The Future of Replication

publications. By following the recommendations outlined here we can make progress in speeding up manuscript reviews by eliminating some revise-and-resubmit decisions, providing better information for referees to make more decisive recommendations to editors. Additionally, we can reduce errors in published research and facilitate replication tests designed to probe the robustness of findings. All of these changes will improve the quality and impact of research and so should be adopted as soon as possible.

Bruce Bueno de Mesquita
New York University and Stanford University

The Future of Replication[16]

Since the replication standard was proposed for political science research, more journals have required or encouraged authors to make data available, and more authors have shared their data. The calls for continuing this trend are more persistent than ever, and the agreement among journal editors in this Symposium continues this trend. In this article I offer a vision of a possible future of the replication movement. The plan is to implement this vision via the Virtual Data Center project, which—by automating the process of finding, sharing, archiving, subsetting, converting, analyzing, and distributing data—may greatly facilitate adherence to the replication standard.

In King (1995), I proposed that political scientists try to meet *The Replication Standard*, which holds that

> *Sufficient information exists with which to understand, evaluate, and build upon a prior work if a third party can replicate the results without any additional information from the author.*

Meeting the replication standard does not require any work to be replicated. It only requires the provision of enough information so that it could be replicated in principle. Actual replication will often produce other benefits, but the purpose of the standard is only to provide a way to judge the adequacy of the information provided in a published work or a "replication data set" that accompanies the work.

Without information sufficient to make replication possible in principle, the "chain of evidence" from the empirical world to the article's conclusions is broken, and so the basis for conclusions about the world remains unjustified, at least without taking the word of the author. Although trust is a good thing in personal relationships, science requires that answers depend on public knowledge, not private understandings about individuals and their personalities. The contribution of a work should be represented in that work, not only in the memory or file cabinet of the person who wrote it originally. What counts for the community of political scientists and for our contribution to knowledge and society is only what is on the public record. If they had been academics, Christopher Columbus would have been given tenure early, while Leif Erikson would have been given an extra year to get his book out.

In King (1995) I also proposed some policies for graduate programs, funding agencies, journals, and book presses to encourage or require scholars to follow the standard. Other contributions in this volume report on current practices and some progress at achieving consensus at developing standards. The empirical result in Gleditsch and Metelits (2003)—showing that articles that make data available have twice the impact, and hence make twice the contribution, as those that do not—is

[16]Gary King is the David Florence Professor of Government, Harvard University and Senior Science Advisor, World Health Organization (Center for Basic Research in the Social Sciences, 34 Kirkland Street, Harvard University, Cambridge, MA 02138; http://GKing.Harvard.Edu, King@Harvard.Edu, (617) 495–2027). The Virtual Data Center project described herein is joint with Micah Altman and Sidney Verba.

particularly striking. Put differently, an author who makes data available has as much impact as two authors who do not. Since we hire, promote, and pay faculty on the extent of their scholarly contribution, it seems likely that those who make data available to others probably also have better jobs and higher salaries.

Science Requires More Than Being Scientific

The training we receive in various methods during graduate school and afterwards is all about making our individual work more scientific, but for a lot of scholars being individually scientific does not add up to a science. Science also requires community, interaction, a common language, the pursuit of common scholarly goals, communication, and a collection of individuals cooperating and competing to contribute to our common and publicly available knowledge base. The tremendous advances in technology, medicine, basic research in the natural sciences, and to some extent research in the social sciences—and the huge increase in the rate of learning in these areas—have come from intense interactions, from checking each other's work, and from individuals building on the work of others.

The famous cold fusion debacle of some years ago involved two researchers, with good reputations, being apparently scientific to the best of their abilities. Fraud was not an issue. Their efforts obviously failed, but the process worked and the world knew the truth within a few weeks. If the same "finding" had been claimed several hundred years ago before modern science (or in a nonscientific field today), society and these individual researchers would have been led astray and little would have been learned. Similarly, when one of us makes an important and accurate discovery about the social or political world, others cannot build on this discovery unless the original author followed the replication standard.

The almost fantastic increase in the rate of mankind's increase in knowledge since the development of modern science is well known. What we may not all appreciate is how recent this all is. Taking the (very!) long term view, the first version of learning on this planet started about 3.8–3.7 billion years ago when the first protocell developed and evolution began (de Dueve, 1995). Evolution, which requires variation plus (natural) selection, produces a type of knowledge as succeeding generations produced organisms that fit various niches and adapted in various ways. The method of learning works, but it is entirely unplanned, undirected, and very slow.

Somewhere around 5–7 million years ago, the last ancestor (the woodland ape) we have in common with our closest living relatives (the chimpanzee) lived. Given the rate at which they learned and their brains expanded, it took until about 1.5 million years ago to tame fire. Creatures then "learned" by evolution, but sometime back then they also started to learn by demonstration. Even today, chimpanzees learn how to use primitive tools and they pass the knowledge to other chimps through demonstration (Wrangham and Peterson, 1996). Most of the knowledge, however, must be learned anew each generation. Around 40,000 years ago (only 2,000 generations) the evidence seems to suggest that language became anatomically possible and subsequently developed. Whatever we learned could now be passed to larger groups without physical demonstration. The result was the onset of extremely rapid cultural development. The form of learning was Lamarkian—the heredity of acquired characteristics through culture—and, relative to evolution, it was very fast and could be intentional and directed. The invention of written language, which permits the transmission of knowledge even if no one presently alive possesses it, dates to 3100 B.C. (or only 255 generations ago).

In contrast, the invention of modern science—with scholars competing and cooperating in the pursuit of common goals on the public record—dates to only about the last 400 years, around 20 generations. It has not yet spread to all fields of

inquiry (e.g., only about a third of the advice you get in your doctor's office is what medical researchers call "evidence-based"), and yet the progress that has resulted far exceeds any other method of learning yet devised (see Haynes, 2002, and the citations therein).

Science includes numerous individual components, and contributions to the infrastructure and definition of science happen all the time. Research libraries date to only about a century ago. Humanity did not know how to run a controlled and randomized experiment, with which one could reliably learn about the world, until the 1930s. Replication and data sharing are even more recently developed norms, and are still spreading across the fields of inquiry. Data sharing in political science is far ahead of medicine and public health but nowhere near some areas of DNA sequencing, where data must be submitted in fixed formats to the same repository as a condition of publication. It is interaction among scholars pursuing common goals that makes science work, and meeting the replication standard clearly advances science.

A Future for Learning from Numerical Data: The Virtual Data Center

Contributing data along with scholarly articles and analyses is obviously in the interest of individual scholars. Much evidence of this point is provided in the other articles in this symposium (and King, 1995). Yet, convincing individuals to do things that are in their long-term interest is often difficult (e.g., think of exercise, recycling, etc.). Helping to develop community-wide norms, instituting requirements when appropriate, and acknowledging and rewarding the data contributions of others can help enormously and should undoubtedly be pursued in every venue available.

In this section, I report on a large software project that attacks the same problem from another angle—with the goal of making it easier for researchers, journal editors, and others to meet the replication standard and improve the product that the discipline of political science provides. Roughly speaking, our project is a hierarchical, structured, and legal version of Napster, but for social science data. The spirit in which we offer this is to make our research lives easier and simultaneously to contribute to the replication movement. In similar fashion, we would all agree to recycle daily if our newspapers and bottles could be separated from our garbage automatically at the landfill without our participation.

The Present and Near Future

The Virtual Data Center, or VDC, is a software project being created by Micah Altman, Sidney Verba, and me (for details see http://TheData.org and Altman et al., 2001a, 2001b). It is a project of the Harvard-MIT Data Center and the Harvard University Libraries and is supported by a number of federal grants through the digital library initiative and the political science program at NSF (funding agencies include NSF, NASA, DARPA, NLM, LoC, NEH, and FBI) and part of a separate grant at NIA. I explain what is, what is under way, and what is planned, sequentially.

An operational version of a portion of the VDC is now running at Harvard and MIT and selected other test sites. With this system, faculty, staff, and students knowing only how to point and click can

- search (using a highly focused Google-like system) for data available locally or at various international archives (such as the ICPSR, Roper, NCHS, etc.).
- read abstracts of data sets.
- choose a data set, whether it is local or far away.
- peruse documentation.

- subset the data set by choosing variables and rows (e.g., only women 18–24 who voted for Clinton in 1996).
- perform basic statistical analyses (such as descriptive statistics, cross-tabulations, regressions, etc.).
- download code to replicate the analysis automatically.
- convert the data set format to any statistical package, database program, spread-sheet, etc.
- download the converted and subsetted data for further analyses.
- ensure that the user is authenticated and authorized to have access to the data set chosen.

We can guarantee that VDC software will always be available since it is legally and permanently "open source," which means that if our group vanishes the software will be on the public record and anyone else will have the right to modify it as they see fit. If they make any improvements to it and offer it for distribution, the license requires them to also provide the source code without charge.

Although many of us grew up specializing in how to extract data from tapes in arcane formats like binary coded decimal or EPCDIC, researchers now call to ask whether they can put their data in our system so they can then take it out in a more convenient format. The usage of the Harvard-MIT Data Center by faculty and students, and for research and in classes, has increased exponentially since its introduction here.

That's the present. The version to be released shortly will add some critical features. Most important, the same software will be installable anywhere. We have been working closely with the ICPSR and the U.S. Bureau of Labor Statistics, and have had expressions of interest from many universities, governments, and international organizations. When version 1 of the software is released, we expect that many VDC nodes will be installed. This version will make it possible for a user at one site to search across local data sets at their university, data sets at any national archives to which they have access, and data sets at other local VDC nodes. If it works as planned, the system will spread to an ever-increasing user base, and encompass a larger and larger fraction of available social science data. With every additional VDC node, the system will become more useful to everyone.

Others will be able to contribute modules (or add the modules themselves) that snap into the VDC infrastructure and perform specialized tasks, so that many people will no longer need to re-create the wheel at many different sites. Data librarians will be able to build virtual curator's pages that provide hierarchical organizations and guides for groups of studies in specialized areas, and curators at one site can share the knowledge with others around the world.

As important from the perspective of replication, depositing data will be much easier. We plan to put no hurdles in the way of authors willing to share their data, but the more information ("metadata") they provide about their data, the more services the VDC will be able to provide users of those data. Perhaps at first data depositors will only drop in zip files with ascii documentation, but we think that when the advantages of all the services of the VDC become more widely known, authors will want to provide more metadata too.

Finally, each data set will have a permanent name associated with it. The name will look like a URL and work in a browser, but it will persist and so will work indefinitely even if the original link vanishes. With this system, checking an author's claim that data have been deposited will be equivalent to giving a citation to the data. Most journals and copy editors are obsessive about the precise formats of citations to printed matter; this system will make reliable citations to data possible as well.

The Vision: Replication and Deep Citation

The first or second major release of the VDC will also have something like a digital signature associated with each data set and linked to its permanent URL-like name. A digital signature is one number that can be easily calculated from the data set and summarizes all the information in the data set. If any number in the data set changes, then this one number would change too. (An intuitive but bad version of a digital signature, because it is easy to defeat, would be to add up all the numbers in the database.) The advantage of digital signatures is that future researchers could make one easy calculation and then determine immediately and essentially for certain whether they possess the identical version of the data used by the author of the original article. Then we would also be able to verify who created and provided the data. This addition should eliminate an enormous amount of wasted time in building on existing research: a researcher would merely copy down the official name of the data set, type it into a browser, and download the identical data used in the article or book.

Once the names become more widely used, automatic forward citation will be possible. That is, if a researcher comes upon a data set, he or she can instantly find all articles that have subsequently been written using that data set. Ultimately, our goal is "deep citation": formal ways of citing or referring to particular variables, transformations, or individual cells in the data set. This way, if an author says he regressed income on education, some future researcher will not have to waste a day determining which of the twelve education-related variables were used, and which of the infinite array of transformations that could have been applied were in fact applied. This kind of citation must also be independent of the storage medium, so that if the data set is transferred from SPSS to Stata, the user should still be able to make the right connection transparently and automatically.

Over the years, I have written sample editorial policies for journals that try to help authors meet the replication standard while also trying to accommodate the critics of such policies. The VDC will change these policies considerably. For example, here is one policy that could now be (and in fact already has been) implemented, without the VDC:

> Authors of quantitative articles in this journal [or books at this press, or dissertations in this department] must address the issue of data availability and replication in their first footnote. Authors are ordinarily expected to indicate in this footnote which public archive they will deposit the information necessary to replicate their numerical results, and the date when it will be submitted. (The information deposited should include items such as original data, specialized computer programs, lists of computer program recodes, extracts of existing data files, and an explanatory file that describes what is included and explains how to reproduce the exact numerical results in the published work. Authors may find the "Publication-Related Archive" of the Inter-university Consortium for Political and Social Research a convenient place to deposit their data.) Statements explaining the inappropriateness of sharing data for a specific work (or of the necessity for indeterminate periods of embargo of the data or portions of it) may fulfill the requirement. Peer reviewers will be asked to assess the footnote as part of the general evaluative process, and to advise the editor accordingly. Authors of works relying upon qualitative data should submit a comparable footnote that would facilitate replication where feasible. As always, authors are advised to remove information from their data sets that must remain confidential, such as the names of survey respondents.

When the VDC is operational, the policy could be much simpler:

> Authors of articles in this journal [or books at this press, or dissertations in this department] that use quantitative data must cite the data by including the Virtual Data Center name for their data and any accompanying replication information.

> See your local data center, the ICPSR, the U.S. Census Bureau, or http://TheData.org for details of how to do this.

That is, authors merely cite the data in the correct format, just as they are required to cite printed books and articles in the correct format, and the VDC will then automatically take care of the rest.

Concluding Remarks

A warning: beware of vaporware. Although a version of the VDC is now operational at Harvard and MIT, the source code is open and on the Web, and many of the features I have discussed here are already implemented at least in part, much of the rest is presently just promises. Some aspects, such as deep citation, require considerable additional research on our part to address several important unsolved problems. I have enormous confidence in our team of researchers and programmers, but we will have to see.

Whatever the future progress of the VDC, the replication movement will continue. Changing norms and practices is sometimes slow going, but the benefits to individual scholars, to political science as a discipline, and to society at large should keep the movement on its path.

Gary King
Harvard University

Editors' Joint Statement: Minimum Replication Standards for International Relations Journals

Authors of quantitative empirical articles must make their data available for replication purposes. A statement of how that is done should appear in the first footnote of the article. Required material would include all data, specialized computer programs, program recodes, and an explanatory file describing what is included and how to reproduce the published results. This material must be posted by the month of publication, except when, with agreement of the Editor, the deadline is extended to accommodate special need of an author to employ the data for subsequent publications. Information that must remain confidential—such as that which would identify survey respondents—should be removed. All files should be sent electronically to the Managing Editor for posting on a website maintained by the journal for the purpose. In addition, authors may send the material to www.icpsr.umich.edu, and any other sites they wish to use.

We urge other editors to join us in enforcing these minimum guidelines.

Nils Petter Gleditsch, Editor of *Journal of Peace Research*
Patrick James, Co-editor of *International Studies Quarterly*
James Lee Ray, Editor of *International Interactions*
Bruce Russett, Editor of *Journal of Conflict Resolution*

References

ALTMAN, M., L. ANDREEV, M. DIGGORY, G. KING, D. L. KISKIS, E. KOLSTER, M. KROT, AND S. VERBA (2001a) "A Digital Library for the Dissemination and Replication of Quantitative Social Science Research: The Virtual Data Center." *Social Science Computer Review* **19**:458–470. Reprint at http://gking.harvard.edu/files/abs/vdcwhitepaper-abs.shtml

ALTMAN, M., L. ANDREEV, M. DIGGORY, G. KING, D. L. KISKIS, E. KOLSTER, M. KROT, AND S. VERBA (2001b) "An Overview of the Virtual Data Center Project and Software." JCDL '01: First Joint Conference on Digital Libraries. ACM. Reprint http://gking.harvard.edu/files/abs/jcdl01-abs.shtml.

ANDREWS, N. (2001) "Evidence." In *Reader's Guide to the Social Sciences*, vol. 1, edited by J. Michie. London and Chicago: Fitzroy Dearborn.

BARBIERI, K. (1996) "Economic Interdependence: A Path to Peace or a Source of Interstate Conflict?" *Journal of Peace Research* **33**(1):29–49.

BENNETT, D. S., AND A. STAM (2000) "EUGene: A Conceptual Manual." *International Interactions* **26**(2):179–204.

BUENO DE MESQUITA, B. (2003) "Getting Firm on Replication." *International Studies Perspectives* **4**(1): 98–100.

CIOFFI-REVILLA, C. (1998) *Politics and Uncertainty: Theory, Models and Applications*. Cambridge: Cambridge University Press.

CIOFFI-REVILLA, C., AND H. STARR (1995) "Opportunity, Willingness, and Political Uncertainty: Theoretical Foundations of Politics." *Journal of Theoretical Politics* **7**(4):447–476.

CIOFFI-REVILLA, C., AND T. LANDMAN (1999) "Evolution of Maya Polities in the Ancient Mesoamerican System." *International Studies Quarterly* **43**(4):559–598.

DE DUEVE, C. (1995) *Vital Dust*. New York: Basic Books.

FOWLER, L. L. (1995) "Replication as Regulation." *PS. Political Science and Politics* **28**(3):478–481.

GIBSON, J. L. (1995) "Cautious Reflections on a Data-Archiving Policy for Political Science." *PS. Political Science and Politics* **28**(3):473–476.

GLEDITSCH, N. P. (1993) "The Most Cited Articles in *JPR*." *Journal of Peace Research* **30**(4):445–449.

GLEDITSCH, N. P., AND C. METELITS (2003) "Replication in International Relations Journals: Policies and Practices." *International Studies Perspectives* **4**(1):72–79.

GLEDITSCH, N. P., C. METELITS, AND H. STRAND (2003) "Posting Your Data: Will You Be Scooped or Will You Be Famous?" *International Studies Perspectives* **4**(1):89–97.

GLEDITSCH, N. P., T. V. LARSEN, AND H. HEGRE (1994) *Citations to Articles in JPR and by Johan Galtung*. Oslo: Journal of Peace Research.

HAMILTON, D. P. (1991) Research Papers: Who's Uncited Now? *Science* **4** (Jan.):25.

HAYNES, R. B. (2002) "What Kind of Evidence Is It That Evidence-Based Medicine Advocates Want Health Care Providers and Consumers to Pay Attention To?," BMC Health Services Research 2. http://www.biomedcentral.com/1472-6963/2/3.

HERRNSON, P. S. (1995) "Replication, Verification, Secondary Analysis, and Data Collection in Political Science." *PS. Political Science and Politics* **28**(3):452–455.

INTEGRATED NETWORK FOR SOCIETAL CONFLICT RESEARCH, UNIVERSITY OF MARYLAND (2002) http://www.bsos.umd.edu/cidcm/inscr/.

INTER-UNIVERSITY CONSORTIUM OF POLITICAL AND SOCIAL RESEARCH, UNIVERSITY OF MICHIGAN(2002) http://www.icpsr.umich.edu/.

JAGGERS, K., AND T. R. GURR (1995) "Tracking Democracy's Third Wave with the Polity III Data." *Journal of Peace Research* **32**(4):469–482.

JAMES, P. (2002) *International Relations and Scientific Progress: Structural Realism Reconsidered*. Columbus: Ohio State University Press.

JAMES, P. (2003) "Replication Policies and Practices at *International Studies Quarterly*." *International Studies Perspectives* **4**(1):85–88.

KING, G. (1995) "Replication, Replication." *PS. Political Science and Politics* **28**(3):443–499. Reprint at http://gking.harvard.edu/files/abs/replication abs.shtml..

KING, G. (2003) "The Future of the Replication Movement." *International Studies Perspectives* **4**(1): 100–105.

MEIR, K. J. (1995) "Replication: A View from the Streets." *Political Science and Politics* **28**(3):456–459.

MICHIE, J., ED. (2001) *Reader's Guide to the Social Sciences*, **vol. 1**. London and Chicago: Fitzroy Dearborn.

ONEAL, J. R., F. H. ONEAL, Z. MAOZ, AND B. M. RUSSETT (1996) "The Liberal Peace: Interdependence, Democracy, and International Conflict, 1950–85." *Journal of Peace Research* **33**(1):11–28.

PEACE SCIENCE SOCIETY–INTERNATIONAL (2002) http://pss.la.psu.edu/.

PENDLEBURY, D. (1994) (Institute for Scientific Information) Telephone Conversation with Gary King.

RAY, J. L., AND B. VALERIANO (2003) "Barriers to Replication in Systematic Empirical Research on World Politics." *International Studies Perspectives* **4**(1):79–85.

RUSSETT, B. (2003) " *The Journal of Conflict Resolution's* Policy on Replication." *International Studies Perspective* **4**(1):88–89.

SIEBER, J. E. (1991a) Introduction to *Sharing Social Science Data: Advantages and Challenges*, edited by J. E. SIEBER, pp. 1–18. London: Sage.

SIEBER, J. E. (1991b) "Social Scientists' Concerns About Sharing Data." In *Sharing Social Science Data: Advantages and Challenges*. edited by J E. Sieber, pp. 141–150. London: Sage.

Stata Reference Manual Release(1999) Version 6(3). College Station, TX: Stata Press.

TICKNER, J. -A. (1997) "You Just Don't Understand: Troubled Engagements Between Feminists and IR Theorists." *International Studies Quarterly* **41**(4):611–632.

WRANGHAM, R., AND D. PETERSON (1996) *Demonic Males*. Boston: Houghton Mifflin.