The Case for Due Diligence When Empirical Research is Used in Policy Formation

Bruce McCullough
Drexel University

Ross McKitrick
University of Guelph

First Draft: May 2006 Comments Welcome

Abstract

Academics are well aware that "peer review" at journals is an ambiguous process. For empirical research, it does not, as a rule, provide any warranty that the data and code were published or even correctly described, or that the results are reproducible. We review a series of examples from economics and other disciplines which illustrate surprisingly pervasive barriers to independent replication. In some prominent cases, policy decisions were based on studies where the authors refused to release data and/or code to other researchers. When business and investment decisions are made in the private sector, deliberate non-disclosure of data and methods has serious (and possibly criminal) consequences. For this reason, due diligence procedures to ensure unhindered, independent review of data and calculations are well-established. Yet in public policy formation, far larger decisions are made with no guarantee of disclosure or reproducibility of the empirical basis. As a remedy, we propose an agency to be managed under joint government and academic supervision that would provide an independent, unbiased audit of empirical research papers being used for policy formation, in which the audit would be limited to ensuring the data, methods and results are accurately disclosed, and the results are reproducible.

1 Introduction

In recent years a considerable amount of attention has been paid to mechanisms for ensuring that financial reports from corporations meet standards of "full, true and plain" disclosure. Penalties for failure to meet these requirements are based on the nature of the fiduciary trust at stake when securities are offered for sale. Yet policymaking processes that can easily involve much larger usage of public money provide no comparable guarantees for the research underpinning such decisions. The relevant mechanisms in the private sector fall under the heading of "due diligence". We propose herein an analogous concept for use in the public sector.

This essay arose in part out of our experience in attempting to replicate published empirical research, as well as our observations of the way empirical research is used in public policy formation. The process of academic research must strike a balance between independence and relevance. Where the policymaking process interacts with the research process, the concept of academic freedom does not relieve researchers or policymakers of the need to ensure that empirical research findings are transparent and easily reproducible, and assessments or summary reports provide accurate and balanced surveys of the relevant literature.

These are not assured under the current systems that govern the publication of academic research and its use in policy formation. Few journals require archiving of data, and authors are almost never required to disclose their code. As a result, the time cost of attempting to replicate studies is extremely high, and replication efforts are rare. The few systematic attempts of which we are aware give no grounds for optimism regarding the veracity and reproducibility of much empirical research. And beyond these concerns, the process of canvassing scientific literature into major assessments for policymakers suffers from a lack of objective oversight, conflicts of interest and cherry-picking.

With regard to scientific journals, there is a widespread misperception that "peer review" constitutes a careful check of underlying data and findings. This is not true, nor do journals claim it to be true. The public misconception has come to public attention most recently in light of the discovery of a major fraud in the South Korean stem cell laboratory 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 a recent interview why the fraudulent results weren't detected prior to publication. He rejected the suggestion that journal peer review should 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.[1]

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

[1] http://www.pbs.org/newshour/bb/science/july-dec05/scandal_12-27.html. Accessed January 6, 2005.

> 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.[2]

This is a crucial point. Journal peer review does not guarantee the correctness or validity of research results, 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 formal scientific process to actually happen, the data and methods must be published. As well, qualified researchers must be willing to undertake the validation and replication work. There is no guarantee of either condition holding true. It is rare for journals to require authors to archive their data, and rarer still to require publication of computational code sufficient to permit replication. It is also quite rare for scientists to replicate one another's work.

As to the withholding of data, in a recent 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.[3]

The same authors had earlier surveyed over 1800 life scientists, and in a 1998 paper reported similar 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

[2] Scadden, op. cit.

[3] http://www.massgeneral.org/news/releases/012506campbell.html

genetics, the geneticists were no more likely to report denial of their requests than were the non-geneticists.[4]

Numerous examples of data secrecy could be cited at this point, from across scientific disciplines. The above examples concern medicine and the life sciences, but the problems are pervasive. The next section presents some more examples.

Since transparency and reproducibility are essential to the reliability of scientific results, especially those on which major policy decisions may be based, and since neither journals nor funding agencies seem to be in a position to address this troublesome situation, we propose that a mechanism be established by the major users of scientific research, in particular the policymaking community. The mechanism must provide a means for efficient, routine and transparent due diligence audit, while in no way threatening the independence of the scientific enterprise. Our proposal is outlined in Section 3.

2 Examples of published studies or reports that failed disclosure and replication checks

2.1 *The Journal of Money, Credit and Banking* (JMCB) Project

Dewald, Thursby and Anderson [4] (hereafter, "DTA") obtained an NSF 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 responded in the negative, saying that the could not (for various and sundry reasons) supply their data and code – only two of whom cited confidentiality of the data. Twenty-two submitted either data or code (we conjecture that fewer than 22 supplied *both* data *and* code, but DTA did not break these data out – the reason we make this distinction is that, as shall become apparent, data alone is insufficient to permit replication). Of the 65 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 95 in the experimental group, 75 responded to the letter, of whom 68 supplied data *or* code. These authors were made aware, at the time of submission, that their data and code would be requested. Of the 95 articles in this group, 68 responded favorably to the request. A key finding of this experiment is that authors are more inclined to respond favorably to such requests prior to the publication of the article.

[4] http://www.massgeneral.org/news/releases/012202data.htm

Yet, that data and/or code was supplied does not imply that the research was replicable. To assess this question, DTA carefully inspected the first 54 datasets, seeking to determine whether the data were sufficiently complete that it might be possible to replicate the published results. Only eight of the datasets were sufficiently complete. There were many reasons for this incompleteness, but DTA ([4], p. 592) 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." 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. This is because the article cannot possibly describe every single thing that is 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 ([4], p. 601) recommended that each journal establish an archive of data and code. An archive avoids the obvious incentive problems of relying on the honor system (*i.e.* commitment to the scientific method). Researchers' disincentives to supply data and code are described in detail in [13] and [5]. 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 success of this archive, and the extent to which it was utilized by other researchers is discussed in [1]. In sharp contrast, 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 [2]. The policy notably lacked any enforcement provision (would authors who refused to supply data and code endure any sanction? No.) and could reasonably be characterized as a triumph of form over substance.

2.2 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 June 2002 edition of the *American Economic Review* [11]. 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 procure 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. 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 [15] requiring authors, as a precondition for publication, to archive on the journal's web site their data and code. According to this policy, the data and code should be sufficient to reproduce the published results. This raises the question of whether data and code archives are sufficient to ensure replicable research.

Journal	Asked to supply data and code	Actually supplied data and code
IJIO	3	1
JIE	4	1
AER	8	4
total	15	6

Table 1: Do Authors Honor Replication Policies? Results from One Issue of Each Journal

2.3 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 clearly 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 [8] attempted to use the archive to determine whether a mandatory data/code archive produces replicable research.

The first thing they found was the 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. Thus, the journal failed to collect data and code for nearly 2/3 of the empirical articles, despite the journal's clear policy. Obviously, no one on the editorial board was checking to see whether the policy was being followed. Indeed, eleven archvie entries contained only data and no code.

Of the 69 archive entries, they could not attempt replication for seven because they did not have the necessary software, and so could not assess these articles. Of the 62 articles they could assess, they encountered many of the experiences enumerated by DTA: shoddy programming, incomplete data, incomplete code, etc. In a remarkable display of hubris, one author who supplied incomplete data and incomplete code nonetheless pointed out that "since there was considerable experimenation, the version of the [program in the archive] is not necessarily one that produced the results in the paper." In all, only fourteen 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.

2.4 The Federal Reserve Bank of St. Louise *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 [9] turned their attention to the data/code archive for the *Review* published by the Federal

Journal	Art	Ent	Compl
JAE	292	290	99
FSLR	244	137	56
JMCB	193	69	36
JBES	342	121	35
MD	143	20	14

Table 2: Lifetime compliance. "Art" is number of articles that should have archive entries, "Ent" is the actual number of entries,"Compl" is the percent of compliant articles

Reserve Bank of St. Louis. This archive has been in existence since 1993. Of the 384 articles published during the period, 245 of them were empirical and should have had an archive entry. There were only 138 archive entries. This is a much better compliance rate than the *JMCB*, but still barely exceeds 50%. Of the 138 articles, they could not examine 29 because they did not have the necessary software or the articles employed proprietary data. Of the remaining 117, only nine could be replicated.

2.5 Other Archives

The *Journal of Business and Economic Statistics* and the *Journal of Applied Econometrics* have long had data-only archives. As noted, "data only" is insufficient to support replication. One has to wonder why these journals do not simply require the authors to supply code, too. *Macroeconomic Dynamics* has had a data/code archive. These 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.

The above table 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 not the fault of the archive manager. The two articles which lack archive entries were in special issues, over which the archive manager had no control.

In sum, there is no evidence that any economics journal regularly produces replicable research, and there is much evidence to the contrary. We note in passing 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: [3], [14], [10], [6], [7], [34].

2.6 The Harvard Six Cities Study

In 1993 a team of researchers led by D.W. Dockery and C.A. Pope III published a study in the *New England Journal of Medicine* [16] apparently showing a statistically significant correlation between atmospheric fine particulate levels

and premature mortality in six U.S. cities. The "Harvard Six Cities" (HSC) study, as it came to be called, attracted considerable attention and has since been heavily 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 pollution control measures.

In the US, the Environmental Protection Agency (EPA) announced plans to tighten the existing fine particle standards based in part on the HSC findings as well as a follow-up study by the same authors for the American Cancer Society (ACS). However, other researchers objected that the data were not available for independent inspection, and there were doubts whether the results were robust to controls for smoking and educational status. 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 their response the EPA admitted they did not have the data for the HSC study, since the authors had not released it [17]

The House Commerce Committee asked the EPA to obtain the data, but an EPA official responded that since the study was published in a peer-reviewed journal there was no need to do so. Finally, after continuing pressure, Dockery and Pope agreed to release their data to a US research group called the Health Effects Institute (HEI) to conduct a reanalysis. In 2000, fully six years after the CASAC request, the HEI completed its reanalysis. Their audit of the HSC data [18] found no material problems, though there were a few coding errors. They were able to replicate the original results reasonably closely using the same statistical model. Sensitivity analyses were generally supportive of the original results, but they concluded educational attainment was an important explanatory factor not fully accounted for in the original publications, and that there were simultaneous effects of different pollutants that needed to be included in the analysis to obtain more accurate results.

A key lesson here is that secondary audits of influential studies can, in principle, be done, but there is no established mechanism to ensure it happens when it needs to.

2.7 The Hockey Stick Graph

The Mann et al. ([19], [20]) "hockey stick" graph was a key piece of evidence used by the Intergovernmental Panel on Climate Change (IPCC, [35]) in its 2001 Third Assessment Report (TAR) to conclude that humans are causing climate change. The graph has a striking visual effect, suggesting the Earth's climate was stable for nine centuries prior to industrialization, then underwent a rapid, continuing warming in the 20th century. The hockey stick graph appears five times in the TAR and each time it is presented in an unusually large and colorful format compared to other data series. One of the key selling points of the hockey stick graph (as emphasized in the IPCC Report) was its supposedly robust methodology and high level of statistical significance in multiple tests.

However, in the underlying paper itself ([19]), while mention was made of R^2 and RE statistics, only the latter was reported. The RE (Reduction of Error) score is a somewhat obscure test statistic with no exact or asymptotic distribution: critical values must be generated using Monte Carlo analysis.

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, which were not available on-line. Mann initially said that the data were not all in one place and would take a while to gather together, but eventually 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, and concluded that the original results could not be reproduced. After several attempts to clarify methodological points Mann cut off all further inquiries. After a paper was published in late 2003 detailing numerous problems in the data and methods ([22]) Mann released a new version of his data set and released some hitherto undisclosed details of his methodology. However the new version of the data set conflicted with the original description as published. McIntyre and McKitrick filed a materials complaint with Nature, which was upheld after review. Mann et al. were instructed to publish a Corrigendum and provide a new data archive ([21]) but were not required, and indeed refused, to supply their statistical estimation procedures, despite the claim ([19], p. 779) that the principal contribution of their study was the new statistical approach they had developed to analyze global climatic patterns.

McIntyre and McKitrick then published two more studies ([23],[24]) demonstrating that the underlying results, to the extent they could be emulated based on published descriptions, did not exhibit the robustness the authors had claimed. They showed that, prior to computing principal components, the proxy data were decentered in such as way as to inflate the weight assigned to proxies with upward slopes in the 20th century in the first principal component, and substantially inflate the explained variance attributable to them. This nonstandard PC method was not accounted for in the Monte Carlo algorithm, yielding an incorrect critical value for the RE score. Based on a corrected critical value [23] showed that the reconstruction was not statistically significant in its earliest portion. Also, [23] reported that the R^2 value was only 0.02 in the earliest portion of the reconstruction, confirming the inference from the revised RE score. They also showed that the characteristic hockey stick shape depended on the inclusion of a small but controversial series of tree ring records (the "bristlecone pines", 16 of over 400 proxy series) which earlier researchers had warned were invalid for the study of past climate changes. They reported that the exact form of the hockey stick graph could not be replicated, and since Mann refused to release most of his computer code there was no way to explain why not.

In June 2005 the US House Committee on Energy and Commerce intervened to demand Mann release his code. He released an incomplete portion of it in July 2005, seven years after his paper was published. Among other things the code revealed that Mann et al. had calculated a verification R^2 statistics of

0.02, but failed to report it. Among other files found on Mann's FTP site was a folder showing that he had re-run his analysis without the bristlecone pines and had demonstrated that the hockey stick shape disappeared from the results, overturning the conclusions, yet he had withheld this finding.

2.8 The US Obesity Epidemic

In March 2004 the *The Journal of the American Medical Association* published a paper [25] by Dr. Julie Gerberding, Director of the US Centres for Disease Control (CDC), and three other staff scientists, claiming that being overweight caused the deaths of 400,000 Americans annually, up from 300,000 in 1990. 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, 20004 press release announcing a major new public policy initiative on obesity (http://www.hhs.gov/news/press/2004pres/20040309.html).

However, questions began being raised almost immediately about the data and methodology used in the study, and whether it had gone through 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 [26]. 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 3 months later [28] 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.

The CDC found itself under intense criticism over the chaotic statistics and the issue of whether internal dissent was suppressed. They appointed an internal review panel to investigate the matter, 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 [27]. 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 changes or 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.

2.9 The Arctic Climate Impacts Assessment

In late 2004 a summary report entitled "The Arctic Climate Impact Assessment" [29] (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. The Highlights document (http://amap.no/acia/Highlights.pdf) was released to considerable inter-

national media fanfare prior to the full report being made available, which was not published until late August, 2005.

Among other things the panel concluded in the highlights document 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 15-20% over the past 30 years and that the area of Greenland susceptible to melting had increased by 16% in the past 30 years.

Almost immediately after its publication, critics started noting on web sites that the summary graph (page 4 of the Highlights Document) showing unprecedented warmth in the Arctic contradicted the major published Arctic climate histories in the peer-reviewed literature, yet provided no citation to a source nor an explanation of how it was prepared ([30],[31]). When the final report was released some 8 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 at present than it was in the late 1930s ([33]). But while these studies were cited in the full report their findings were not mentioned as a caveat against the dramatic conclusions of the ACIA nor were their data sets presented graphically. Also, recent publications had shown that Greenland was, on average, undergoing a downward trend in average temperature, and sea ice extent was close to its long term mean, and subject to such large natural variability as to rule out finding the current measurements of sea ice coverage unusual ([30],[32]). These studies were not presented in such a way as to balance the presentation. And since the ACIA provided no citations to publications or data sources none of its claims to the contrary could be backed up.

3 A Mechanism for Research Audits

3.1 Journal Articles

The following items are the basic "checklist" that an audit should verify with regards to an empirical journal article.

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 on the basis of 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 is acknowledged and an explanation is offered to reconcile them to the conclusions.

One hopes, indeed assumes, that all these things are true of a scientific article. But the point to emphasize is that they are not, as a rule, verified during the peer review process. If any or all of these conditions fail to hold, readers typically have no way of knowing without doing a great deal of work checking into such things themselves. And if the data and methods are not published, such efforts are effectively stymied.

The implicit replication standard of science journals is to assume (or, perhaps, pretend?) that the quality of the research underlying its articles is of sufficiently high quality that another researcher could, if desired, replicate the articles' results. In fact, 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. But, 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. Can such "advertising' be misleading? And, if so, does there exist a mechanism for enforcing "truth in advertising"?

In most cases a particular scientific article is of limited interest. It may not be of any consequence that conditions (a—f) 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 sometimes studies are published which are 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—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 setting up a public office to check these things. Public corporations are routinely audited without

presuming guilt. Checks and balances are entirely proper where a major public trust is exercised.

3.2 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 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. But, 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 critical comments 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—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 credible explanation is provided as to why one side is emphasized.

If we knew that an Assessment Report failed to meet of 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 is necessary to have a process that specifically checks if the conditions are met.

3.3 How the Process Would Work

We propose the creation of an Office of Research Audit (ORA), to be housed in the federal government. In Canada the most likely place to put it would

be Industry Canada, which has responsibility for granting councils and other science oversight functions. In the US the ORA might be housed in the Office of Management and Budget.

In order to allay fears of politicization or abuse of process, it would be worth considering having the ORA managed by a board appointed by a combination of groups including the government itself, academic organizations (such as the Royal Society of Canada, the US National Academy of Sciences, and the American Statistical Association), and industry, including associations representing the accounting and audit professions such as the Association of Chartered Certified Accountants. The ORA would be have one or more professional statisticians on staff, as well as managerial, legal and secretarial support.

The ORA would respond to qualified requests for audit of scientific studies. A "qualified" request would be one originating from a legislator: in Canada, an MP or Senator; in the US a member of Congress or the Administration. The requestor would need to show that the study is relevant and influential on an ongoing policy debate.

Upon receiving a journal article for audit, the ORA would send a letter to the author(s) indicating the opening of an audit at the request of a policymaker. The letter would request the author supply all data and code sufficient to permit replication of the results within a specified time period. The letter would indicate the exact conditions to be examined (a–f), and would state that if any discrepancies are discovered or problems arise in the replication work, the authors will be notified confidentially and given full opportunity to respond and, if necessary, correct the publication record prior to the completion of the audit.

If the authors are unable or unwilling to supply data and/or code, the audit would end with submission of a report indicating that the results cannot be verified or reproduced and the study should not be used for policymaking purposes.

Assuming the authors supply workable data and code, the audit would not offer any judgment about the correctness of the study or its conclusions, it would only verify conditions (a–f). This would involve checking data and cited sources, which may or may not always be possible, especially if it is a unique or novel data set (e.g. survey data). Where a data set is assembled drawing on known archives (e.g. CANSIM) the auditors would attempt to verify the provenance of the data. Otherwise the audit would check what can, feasibly, be checked, and report on the results.

The ORA would either have sufficient statistical staff internally, or would rely on a roster of qualified consultants, to run code and verify the output against the results reported in the paper. For item (f), the auditors would examine conventional, default statistics to see if any obviously adverse results were obtained but not reported. For instance, if time series data are used and the code produces a Durbin-Watson statistic of (say) 0.2, but this is not reported, it would be a failure on condition (f). Likewise if a standard significance test (e.g. F) in a regression model is generated in the default settings and shows insignificance, but this is not reported, then condition (f) would not be met. Some judgment would be required in this case, however, since there might be

test statistics that a researcher could legitimately claim not to know about, and hence not to have checked during specification testing. This would go beyond the function of auditing, and would properly fall into the domain of professional methodological debate.

The case of an assessment report would be somewhat more complex. The report would need to be read through so as to compile a list of journal article citations that support all the main conclusions. This would require some judgment, and might involve interacting with the referring legislator to ensure the important conclusions are identified.

The list of citations might be short enough that all could be audited. If the list is too long, a random sampling could be done, or a triage list generated with only the top dozen or so selected.

For checking condition (i), a notice could be published indicating the list of citations, and calling for open public comment on whether contradictory evidence was available in comparable journals, at the time the report was being prepared. Condition (i) does not imply that an assessment report must always avoid taking a position on controversial debates, it only stipulates that it cannot cherry-pick, or omit material evidence and thereby deprive the reader of an accurate sense of an issue.

The result of an audit will be a report indicating which of conditions (a–i) were met. If substantial problems arise the report would need to conclude that the study should not be used for policy planning and debates. If minor problems arise, these should be noted with a caution (if necessary) issued regarding the study. If no problems are found the report would not endorse the study in question, it would merely indicate that no problems were encountered in auditing these specific conditions.

References

[1] 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* **76**, 79-83

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

[3] 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

[4] 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

[5] Feigenbaum, S. and D. Levy. (1993) "The Market for (Ir)Reproducible Econometrics." *Social Epistemology*, 7:3, 215-32.

[6] 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

[7] McCullough, B. D. (1999), "Assessing the Reliability of Statistical Software: Part II," *The American Statistician*, **53** (2), 149-159

[8] McCullough, B. D., Kerry Anne McGeary and Teresa D. Harrison (2006) "Lessons from the JMCB Archive" *Journal of Money, Credit and Banking* (to appear)

[9] McCullough, Bruce D., Kerry Anne McGeary and Teresa D. Harrison (2005). "Lessons from the JMCB Archive." J*ournal of Money, Credit and Banking*, forthcoming.

[10] 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

[11] 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.

[12] McCullough, B. D. and H.D. Vinod (2004). "Reply to Comment." *American Economic Review* 94(1) 391—396.

[13] Mirowski, Philip and Steven Sklivas. (1991) "Why Econometricians Don't Replicate (Although They Do Reproduce)." *Review of Political Economy*, 3:2, 146-163.

[14] Newbold, P., C. Agiakloglou and J. Miller, "Adventures with ARIMA Software," *International Journal of Forecasting* **10**, 573-581

[15] Editorial Statement, *American Economic Review*,March 2004 94(1) p. 404.

[16] Dockery D.W., C. Arden Pope, et. al (1993) "An Association Between Air Pollution And Mortality In Six U.S. Cities.' " *New England Journal of Medicine* 329(24) 1753-1759.

[17] Fumento, Michael (1997) "Polluted Science" *Reason Magazine* August-September

[18] Health Effects Institute (2000) "Reanalysis of the Harvard Six Cities Study and the American Cancer Society Study of Particulate Air Pollution and Mortality. " http://www.healtheffects.org/Pubs/Rean-ExecSumm.pdf.

[19] Mann, Michael, Raymond Bradley and Malcolm Hughes (1998) "Global-scale Temperature Patterns and Climate Forcings over the Past Six Centuries. " *Nature* 392, 779-787. 391—396.

[20] 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.

[21] Mann, Michael, Raymond Bradley and Malcolm Hughes (2004) "Corrigendum " *Nature* 430, July 1 2004, 105. 391—396.

[22] 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.

[23] McIntyre, Stephen and Ross McKitrick (2005a) "Hockey Sticks, Principal Components and Spurious Significance. " *Geophysical Research Letters* Vol. 32, No. 3, L03710 10.1029/2004GL021750 12 February 2005.

[24] 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.

[25] 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.

[26] 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.

[27] Couzin, J. (2004) "A Heavyweight Battle Over CDC's Obesity Forecasts " *Science* 308, 6 May 2005, 770—771.

[28] 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.

[29] Arctic Council (2004) *Arctic Climate Impact Assessment* Highlights Report available at http://amap.no/acia/Highlights.pdf (accessed May 2, 2006) Final Report Chapter 2 available at http://www.acia.uaf.edu/(accessed May 2, 2006)

[30] Taylor, G. "What's Going on with the Arctic? " available at http://www.techcentralstation.com/112204A.html (accessed May 2, 2006)

[31] Soon, W, S. Baliunas, D. Legates and G. Taylor (2004). "What Defines the Arctic? A Discussion of the Arctic Climate Impact Assessment. " available at http://www.techcentralstation.com/122004F.html (accessed May 2, 2006)

[32] Pielke, Roger A. Sr. (2004) "Is Arctic Sea Ice Melting? " available at http://ccc.atmos.colostate.edu/blog/?p=7. (accessed May 2, 2006)

[33] Polyakov, I., et al. (2002) "Trends and Variations in Arctic Climate Systems. " *Eos, Transactions* American Geophysical Union, 83, 547548.293(15) 1861—1867.

[34] 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

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