NBER WORKING PAPER SERIES

REPLICATION IN ECONOMICS

Daniel S. Hamermesh

Working Paper 13026
http://www.nber.org/papers/w13026

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
April 2007

I thank Dwayne Benjamin, Bernd Fitzenberger and Gerald Oettinger for helpful suggestions on an earlier draft, the authors of several of the studies cited here for useful clarifications of their views on the controversies in which they were involved, and several editors for their experiences at their journals. The views expressed herein are those of the author(s) and do not necessarily reflect the views of the National Bureau of Economic Research.

Replication in Economics
Daniel S. Hamermesh
NBER Working Paper No. 13026
April 2007
JEL No. A14,B41,C59

ABSTRACT

This examination of the role and potential for replication in economics points out the paucity of both pure replication -- checking on others' published papers using their data -- and scientific replication -- using data representing different populations in one's own work or in a Comment. Several controversies in empirical economics illustrate how and how not to behave when replicating others' work. The incentives for replication facing editors, authors and potential replicators are examined. Recognising these incentives, I advance proposals aimed at journal editors that will increase the supply of replication studies, and I propose a way of generating more scientific replication that will make empirical economic research more credible.

Daniel S. Hamermesh
Department of Economics
University of Texas
Austin, TX 78712-1173
and NBER
hamermes@eco.utexas.edu

1. Introduction—What Is Replication?

Economists treat replication the way teenagers treat chastity—as an ideal to be professed but not to be practiced. Why is this? How much replication is done by economic researchers? What are the incentives/disincentives to engage in replication exercises? These are positive questions. Does our current treatment of replication lead to credible knowledge about human behaviour? Does it do so in a manner that is socially optimal? These are normative questions.

Even before these questions can be answered, we need to clarify what we mean by replication in the context of economic research. One dictionary defines replicate as "duplicate, repeat, as in a statistical experiment;" another defines it as "to make or do something again in exactly the same way."[1] This clarity exists only at the level of dictionaries. Indeed, in a recent lawsuit involving economists the definition of replication constituted one of the two central issues, with the judge granting the defendant summary judgment by accepting the dictionary definition.[2] Here I will refer to the dictionary definition as *pure replication.*

In economists' common parlance the idea of replication clearly goes beyond this. A useful taxonomy was provided by the psychologist John Hunter (2001), who described *statistical replication*—different sample, but the identical model and underlying population. Statistical replication is only marginally relevant for us: I cannot imagine, for example, anyone taking a different sample from a major data set to examine the same model that has already been studied using those data. Presumably the original author used the entire data set, subject to excluding cases because of item non-response or disqualifying characteristics. One could, however, repeat the same lab experiment on a different set of undergraduate subjects, and that is the closest we can approach statistical replication. *Scientific replication*—different sample, different population and perhaps similar, but not identical model—appears much more suited in type to our methods of research and, indeed comprises most of what economists view as replication.

In economic research examining the same question and model using the underlying original data set amounts to pure replication (and in one of the controversies discussed below,

Antonovics and Goldberger, 2005, did exactly that). In the context of research by members of our profession scientific replication also merits substantial attention. Using this taxonomy, I examine the history of, facts about and prescriptions for pure replication in Section II. Section III focuses on scientific replication, particularly on internal replication—examining multiple sets of data within one study. Throughout I concentrate on the efforts that have been made to ensure replication and replicability, and I examine the incentives for replication facing the agents in the market for scholarly research in economics.

2. Pure Replication

2.1. Recent History

Pure replication depends on the availability of all the information from the project that is to be replicated. Its potential importance was demonstrated to the profession in the mid-1980s by a project undertaken by the editor of the *Journal of Money, Credit and Banking* (Dewald *et al*, 1986). He sought data sets and documentation from authors of recently published and accepted manuscripts from his *Journal*, received them from a disturbingly small fraction (about one-third), and attempted to replicate the authors' results. In some ways the results were encouraging: Replication uncovered what the authors believed were many minor errors, but only in a few cases were the mistakes severe enough to alter the original work's qualitative conclusions. Of course, one wonders whether those data sets that were provided were not self-selected from the upper tail of quality of empirical work, so that the apparent relative lack of major errors underestimates the general severity of difficulties with published studies.

Even without positive selection, the results of the *JMCB* project should be disturbing to empirical researchers, to those theorists who pay attention to empirical work and, most important, to the public generally, whose inchoate notions about economic phenomena are in the end partly formed by evidence generated through economic research. The findings clearly disturbed the editors of what is by far the most widely-read refereed economics journal, the *American Economic Review*. Shortly after publishing the report on the *JCMB* experiment, the editors

mandated that authors of accepted articles pledge to make their data sets available upon request unless the data were proprietary. Published articles never were required to state explicitly that the data were available, but a generalised editorial statement made that fact clear to the profession.

Beginning with the 2005 volume the *AER* altered the policy to take advantage of technologically induced economies of scale by requiring authors of empirical papers to submit their data sets for inclusion on a special website maintained by the *Review*. Compliance with this requirement has been prompt and complete.[3] Authors can opt out of the requirement if the data set they used was proprietary and/or confidential. Regrettably, very few other journals (the *Journal of Applied Econometrics* being a rare exception) have yet adopted this technological advance.

Other than the *AER* none of the top-ranked general journals in economics explicitly requires provision of the data; and my perusal of the journals more generally suggests such statements are indeed quite rare. In my own subfield of labour economics, however, by the early 1990s two of the three leading publication outlets began requiring authors to include a statement about the availability of their data in the Acknowledgement footnote (beginning with the 1990-91 volume of the *Industrial and Labor Relations Review (ILRR)*, and with the 1991 volume of the *Journal of Human Resources (JHR)*). With the example of the *AER* confronting economic researchers, it is clear that authors today are expected to make their data available and that they can expect to field inquiries about their data.

With explicit statements of the availability of data and the development of the ethos that pure replication should be facilitated by authors of empirical studies, it is worth examining whether in fact these new tools for putting economic research on a more scientific basis have been taken up by economists. To examine this I sent an email survey to authors of all 139 empirical studies published in the *ILRR* and *JHR* in 2002-2004. Each author was asked how many times s/he had received requests for the data used in the published study (and whose availability was advertised on the first page of the article).

The first two rows of Table 1 present the results of the survey of experiences at these two specialised journals. First, one should note that the response rate was unusually high. While unit non-response may be non-random, it is difficult to conjure up arguments why the non-respondents may have been self-selected from the upper tail of articles in terms of requests for data. Absent such arguments, I shall assume that the respondents are taken randomly from the distribution of empirical articles in these journals along the dimension of requests. The mean number of requests for data in each of these two specialised journals was just one. More important, 67 percent of authors of articles in the *ILRR*, and 54 percent of those in the *JHR*, never received a request for data. The histogram in Figure 1 combines both journals' experiences. While most papers received no data requests, a few papers did receive a substantial number.[4] The conclusion from these data is that, despite the effort to make data available, few researchers are availing themselves of the opportunities offered.

Perhaps the low demand for data and the small supply of replication studies suggested by the first two rows of Table 1 arises from the relatively low readership of these journals and their specialised nature. To examine this possibility I sent the same email questionnaire to authors of empirical articles published in 1999 and 2000 in the *AER*, whose circulation is over three times that of any other refereed journal in our field.[5] The response rate here too was excellent—83 percent. A few of the responses may have been exaggerated (one author responded "over 100 times"), so that the mean of nine requests may be biased up; but data from the median article were requested three times, and only 22 percent of the authors received no requests for their data.

A fair conclusion from these little surveys is that, except for articles published in the most visible outlets, data sets are simply not requested for any purpose. One wonders whether even those that are requested are used for scholarly purposes or simply for pedagogy: One of my own data sets (Hamermesh and Biddle, 1994) has been requested over twenty times, but all but one of these have been for use as an example in course work in econometrics or labour economics.

Requests for classroom or textbook use are important for developing future scholars; the data sets are clearly being used, but hardly for pure replication for scientific purposes.

2.2. Incentives in the Market for Replication—Positive Aspects

One might consider the market for replication as containing three types of agents: Editors of scholarly journals, who decide what to publish and what requirements about data to impose on authors of empirical studies; the authors of such studies; and other researchers, especially those who might potentially be interested in replicating a particular published work. Each set of agents has different goals, and changes in the technology of doing research and publishing alter their ability to meet those goals. In this sub-section I examine for each agent how the possibility of replication might affect behaviour and how that interaction has changed with technology.

Journal editors presumably have one simple goal: Publishing articles that by their scholarly novelty and perhaps contemporary relevance maintain the current readership's interest and attract new readers. To achieve this goal the editor needs to make sure that the research is credible. This is less difficult for theoretical work: A properly chosen referee can check the lemmas and theorems and recommend rejection, or at least revision, if there are mistakes. No referee of empirical papers can or will re-estimate the models that the author has written up, so that the editor must rely on the author's credibility as a researcher and the general believability of the results. In almost all instances the editor is thus at the mercy of the author and can only hope that the authors' incentives are compatible with the journal's.

Technological changes, in the form of greater portability of data and the frequent use of readily available standardised data sets, have probably made the editor's job easier. Although such cases are very rare and less common than they should be, electronic communication does enable the editor to insist on verifying, or having an assistant verify some particular calculation to satisfy his/her concerns about the submission.[6] Even with lower costs of verification, however, it is unlikely that the editors will take the time or, in many cases, have the ability to verify the results of an empirical submission.

Every empirical paper that I have published over the last forty years has involved repeated estimation and re-estimation of the underlying models. No study sprang Athena-like from my brow or that of my computer—all involved re-specifications, obtaining additional data to extend the model, and other extra work. Before the advent of the personal computer this meant keeping detailed chronological records of the computer output that I had generated—an econometric equivalent of a bench researcher's lab book. With a personal computer the pile of output has been replaced by electronic files marked by date/time of production. Maintaining these records and the underlying data files has always taken time, and it would be much easier for me to be sloppy (or even sloppier than I am) in doing so. Nonetheless, the cost of keeping records has dropped over time.

The benefits of careful record-keeping and maintaining one's data files appear to have risen over time. As Ellison (2002) has shown, the number of rounds through which an economics paper goes before publication has increased over the past thirty years, and the time between initial submission and eventual acceptance at the leading journals has more than doubled. That being the case, knowing that multiple revisions will be requested, and recognising that resurrecting a particular result from the distant, often multi-year past may be difficult, empirical researchers today have greater incentives to maintain careful records.

Publishing a mistake has always left researchers open to general opprobrium from their fellows, and if the mistake is viewed as deliberate—as reflecting falsified data—can lead to the scholarly equivalent of Mennonite shunning (Kevles, 1998), including the loss of funding and position. With instant gossip through email and blogs, the ease and speed with which one's mistakes might subject one to sanctions have increased. With any reasonable loss function, and most scholars are particularly risk averse along this dimension, changing technology has increased the benefits of careful documentation and maintenance of one's data sets for this reason too.

For most scholars the payoff is in the influence of one's ideas—having other scholars base their work on those ideas, having students learn from them, and (an old man's perhaps vain hope) having public policy influenced by them (perhaps particularly important in the case of economics). The incentives we face here are clear: Our ideas are unlikely to be taken seriously if our empirical research is not credible, so that the likelihood of positive payoffs to our research is enhanced if we maintain our data and records and ensure the possibility of replication. Technology has not altered these incentives—they have always been there. Here too, however, the greater ease of communication worldwide may have enhanced these returns, particularly in the areas of influencing policy and stimulating students.

The third agent in the market for any empirical research is the worldwide set of researchers who might seek to replicate the study. For any individual the costs of replication are substantial—obtaining the data set and/or computer code and understanding what the author did. More important is the opportunity cost—with teaching and other duties, time for research is limited, so why spend it on replication if one has one's own ideas and data on which to test them? For most of us the return to spending time on replication is not great. First, most of the empirical studies are going to have at most only marginal impacts on economists' thinking. That is perhaps why we see such a difference between the experiences reflected in the first two rows of Table 1 and the third row. Second, assuming authors are careful, the likelihood of the replication providing substantial new knowledge rather than minor corrections is small—but the counter-example of Feldstein (1974; 1982) and Leimer and Lesnoy (1982), probably the best-known replication issue in this profession in the past forty years, suggests that this is not always so. As that example showed, and as theory would indicate, the likelihood of somebody attempting replication rises with the visibility of the published study and its author, and decreases with the visibility of the potential replicating author. Under those circumstances the benefits of replicating are greater, and the costs are lower.

Technology has diminished the costs of providing the materials necessary for replication at the same time that changes in the publication process in economics have increased the benefits to authors of maintaining the records that might make replication possible. Even without journals requiring posting of data sets, or acknowledging the availability of data, I have no doubt that there would be greater possibilities for pure replication today than in the 1970s. So were the requirements for making data available necessary when they were imposed? More important, are they still necessary?

I had a paper accepted by the *AER* shortly after it instituted its policy of accepting empirical articles conditional on the author acknowledging his/her willingness to make the data available. I was really angry about this new requirement, as I felt it increased the cost of doing empirical relative to theoretical work at a time when, in my view, too much useless theory was already being published. Why not impose a similar tax on theorists, e.g., require submission of all the scribbled notes that led up to the work, or at least every last detail of proofs? Require theorists to submit photographs of blackboards on which they chalked down their ideas in front of colleagues! I calmed down after a bit, recognising both that I could not do anything about the requirement and that the rigors it imposed on me might improve the quality of my work. Today I have no doubt that this requirement and similar subsequent ones hastened the move toward an equilibrium in which empirical researchers behave in a way that is based on the expectation, or at least the fear, of having their work replicated.

I doubt that removing the current strictures that make replication easier would return us to the old equilibrium in which authors who potentially might replicate published work could not expect to obtain the data they desire. The cost of maintaining records is small. Also, I hope empirical researchers now recognise that it is in their own interest to maintain their data sets and make them available when asked. If nothing else, a few researchers' continuing willingness to make data available spills over onto others' behaviour, so that it is unlikely that removing requirements would shift the current equilibrium. Nonetheless, increasing the possibility of

replication by maintaining these requirements or, better still, providing a centralised depository for data as in the new *AER* policy, seems especially desirable.

Does this availability mean that pure replication will take place? The survey evidence provided above suggests that in most cases replication is a threat that might keep potential cheaters honest rather than a common practice. Indeed, two attempts to institutionalise replication work with which I am familiar died from neglect. The *Journal of Political Economy* ran a section entitled "Confirmations and Contradictions" from 1977 to 1999, occasionally publishing what were in most cases comments using new data, not pure replications. *Labour Economics*, which began publication in 1993, stated in its opening issue that it explicitly welcomed replication studies. After a few years, however, the invitation was dropped, despite the Editors' commissioning a short piece (Hamermesh, 1997) touting the virtues of replication. As the then-Editor wrote, there was a, "…lack of interest: we simply got no submissions. There is a structural lack of interest in replication."[7] A similar lack of response is reported by the Editor of the Replications section of *Empirical Economics*.

An optimistic explanation for the apparent disinterest is that major errors (as opposed to the minor mistakes found by Dewald *et al*, 1986) are relatively infrequent. Editors have a clear bias in favor of publishing papers in which maintained hypotheses are refuted (DeLong and Lang, 1992). Perhaps authors' interest in writing up the results of replications is only spurred when a central qualitative result of the original study is contradicted; and perhaps editorial decisions are motivated by the "gotcha" mentality that pervades such diverse aspects of modern life as Presidential politics and celebrity-watching.

2.3. Some Normative Conclusions Based on Replication Controversies

Economists cannot use the social-science equivalent of genetically identical laboratory rats: People do not behave that way, and, in any event, we are looking for social behaviour not the purely biological responses of organisms. Nonetheless, making replication possible offers the social virtue of allowing those findings that seem most important to be verified or refuted directly

on the particular data set used to generate them rather than examined later on a different and perhaps less appropriate set of data. If a finding is specious, better to have its props knocked out from under immediately than have it chipped away at slowly. This is especially important given how avidly the media pick up unusual findings (and how avidly some of us seek to publicize those findings, even highly preliminary ones, in the media). Given the incentives outlined above, I doubt that much pure replication will take place; but anything that aids it is praiseworthy.

To facilitate replication, if one receives a request for one's data sets or code, one should comply fully and speedily, and should make the data available in a readily usable fashion. Among the few requests for data that I have made over the years, most were handled expeditiously and completely; but one was complied with in name only, with the data being essentially useless to anyone except the original author.

Despite the paucity of replication studies, a number of Comment-Reply pairs involving at least in part pure replications have been published, and the controversies implicit in them have gripped the attention of many members of the profession (appealing, perhaps, to the same prurient interest as mud-wrestling). I briefly review four recent examples, with the sole purpose of providing instruction on how to behave in such situations, not to judge who was right (or even if there was a "right" in the controversy). In reviewing them one should remember that the replicating author usually views him/herself as on the attack. Usually too, as noted above, the replicating author is more junior and sees the opportunity to make his/her reputation through discrediting a more senior economist's highly visible work. With these incentives it is all the more important to take a gentle, restrained, professional tone in the comment.

Leimer and Lesnoy (1982) made the bald statement, "This paper presents new evidence that casts considerable doubt on Feldstein's conclusion." While measured, but perhaps stronger than the work merited (although they did catch an important error), the statement is neither gratuitous nor *ad hominem*. In their long comment on Card and Krueger's (1994) evidence suggesting the absence of negative effects of higher state minimum wage rates on employment,

Neumark and Wascher (2000) used both the original and new data, and concluded that, "… the payroll data raise serious doubts about the conclusions CK drew from their data…." Antonovics and Goldberger (2005) engaged in pure replication to examine a study of the role of mother-child transfers of human capital (Behrman and Rosenzweig, 2002) and concluded, "We have seen that the results are not robust and that the policy inference may be misguided." Again, this is a fairly mild conclusion given the force of the arguments the authors had presented about problems with the underlying data.

Perhaps the best recent example of how to write a Comment based on a replication is Easterly *et al* (2004), commenting on a recent paper (Burnside and Dollar, 2000), whose data set was the third most frequently requested among those *AER* studies included in the sample reported in Table 1. The Comment extended the earlier data set to some additional countries and a few additional years, so that it falls somewhere between pure replication and scientific replication. Its main result was to demonstrate that the original finding—that the amount of foreign aid a developing nation receives interacts with good macroeconomic policy to induce growth but does nothing absent such policy—did not seem to be robust to the addition of relatively few data points.

If one complies with a request for data and shortly thereafter finds an article circulating that claims to have destroyed the principle findings of one's paper by challenging or at least casting considerable doubt upon them, what should one do? If a mistake was actually made, admit it honestly and immediately and move on to set out the importance of the error for your fundamental conclusions, thus "limiting the damage" while also advancing closer to understanding the phenomena under consideration. The best model for this admission is Feldstein's (1982) "Reply," the first sentence of which was, "I am embarrassed by the programming error that Dean Leimer and Selig Lesnoy uncovered but grateful to them for the care with which they repeated my original study."

Following in order the Comments discussed above, consider the Replies. Card and Krueger (2000) responded by analysing the new Neumark-Wascher data and obtaining additional data in a different manner from their original data set (whose method of collection was the replicating authors' main concern) and concluded that, "… the non-representative sample… produced [Neumark and Wascher's] anomalous results." Behrman and Rosenzweig (2005) reworked the original data set and examined a new one, but concluded, using even stronger terminology than in the Comment, "We show that the focus of AG's comments is misguided, given the policy issues." The Burnside and Dollar (2004) Reply is a modern model of scholarship; but like the other Replies it concluded that additional re-specifications lent stronger support to the original work.

By the time final versions of the Comment and Reply are in print and belong to the ages, replicating and original authors in these major controversies usually rise to appropriate standards of scientific discourse (whether voluntarily or at the insistence of an editor is unknown). When these significant scientific controversies play out in the press or in unpublished drafts, however, the comments and replies often are phrased in more strident tones, One example comes from a highly publicised (*Wall Street Journal*, October 24, 2005, page 1) controversy about Hoxby (2000), on which in an advanced, but as yet unpublished draft, Rothstein (2005) wrote, "… Hoxby has not provided the precise data set from which her published results were derived." Hoxby (2005) countered and made a crucial point about the process of empirical research today:

> "The original data simply do not exist and for a very good reason. I was trained to … [write] code that takes research … from the raw data to estimation. Such code may create intermediate datasets…, but they are replaced every time the code is run. …. Obsolete datasets are not left sitting around to be used later, accidentally. The procedure prevents the unwitting propagation of erroneous or superseded data and code."

In the four more recent controversies touched upon here questions of simple mistakes did not arise. Rather, the issues involved sample selection, choice of instrumental variables, questions about model specification and other legitimate econometric concerns. In all but one case both

new data were collected and the original data were reworked, so that most comments amounted to a mixture of pure and scientific replication. Readers of each should have been able to take away from the exchanges an assessment of each side's merit. If nothing else, the exchanges should have the additional benefit of offering students a better feel for the care needed in empirical work, particularly the importance of choice of sample and construction of variables. Both of these goals are enhanced if replicating and original authors temper their comments, not just in the published versions (as is true in all the examples discussed here) but also in the early drafts, so that they, and bystanders, including those who may be referees, can concentrate on the technical merit and the importance of the arguments.

One should remember that professional and public perceptions of facts do not in the end rest solely on the validity of one particular empirical study. Any alleged fact that affects the way that we think about a phenomenon will be hardened diamond-like by the heat and pressure of repeated empirical examination before it becomes background to our beliefs about how the world works. Nevertheless, given the media interest in reporting novel or titillating empirical findings, and politicians' desires to robe their proposals in scientific empirical cloth, however novel or inconsistent with prior research, it is crucial that as a profession we ensure that replication, or at least fear of replication, is our norm.[8] Empirical economics is never going to become a laboratory science; but recognising the role of replication can move us slightly in that direction by preventing us from propagating erroneous results.

2.4. A Modest Proposal

As the discussion in this Section shows, replication is rare, mainly because the incentives for empirical researchers to perform replication studies are weak. Feldstein's (1982) prediction, "… replication studies … should become increasingly important," has not proven to be accurate. Clearly, reductions in the cost of replication, and even the further reduction in cost created by central archives like that created by the *AER*, are unlikely to move the equilibrium quantity of replication studies very far from zero. Absent incentives on the supply side, what are needed are

additional demand-side incentives for potential replicating authors. Given the reward structure in the economics profession, however, we cannot expect junior or even mid-level researchers to undertake replication studies. Even very senior economists, whose age-earnings and age-status profiles are at best flat, are unlikely to undertake replications without incentives beyond those that have been provided so far (editors' general expressions of interest in publishing such studies). A more proactive approach is required of editors of scholarly journals.

One arrangement consistent with the incentives I have discussed and that would generate additional replications would be for journal editors to commission leading senior empirical researchers to undertake a replication study of a paper of their choice, one that had previously been published in the journal. If editors of each of the three leading general journals commissioned two replication studies per year, with publication guaranteed subject to refereeing (NOT by the author of the original study) to assure some minimum quality level, more replications would be undertaken. Original authors would be expected to write a short reply to the final version of the replication study.

The cost of this proposal (the opportunity costs of the time of some very senior economists and of the journal pages that would otherwise be filled with what one hopes are the marginally acceptable papers) seems fairly small. The benefits, in terms of nudging the profession toward additional legitimacy and providing guidelines for researchers on how to conduct empirical research, seem substantial. Without this direct subsidy, however, I am certain that, while we will continue to see sporadic calls for more replication, the supply of replication studies will hover near zero.

In these studies the replicating author should expect reasonable cooperation from the original author. If the original author has provided documentation showing how s/he chose the samples and constructed the variables that underlay the published tabular material, s/he should not be required to do much more to help the replicating author. Nor should the replicating author

expect much more: The original author cannot be expected to co-author a commissioned replication of his/her own work!

3. Scientific Replications

3.1. The Sad State of Scientific Replication in Economics

Most of what economists view as replication represents scientific replication—re-examining an idea in some published research by studying it using a different data set chosen from a different population from that used in the original paper. The recent Comments discussed in Section II may have started as attempts at pure replication, but in most cases soon transformed themselves into hybrids of pure and scientific replication. There are a variety of reasons why scientific replication is likely to be more important in economics (and social sciences more generally) than in the natural sciences. With laboratory experiments one might argue that scientific replication is not very useful (but are the usual subjects of economic experiments—college students at selective American universities—credibly representative of the average consumer in Bourkina Faso?). The argument might be applied, albeit with less force, to field experiments, increasingly in vogue in the literatures in policy evaluation and mechanism design (see Levitt and List, 2007). But for most empirical work we need to replicate analyses obtained on one data set many times over before the underlying and hopefully general economic point is to be believed.

By far the most important justification for scientific replication in non-experimental studies is that one cannot expect econometric results produced for one time period or one economy to carry over to another. Temporal change in econometric structure may alter the size and even the sign of the effects being estimated, so that the hypotheses we are testing might fail to be refuted with data from another time. This alteration might occur because institutions change, because incentives that are not accounted for in the model change and are not separable from the behaviour on which the model focuses, or crucially that even without these changes the behaviour is dependent on random shocks specific to the period over which an economy is observed.

The same types of differences can generate non-robust estimates in studies based on one cross-section of data (from a particular area or economy). Institutions differ across and even within economies, and different random shocks buffet different cross-sections at any particular time. Even more of a problem arises from the nature of the social interactions that might generate, or at least affect the behaviour to which our hypotheses are addressed and for which we do not, and perhaps even cannot account in our models. While social norms and ethnic capital have become an important focus of economic analysis in the past fifteen years, their likely impact goes far beyond the few areas of behaviour in which they have been studied.

Underlying the importance of scientific replication is the dominance of American economists, and American data, in the study of economic phenomena. It is not the case that American journals have a bias toward publishing studies based on American data (Hamermesh, 2002). Rather, researchers at American universities have for many years been pre-eminent in the profession; and U.S.-based journals have circulations that far exceed those of journals produced elsewhere. If our theories are intended to be general, to describe the behaviour of consumers, firms or markets independent of the social or broader economic context, they should be tested using data from more than just one economy.

There is nothing wrong with our research jingoism, provided we make it clear that the statistical evidence we adduce about our hypotheses may well only be applicable to the United States, so that the verification of the idea is quite limited. One might argue that many such studies do not claim to be general—they deal with a specifically American problem—and that, in any case, U.S.-based journals should be dealing with specifically American problems. Unfortunately many studies do not stake out such a narrow area; and too often modesty about the broader applicability of U.S.-based results is missing.

Consider the results shown in Table 2, which presents statistics describing studies published in regular issues of the *AER*, in 2005 and 2006, and in the *JPE* and *Quarterly Journal of Economics* (*QJE*) from 2004-2006, that use or are based on data. (This compilation thus

includes the calibration literature.) I exclude comments, replies, Presidential and Nobel lectures. I first divide the studies into two groups, those that could not be interpreted (and that the authors do not interpret) as dealing with anything other than a country-specific issue, and those that examine a general hypothesis or claim that the idea is generally applicable to economic behaviour. I further divide the latter group into studies that deal with international economics and those dealing with other topics. Finally, within each of these three categories I sub-divide the studies into those based solely on American data, those based on some other single country's data (or on EU or Latin American data) and those that use data from at least two countries. It is only the last group of studies that demonstrates scientific replication within the same study; and those studies alone are ones that have any hope of vitiating concerns about the broader applicability of our empirical work.

That the country-specific studies are disproportionately based on U.S. data is unsurprising: After all, these are U.S.-based journals. Also unsurprising is the tremendous concentration of studies in international economics that use multi-country data: After all, their topic is international economics. The bulk of the 292 studies that are tabulated (85 percent) deal with general issues other than those in international economics, however, and of these 60 percent are based solely on U.S. data. Of the studies that deal with general issues outside international economics, only 16 percent use data from more than one country. Moreover, about half of those deal with issues in political economy where data from many countries are used to examine how institutional differences affect political and market outcomes. The conclusion from this brief examination of the pinnacle of the scholarly literature in economics is that it presents empirical research on general economic ideas that is to a great extent based on data representing one country, by far most often the United States.[9]

Are we economists are particularly guilty of jingoistic testing of general ideas on American data? Not being knowledgeable about the categories/distinctions in other disciplines, I cannot answer this question broadly, and even a narrow answer is fraught with problems. To

attempt an answer, however, I tried to classify two years of empirical articles in the *American Journal of Sociology* (*AJS*) (like the *JPE*, published by the University of Chicago Press) and the *American Sociological Review* (*ASR*) (the Association journal), the two leading journals in that field. I was unable to distinguish between General/International and other General research, so that I present the results for the two types combined. While categorising studies in a field outside one's own is problematic, many of the topics and, to a lesser extent, the approaches in sociology are similar to those in parts of empirical economics.

I present the results of this little exercise in the bottom row of Table 2. The distribution looks remarkably like that for the aggregate of economics journals.[10] More than half of all the papers can be classified as claiming general implications based on empirical analyses performed solely on U.S. data. The classifications suggest that economists are no more guilty of intellectual ethnocentrism than practitioners of this related social science.

Now there may be reasons for immodesty and for the absence of much within-study scientific replication. First, and most obviously, the profession puts a premium on the creativity and generality of the idea, not on verifying the breadth of its applicability. Also, the academic world offers us a variety of incentives to generate a coterie of followers of our research ideas. Engaging in scientific replication within an original study reduces the opportunity for others to publish scientific replications of our work and thus reduces the potential size of a scholar's group of disciples. One might argue that many of the publications in second-level general journals and in many specialised journals too are essentially scientific replications (on similar data sets for a different country, or different data sets for the same country) of a publication in a major general journal. Those publications enhance the original scholar's professional standing, and we can maximise them by producing a novel and broadly applicable idea that seems supported by the one data set we used but that leaves wide room for additional confirmation/contradiction/extension.

I doubt that many (any?) empirical researchers consciously engage in such gaming. The rarity of scientific replication within studies may well arise from the extra data work it would

involve. In order to perform a within-study replication one must become familiar with a different data set, often one that is not coded in English. It is also the case, however, that the incentives for doing within-study scientific replication are non-existent.

One might dismiss this concern by saying that adding a second sampling base to one's study will merely increases the reader's confidence in one's findings (assuming the results are similar) in proportion to √2. Whether one views it approvingly as Baconian scientific method or disapprovingly as specification search (Leamer, 1983), testing an idea on a single set of data is likely to lead to the conclusion of most interest to the author(s) (and to journal editors). A second data set, one that does not merely sample from the same population as the first, should enhance a reader's confidence in the validity of the results by a proportion far above √2.

3.2. What Is To Be Done?

Given the incentives in the academic market, the outcomes suggested by the compilation reported in Table 2 are unsurprising: The rewards to within-study scientific replication are small. The compilation does not clarify whether editors do not care much about within-study scientific replication, or whether the supply of such studies is tiny. I should imagine, however, that editorial demand will create its own supply, so that the crucial issue is altering editorial recognition of the importance of scientific replication. No editor of a major journal is likely to publish replications of previous original pieces. Also, why should an editor wish to burden empirical researchers with more work, and why should the relative burden compared to writing purely theoretical papers be raised still further?

The answer must be the unappealing equivalent of the moral imperative: Because it is right for the profession. The best empirical economists are quite adept at writing down clever, novel models that are wonderfully consistent internally, that are intellectually beautiful and are supported by the single set of data on which they are "tested." As I noted above, however, the validity of these tests is questionable; and adding a second data set will enhance their validity more than proportionately. The most credible studies I have done are those that have proposed an

idea (hopefully a fairly novel one) and then tested it either on data sets from different economies or at least from the U.S. at different points in time. Policy makers and the general public will take the work that editors choose to publish (and the unpublished studies that, given editorial lags, circulate widely pre-publication) much more seriously if they do include scientific replications.

One cannot expect editors to require all authors to include within-study scientific replications: In some cases data sets are unique to a particular idea, even though the model may be universally applicable. In other cases, however, the data set is just one of many, equally useful and generally available sources. For example, why the concentration on the American PSID when the equally readily available and quite comparable Australian HILDA, British BHPS, Canadian SLID and German SOEP are readily at hand? Editors and referees should be sufficiently aware of these cases to require authors of papers that appear acceptable for publication to broaden their empirical focus beyond the single (usually American) data set they have used.

One might take the Alfred E. Newman approach to scientific replication: "What, me worry?" about studies based on just one set of data. Eventually an idea will be tested in the market by subsequent empirical studies that are usually published further down the food chain of journals. Also, survey articles and meta-analyses of the leading and following studies can eventually tease out general facts about the particular phenomena (assuming that such facts exist).

One difficulty with this Panglossian view is that these follow-ups are usually not brought to fruition until many years after the initial original empirical piece is published. During that hiatus perceptions of the truth can easily crystallize around results that are not generalisable and perhaps not even essentially correct. Also, academic incentives are such that the people undertaking the surveys and meta-analyses do not usually have the same professional visibility or credibility as the author of the original study (and may not be viewed as being as competent professionally). For these reasons it is crucial that editors of the leading journals tilt the publishing process a bit more in favor of within-study scientific replication.

4. Conclusions

The cost of pure and scientific replication in economic research has diminished over the past forty years. At the same time our verbal nods toward the need for replication have increased (as shown by the authors' common practice in the Replies discussed in Section II of thanking the authors of the Comments and stressing the great scientific benefits that result from replication). Despite the declining costs and despite our expressions of preferences, both pure and scientific replications are very rare in leading journals; and even within-study scientific replication is unusual.

Our expressions of preference are cheap talk. The profession provides few incentives for most active economists to produce replications of others' research, and similarly few to increase the believability of one's own research by testing ideas on multiple sets of data. Any demand for replication must arise from the profession as a whole, as intermediated by the actions and decisions of the editors of the most visible professional journals. Editors need to take the lead by providing sufficient incentives for top-flight authors of empirical work to engage in pure and scientific replication, and by insisting on within-study scientific replication in many more articles. Without these changes occasional paeans to the virtues of replication are as likely to enhance the scientific soundness of empirical research in economics as programmes that urge abstinence are to reduce teenagers' sexual activity.

REFERENCES

Antonovics, Kate, and Arthur Goldberger (2005) 'Does Increasing Women's Schooling Raise the Schooling of the Next Generation? Comment,' *American Economic Review* 95, 1738-1744

Behrman, Jere, and Mark Rosenzweig (2002) 'Does Increasing Women's Schooling Raise the Schooling of the Next Generation?' *American Economic Review* 92, 323-334

---------- and ---------- (2005) 'Does Increasing Women's Schooling Raise the Schooling of the Next Generation? Reply,' *American Economic Review* 95, 1745-1751

Burnside, Craig, and David Dollar (2000) 'Aid, Policies and Growth,' *American Economic Review* 90, 847-868

---------- and ---------- (2004) 'Aid, Policies and Growth: Reply,' *American Economic Review* 94, 781-784

Card, David, and Alan Krueger (1994) 'Minimum Wages and Employment; A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania,' *American Economic Review* 84, 772-793

---------- and ---------- (2000) 'Minimum Wages and Employment; A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania: Reply,' *American Economic Review* 90, 1397-1420

De Long, J. Bradford, and Kevin Lang (1992) 'Are All Economic Hypotheses False?' *Journal of Political Economy* 100, 1257-1272

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

Easterly, William, Ross Levine and David Roodman (2004) 'Aid, Policies and Growth: Comment' *American Economic Review* 94, 774-780

Ellison, Glenn (2002) 'The Slowdown in the Economics Publishing Process,' *Journal of Political Economy* 110, 990-1034

Feldstein, Martin (1974) 'Social Security, Induced Retirement and Aggregate Capital Accumulation,' *Journal of Political Economy* 82, 905-926

---------- (1982) 'Social Security and Private Saving: Reply,' *Journal of Political Economy* 90, 630-642

Hamermesh, Daniel (1997) 'Some Thoughts on Replications and Reviews,' *Labour Economics* 4, 107-109

---------- (2002) 'International Labor Economics,' *Journal of Labor Economics* 20, 709-732

---------- and Jeff Biddle (1994) 'Beauty and the Labor Market,' *American Economic Review* 84, 1174-1194

---------- and Peter Schmidt (2003) 'The Determinants of Econometric Society Fellows Elections,' *Econometrica* 71, 399-407

Hoxby, Caroline (2000) 'Does Competition Among Public Schools Benefit Students and Taxpayers?' *American Economic Review* 90, 1209-1238

---------- (2005) 'Does Competition Among Public Schools Benefit Students and Taxpayers? A Reply to Rothstein,' National Bureau of Economic Research, Working Paper No. 11216

Hunter, John (2001) 'The Desperate Need for Replications,' *Journal of Consumer Research* 28, 149-158

Kevles, Daniel (1998) *The Baltimore Case*. (New York: Norton)

Leamer, Edward (1983) 'Let's Take the 'Con' Out of Econometrics,' *American Economic Review* 73, 31-43

Leimer, Dean, and Selig Lesnoy (1982) 'Social Security and Private Saving: New Time-Series Evidence,' *Journal of Political Economy* 90, 606-629

Levitt, Steven, and John List (2007) 'On the Generalisability of Lab Behavior to the Field,' this *Journal* 40,

Neumark, David, and William Wascher (2000) 'Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania: Comment,' *American Economic Review* 90, 1362-1396

Rothstein, Jesse (2005) 'Does Competition Among Public Schools Benefit Students and Taxpayers? Comment,' National Bureau of Economic Research, Working Paper No. 11215

FOOTNOTES

[1]Merriam-Webster Online Dictionary; Cambridge Advanced Learner's Dictionary.

[2]*Lott v. Levitt*, Northern District of Illinois, ruling of January 11, 2007, Judge Ruben Castillo.

[3]Report of the Managing Editor, American Economic Association, *Papers and Proceedings*, May 2006, Table 8, May 2007, Table 8. The Editor of a specialised journal mentioned that he does not get involved in requests for data unless the author is slow to respond, which he notes has occurred only twice in five years.

[4]In statistical terms, the distribution is highly over-dispersed: A negative binomial approximates it far better than a simple Poisson distribution. One might worry that using examples from 2002-2004 fails to allow enough time for requests to have been made. Indeed, time since publication for the *ILRR* is significantly positively related to the number of requests in a negative binomial regression; but for the *JHR* the opposite is surprisingly true.

[5]Circulation of the *ILRR* is currently 2400, that of the *JHR* is 2200. The circulation of the *AER* is over 21,000.

[6]One such occurrence involved a paper of mine (Hamermesh and Schmidt, 2003), for which the journal Editor insisted upon receiving the data set and presumably re-checking some of the estimates before agreeing to publish the piece. Initially I was extremely annoyed that the Editor of this predominantly theoretical journal, himself a theorist, had the temerity to question my competence and/or honesty. After some thought, however, I was less bothered, wishing only that more people were subjected to this but wondering whether the Editor might not have had something better to do with his time.

[7]Email communication, Joop Hartog, February 2, 2007.

[8]The best example is President Clinton's seizing on the Card-Krueger results to justify a proposal to raise the federal minimum wage (State of the Union Message, January 24, 1995.).

[9]While the research disproportionately uses the U.S. as the "guinea pig," a remarkable variety of other countries' data sets form the basis for the research published in these outlets. Data sets from 27 different countries, the EU and Latin America underlie the empirical research published recently in these journals. As a bibliometric note, a test of the hypothesis that the patterns of publication by type across the three journals are independent cannot reject that hypothesis ($\chi^2(14) = 9.28$, not significantly different from zero at any conventional level). Implicitly these three general journals publish remarkably similar types of research based on remarkably similar types of data (classified by country of origin).

[10]Comparing the distribution of all economic studies except those classified as general-international to that of the sociological studies, a test that the distributions are identical cannot reject the null hypothesis ($\chi^2(4) = 6.89$).

Table 1. Statistics Describing Data Requests, Articles in the *ILRR* 2002-04, *JHR* 2002-04, *AER* 1999-2000

Journal	Mean (Std. Error of Mean)	Median	Response Rate	N
ILRR	0.78 (0.24)	0	0.87	69
JHR	0.95 (0.21)	0	0.87	70
AER	8.88 (2.53)	3	0.83	60

Table 2. The Distribution of Empirical Studies by Type of Problem and Source of Data Used, Recent Volumes, *AER*, *JPE* and *QJE*, and *AJS*/*ASR*

	Type of Issue							
	General			General (Intl.)			Country-Specific	
	US	Other	Many	US	Other	Many	US	Other
AER (N=115) 2005-2006	54	26	13	2	1	8	8	3
JPE (N=82) 2004-2006	49	12	10	0	1	3	5	2
QJE (N=95) 2004-2006	46	16	18	2	1	4	5	3
All three (N=292) journals	149	54	41	4	3	15	18	8
AJS, *ASR* (N = 122) (2003-2004)	75	14	18				7	8

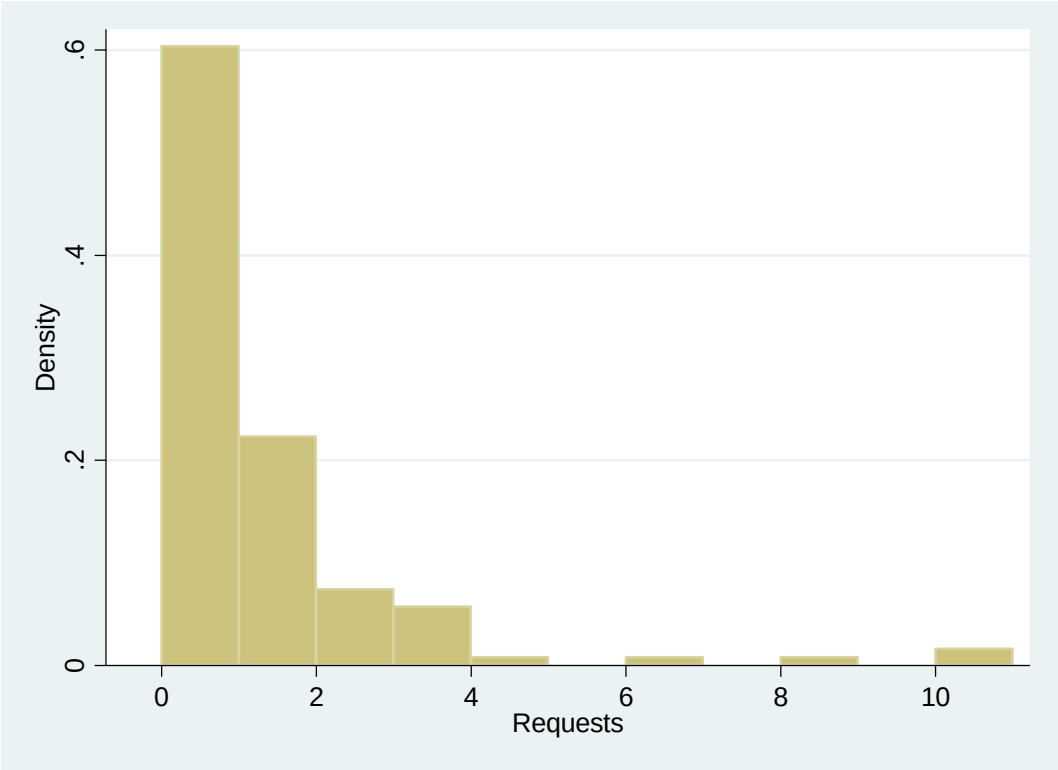

Figure 1. Distribution of Requests for Data, *Industrial and Labor Relations Review* and *Journal of Human Resources*, 2002-2004