <http://www.techcentralstation.com/122004F.html>, as of November 19, 2008.

Sorace, James M., and Min Zhan (2003). A data review and re-assessment of ovarian cancer serum proteomic profiling. *BMC Bioinformatics* 4(24).

Sternstein, Jerome (2002a, April 9). Are Michael Bellesiles's critics afraid to say what they really think? *History News Network.*

Sternstein, Jerome (2002b, May 20). "Pulped" fiction: Michael Bellesiles and his yellow note pads. *History News Network*.

Stokes, Houston (2004). On the advantage of using two or more econometric software systems to solve the same problem. *Journal of Economic and Social Measurement* 29(1-3): 307-320.

Taylor, G. (2004). What's going on with the arctic? Tech Central Station. TCS Daily web site (November 22). <http://www.techcentralstation.com/112204A.html>, as of November 19, 2008.

Working Group I, Intergovernmental Panel on Climate Change (IPCC) (2001). *Climate Change 2001: The Scientific Basis*. Cambridge University Press for the Intergovernmental Panel on Climate Change <http://www.grida.no/climate/ipcc_tar/wg1/figspm-1.htm>, as of November 18, 2008.

US Department of Health and Human Services (2004). Citing "dangerous increase" in deaths, HHS launches new strategies against overweight epidemic. News release (March 9). US Department of Health & Human Services. <http://www.hhs.gov/news/press/2004pres/20040309.html>, as of November 19, 2008.

US Environmental Protection Agency (2008). Particulate matter: PM standards. US EPA. http://www.epa.gov/air/particles/standards.html, as of November 17, 2008.

Wegman, Edward et al. (2006). Ad hoc report on the "hockey stick" global climate reconstruction. US House of Representatives Committee on Energy and Commerce (July).

About the authors

Bruce D. McCullough is a Professor of Decision Sciences at Drexel University in Philadelphia, with a courtesy appointment to the Department of Economics. He specializes in the accuracy of statistical and econometric software, and the replicability of economic research. He has authored or co-authored over fifty scholarly publications, and holds associate editor positions at five journals in the fields of statistics and economics. His work is largely responsible for convincing several top economics journals to adopt mandatory data and code archives in recent years.

Ross McKitrick is an Associate Professor of Economics at the University of Guelph, where he focuses on environmental economics, and a Senior Fellow of the Fraser Institute. He has published scholarly papers on topics ranging from the economic theory of pollution regulation to statistical methods in paleoclimatology. In 2003 his (coauthored) book *Taken By Storm: The Troubled Science, Policy and Politics of Global Warming* won the $10,000 Donner Prize for the best book on Canadian Public Policy. Professor McKitrick has testified before the US Congress and the Canadian Parliamentary Finance and Environment Committees. In 2006 he was one of 12 experts from around the world invited to give a briefing to a panel of the US National Academy of Sciences on paleoclimate reconstruction methodology.

Acknowledgements

The authors would like to thank Andrew Miall, Gilles Paquet, William Leiss, James MacKinnon, Bill Robson, Cliff Halliwell, Finn Poschmann, Willie Soon, John Sessions, participants at a 2005 workshop on data disclosure at the Canadian Economics Association (organized by Bill Robson of the CD Howe Institute), and numerous other readers and anonymous referees whose comments, critiques, and input helped improve the paper. All remaining errors or omissions are our responsibility.

Publishing information

Fraser Institute digital publications are published from time to time by the Fraser Institute to provide, in a format easily accessible online, timely and comprehensive studies of current issues in economics and public policy.

Distribution

These publications are available from http://www.fraserinstitute.org in Portable Document Format (PDF) and can be read with Adobe Acrobat® or with Adobe Reader®, which is available free of charge from Adobe Systems Inc. To down-load Adobe Reader, go to this link: http://www.adobe.com/products/acrobat/ readstep2.html with your Browser. We encourage you to install the most recent version.

Ordering publications

For information about ordering the Fraser Institute's printed publications, please contact the publications coordinator

- e-mail: sales@fraserinstitute.org.
- telephone: 604.688.0221 ext. 580 or, toll free, 1.800.665.3558 ext. 580
- fax: 604.688.8539

Media

For media enquiries, please contact our Communications Department

- telephone: 604.714.4582
- e-mail: communications@fraserinstitute.org

Disclaimer

Copyright

ISSN

1918-8323 (online version)

Date of issue

February 2009

Editing, design, and production

Kristin McCahon and Lindsey Thomas Martin

Cover

Design by Bill Ray. Cover images: Sous les dossiers © Karl Bolf, Fotolia; red tape © Stockbyte, Punchstock; Balance © Alon Othnay, Fotolia; Question mark © Stephen Coburn, Fotolia; Treehugger © Leah-Anne Thompson, Fotolia; Stop © Alexander Samoilov, Fotolia

About the Fraser Institute

Our vision is a free and prosperous world where individuals benefit from greater choice, competitive markets, and personal responsibility. Our mission is to measure, study, and communicate the impact of competitive markets and government interventions on the welfare of individuals.

Founded in 1974, we are an independent research and educational organization with locations throughout North America and international partners in over 70 countries. Our work is financed by tax-deductible contributions from thousands of individuals, organizations, and foundations. In order to protect its independence, the Institute does not accept grants from government or contracts for research.

菲沙研究所的願景乃一自由而昌盛的世界，當中每個人得以從更豐富的選擇、具競爭性的市場及自我承擔責任而獲益。我們的使命在於量度、研究並使人知悉競爭市場及政府干預對個人福祉的影響。

Nous envisageons un monde libre et prospère, où chaque personne bénéficie d'un plus grand choix, de marchés concurrentiels et de responsabilités individuelles. Notre mission consiste à mesurer, à étudier et à communiquer l'effet des marchés concurrentiels et des interventions gouvernementales sur le bien-être des individus.

تتمثل رؤيتنا في وجود عالم حر ومزدهر يستفيد فيه الأفراد من القدرة على الاختيار بشكل أكبر، والأسواق التنافسية، والمسؤولية الشخصية. أما رسالتنا فهي قياس، ودراسة، وتوصيل تأثير الأسواق التنافسية والتدخلات الحكومية المتعلقة بالرفاه الاجتماعي للأفراد.

Nuestra visión es un mundo libre y próspero donde los individuos se beneficien de una mayor oferta, la competencia en los mercados y la responsabilidad individual. Nuestra misión es medir, estudiar y comunicar el impacto de la competencia en los mercados y la intervención gubernamental en el bienestar de los individuos.

Editorial Advisory Board