The Political Methodologist

Newsletter of the Political Methodology Section
American Political Science Association
Volume 22, Number 1, Fall 2014

Editor:
Justin Esarey, Rice University
justin@justinesarey.com

Associate Editors:
Randolph T. Stevenson, Rice University
stevenso@rice.edu

Editorial Assistant:
Ahra Wu, Rice University
ahra.wu@rice.edu

Rick K. Wilson, Rice University
rkw@rice.edu

Contents

Notes From the Editors

In this issue of *The Political Methodologist*, the editorial staff is proud to present a collection of articles addressing the timely and important matter of transparent and replicable research in political science. One of our Associate Editors, Rick Wilson, introduces the issue by announcing *TPM*'s signing of the Data Access and Research Transparency (DA-RT) Statement; among other things, this commits the newsletter to including replication data and code with any of its future articles that involve computation or analysis. Rick also discusses reasons *why* including replication materials is important to ensure the integrity of scientific work and increase the degree to which errors are self-correcting. Nick Eubank and Mike Alvarez discuss the experiences of *QJPS* and *Political Analysis* in implementing DA-RT style standards at the journal, including (for *QJPS*) the result of closely auditing of replication code by the journal staff; these articles provides an interesting insight into what the discipline can expect to gain by increasing the transparency and replicability of its work and the challenges it will face to make the system work. Andrew Moravcsik describes a plan for applying the principles of replicability and transparency for qualitative research, discussing how the implementation of such a plan requires re-thinking certain aspects of the usual strategy applied to quantitative research. Finally, Jonathan Rogers and Wendy Martinek provide reflections on the experience of using replication as an educational tool, arguing that this approach provides students with valuable hands-on experience and helps them to develop a deep understanding of the relationship between theory and empirical observation. There is also a contribution from the *TPM* editor about the way the scientific community should interpret a failed replication, with a simple simulation illustrating its central point: a single failed replication can be as fragile as a single positive result.

This issue also contains two regular *TPM* contributions beyond the theme of our special edition, each of which deaks with a practical aspect of the practice of empirical research in political science. Dvora Yanow and Peregrine Schwartz-Shea provide advice to political scientists in obtaining consent for empirical research from your local Institutional Review Board, including a primer on how different kinds of

research require different levels of review and the criteria that the IRB uses for this and other decisions. Jacob Montgomery and Ryan Moore illustrate how R code can be consolidated into a library and hosted in a public archive using two time-saving tools in the R code base.

Although we are happy to receive and publish articles on a variety of subjects at *TPM*, we think that thematic special issues have been especially successful at bringing together multiple viewpoints on an issue to create discussion and provide a resource to start conversations in seminars and classrooms all over the world. We hope that you enjoy this special issue on replication, and encourage you to submit your own ideas for special issues to the editorial staff at *TPM*! **-The Editors**

Reproducibility and Transparency

Rick K. Wilson
Rice University
rkw@rice.edu

The Political Methodologist is joining with 12 other political science journals in signing the Data Access and Research Transparency (DA-RT) joint statement.

The social sciences receive little respect from politicians and segments of the mass public. There are many reasons for this, including:

- we are not very good about translating our work to the public
- many believe that our research is just common sense and so all we offer is opinion
- and many distrust us as pointy-headed academics.

A partial solution to building trust is to increase the transparency of our claims and this is why *The Political Methodologist* is signing on to DA-RT.

As researchers we need to ensure that the claims we make are supported by systematic argument (either formal or normative theory) or by marshaling empirical evidence (either qualitative or quantitative). I am going to focus on empirical quantitative claims here (in large part because many of the issues I point to are more easily solved for quantitative research). This idea of DA-RT is simple and has three elements. First, an author should ensure that data are available to the community. This means putting it in a trusted digital repository. Second, an author should ensure that the analytic procedures on which the claims are based are public record. Third, data and analytic procedures should be properly cited with a title, version and persistent identifier. Interest in DA-RT extends beyond political science. From November 3-4, 2014 the Center for Open Science cosponsored a workshop designed to produce standards for data accessibility, transparency and reproducibility. At the table were journal editors from the social sciences and Science. The latter issued a rare joint editorial with *Nature* detailing standards for the biological sciences to ensure reproducibility and transparency. Science magazine aims at doing the same for the social sciences.

Ensuring that our claims are grounded in evidence may seem non-controversial. Foremost, the evidence used to generate claims needs to be publicly accessible and interpretable. For those using archived data (e.g., COW or ANES) this is relatively easy. For those collecting original data it may be more difficult. Original data require careful cataloging in a trusted digital repository (more on this in a bit). It means that the data you have carefully collected will persist and will be available to other scholars. Problematic are proprietary data. Some data may be sensitive, some may be protected under Human Subjects provisions and some may be privately owned. In lieu of providing such data, authors have a special responsibility to carefully detail the steps that could, in principle, be taken to access the data. Absent the data supporting claims, readers should be skeptical of any conclusions drawn by an author.

Surprisingly, there are objections to sharing data. Many make the claim that original data is proprietary. After all, the researcher worked hard to generate them and doesn't need to share. This is not a principled defense. If the researcher chooses not to share data, I see no point in allowing the researcher to share findings. Both can remain private. A second claim to data privacy is that the data have not yet been fully exploited. Editors have the ability to embargo the release of data, although this should happen under rare circumstances. It seems odd that a researcher would request an embargo, given that the data of concern is that which supports the claims of the researcher. Unless the author intends to use exactly the same data for another manuscript, there is no reason to grant an embargo. If the researcher is intending to exactly use the same data, editors should be concerned about self-plagiarism. Reproducible data should focus on the data used to make a claim.

The second feature of reproducibility and transparency involves making the analytic procedures publicly available. This gets to the key element of transparency. The massaged data that are publicly posted have been generated through numerous decisions by the researcher. A record of those decisions is critical for understanding the basis of empirical claims. For most researchers, this means providing a complete listing of data transformation steps. All statistical programs allow for some form of a log file that document what a researcher did. More problematic may be detail-

ing the instruments that generated some of the data. Code used for scraping data from websites, videos used as stimuli for an experiment or physical recoding devices all pose problems for digital storage. However, if critical for reaching conclusions, a detailed record of the steps taken by a researcher must be produced. The good news is that most young scholars are trained to do this routinely.

There are objections to providing this kind of information. Typically it has to do with it being too difficult to recreate what was done to get to the final data set. If true, then it is likely that the data are problematic. If the researcher is unable to recreate the data, then how can it be judged?

The final element of transparency deals with the citation of data and code. This has to be encouraged. Assembling and interpreting data is an important intellectual endeavor. It should be rewarded by proper citation – not just by the researcher, but by others. This means that the record of the researcher must have a persistent and permanent location. Here is where trusted digital repositories come into play. These may be partners in the Data Preservation Alliance for the Social Sciences (Data-PASS)[1] or institutional repositories. They are not an author's personal website. If you're like me, my website is outdated, and I should not be trusted to maintain it. The task of a trusted data repository is to ensure that the data are curated and appropriately archived. Repositories do not have the responsibility for documenting the data and code – this is the responsibility of the researcher. All too often stored data have obscure variable names that are only meaningful to the researcher and there is little way to match the data to what the researcher did in a published article.

The aim for transparency, of course, is to ensure that claims can be subject to replication. Replication has a troubled history in that it often looks like "gotcha" journalism. There is bias in publication in that replications overturning a finding are much more likely to be published. This obscures the denominator and raises the question of how often findings are confirmed, rather than rejected. We have very few means for encouraging the registration of replications. It is a shame, since we have as much to learn from instances where a finding appears to be confirmed as when it may not. If journals had unlimited resources, no finding would be published unless independently replicated. This isn't going to happen. However, good science should ensure that findings are not taken at face value, but subjected to further test. In this age of electronic publication it is possible to link to studies that independently replicate a finding. Journals and Associations are going to have to be more creative about how claims published in their pages are supported. Replication is going to have to be embraced.

It may be that authors will resist data sharing or making their analytic decisions public. However, resistance may be futile. The journals, including *The Political Methodologist*, are taking the high road and eventually will require openness in science.

One Norm, Two Standards: Realizing Transparency in Qualitative Political Science

Andrew Moravcsik
Princeton University
amoravcs@princeton.edu

Quantitative and qualitative political scientists are currently working closely together, and with their counterparts in other disciplines, to render social science research more transparent. The Data Access and Research Transparency (DA-RT) initiative launched by the American Political Science Association (APSA) has resulted in the new professional responsibility norms mandating transparency. To realize this goal, APSA co-hosted a workshop several months ago assembling thirteen editors of top political science journals of all types – including *American Political Science Review*, *International Security*, *PS: Political Science & Politics*, *American Journal of Political Science*, and *Comparative Political Studies*. The editors issued in a joint public statement committing their journals to implement new qualitative and quantitative transparency standards by January 2016. It is already gaining new adherents from journals outside the original group.[1] A number of other institutions, including the Qualitative Data Repository at the Institute for Qualitative and Multi-Method Research (IQMR), National Science Foundation (NSF), Social Science Research Council (SSRC), Berkeley Initiative for Transparency in the Social Sciences (BITSS) have are mounting efforts in this area as well.[2] Social science transparency is an idea whose time has come.

This paper addresses one perceived constraint on this trend: the lack of broad understanding (outside of those closely involved) about exactly what transparency implies for *qualitative* political science. Quantitative analysts have more self-awareness in this domain. They generally understand enhanced transparency to mean "more of the same,"

[1] http://www.data-pass.org/
[1] http://www.dartstatement.org/
[2] http://bitss.org/; https://qdr.syr.edu/;

that is, wider adoption of rules (already in place at some journals as a precondition for review or publication) requiring that data be placed in database at a trusted third-party repository with a digitally automated analytical protocol.

Yet what is the equivalent default transparency scheme for qualitative researchers, most of whom conduct process-tracing analyses of case studies? Finding an effective means to promote qualitative transparency is likely to have a deeper impact on political science than improving quantitative transparency – or, indeed, any other immediate methodological reform. This is true for three reasons. First, the majority of researchers in political science – over 90% in a sub-discipline like International Relations – employ qualitative methods (Maliniak, Peterson, and Tierney 2012, charts 28-30, 42, 57). Second, little has been done so far to promote qualitative transparency, so any change is likely to have a greater marginal impact. Third, for any changes at the disciplinary level to succeed, the active involvement of non-quantitative scholars will be required.

The answer may seem straightforward. Qualitative transparency ought to be enhanced by familiar means: either by generalizing conventional citation currently employed by qualitative analysts, or by introducing the same centralized data archiving employed by quantitative analysts. Yet scholars have recently come to understand – for reasons this essay will discuss in detail – that for most qualitative political science, neither of these solutions offers a practical means to enhance transparency. More innovative use of digital technology is required. A consensus is forming that the most workable and effective default transparency standard for qualitative political science is Active Citation: a system of digitally-enabled citations linked to annotated excerpts from original sources.

To explain how scholars reached these conclusions, this essay proceeds in four sections. The first section presents a consensual definition of research transparency and why almost all political scientists accept it. The second sets forth essential criteria for judging how this transparency norm should be implemented, realized and applied within specific research communities. The third explains why the best-known approaches – conventional citation, hyperlinks and digital archiving – fail to meet these criteria for most qualitative political science today. The fourth section concludes by describing Active Citation in detail and explains why it best fulfills the essential criteria of an applied standard.

1. What is Research Transparency?

"Research transparency" designates a disciplinary norm whereby scholars publicize the process by which they generate empirical data and analysis (King 1995, 444). It obliges scholars to present the evidence, analysis and methodological choices they use to reach research conclusions about the social world – in plain English, to "show their work." Recent scholarship has refined this definition by isolating three dimensions of research transparency.[3] The first, *data transparency*, obliges social scientists to publicize the evidence on which their research rests. The second dimension, *analytic transparency*, obliges social scientists to publicize how they measure, code, interpret, and analyze that data. The third dimension, *process transparency*, obliges social scientists to publicize the broader set of research design choices that gave rise to the particular combination of data, theories, and methods they employ.

Unlike almost any other methodological ideal, transparency unifies rather than divides social scientists across the full range of disciplines, epistemologies, methods, theories and substantive interests. Transparency enjoys this consensual status because it constitutes social science as a legitimate collective activity. To publicize data, theory and methods is to fulfill a basic ethical responsibility to act toward other scholars with openness and honesty. Underlying this responsibility is the realization that scholars are humans whose research choices are inevitably subject, sometimes unconsciously, to arbitrary interests, commitments and biases. To be part of a common scholarly conversation requires, therefore, that one admit fallibility and pay respect to readers and potential interlocutors by opening the choices one makes to meaningful public discussion and debate (Nosek, Spies, and Motyl 2012, 615-631). The Ethics Committee of the American Political Science Association (APSA) – which represents a very broad range of methods and interests – recently recognized this by expanding the professional responsibilities political scientists share to include data access and research transparency.[4]

Social scientific transparency is essential for a more pragmatic reason as well, namely that the legitimacy and credibility of academic findings (inside and outside of academia) rests in large part on the belief that they result from well-informed scholarly discussion and debate.[5] The recent discovery of a large number of non-replicable positive results in the natural and social sciences has called this legitimacy and credibility into question. Transparency offers one check on this tendency by inviting scholars to investigate, replicate, critique, extend and reuse the data, analysis and methods that their colleagues have published. When new work appears, other scholars are inspired to accept, reject or improve its findings. Citizens, private organizations, sponsoring bodies, and public decision makers can evaluate and apply the results, feeding back their experiences to researchers as new data and questions. Scholars are trained to contribute to the advancement of this collective enterprise and

[3] This tripartite distinction revises that found in Lupia and Elman (2014), appendices A and B.
[4] APSA, *Guidelines for Qualitative Transparency* (2013), see fn. 6.
[5] For a more complete discussion of this topic and others in this essay, see Moravcsik 2010, 2012, 2014a, 2014b.

are recognized and rewarded for doing so well. They challenge, extend, or borrow from prior data, analysis and methods to move in innovative directions, thereby renewing the flow of research and starting anew. Transparency not only permits this cycle of research to take place, but displays it publicly, enhancing its credibility and legitimacy – thereby ultimately justifying society's investment in it.

2. Principled and Pragmatic Criteria for Implementing Qualitative Transparency

Almost all social scientists recognize transparency as an abstract good. Yet political science is divided into diverse research communities with different methods, theories and substantive interests. It would be inappropriate for all to seek to realize transparency by applying the same concrete rules. Different styles of research require different procedures (Lupia and Elman 2014, fn.6). To understand how contemporary qualitative political scientists can best achieve transparency, we must understand two aspects of qualitative political science. One is the basic "process-tracing" epistemology most qualitative political scientists employ. The other is the set of everyday practical constraints they face.

A. The "Process-Tracing" Epistemology of Most Qualitative Political Science

Most qualitative political science researchers employ some form of "process-tracing" to investigate a few cases intensively. Comparing across cases (let alone analyzing a large number of qualitative observations in a database) plays a secondary role. Process-tracing analysts focus primarily within single cases, citing individual pieces of textual and factual evidence to infer whether initial conditions and subsequent events actually took place, and to establish the existence of (and relative importance of) descriptive or causal processes that may link cause and effect. Evidence appears as a set of heterogeneous insights or pieces of data about a causal mechanism and its context ("causal process observations") linked to specific successive points in a narrative or causal sequence. This inferential structure differs from that found in most quantitative political science, where data is typically analyzed as "part of a larger, systematized array of observations" ("dataset observations") describing cross-case variation.[6]

There are many reasons – ranging from scholarly, normative or policy interest in a particular case to unit heterogeneity or uniqueness – why a scholar might legitimately wish to focus on internal validity in this way rather than average outcomes across a larger population.[7] One methodological advantage of the process-tracing mode of qualitative analysis is that researchers can exercise a unique degree of interpretive subtlety in assigning an appropriate role, weight, and meaning to each piece of causal process evidence, depending on its position in the underlying temporal narrative or causal sequence and on its basic reliability in context. For example, they may view a particular fact as a necessary condition (a "hoop test"), a sufficient condition (a "smoking gun" test), INUS (an insufficient but necessary part of a condition which is itself unnecessary but sufficient for the result), SUIN (a sufficient but unnecessary part of a factor that is insufficient but necessary for an outcome), a probabilistic link (a "straw in the wind"), or another type of condition (Van Evera 1997, 32; Mahoney 2012; Collier 2011; Hall 2006). Within each category, a test may be viewed as more or less probative, based on the nature of the test, the particulars of the text, the reliability of the evidence, the broader political, social and cultural context, and the Bayesian priors in the literature. This nuanced but rigorous, discursively contextualized, source-by-source interpretation, which is highly prized in fields such as history and law, contrasts with the norm of imposing uniform rules that weight all data in a general database equally or randomly, which is more generally found in quantitative political science.

B. Pragmatic Constraints on Process-Tracing Research

Perfect research transparency seems attractive in theory. Yet five practical considerations limit the feasibility and desirability of any effort to maximize the evidentiary, analytic and procedural information that any scholar, qualitative or quantitative, reveals.

1. *Intellectual property law* imposes limits on the secondary use of published material: in most countries a modest quotation can be employed for non-profit or scholarly purposes, but entire published documents cannot be reproduced at will. Even sources formally in the public domain (such as archival documents) are often subject to informal restrictions, and any scholar who disseminated them wholesale might find it difficult to research a related topic again in the future.

2. *Confidentiality for human subject protection* comes into play when scholars agree (either informally or as part of a formal research design or Institutional Review Board-approved arrangement) to keep information confidential. Some such information cannot be cited at all; some has to be sanitized.

[6] This distinction was originally introduced by Henry Brady, David Collier and Justin Seawright, subsequently adopted by James Mahoney, Gary Goertz and many others, and has now become canonical. The causal process mode of inferring causality, in Brady and Collier's words, "contributes a different kind of leverage in causal inference... A causal-process observation may [give] insight into causal mechanisms, insight that is essential to causal assessment and is an indispensable alternative and/or supplement to correlation-based causal inference" (Brady and Collier 2010, 252-3).

[7] See Moravcsik 2014b for a discussion.

3. An *unreasonable logistical burden* may arise if scholars are obliged to deposit, sanitize or reproduce massive quantities of data in qualitative form, particularly if they did not expect to do so when collecting the data.

4. Scholars should expect to enjoy a reasonable *first-use right to exploit newly collected data and analysis* before other scholars can access it. This helps incentivize and fairly reward scholars, particularly younger ones, who collect and analyze data in new ways.

5. *Publication procedures and formats* are often costly to change, especially in the short-term.[8]

C. What Should We Look for in an Applied Transparency Standard?

For qualitative process-tracing to be intelligible, credible and legitimate, data and analysis must be transparent and the major research choices must be justified. The unique interpretive flexibility qualitative scholars enjoy only serves to increase transparency requirements. Any appropriate and workable standard of qualitative transparency be suited to this type of research. It must preserve the basic narrative "process-tracing" structure of presentation. Scholars should provide readers with the data and analytical interpretation of each piece of evidence in context, and a methodological justification for their selection. Readers must be able to move efficiently, in real time, from a point in the main narrative directly to the source and its analysis, and back again – a function traditionally carried out by footnotes and endnotes. Third, analytic transparency provisions must permit scholars to explain the interpretive choices they have made with regard to each piece of evidence. All this must take place within the real-world constraints set by intellectual property law, human subject protection, logistics, first use rights and existing publication formats.

3. Can Existing Transparency Instruments Enhance Qualitative Research Transparency?

This section examines three relatively familiar options for enhancing social scientific transparency: conventional citations, hyperlinks to external web sources and data archiving. Though all are useful in specific circumstances, none offers a workable default standard for qualitative political scientists given the constraints discussed above.

A. Conventional Citation

The transparency standard employed today in political science, conventional citation, is manifestly unable to provide even minimal qualitative research transparency.[9] To be sure, the potential exists. Legal academics and historians employ "best practices," such as ubiquitous discursive footnotes containing long, annotated quotations and a disciplinary discourse in which other scholars often challenge textual claims. Law reviews permit readers to scan at a glance the main argument, quotations from sources and the basis of the interpretation. Historical journals note the exact location, nature and interpretation of a document, often backed by quotations.

Yet in political science qualitative transparency has been rendered all but impossible by recent trends in formatting journals and books: notably ever tighter word limits and so-called scientific citations designed to accommodate references only to secondary literature rather than data. Political science has regressed: decades ago, discursive long-form citations permitted political scientists to achieve greater research transparency than they can today. This trend is unlikely to reverse. The results have been profound, extending far beyond format. Political science has seen a decay in the humanistic culture of appreciation, debate, reuse, extension and increasing depth of text-based empirical understanding – and an atrophy in the skills to engage in such debates, such as linguistic, functional, regional area, policy and historical knowledge remain prized, skills that remain commonplace in legal academia, history and other disciplines. Only in small pockets of regional area, historical or policy analysis linked by common linguistic, functional or historical knowledge does qualitative transparency and depth of debate persist. This is a major reason for the crisis in qualitative political science today.

B. Two Digital Alternatives: External Hyperlinks and Data Archiving

If fundamental reform of conventional citation practice is unlikely, then effective enhancement of research transparency will require intensive application of digital technology. In recent years, political scientists have considered the digital alternatives. Two possibilities mentioned most often – external hyperlinks and data archiving – prove on close inspection to be too narrowly and selectively applicable to be broad default instruments to enhance qualitative transparency.

The first digital option is to ***hyperlinks to external web-based sources***. Hyperlinks are now ubiquitous in journalism, policy analysis, social media, government reports and law, as well as scholarly disciplines (such as medicine) in which researchers mainly reference other scholars, not data. A few areas of political science are analogous. Yet most sources political scientists cite – including the bulk of archival materials, other primary documents, pamphlets

[8] For previous discussions of these factors, see Moravcsik 2010, 2012, 2014a, 2014b.
[9] See Moravcsik 2010 for a lengthier discussion of this point.

and informally published material, books published in the past 70 years, raw interview transcripts, ethnographic field materials, photographs, diagrams and drawings – are unavailable electronically. Even when sources exist on-line, external hyperlinks offer only data access, not analytic or process transparency. We learn what a scholar cited but not why or how. In any case, as we shall see below, other transparency instruments subsume the advantages of hyperlinks.

The second digital option is to ***archive_data*** in a single unified database. Data archiving has obvious advantages: it is consistent with common practice in quantitative analysis, builds a procedural bulwark against selection bias (cherry-picking) of favorable quotations and documents; and, with some further technological innovation, and might even permit specific textual passages to be linked conveniently to individual citations in a paper. Certainly data archiving is essential to preserve and publicize complete collections of new and unique field data, such as original interviews, ethnographic notes, primary document collections, field research material, and web scrapes – an important responsibility for social scientists. Data archiving is also the best means to present qualitative data that is analyzed as a single set of "dataset observations," as in some content, ethnographic or survey analyses. Databases such as Atlas, Access, Filemaker, and Endnote now offer promising and innovative way to store, array and manipulate textual data particularly useful in research designs that emphasize macro-comparative inquiry, systematic coding, content analysis, or weighing of a large number of sources to estimate relatively few, carefully predefined variables – often in "mixed-method" research designs (Lieberman 2010). These important considerations have recently led political scientists to launch data repositories for textual material, notably the Qualitative Data Repository established with NSF funding at Syracuse University.[10]

Yet, however useful it may be for these specific purposes, data archiving is not an efficient format for enhancing general qualitative research transparency. This is because it copes poorly with the modal qualitative researcher, who employs process tracing to analyze causal process observations. The reasons are numerous. First, intellectual property law imposes narrow *de jure* and *de facto* limits on reproducing textual sources, including almost all secondary material published in the seventy years and a surprising portion of archival government documents, private primary material, web text, commercial material and artistic and visual products. Second, many sources (even if legally in the public domain) cannot be publicized due to confidentiality agreements and human subject protection enforced by Institutional Review Boards (IRBs). These include many interview transcripts, ethnographic observation, and other unpublished material. Even when partial publication is possible, the cost and risk of sanitizing the material appropriately can be prohibitively high, if possible at all.[11] Third, the logistical burden of data archiving imposes additional limits. In theory addressing "cherry picking" by archiving all qualitative data a scholar examines seems attractive, yet it is almost always impractical. Serious primary source analysts sift through literally *orders of magnitude* more source material than they peruse intensively, and peruse *orders of magnitude* more source material than are important enough to cite. Almost all the discarded material is uninteresting. Sanitizing, depositing and reading tens or hundreds of times more material than is relevant would create prohibitive logistical burdens for scholar and reader alike. Finally, even if all or some relevant data is archived efficiently, it does little to enhance analytic transparency. As with hyperlinking, we learn what a scholar cited but not why. Overall, these limitations mean that data archiving may be a first best transparency approach for qualitative evidence that is arrayed and analyzed in the form of database observations, but it is sub-optimal for the modal process-tracing form of analysis.

4. A Practical yet Effective Standard: Active Citation

Conventional citation, hyperlinks and data archiving, we have seen, have inherent weaknesses as general transparency standards. A new and innovative approach is required. Scholars are converging to the view that the most promising and practical default standard for enhancing qualitative transparency is *Active Citation (AC)*: a system of digitally-enabled citations linked to annotated excerpts from original sources.

In the AC format, any citation to a contestable empirical claim is hyperlinked to an entry in an appendix appended to the scholarly work (the "Transparency Appendix" (TRAX)).[12] Each TRAX entry contains four elements, the first three required and the last one optionally:

1. a short excerpt from the source (presumptively 50-100 words long);
2. an annotation explaining how the source supports the underlying claim in the main text (of a length at the author's discretion);
3. the full citation;

[10] For link, see fn. 2.

[11] It is striking, for example, that even quantitative political science journals that pride themselves on a rigorous standard of "replicability" often apply that standard only to quantitative data, and resolutely refuse to publicize the qualitative data that is coded to create it – precisely for these reasons.

[12] For a longer discussion of AC, with examples, see Moravcsik 2014b.

4. *optionally*, a scan of or link to the full source.

The first entry of the TRAX is reserved exceptionally for information on general issues of production transparency. Other TRAX entries can easily be adapted to presenting sources in visual, audio, cinematic, graphic, and other media. AC can be employed in almost any form of scholar work: unpublished papers, manuscripts submitted for publication, online journal articles, and e-books, or as separate online appendices to printed journals and books. Examples of AC can be found on various demonstration websites and journal articles.

AC is an innovative yet pragmatic standard. It significantly enhances research transparency while respecting and managing legal, administrative, logistical and commercial constraints. The relatively short length of the source excerpt assures basic data transparency, including some interpretive context, while remaining avoiding most legal, human subject, logistical, and first-use constraints.[13] (If such constraints remain, they trump the requirement to publicize data, but scholars may employ an intermediate solution, such as providing a brief summary of the data.) The optional scan or link offers the possibility of referencing a complete source, when feasible – thereby subsuming the primary advantage of hyperlinking and data archiving. The annotation delivers basic analytic transparency by offering an opportunity for researchers to explain how the source supports the main text – but the length remains at the discretion of the author. The exceptional first entry enhances process transparency by providing an open-ended way of addressing research design, selection bias, and other general methodological concerns that remain insufficiently explained in the main text, footnotes, empirical entries, or other active citations. The full citation assures that each TRAX entry is entirely self-contained, which facilitates convenient downloading into word-processing, bibliographical and database software. This contributes over time to the creation of a type of networked database of annotated quotations, again subsuming an advantage of data archiving at lower logistical cost. Finally, AC achieves these goals while only minimally disrupting current publishing practices. Except for the hyperlinks, existing formats and word limits remain unchanged (the TRAX, like all appendices, lies outside word limits). Journals would surely elect to include active footnotes only in electronic formats. Editors need not enforce *ex post* archiving requirements.

AC is a carefully crafted compromise reflecting years of refinement based on numerous methodological articles, workshops, discussions with editors, interdisciplinary meetings, consultations with potential funders, instructional sessions with graduate students, and, perhaps most importantly, consultation sessions with researchers of all methodological types. Underneath the digital technology and unfamiliar format lies an essentially conservative and cautious proposal. Interpretivists and traditional historians will note that it recreates discursive citation practices they have long employed in a digital form. Quantitative researchers will note that it adheres to the shared norm of placing data and analysis in a single third-party "database," though it interprets that term in a very different manner than conventional statistical researchers do.

For these reasons, AC has emerged as the most credible candidate to be the default transparency standard for qualitative papers, articles and, eventually, books in political science. It is rapidly sinking disciplinary roots. AC has been elaborated and taught in published articles, at disciplinary and interdisciplinary conferences, at training institutes and at graduate seminars. The NSF-funded Qualitative Data Repository has commissioned ten "active citation compilations" by leading scholars of international relations and comparative politics, who are retrofitting classic and forthcoming articles or chapters to the active citation format. Software developers are creating tools to assist in preparing TRAXs, in particular via software add-ons to automate the formatting of transparency appendices and individual entries in popular word processing programs. The new APSA ethics and professional responsibility rules recommend AC, as do, most importantly, the APSA-sponsored common statement in which 15 journals (so far) jointly committed to move to enhanced transparency in January 2016.[14]

5. Conclusion: A Transparent Future

At first glance, transparency may appear to be an unwelcome burden on qualitative researchers. Yet in many respects it should be seen instead as a rare and valuable opportunity. The cost of implementing active citation is relatively low, particularly when one knows in advance that it is expected.[15] After all, legal scholars and many historians – not to mention political scientists in generations past – have long done something similar as a matter of course, and without the assistance of word processers, hand-held cameras and the web, which ease archival and textual preservation and recognition. More importantly, greater transparency offers political scientists large individual and collective benefits. It provides an opportunity to demonstrate clearly, and to be rewarded for, scholarly excellence. This in turn is likely

[13] In the US, fifty to one hundred words for non-profit scholarly use lies within the customary "fair use" exception, with the exception of artistic products. Most other countries have similar legal practices. This length is also relatively easy to scan, logistically feasible to sanitize for human subject protection, and poses less of a threat to "first use"exploitation rights of new data.

[14] For link, see fn 2.

[15] For a detailed argument and evidence that the costs are low, see Moravcsik 2012 and 2014b. Even those who point out that retrofitting articles to the standard is time-consuming point out that with advance knowledge, the workload will diminish considerably. See Saunders 2014.

to incentivize scholars to invest in relevant skills, which include interpretive subtlety in reading texts, process-tracing and case selection techniques, and deep linguistic, area studies, historical, functional and policy knowledge. These skills will be in greater demand not just to conduct research, but to referee and debate it. Perhaps most important, the networked pool of transparent data, analysis and methods in active citations will constitute an ever-expanding public good for researchers, who can appreciate, critique, extend and reuse that data and analysis, just as quantitative scholars make low-cost use of existing datasets and analytical techniques.

The trend toward more open research is not just desirable; it is inevitable. In almost every walk of life – from science and government to shopping and dating – the wave of digital transparency has proven irresistible. The question facing qualitative political scientists is no longer whether they will be swept along in this trend. The question is when and how. Our responsibility is to fashion norms that acknowledge and respect the epistemological priorities and practices qualitative political scientists share.

References

Brady, Henry and David Collier. 2010. *Rethinking Social Inquiry: Diverse Tools, Shared Standards*. Lanham, MD: Rowman & Littlefield.

Collier, David. 2011. "Understanding Process Tracing." *PS: Political Science & Politics* 44(4): 823-30.

Hall, Peter A. 2006. "Systematic Process Analysis: When and How to Use It." *European Management Review* 3: 24-31.

Maliniak, Daniel, Susan Peterson, and Daniel J. Tierney. 2012. *TRIP around the World: Teaching, Research and Policy Views of International Relations Faculty in 20 Countries*. Williamsburg, Virginia: Institute for the Theory and Practice of International Relations, College of William and Mary. `http://www.wm.edu/offices/itpir/_documents/trip/trip_around_the_world_2011.pdf` (February 1, 2015).

Lupia, Arthur, and Colin Elman. 2014. "Openness in Political Science: Data Access and Research Transparency." *PS: Political Science & Politics* 47(1): 19-42.

King, Gary. 1995. "Replication, Replication." *PS: Political Science & Politics* 28(3): 444-52.

Lieberman, Evan S. 2010. "Bridging the Qualitative-Quantitative Divide: Best Practices in the Development of Historically Oriented Replication Databases." *Annual Review of Political Science* 13: 37-59.

Nosek, Brian A. Jeffrey R. Spies, and Matt Motyl. 2012. "Scientific Utopia: II. Restructuring Incentives and Practices to Promote Truth over Publishability." *Perspectives on Psychological Science* 7(6): 615-31.

Mahoney, James. 2012. "The Logic of Process-Tracing Tests in the Social Sciences." *Sociological Methods & Research* 41(4): 570-97.

Moravcsik, Andrew. 2010. "Active Citation: A Precondition for Replicable Qualitative Research." *PS: Political Science & Politics* 43(1): 29-35.

Moravcsik, Andrew. 2012. "Active Citation and Qualitative Political Science." *Qualitative & Multi-Method Research* 10(1):33-7.

Moravcsik, Andrew. 2014a. "Trust, yet Verify: The Transparency Revolution and Qualitative International Relations." *Security Studies* 23(4): 663-88.

Moravcsik, Andrew. 2014b. "Transparency: The Revolution in Qualitative Political Science." *PS: Political Science & Politics* 47(1): 48-53.

Saunders, Elizabeth. 2014. "Transparency without Tears: A Pragmatic Approach to Transparent Security Studies Research." *Security Studies* 23(4): 689-98.

Van Evera, Stephen. 1997. *Guide to Methods for Students of Political Science*. Ithaca, NY: Cornell University Press.

Improving Research Transparency in Political Science: Replication and Political Analysis

R. Michael Alvarez
California Institute of Technology
rma@hss.caltech.edu

Recently, the American Political Science Association (APSA) launched an initiative to improve research ethics. One important outcome of this initiative is a joint statement that a number of our discipline's major journals have signed: the Data Access and Research Transparency statement. These journals include *Political Analysis*, the *American Political Science Review*, the *American Journal of Political Science*, and at present a handful of other journals.

The joint statement outlines four important goals for these journals:

1. Require that authors provide their datasets at the time of publication in a trusted digital repository.

2. Require that authors provide analytic procedures so that the results in their publication can be replicated.

3. Develop and implement a data citation policy.

4. Make sure that journal guidelines and other materials delineate these requirements.

We are happy to report that *Political Analysis* has complied with these requirements and in the case of our policies and procedures on research replication, our approach has provided an important example for how replication can be implemented by a major journal. Our compliance with these requirements is visible in the new instructions for authors and reviewers that we recently issued.

Political Analysis has long had a policy that all papers published in our journal need to provide replication data. Recently we have taken steps to strengthen our replication policy and to position our journal as a leader on this issue. The first step in this process was to require that before we send an accepted manuscript to production that the authors provide the materials necessary to replicate the results reported in the manuscript. The second step was for us to develop a relatively simple mechanism for the archiving of this replication material the development of the Political Analysis Dataverse. We now have over 200 replication studies in our Dataverse, and these materials have been downloaded over 15,000 times.

Exactly how does this work? Typically, a manuscript successfully exits the review process, and the editors conditionally accept the manuscript for publication. One of the conditions, of course, is that the authors upload their replication materials to the journal's Dataverse, and that they insert a citation to those materials in the final version of their manuscript. Once the materials are in the journal's Dataverse, and the final version of the manuscript has been returned to our editorial office, both receive final review. As far as the replication materials go, that review usually involves:

1. An examination of the documentation provided with the replication materials.

2. Basic review of the provided code and other analytic materials.

3. A basic audit of the data provided with the replication materials.

The good news is that in most cases, replication materials pass this review quickly authors know our replication requirement and most seem to have worked replication into their research workflow.

Despite what many may think (especially given the concerns that are frequently expressed by other journal editors when they hear about our replication policy), authors do not complain about our replication policy. We've not had a single instance where an author has refused or balked about complying with our policy. Instead, the vast majority of our authors will upload their replication materials quickly after we request them, which indicates that they have them ready to go and that they build the expectation of replication into their research workflow.

The problems that we encounter generally revolve around adequate documentation for the replication materials, clarity and usability of code, and issues with the replication data itself. These are all issues that we are working to develop better guidelines and policies regarding, but here are some initial thoughts.

First, on documentation. Authors who are developing replication materials should strive to make their materials as usable as possible for other researchers. As many authors already know, by providing well-documented replication materials they are increasing the likelihood that another scholar will download their materials and use them in their own research which will likely generate a citation for the replication materials and the original article they come from. Or a colleague at another university will use well-documented replication materials in their methods class, which will get the materials and the original article in front of many students. Perhaps a graduate student will download the materials and use them as the foundation for their dissertation work, again generating citations for the materials and the original article. The message is clear; well-documented replication materials are more likely to be used, and the more they are used the more attention the original research will receive.

Second, clarity and usability of code. For quantitative research in social science, code (be it R, Stata, SPSS, Python, Perl or something else) is the engine that drives the analysis. Writing code that is easy to read and use is critical for the research process. Writing good code is something that we need to focus more attention on in our research methodology curriculum, and as a profession we need more guidelines regarding good coding practices. This is an issue that we are *Political Analysis* will be working on in the near future, trying to develop guidelines and standards for good coding practices so that replication code is more usable.

Finally data. There are two primary problems that we see with replication data. The first is that authors provide data without sufficiently clearing the data of "personally identifying information" (PII). Rarely is PII necessary in replication data; again, the purpose of the replication material is to reproduce the results reported in the manuscript. Clearly there may be subsequent uses of the replication data, in which another scholar might wish to link the replication materials to other datasets. In those cases we urge the producer of the replication materials to provide some indication in their documentation about how they can be contacted to assist in that linkage process.

The second problem we see regards the ability of authors to provide replication data that can be made freely available. There are occasions where important research uses proprietary data; in those situations we encourage authors to let the editors know that they are using proprietary or

restricted data upon submission so that we have time to figure out how to recommend that the author comply with our replication requirement. Usually the solution entails having the author provide clear details about how one would reproduce the results in the paper were one to have access to the proprietary or restricted data. In many cases, those who wish to replicate a published paper may be able to obtain the restricted data from the original source, and in such a situation we want them to be able to know exactly each step that goes from raw data to final analysis.

Recently we updated our replication policies, and also developed other policies that help to increase the transparency and accessibility of the research that is published in *Political Analysis*. However, policies and best practices in these areas are currently evolving and very dynamic. We will likely be updating the journal's policies frequently in the coming years, as *Political Analysis* is at the forefront of advancing journal policies in these areas. We are quite proud of all that we have accomplished regarding our replication and research policies at *Political Analysis*, and happy that other journals look to us for guidance and advice.

The Use of Replication in Graduate Education and Training

Wendy L. Martinek
Binghamton University
martinek@binghamton.edu

Writing almost 20 years ago as part of a symposium on the subject,[1] King (1995) articulated a strong argument in favor of the development of a replication standard in political science. As King wrote then, "Good science requires that we be able to reproduce existing numerical results, and that other scholars be able to show how substantive findings change as we apply the same methods in new contexts" (King 1995, 451). Key among the conditions necessary for this to occur is that the authors of published work prepare replication data sets that contain everything needed to reproduce reported empirical results. And, since a replication data set is not useful if it is not accessible, also important are authors' efforts to make replication data sets easily available. Though his argument and the elements of his proposed replication standard were not tied to graduate education per se, King did make the following observation: "Reproducing and then extending high-quality existing research is also an extremely useful pedagogical tool, albeit one that political science students have been able to exploit only infrequently given the discipline's limited adherence to the replication standard" (King 1995, 445). Given the trend towards greater data access and research transparency, King (2006) developed a guide for the production of a publishable manuscript based on the replication of a published article.

With this guide in hand, and informed by their own experiences in teaching graduate students, many faculty members have integrated replication assignments into their syllabi. As Herrnson has observed, "Replication repeats an empirical study in its entirety, including independent data collection" (1995, 452). As a technical matter, then, the standard replication assignment is more of verification assignment than a true replication assignment. Regardless, such assignments have made their way onto graduate syllabi in increasing numbers. One prominent reason – and King's (2006) motivation – is the facilitation of publication by graduate students. The academic job market is seemingly tighter than ever (Jaschik 2009) and publications are an important element of an applicant's dossier, particularly when applying for a position at a national university (Fuerstman and Lavertu 2005). Accordingly, incorporating an assignment that helps students produce a publishable manuscript whenever appropriate makes good sense. Well-designed replication assignments, however, can also serve other goals. In particular, they can promote the development of practical skills, both with regard to the technical aspects of data access/manipulation and with regard to best practices for data coding/maintenance. Further, they can help students to internalize norms of data accessibility and research transparency. In other words, replication assignments are useful vehicles for advancing graduate education and training.

A replication assignment that requires students to obtain the data set and computer code to reproduce the results reported in a published article (and then actually reproduce those results) directs student attention to three very specific practical tasks. They are tasks that require skills often taken for granted once mastered, but which most political science graduate students do not possess when starting graduate school (something more advanced graduate students and faculty often forget). Most basically, it requires students to work out how to obtain the data and associated documentation (e.g., codebook). Sometimes this task turns out to be ridiculously easy, as when the data is publicly archived, either on an author's personal webpage or through a professional data archive (e.g., Dataverse Network, ICPSR). But that is certainly not always the case, much to students' chagrin and annoyance (King 2006, 120; Carsey 2014, 74-5). To be sure, the trend towards greater data accessibility is reflected in, for example, the editorial policies of many

[1] The symposium appeared in the September 1995 issue of *PS: Political Science & Politics*.

political science journals[2] and the data management and distribution requirements imposed by funding agencies like the National Science Foundation.[3] Despite this trend, students undertaking replication assignments not infrequently find that they have to contact the authors directly to obtain the data and/or the associated documentation. The skills needed for the simple (or sometimes not-so-simple) task of locating data may seem so basic as to be trivial for experienced researchers. However, those basic skills are not something new graduate students typically possess. A replication assignment by definition requires inexperienced students to plunge in and acquire those skills.

The second specific task that such a replication assignment requires is actually figuring out how to open the data file. A data file can be in any number of formats (e.g., .dta, .txt, .xls, .rda). For the lucky student, the data file may already be available in a format that matches the software package she intends to use. Or, if not, the student has access to something like Stat/Transfer or DBMS-Copy to convert the data file to a format compatible with her software package. This, too, may seem like a trivial skill to an experienced researcher. That is because it is a trivial skill to an experienced researcher. But it is not trivial for novice graduate students. Moreover, even more advanced graduate students (and faculty) can find accessing and opening data files from key repositories such as ICPSR daunting. For example, students adept at working with STATA, SAS, and SPSS files might still find it less than intuitive to open ASCII-format data with setup files. The broader point is that the mere act of opening a data file once it has been located is not necessarily all that obvious and, as with locating a data file, a replication assignment can aid in the development of that very necessary skill.

The third specific task that such a replication assignment requires is learning how to make sense of the content of someone else's data file. In an ideal world (one political scientists rarely if ever occupy), the identity of each variable and its coding are crystal clear from a data set's codebook alone. Nagler outlines best practices in this regard, including the use of substantively meaningful variable names that indicate the subject and (when possible) the direction of the coding (Nagler 1995, 490). Those conventions are adhered to unevenly at best, however, and the problem is exacerbated when relying on large datasets that use either uninformative codebook numbers or mnemonics that make sense but only to experienced users. For example, the General Social Survey (GSS) includes the SPWRKSTA variable. Once the description of the variable is known ("spouse labor force status") then the logic of the mnemonic makes some sense: SP = spouse, WRK = labor force, STA = status. But it makes no sense to the uninitiated and even an experienced user of the GSS might have difficulty recalling what that variable represents without reference to the codebook. There is also a good deal of variation in how missing data is coded across data sets. Not uncommonly, numeric values like 99 and -9 are used to denote a missing value for a variable. That is obviously problematic if those codes are used as nonmissing numeric values for the purposes of numeric calculations. Understanding what exactly "mystery name" variables reference and how such things as missing data have been recorded in the coding process are crucial for a successful replication. The fact that these things are so essential for a successful replication forces students to delve into the minutia of the coded data and become familiar with it in a way that is easier to avoid (though still unadvisable) when simply using existing data to estimate an entirely new model.

Parenthetically, the more challenges students encounter early on when learning these skills, the better off they are in the long run for one very good reason. Students receive lots of advice and instruction regarding good data management and documentation practices (e.g., Nagler 1995). But there is nothing like encountering difficulty when using someone else's data to bring home the importance of relying on best practices in coding one's own data. The same is true with regard to documenting the computer code (e.g., STATA do-files, R scripts). In either case, the confusions and ambiguities with which students must contend when replicating the work of others provide lessons that are much more visceral and, hence, much more effective in fostering the development of good habits and practices than anything students could read or be told by their instructor.

These three specific tasks (acquiring a data set and its associated documentation, then opening and using that data set) require skills graduate students should master very early on in their graduate careers. This makes a replication assignment especially appealing for a first- or second-semester methods course. But replication assignments are also valuable in more advanced methods courses and substantive classes. An important objective in graduate education is the training and development of scholars who are careful and meticulous in the selection and use of methodological tools. But, with rare exception, the goal is not methodological proficiency for its own sake but, rather, methodological proficiency for the sake of advancing theoretical understanding of the phenomena under investigation. A replication assignment is ideal for grounding the development of methodological skills in a substantively meaningful context, thereby helping to fix the notion in students' minds of methodological tools as in the service of advancing theoretical understanding.

Consider, for example, extreme bounds analysis (EBA), a useful tool for assessing the robustness of the relationship between a dependent variable and a variety of possible

[2]See, for example, `http://ajps.org/guidelines-for-accepted-articles/` (November 15, 2014).
[3]See `http://www.nsf.gov/sbe/SBE_DataMgmtPlanPolicy.pdf` (November 15, 2014).

determinants (Leamer 1983). The basic logic of EBA is that, the smaller the range of variation in a coefficient of interest given the presence or absence of other explanatory variables, the more robust that coefficient of interest is. It is easy to imagine students focusing on the trivial aspects of determining the focus and doubt variables (i.e., the variables included in virtually all analyses and the variables that may or may not be included depending upon the analysis) in a contrived class assignment. A replication assignment by its nature, however, requires a meaningful engagement with the extant literature to understand the theoretical consensus among scholars as to which variable(s) matter (and, hence, which should be considered focus rather than doubt variables). Matching methods constitute another example. Randomized experiments, in which the treatment and control groups differ from one another only randomly vis-à-vis both observed and unobserved covariates, are the gold standard for causal inference. However, notwithstanding innovative resources such as Time-Sharing Experiments for the Social Sciences (TESS) and Amazon's Mechanical Turk and the greater prevalence of experimental methods, much of the data available to political scientists to answer their questions of interest are observational. Matching methods are intended to provide leverage for making causal claims based on observational data through the balancing of the distribution of covariates in treatment and control groups regardless of the estimation technique employed post-matching (Ho et al. 2007). Considering matching in the context of a published piece of observational research of interest to a student necessitates that the student is thinking in substantive terms about what constitutes the treatment and what the distribution of covariates looks like. As with EBA, a replication assignment in which students are obligated to apply matching methods to evaluate the robustness of a published observational study would insure that the method was tied directly to the assessment (and, hopefully, advancement) of theoretical claims rather than as an end to itself.

Though there remain points of contention and issues with regard to implementation that will no doubt persist, there is currently a shared commitment to openness in the political science community, incorporating both data access and research transparency (DA-RT). This is reflected, for example, in the data access guidelines promulgated by the American Political Science Association (Lupia and Elman 2014). The training and mentoring provided to graduate students in their graduate programs are key components of the socialization process by which they learn to become members of the academic community in general and their discipline in particular (Austin 2002). Replication assignments in graduate classes serve to socialize students into the norms of DA-RT. As Carsey notes, "Researchers who learn to think about these issues at the start of their careers, and who see value in doing so at the start of each research project, will be better able to produce research consistent with these principles" (Carsey 2014, 75). Replication assignments serve to inculcate students with these principles. And, while they have obvious value in the context of methods courses, to fully realize the potential of replication assignments in fostering the development of these professional values in graduate students they should be part of substantive classes as well. The more engaged students are with the substantive questions at hand, the easier it should be to engage their interest in understanding the basis of the inferences scholars have drawn to answer those questions and where the basis for those inferences can be improved to the betterment of theoretical understanding. In sum, the role of replication in graduate education and training is both to develop methodological skills and enhance theory-building abilities.

References

Austin, Ann E. 2002. "Preparing the Next Generation of Faculty: Graduate School as Socialization to the Academic Career." *Journal of Higher Education* 73(1): 94-122.

Carsey, Thomas M. 2014. "Making DA-RT a Reality." *PS: Political Science & Politics* 47(1): 72-77.

Fuerstman, Daniel and Stephen Lavertu. 2005. "The Academic Hiring Process: A Survey of Department Chairs." *PS: Political Science & Politics* 38(4): 731-736.

Herrnson, Paul S. 1995. "Replication, Verification, Secondary Analysis, and Data Collection in Political Science." *PS: Political Science & Politics* 28(3): 452-5.

Ho, Daniel E., Kosuke Imai, Gary King, and Elizabeth A. Stuart. 2007. "Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference." *Political Analysis* 15(3): 199-236.

Jaschik, Scott. 2009. "Job Market Realities." *Inside Higher Ed*, September 8. `https://www.insidehighered.com/news/2009/09/08/market` (November 29, 2014).

King, Gary. 1995. "Replication, Replication." *PS: Political Science & Politics* 28(3): 444-52.

King, Gary. 2006. "Publication, Publication." *PS: Political Science & Politics* 39(1): 119-125.

Leamer, Edward. 1983. "Let's Take the 'Con' Out of Econometrics." *American Economic Review* 73(1): 31-43.

Lupia, Arthur and Colin Elman. 2014. "Openness in Political Science: Data Access and Research Transparency." *PS: Political Science & Politics* 47(1): 19-42.

Nagler, Jonathan. 1995. "Coding Style and Good Computing Practices." *PS: Political Science & Politics* 39(1): 488-492.

On the Replication of Experiments in Teaching and Training

Jonathan Rogers
New York University Abu Dhabi
jonathan.rogers@nyu.edu

Introduction

Students in the quantitative social sciences are exposed to high levels of rational choice theory. Going back to Marwell and Ames (1981), we know that economists free ride, but almost no one else does (in the strict sense anyway). In part, this is because many social science students are essentially taught to free ride. They see these models of human behavior and incorrectly take the lesson that human beings *should* be rational and free ride. To not free ride would be irrational. Some have difficulty grasping that these are models meant to predict, not prescribe and judge human behavior.

Behaviorally though, it is well established that most humans are not perfectly selfish. Consider the dictator game, where one player decides how much of her endowment to give to a second player. A simple Google Scholar search for dictator game experiments returns nearly 40,000 results. It is no stretch to posit that almost none of these report that every first player kept the whole endowment for herself (Engel 2011). When a new and surprising result is presented in the literature, it is important for scholars to replicate the study to examine its robustness. Some results however, are so well known and robust, that they graduate to the level of empirical regularity.

While replication of surprising results is good for the discipline, replication of classic experiments is beneficial for students. In teaching, experiments can be used to demonstrate the disconnect between Nash equilibrium and actual behavior and to improve student understanding of the concept of modeling. Discussions of free-riding, the folk theorem, warm glow, and the like can all benefit from classroom demonstration. For graduate students, replication of experiments is also useful training, since it builds programming, analysis, and experimenter skills in an environment where the results are low risk to the grad student's career. For students of any type, replication is a useful endeavor and one that should be encouraged as part of the curriculum.

Replication in Teaching

Budding political scientists and economists are virtually guaranteed to be introduced, at some level, to rational choice. Rational choice is characterized by methodological individualism and the maximization of self interest. That is, actors (even if the actor of interest is a state or corporation) are assumed to be individuals who make choices based on what they like best. When two actors are placed in opposition to one another, they are modeled as acting strategically to maximize their own payoffs and only their own payoffs.

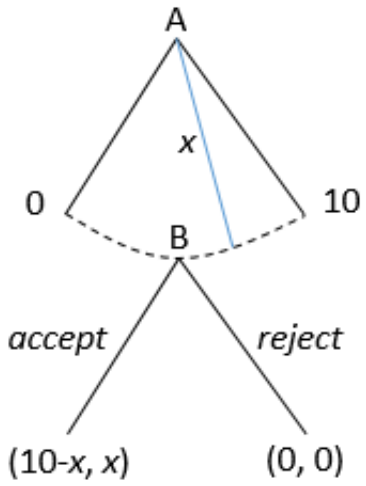

Figure 1: Ultimatum Game

Consider the classic ultimatum game, presented in Figure 1. Player A is granted an endowment of 10 tokens and is tasked with choosing how much to give to player B. Player B can then decide to either accept or reject the offer. If she accepts, then the offer is enforced and subjects receive their payments. If she rejects the offer, then both players receive nothing. In their game theory course work, students are taught to identify the Nash equilibrium through backward induction. In the second stage, player B chooses between receiving 0 and receiving the offer x, with certainty. Since she is modeled as being purely self interested, she accepts the offer, no matter how small. In the first stage, player A knows that player B will accept any offer, so she gives the smallest $\epsilon > 0$ possible. This yields equilibrium payoffs of $(10 - \epsilon, \epsilon)$.

Students are taught to identify this equilibrium and are naturally rewarded by having test answers marked correct. Through repeated drilling of this technique, students become adept at identifying equilibria in simple games, but make the unfortunate leap of seeing those who play the rational strategy as being smarter or better. A vast literature reports that players rarely make minimal offers and that such offers are frequently rejected (Oosterbeek, Sloof, and van de Kuilen 2004). Sitting with their textbooks however, students are tempted to misuse the terminology of rational choice and deem irrational any rejection or non-trivial offer. Students need to be shown that Nash equilibria are sets of strategy profiles derived from models and not inherently predictions in and of themselves. Any model is an

abstraction from reality and may omit critical features of the scenario it attempts to describe. A researcher may predict that subjects will employ equilibrium strategies, but she may just as easily predict that considerations such as trust, reciprocity, or altruism might induce non-equilibrium behavior. The Nash Equilibrium is a candidate hypothesis, but it is not necessarily unique.

This argument can be applied to games with voluntary contribution mechanisms. In the public goods game for example, each player begins with an endowment and chooses how much to contribute to a group account. All contributions are added together, multiplied by an efficiency factor, and shared evenly among all group members, regardless of any individual's level of contribution. In principal, the group as a whole would be better off, if everyone gave the maximum contribution. Under strict rationality however, the strong free rider hypothesis predicts 0 contribution from every player. Modeling certain situations as public goods games then leads to the prediction that public goods will be under-provided. Again however, students are tempted to misinterpret the lesson and consider the act of contribution to be inherently irrational. Aspects of other-regarding behavior can be rational, if they are included in the utility function (Andreoni 1989).

In each of the above circumstances, students could benefit from stepping back from their textbooks and remembering the purpose of modeling. Insofar as models are neither true nor false, but useful or not (Clarke and Primo 2012), they are meant to help researchers predict behavior, not prescribe what a player *should* do, when playing the game. Simple classroom experiments, ideally before lecturing on the game, combined with post experiment discussion of results, help students to remember that while a game may have a pure strategy Nash equilibrium, it's not necessarily a good prediction of behavior. Experiments can stimulate students to consider why behavior may differ from the equilibrium and how they might revise models to be more useful.

Returning to voluntary contribution mechanisms, it is an empirical regularity in repeated play that in early rounds contributions are relatively high, but over time tend to converge to zero. Another regularity is that even if contributions have hit zero, if play is stopped and then restarted, then contributions will leap upward, before again trending toward zero. Much of game theory in teaching is focused on identifying equilibria without consideration of how these equilibria (particularly Nash equilibria) are reached. Replication of classic experiments allows for discussion of equilibrium selection, coordination mechanisms, and institutions that support pro-social behavior.

One useful way to engage students in a discussion of modelling behavior is to place them in a scenario with solution concepts other than just pure strategy Nash equilibrium. For instance, consider k-level reasoning. The beauty contest game takes a set of N players and gives them three options: A, B, and C. The player's task is to guess which of the three options will be most often selected by the group. Thus, players are asked not about their own preferences over the three options, but their beliefs on the preferences of the other players. In a variant of this game, Rosmarie Nagel (1995) takes a set of N players and has them pick numbers between one and one hundred. The player's task is to pick a number closest to what they believe will be the average guess, times a parameter p. If $p = 0.5$, then subjects are attempting to guess the number between one and one hundred that will be half of the average guess. The subject with the guess closest to the amount wins.

In this case, some players will notice that no number $x \in (50, 100]$ can be the correct answer, since these numbers can never be half of the average. A subject who answers 50 would be labeled level-0, as she has avoided strictly dominated strategies. Some subjects however, will believe that all subjects have thought through the game at least this far and will realize that the interval of viable answers is really $(0, 50]$. These level-1 players then respond that one half of the average will be $x = 25$. The process iterates to its logical (Nash Equilibrium) conclusion. If all players are strictly rational, then they will all answer 0. Behaviorally though, guesses of 0 virtually never win.

In a classroom setting, this game is easy to implement and quite illustrative.[1] Students become particularly attentive, if the professor offers even modest monetary stakes, say between \$0.00 and \$10.00, with the winning student receiving her guess as a prize. A class of robots will all guess 0 and the professor will suffer no monetary loss. But all it takes is a small percentage of the class to enter guesses above 0 to pull the winning guess away from the Nash Equilibrium. Thus the hyper-rational students who guessed 0 see that the equilibrium answer and the winning answer are not necessarily the same thing.[2]

In each of the above settings, it is well established that many subjects do not employ the equilibrium strategy. This is surprising to no one beyond those students who worship too readily at the altar of rational choice. By replicating classic experiments to demonstrate to students that models are not perfect in their ability to predict human behavior, we demote game theory from life plan to its proper level of mathematical tool. We typically think of replication as a check on faulty research or a means by which to verify the robustness of social scientific results. Here, we are using replication of robust results to inspire critical thinking about social science itself. For other students however, replication has the added benefit of enabling training in skills needed

[1] Thanks are due to David Cooper, who first introduced me to this game in his class, when I was a graduate student.

[2] The 11-20 money request game by Arad and Rubinstein (2012) is an interesting variant of this without a pure strategy Nash equilibrium at all.

to carry out more advanced experiments.

Replication in Training

To some extent, the internet era has been a boon to the graduate student of social sciences, providing ready access to a wide variety of data sources. Responsible researchers make their data available on request at the very least, if not completely available online. Fellow researchers can then attempt to replicate findings to test their robustness. Students, in turn, can use replication files to practice the methods they've learned in their classes.

The same is true of experimental data sets. However the data analysis of experiments is rarely a complex task.[3] For the budding experimentalist, replication of data analysis is a useful exercise, but one not nearly as useful as the replication of experimental procedures. Most data generating processes are, to some extent, sensitive to choices made by researchers. Most students however, are not collecting their own nationally representative survey data. Particularly at early levels of development, students may complete course work entirely from existing data. The vast majority of their effort is spent on the analysis. Mistakes can be identified and often corrected with what may be little more than a few extra lines of code.

For experimentalists in training though, the majority of the work comes on the front end, as does the majority of the risk. From writing the experimental program in a language such as zTree (Fischbacher 2007), which is generally new to the student, to physically running the experimental sessions, a student's first experiment is an ordeal. The stress of this endeavor is compounded, when its success or failure directly relates to the student's career trajectory and job market potential. It is critical for the student to have solid guidance from a well trained advisor.

This is, of course, true of all research methods. The better a student's training, the greater their likelihood of successful outcomes. Data analysis training in political science graduate programs has become considerably more sophisticated in recent years, with students often required to complete three, four, or even more methods courses. Training for experimentalists however, exhibits considerably more variance and formal training may be unavailable. Some fortunate students are trained on the job, assisting more senior researchers with their experiments. But while students benefit from an apprenticeship with an experimentalist, they suffer, ironically enough, from a lack of experimentation.

Any student can practice working with large data. Many data sets can be accessed for free or via an institutional license. A student can engage in atheoretical data mining and practice her analysis and interpretation of results. She can do all of this at home with a glass of beer and the television on. When she makes a mistake, as a young researcher is wont to do, little is lost and the student has gained a valuable lesson. Students of experiments, however, rarely get the chance to make such mistakes. A single line of economic experiments can cost thousands of dollars and a student is unlikely to have surplus research funds with which to gain experience. If she is lucky enough to receive research funding, it will likely be limited to subject payments for her dissertation's experiment(s). A single failed session could drain a meaningful portion of her budget, as subjects must be paid, even if the data is unusable.[4]

How then is the experimentalist to develop her craft, while under a tight budget constraint? The answer lies in the empirical regularities discussed earlier. The size of financial incentives in an experiment does matter, at least in terms of salience (Morton and Williams 2010), but some effects are so robust as to be present in experiments with even trivial or non-financial incentives. In my own classroom demonstrations, I have replicated prisoner's dilemma, ultimatum game, public good game, and many other experiments, using only fractions of extra credit points as incentives and the results are remarkably consistent with those in the literature.[5] At zero financial cost, I gained experience in the programming and running of experiments and simultaneously ran a lesson on end game effects, the restart effect, and the repeated public goods game.

Not all graduate students teach courses of their own, but all graduate students have advisors or committee members who do. It is generally less of an imposition for an advisee to ask a faculty member to grant their students a few bonus points than it is to ask for research funds, especially funds that would not be directly spent on the dissertation. These experiments can be run in every way identical to how one would be run with monetary incentives, but without the cost or risk to the student's career. This practice is all the more important at institutions without established laboratories, where the student is responsible for building an ad hoc network.

Even for students with experience assisting senior researchers, independently planning and running an experiment from start to finish, without direct supervision, is invaluable practice. The student is confronted with the dilemma of how she will run the experiment, not how her advisor would do so. She then writes her own program and

[3] Indeed, the technical simplicity of analysis is one of the key advantages of true experiments.

[4] The rule at many labs is that subjects in failed sessions must still receive their show up fees and then additional compensation for any time they have spent up to the point of the crash. Even with modest subject payments, this could be hundreds of dollars.

[5] Throughout the course, students earn "Experimental Credit Units." The top performing student at the end of the semester receives five extra credit points. All other students receive extra credit indexed to that of the top performer. I would love to report the results of the experiments here, but at the time I had no intention of using the data for anything other than educational purposes and thus did not apply for IRB approval.

[6] A well-run experiment is the result of not only a properly written program, but also of strict adherence to a set of physical procedures such as

instructions, designs her own physical procedures, and plans every detail on her own.[6] She can and should seek advice, but she is free to learn and develop her own routine. The experiment may succeed or fail, but the end product is similar to atheoretical playing with data. It won't likely result in a publication, but it will prove to be a valuable learning experience.

Discussion

Many of the other articles in this special issue deal with the replication of studies, as a matter of good science, in line with practices in the physical sciences. But in the physical sciences, replication also plays a key role in training. Students begin replicating classic experiments often before they can even spell the word science. They follow structured procedures to obtain predictable results, not to advance the leading edge of science, but to build core skills and methodological discipline.

Here though, physical scientists have a distinct advantage. Their models are frequently based on deterministic causation and are more readily understood, operationalized, tested, and (possibly) disproved. To the extent that students have encountered scientific models in their early academic careers, these models are likely to have been deterministic. Most models in social science however, are probabilistic in nature. It is somewhat understandable that a student in the social sciences, who reads her textbook and sees the mathematical beauty of rational choice, would be enamored with its clarity. A student, particularly one who has self selected into majoring in economics or politics, can be forgiven for seeing the direct benefits of playing purely rational strategies. It is not uncommon for an undergraduate to go her entire academic career without empirically testing a model. By replicating classic experiments, particularly where rational choice fails, we can reinforce the idea that these are models meant to predict behavior, not instructions for how to best an opponent.

In contrast, graduate students explicitly train in designing and testing models. A key component of training is the ability to make and learn from mistakes. Medical students learn by practicing on cadavers who cannot suffer. Chemists learn by following procedures and comparing results to established parameters. Large-n researchers learn by working through replication files and testing for robustness of results. In the same spirit, experimentalists can learn by running low risk experiments based on established designs, with predicable results. In doing so, even if they fail, they build competence in the skills they will need to work independently in the future. At any rate, while the tools employed in the social sciences differ from those in the physical sciences, the goal is the same: to improve our understanding of the world around us. Replicating economic experiments aids some in their study of human behavior and others on their path to learn how to study human behavior. Both are laudable goals.

References

Andreoni, James. 1989. "Giving with Impure Altruism: Applications to Charity and Ricardian Equivalence." *The Journal of Political Economy* 97(6): 1447-58.

Arad, Ayala and Ariel Rubinstein. 2012. "The 11-20 Money Requestion Game: A Level-k Reasoning Study." *The American Economic Review* 102(7): 3561-73.

Clarke, Kevin A. and David M. Primo. 2012. *A Model Discipline: Political Science and the Logic of Representations.* New York, NY: Oxford University Press.

Engel, Christoph. 2011. "Dictator Games: A Meta Study." *Experimental Economics* 14: 583-610.

Fischbacher, Urs. 2007. "z-Tree: Zurich Toolbox for Ready-made Economic Experiments." *Experimental Economics* 10(2): 171-8.

Marwell, Gerald and Ruth E. Ames. 1981. "Economists Free Ride, Does Anyone Else? Experiments on the Provision of Public Goods, IV." *Journal of Public Economics* 15(3): 295-310.

Morton, Rebecca B., and Kenneth C. Williams. 2010. *Experimental Political Science and the Study of Causality.* New York, NY: Cambridge University Press.

Nagel, Rosemarie. 1995. "Unraveling in Guessing Games: And Experimental Study." *The American Economic Review* 85(5): 1313-26.

Oosterbeek, Hessel, Randolph Sloof, and Gijs van de Kuilen. 2004. "Cultural Differences in Ultimatum Game Experiments: Evidence from a Meta-Analysis." *Experimental Economics* 7(2): 171-88.

(among many others) how to seat subjects, how to convey instructions, and how to monitor laboratory conditions. A program can be vetted in a vacuum, but the experimenter's procedures are subject to failure in each and every session, thus practice is crucial.

A Decade of Replications: Lessons from the Quarterly Journal of Political Science

Nicholas Eubank
Stanford University
nicholaseubank@stanford.edu

The success of science depends critically on the ability of peers to interrogate published research in an effort not only to confirm its validity but also to extend its scope and probe its limitations. Yet as social science has become increasingly dependent on computational analyses, traditional means of ensuring the accessibility of research – like peer review of written academic publications – are no longer sufficient. To truly ensure the integrity of academic research moving forward, it is necessary that published papers be accompanied by the code used to generate results. This will allow other researchers to investigate not just whether a paper's methods are theoretically sound, but also whether they have been properly implemented and are robust to alternative specifications.

Since its inception in 2005, the *Quarterly Journal of Political Science (QJPS)* has sought to encourage this type of transparency by requiring all submissions to be accompanied by a replication package, consisting of data and code for generating paper results. These packages are then made available with the paper on the *QJPS* website. In addition, all replication packages are subject to internal review by the *QJPS* prior to publication. This internal review includes ensuring the code executes smoothly, results from the paper can be easily located, and results generated by the replication package match those in the paper.

This policy is motivated by the belief that publication of replication materials serves at least three important academic purposes. First, it helps directly ensure the integrity of results published in the *QJPS*. Although the in-house screening process constitutes a minimum bar for replication, it has nevertheless identified a remarkable number of problems in papers. In the last two years, for example, 13 of the 24 empirical papers subject to in-house review were found to have discrepancies between the results generated by authors' own code and the results in their written manuscripts.

Second, by emphasizing the need for transparent and easy-to-interpret code, the *QJPS* hopes to lower the costs associated with other scholars interrogating the results of existing papers. This increases the probability other scholars will examine the code for published papers, potentially identifying errors or issues of robustness if they exist. In addition, while not all code is likely to be examined in detail, it is the hope of the *QJPS* that this transparency will motivate submitting authors to be especially cautious in their coding and robustness checks, preventing errors before they occur.

Third and finally, publication of transparent replication packages helps facilitate research that builds on past work. Many papers published in the *QJPS* represent methodological innovations, and by making the code underlying those innovations publicly accessible, we hope to lower the cost to future researchers of building on existing work.

1. In-House Replication

The experience of the *QJPS* in its first decade underscores the importance of its policy of in-house review. Prior to publication, all replication packages are tested to ensure code runs cleanly, is interpretable, and generates the results in the paper.

This level of review represents a sensible compromise between the two extremes of review. On the one hand, most people would agree that an *ideal* replication would consist of a talented researcher re-creating a paper from scratch based solely on the paper's written methodology section. However, undertaking such replications for every submitted paper would be cost-prohibitive in time and labor, as would having someone check an author's code for errors line-by-line. On the other hand, direct publication of replication packages without review is also potentially problematic. Experience has shown that many authors submit replication packages that are extremely difficult to interpret or may not even run, defeating the purpose of a replication policy.

Given that the *QJPS* review is relatively basic, however, one might ask whether it is even worth the considerable time the *QJPS* invests. Experience has shown the answer is an unambiguous "yes." Of the 24 empirical papers subject to in-house replication review since September 2012,[1] only 4 packages required no modifications. Of the remaining 20 papers, 13 had code that would not execute without errors, 8 failed to include code for results that appeared in the paper,[2] and 7 failed to include installation directions for software dependencies. Most troubling, however, 13 (54 percent) had results in the paper that differed from those generated by the author's own code. Some of these issues were relatively small – likely arising from rounding errors during transcription – but in other cases they involved incorrectly signed or mis-labeled regression coefficients, large errors in observation counts, and incorrect summary statistics. Frequently, these discrepancies required changes to full columns or tables of results. Moreover, Zachary Peskowitz, who served as the *QJPS* replication assistant from 2010 to 2012, reports similar levels of replication errors during his tenure as well. The extent of the issues – which occurred despite authors having been informed their packages would be subject to review – points to the necessity of this type of

[1] September 2012 is when the author took over responsibility for all in-house interrogations of replication packages at the *QJPS*.
[2] This does not include code which failed to execute, which might also be thought of as failing to replicate results from the paper.

in-house interrogation of code prior to paper publication.

2. Additional Considerations for a Replication Policy

This section presents an overview of some of the most pressing and concrete considerations the *QJPS* has come to view as central to a successful replication policy. These considerations – and the specific policies adopted to address them – are the result of hard-learned lessons from a decade of replication experience.

2.1. Ease of Replication

The primary goal of *QJPS* policies is ensuring replication materials can be used and interpreted with the greatest of ease. To the *QJPS*, ease of replication means anyone who is interested in replicating a published article (hereafter, a "replicator") should be able to do so as follows:

1. Open a README.txt file in the root replication folder, and find a summary of all replication materials in that folder, including subfolders if any.
2. After installing any required software (see Section 2.4. on Software Dependencies) and setting a working directory according to directions provided in the README.txt file, the replicator should be able simply to open and run the relevant files to generate every result and figure in the publication. This includes all results in print and/or online appendices.
3. Once the code has finished running, the replicator should be able easily to locate the output and to see where that output is reported in the paper's text, footnotes, figures, tables, or appendices.

2.2. README.txt File

To facilitate ease of replication, all replication packages should include a README.txt file that includes, at a minimium:

1. **Table of Contents:** a brief description of every file in the replication folder.
2. **Notes for Each Table and Figure:** a short list of where replicators will find the code needed to replicate all parts of the publication.
3. **Base Software Dependencies:** a list of all software required for replication, including the version of software used by the author (e.g. Stata 11.1, R 2.15.3, 32bit Windows 7, OSX 10.9.4).
4. **Additional Dependencies:** a list of all libraries or added functions required for replication, as well as the versions of the libraries and functions that were used and the location from which those libraries and functions were obtained.
   - **R**: the current R versions can be found by typing `R.Version()` and information on loaded libraries can be found by typing `sessionInfo()`.
   - **STATA:** Stata does not specifically "load" extra-functions in each session, but a list of all add-ons installed on a system can be found by typing `ado.dir`.
5. **Seed locations:** Authors are required to set seeds in their code for any analyses that employ randomness (e.g., simulations or bootstrapped standard errors. For further discussion, see Section 2.5.). The README.txt file should include a list of locations where seeds are set in the analyses so that replicators can find and change the seeds to check the robustness of the results.

2.3. Depth of Replication

The *QJPS* requires that every replication package include the code that computes the primary results of the paper. In other words, it is not sufficient to provide a file of precomputed results along with the code that formats the results for LaTeX. Rather, the replication package must include everything that is needed to execute the statistical analyses or simulations that constitute the primary contribution of the paper. For example, if a paper's primary contribution is a set of regressions, then the data and code needed to produce those regressions must be included. If a paper's primary contribution is a simulation, then code for that simulation must be provided—not just a dataset of the simulation results. If a paper's primary contribution is a novel estimator, then code for the estimator must be provided. And, if a paper's primary contribution is theoretical and numeric simulation or approximation methods were used to provide the equilibrium characterization, then that code must be included.

Although the *QJPS* does not necessarily require the submitted code to access the data if the data are publicly available (e.g., data from the National Election Studies, or some other data repository), it does require that the dataset containing all of the original variables used in the analysis be included in the replication package. For the sake of transparency, the variables should be in their original, untransformed and unrecoded form, with code included that performs the transformations and recodings in the reported analyses. This allows replicators to assess the impact of transformations and recodings on the results.

2.3.1. Proprietary and Non-Public Data

If an analysis relies on proprietary or non-public data, authors are required to contact the *QJPS* Editors before or during initial submission. Even when data cannot be released publicly, authors are often required to provide *QJPS* staff access to data for replication prior to publication. Although this sometimes requires additional arrangements – in the past, it has been necessary for *QJPS* staff to be written in IRB authorizations – in-house review is especially important in these contexts, as papers based on non-public data are difficult if not impossible for other scholars to interrogate post-publication.

2.4. Software Dependencies

Online software repositories – like CRAN or SSC – provide authors with easy access to the latest versions of powerful add-ons to standard programs like R and Stata. Yet the strength of these repositories – their ability to ensure authors are always working with the latest version of add-ons – is also a liability for replication.

Because online repositories always provide the most recent version of add-ons to users, the software provided in response to a given query actually changes over time. Experience has shown this can cause problems when authors use calls to these repositories to install add-ons (through commands like `install.packages("PACKAGE")` in R or `ssc install PACKAGE` in Stata). As scholars may attempt to replicate papers months or years after a paper has been published, changes in the software provided in response to these queries may lead to replication failures. Indeed, the *QJPS* has experienced replication failures due to changes in the software hosted on the CRAN server that occurred between when a paper was submitted to the *QJPS* and when it was reviewed.

With that in mind, the *QJPS* now requires authors to include copies of all software (including both base software and add-on functions and libraries) used in the replication in their replication package, as well as code that installs these packages on a replicator's computer. The only exceptions are extremely commonly tools, like R, Stata, Matlab, Java, Python, or ArcMap (although Java- and Python-based applications must be included).[3]

2.5. Randomizations and Simulations

A large number of modern algorithms employ randomness in generating their results (e.g., the bootstrap). In these cases, replication requires both (a) ensuring that the *exact* results in the paper can be re-created, and (b) ensuring that the results in the paper are typical rather than cherry-picked outliers. To facilitate this type of analysis, authors should:

1. Set a random number generator seed in their code so it consistently generates the exact results in the paper;
2. Provide a note in the README.txt file indicating the location of all such commands, so replicators can remove them and test the representativeness of result.

In spite of these precautions, painstaking experience has shown that setting a seed is not always sufficient to ensure exact replication. For example, some libraries generate slightly different results on different operating systems (e.g. Windows versus OSX) and on different hardware architectures (e.g. 32-bit Windows 7 versus 64-bit Windows 7). To protect authors from such surprises, we encourage authors to test their code on multiple platforms, and document any resulting exceptions or complications in their README.txt file.

2.6. ArcGIS

Although we encourage authors to write replication code for their ArcGIS-based analyses using the ArcPy scripting utility, we recognize that most authors have yet to adopt this tool. For the time being, the *QJPS* accepts detailed, step-by-step instructions for replicating results via the ArcGIS Graphical User Interface (GUI). However, as with the inclusion and installation of add-on functions, the *QJPS* has made available a tutorial on using ArcPy available to authors which we hope will accelerate the transition towards use of this tool.[4]

3. Advice to Authors

In addition to the preceding requirements, the *QJPS* also provides authors with some simple guidelines to help prevent common errors. These suggestions are not mandatory, but they are highly recommended.

1. **Test files on a different computer, preferably with a different operating system:** Once replication code has been prepared, the *QJPS* suggests authors email it to a different computer, unzip it, and run it. Code often contains small dependencies—things like unnoticed software requirements or specific file locations—that go unnoticed until replication. Running code on a different computer often exposes these issues in a way that running the code on one's own does not.
2. **Check every code-generated result against your final manuscript PDF:** The vast majority of replication problems emerge because authors either modified their code but failed to update their manuscript,

[3] To aid researchers in meeting this requirement, detailed instructions on how to include CRAN or SSC packages in replication packages are provided through the *QJPS*.

[4] ArcPy is a Python-based tool for scripting in ArcGIS.

or made an error while transcribing their results into their paper. With that in mind, authors are strongly encouraged to print out a copy of their manuscript and check each result before submitting your final version of the manuscript and replication package.

4. Conclusion

As the nature of academic research changes, becoming ever more computationally intense, so too must the peer review process. This paper provides an overview of many of the lessons learned by the *QJPS*'s attempt to address this need. Most importantly, however, it documents not only the importance of requiring the transparent publication of replication materials but also the strong need for in-house review of these materials prior to publication.

What Does a Failed Replication Really Mean? (or, One Cheer for Jason Mitchell)

Justin Esarey
Rice University
justin@justinesarey.com

A few weeks ago, Jason Mitchell wrote a piece entitled "On the emptiness of failed replications." Mitchell is a professor in Harvard University's department of Psychology studying "the cognitive processes that support inferences about the psychological states of other people and introspective awareness of the self." In practice, this means his lab spends a lot of time doing experiments with fMRI machines.

It is worth saying at the outset that I don't agree with Mitchell's core claim: unlike him, I believe that failed replications can have a great deal of scientific value. However, I believe that there is a grain of truth in his argument that we should consider. Namely, I think that failed replications should be thought about in the same way that we think about the initial successful experiment: *skeptically*. A positive result is not proof of success, and a failed replication is not proof of failure; the interpretation of both must be uncertain and ambiguous at first. Unfortunately, I have observed that most of us (even highly trained scientists) find it hard to make that kind of uncertainty a part of our everyday thinking and communicating.

The thesis of Mitchell's argument is summarized in his opening paragraph:

> Recent hand-wringing over failed replications in social psychology is largely pointless, because unsuccessful experiments have no meaningful scientific value. Because experiments can be undermined by a vast number of practical mistakes, the likeliest explanation for any failed replication will always be that the replicator bungled something along the way. Unless direct replications are conducted by flawless experimenters, nothing interesting can be learned from them.

Why do I believe that Mitchell's claim about the scientific value of failed replication is wrong? Lots of other statisticians and researchers have explained why very clearly, so I will just link to their work and present a brief sketch of two points:

1. Failed replications have proved extremely informative in resolving past scientific controversies.
2. Sampling variation and statistical noise cannot be definitively excluded as explanations for a "successful experiment" without a very large sample and/or replication.

First, the consistent failure to replicate an initial experiment has often proven informative about what we could learn from that initial experiment (often, very little). Consider as one example the Fleischmann – Pons experiment apparently demonstrating cold fusion. Taken at face value, this experiment would seem to change our theoretical understanding about how nuclear fusion works. It would also seem to necessitate the undertaking of intense scientific and engineering study of the technique to improve it for commercial and scientific use. But if we add the fact that no other scientists could ever make this experiment work, despite sustained effort by multiple teams, then the conclusion is much different and much simpler: Fleischmann and Pons' experiment was flawed.

Second, and relatedly, Mitchell seems to admit multiple explanations for a failed replication (error, bias, imperfect procedure) but only one explanation for the initial affirmative result (the experiment produced the observed relationship):

> To put a fine point on this: if a replication effort were to be capable of identifying empirically questionable results, it would have to employ flawless experimenters. Otherwise, how do we identify replications that fail simply because of

> undetected experimenter error? When an experiment succeeds, we can celebrate that the phenomenon survived these all-too-frequent shortcomings. But when an experiment fails, we can only wallow in uncertainty about whether a phenomenon simply does not exist or, rather, whether we were just a bit too human that time around.

The point of conducting statistical significance testing is to exclude another explanation for a successful experiment: random noise and/or sampling variation produced an apparent result where none in fact exists. There is also some [1]evidence to suggest that researchers consciously or unconsciously make choices in their research that improve the possibility of passing a statistical significance test under the null (so-called "p-hacking").

Perhaps Mitchell believes that passing a statistical significance test and peer review definitively rules out alternative explanations for an affirmative result. Unfortunately, that isn't necessarily the case. In joint work with Ahra Wu I find evidence that, even under ideal conditions (with no p-hacking, misspecification bias, etc.), statistical significance testing cannot prevent excess false positives from permeating the published literature. The reason is that, if most research projects are ultimately chasing blind alleys, the filter imposed by significance testing is not discriminating enough to prevent many false positives from being published. The result is one of the forms of "publication bias."

Yet despite all this, I think there is a kernel of insight in Mitchell's argument. I think the recent interest in replication is part of a justifiably greater skepticism that we are applying to new discoveries. But we should also apply that greater skepticism to isolated reports of failed replication–and for many of the same reasons. Allow me to give one example.

One source of publication bias is the so-called "file drawer problem," whereby studies of some phenomenon that produce null results never get published (or even submitted); thus, false positive results (that *do get* published) are never placed into their proper context. But this phenomenon is driven by the fact that evidence in favor of new theories is considered more scientifically important than evidence against theories without a wide following. But if concern about false positives in the literature becomes widespread, then replications that *contradict* a published result may become more scientifically noteworthy than replications that *confirm* that result. Thus, we may become primed to see (and publish) falsifying results and to ignore confirmatory results. The problem is the same as the file drawer problem, but in reverse.

Even if we do our best to publish and take note of all results, we can reasonably expect many replications to be false negatives. To demonstrate this, I've created a simulation of the publication/replication process. First, a true relationship (b) is drawn from the distribution of underlying relationships in a population of potential research studies; this population has pr.null proportion of relationships where $b = 0$. My initial simulation sets $pr.null = 0$ for demonstrative purposes; thus, b comes from the uniform density between $[-2, 1]$ and $[1, 2]$. (I extracted the values between $(-1, 1)$ to remove the possibility of small, noise-dominated relationships; the reason why will become clear once I present the results.) Then, I simulate an estimate produced by a study of this relationship with noise and/or sampling variation ($= b.est$) by adding b and an error term drawn from the normal distribution with mean $= 0$ and standard error $= se.b$, which is set to 0.5 in my initial run. If the resulting coefficient is statistically significant, then I replicate this study by drawing another estimate ($b.rep$) using the same process above.

However, I also allow for the possibility of "biased" replications that favor the null; this is represented by moving the $b.rep$ coefficient a certain number of standard deviations closer to zero. The initial setting for bias is $0.5 * se.b$, meaning that I presume that a motivated researcher can move a replicated result closer to the null by 1/2 of a standard deviation via making advantageous choices in the data collection and analysis. In short, I allow for "p-hacking" in the process, but p-hacking that is designed to handicap the result rather than advantage it. The idea is that motivated researchers trying to debunk a published claim may (consciously or unconsciously) pursue this result.

The code to execute this simulation in R is shown here:

```
set.seed(123456)
rm(list=ls())

se.b <- 0.5     # std. error of est. beta
reps <- 1000    # number of MC runs
bias <- 0.5*se.b # degree of replicator null bias
pr.null <- 0    # prior Pr(null hypothesis)

# where to store true, est., and replicated results
b.store <- matrix(data=NA, nrow=reps, ncol=3)
# where to store significance of est. and replicated betas
sig.store <- matrix(data=NA, nrow=reps, ncol=2)

pb <- txtProgressBar(init=0, min=1, max=reps, style=3)
for(i in 1:reps){

  setTxtProgressBar(pb, value=i)

  # draw the true value of beta
  if(runif(1) < pr.null){
    b <- 0
  }else{
    b <- sign(runif(1, min=-1, max=1))*runif(1, min=1, max=2)
  }

  # simulate an estimated beta
  b.est <- b + rnorm(1, mean=0, sd=se.b)

  # calculate if est. beta is statistically significant
```

[1]See Malhotra, Neil. 2014. "Publication Bias in Political Science: Using TESS Experiments to Unlock the File Drawer." PolMeth XXXI: 31st Annual Summer Meeting.

```
  if( abs(b.est / se.b) >= 1.96){sig.init <- 1}else{sig.init <- 0}

  # if the est. beta is stat. sig., replicate
  if( sig.init == 1 ){

    # draw another beta, with replicator bias
    b.rep <- b + rnorm(1, mean=0, sd=se.b) - sign(b)*bias
    # check if replicated beta is stat. sig.
    if( abs(b.rep / se.b) >= 1.96){sig.rep <- 1}else{sig.rep <- 0}

  }else{b.rep <- NA; sig.rep <- NA}

  # store the results
  b.store[i, ] <- c(b, b.est, b.rep)
  sig.store[i, ] <- c(sig.init, sig.rep)

}
close(pb)

# plot estimated vs. replicated results
plot(b.store[,2], b.store[,3], xlab = "initial␣estimated␣beta", ylab
     = "replicated␣beta")
abline(0,1)
abline(h = 1.96*se.b, lty=2)
abline(h = -1.96*se.b, lty=2)

dev.copy2pdf(file="replication-result.pdf")

# false replication failure rate
1 - sum(sig.store[,2], na.rm=T)/sum(is.na(sig.store[,2])==F)
```

What do we find? In this scenario, about 30% of replicated results are false negatives; that is, the replication study finds no effect where an effect actually exists. Furthermore, these excess false negatives cannot be attributed to small relationships that cannot be reliably detected in an underpowered study; this is why I extracted the values of b between $(-1, 1)$ from the prior distribution of the relationship b.

So: I believe that it is important not to replace one suboptimal regime (privileging statistically significant and surprising findings) with another (privileging replications that appear to refute a prominent theory). This is why many of the people advocating replication are in favor of something like a results-blind publication regime, wherein *no* filter is imposed on the publication process. As Ahra and I point out, that idea has its own problems (e.g., it might creates an enormous unpaid burden on reviewers, and might also force scientists to process an enormous amount of low-value information on null results).

In summary: I think the lesson to draw from the publication bias literature, Mitchell's essay, and the simulation result above is: the prudent course is to be skeptical of *any* isolated result until it has been vetted multiple times and in multiple contexts. Unexpected and statistically significant relationships discovered in new research should be treated as promising leads, not settled conclusions. Statistical evidence against a conclusion should be treated as reason for doubt, but not a debunking.

Encountering Your IRB: What Political Scientists Need to Know

Dvora Yanow
Wageningen University
dvora.yanow@wur.nl

Peregrine Schwartz-Shea
University of Utah
psshea@poli-sci.utah.edu

Pre-script.[1] After we finished preparing this essay, a field experiment concerning voting for judges in California, Montana, and New Hampshire made it even more relevant. Three political scientists – one at Dartmouth, two from Stanford – mailed potential voters about 300,000 flyers marked with the states' seals, containing information about the judges' ideologies. Aside from questions of research design, whether the research passed IRB review is not entirely clear (reports say it did not in Stanford but was at least submitted to the Dartmouth IRB; for those who missed the coverage, see Derek Willis' NYT article and political scientist Melissa Michelson's blog (both accessed November 3, 2014). Two bits of information offer plausible explanations for what have been key points in the public discussion:

1. Stanford may have had a reliance agreement with Dartmouth, meaning that it would accept Dartmouth's IRB's review in lieu of its own separate review;

2. Stanford and Dartmouth may have "unchecked the box" (see below), relevant here because the experiments were not federally funded, meaning that IRB review is not mandated and that universities may devise their own review criteria.

Still, neither explains what appear to be lapses in ethical judgment in designing the research (among others, using the state seals without permission and thereby creating the appearance of an official document). We find this a stellar example of a point we raise in the essay: the discipline's lack of attention to research ethics, possibly due to reliance on IRBs and the compliance ethics that IRB practices have inculcated.

* * *

Continuing our research on US Institutional Review Board (IRB) policies and practices (Schwartz-Shea and

[1]This is a condensed version of an essay appearing in *Qualitative & Multi-Method Research* [Newsletter of the APSA Organized Section for Qualitative and Multi-Method Research] Vol. 12, No. 2 (Fall 2014). The original, which is more than twice the length and develops many of these ideas more fully, is available from the authors: Dvora.Yanow@wur.nl, psshea@poli-sci.utah.edu.

Yanow 2014, Yanow and Schwartz-Shea 2008) shows us that many political scientists lack crucial information about these matters. To facilitate political scientists' more effective interactions with IRB staff and Boards, we would like to share some insights gained from this research. University IRBs implement federal policy, monitored by the Department of Health and Human Services' Office of Human Research Protections (OHRP). The Boards themselves are comprised of faculty colleagues (sometimes social scientists) plus a community member. IRB administrators are often not scientists (of any sort), and their training is oriented toward the language and evaluative criteria of the federal code. Indeed, administering an IRB has become a professional occupation with its own training and certification. IRBs review proposals to conduct research involving "human subjects" and examine whether potential risks to them have been minimized, assessing those risks against the research's expected benefits to participants and to society. They also assess researchers' plans to provide informed consent, protect participants' privacy, and keep the collected data confidential. The federal policy was created to rest on local Board decision-making and implementation, leading to significant variations across campuses in its interpretation. Differences in practices often hinge on whether a university has a single IRB evaluating all forms of research or different ones for, e.g., medical and social science research. Therefore, researchers need to know their own institutions' IRBs. In addition, familiarity with key IRB policy provisions and terminologies will help. We explain some of this "IRB-speak" and then turn to some procedural matters, including those relevant to field researchers conducting interviews, participant-observation/ethnography, surveys, and/or field experiments, whether domestically or overseas.

IRB-speak: A Primer

Part of what makes IRB review processes potentially challenging is its specialized language. Regulatory and discipline-based understandings of various terms do not always match. Key vocabulary includes the following.

- "Research." IRB regulations tie this term's meaning to the philosophically-contested idea of "generalizable knowledge" (CFR §46.102(d)). This excludes information-gathering for other purposes and, on some campuses, other scholarly endeavors (e.g., oral history) and course-related exercises.

- "Human subject." This is a living individual with whom the researcher interacts to obtain data. "Interaction" is defined as "communication or interpersonal contact between investigator and subject" (CFR §46.102(f)). But "identifiable private information" obtained without interaction, such as through the use of existing records, also counts.

- "Minimal risk." This research is when "the probability and magnitude of harm or discomfort anticipated in the research are not greater in and of themselves than those ordinarily encountered in daily life or during the performance of routine physical or psychological examinations or tests" (CFR §46.102(i)). But everyday risks vary across subgroups in American society, not to mention worldwide, and IRB reviewers have been criticized for their lack of expertise in risk assessment, leading them to misconstrue the risks associated with, e.g., comparative research (Schrag 2010, Stark 2012).

- "Vulnerable populations." Six categories of research participants "vulnerable to coercion or undue influence" are subject to additional safeguards: "children, prisoners, pregnant women, mentally disabled persons, or economically or educationally disadvantaged persons" (CFR §46.111(b)). Federal policy enables universities also to designate other populations as "vulnerable," e.g., Native Americans.

- Levels of review. Usually, IRB staff decide a proposed project's level of required review: "exempt," "expedited," or "convened" full Board review. "Exempt" does not mean that research proposals are not reviewed. Rather, it means exemption from full Board review, a status that can be determined only via some IRB assessment. Only research entailing no greater than minimal risk is eligible for "exempt" or "expedited" review. The latter means assessment by either the IRB chairperson or his/her designee from among Board members. This reviewer may not disapprove the proposal, but may require changes to its design. Projects that entail greater than minimal risk require "convened" (i.e., full) Board review.

- Exempt category: Methods. Survey and interview research and observation of public behavior are exempt from full review if the data so obtained do not identify individuals and would not place them at risk of "criminal or civil liability or be damaging to the subjects' financial standing, employability, or reputation" if their responses were to be revealed "outside of the research" (CFR §46.101(b)(2)(ii)). Observing public behaviors as political events take place (think: "Occupy") is central to political science research. Because normal IRB review may delay the start of such research, some IRBs have an "Agreement for Public Ethnographic Studies" that allows observation to begin almost immediately, possibly subject to certain stipulations.

- Exempt category: Public officials. IRB policy explicitly exempts surveys, interviews, and public observation *involving "elected or appointed public officials or candidates for public office*" (45 CFR §46.101(b)(3))

– although who, precisely, is an "appointed public official" is not clear. This exemption means that researchers studying public officials using any of these three methods might – in complete compliance with the federal code – put them at risk for "criminal or civil liability" or damage their "financial standing, employability, or reputation" (CFR §46.101(b)(2)). The policy is consistent with legal understandings that public figures bear different burdens than private citizens.

- Exempt category: Existing data. Federal policy exempts from full review "[r]esearch involving the collection or study of existing data, documents, [or] records, ... if these sources are publicly available or if the information is recorded by the investigator in such a manner that *subjects cannot be identified, directly or through identifiers linked to the subjects*" (§46.101(b)(4)). However, university IRBs vary considerably in how they treat existing quantitative datasets, such as the Inter-University Consortium for Political and Social Research collection (see `http://www.icpsr.umich.edu/icpsrweb/ICPSR/irb/`). Some universities require researchers to obtain IRB approval to use any datasets not on a preapproved list even *if* those datasets incorporate a responsible use statement.

- "Unchecking the box." The "box" in question – in the Federal-wide Assurance form that universities file with OHRP registering their intention to apply IRB regulations to all human subjects research conducted by employees and students, regardless of funding source – when "unchecked" indicates that the IRB will omit from review any research funded by sources other than the HHS (thereby limiting OHRP jurisdiction over such studies). IRB administrators may still, however, require proposals for unfunded research to be reviewed.

Procedural Matters: Non-Experimental Field Research

The experimental research design model informing IRB policy creation and constituting the design most familiar to policy-makers, Board members and staff means that field researchers face particular challenges in IRB review. As the forms and online application sites developed for campus IRB uses reflect this policy history, some of their language is irrelevant for non-experimental field research designs (e.g., the number of participants to be "enrolled" in a study or "inclusion" and "exclusion" criteria, features of laboratory experiments or medical randomized controlled clinical trials). Those templates can be frustrating for researchers trying to fit them to field designs. Although that might seem expeditious, conforming to language that does not fit the methodology of the proposed research can lead field researchers to distort the character of their research.

IRB policy generally requires researchers to inform potential participants – to "consent" them – about the scope of both the research and its potential harms, whether physical, mental, financial or reputational. Potential subjects also need to be consented about possible identity revelations that could render them subject to criminal or civil prosecution (e.g., the unintentional public revelation of undocumented workers' identities). Central to the consent process is the concern that potential participants not be coerced into participating and understand that they may stop their involvement at any time. Not always well known is that federal code allows more flexibility than some local Boards consider. For minimal risk research, it allows: (a) removal of some of the standard consent elements; (b) oral consent without signed forms; (c) waiver of the consent process altogether if the "research could not practicably be carried out without the waiver or alteration" (CFR §46.116(c)(2)).

Procedural Matters: General

IRB review process backlogs can pose significant time delays to the start of a research project. Adding to potential delay is many universities' requirement that researchers complete some form of training before they submit their study for review. Such delay has implications for field researchers negotiating site "access" to begin research and for all empirical researchers receiving grants, which are usually not released until IRB approval is granted. Researchers should find out their campus IRB's turnaround time as soon as they begin to prepare their proposals.

Collaborating with colleagues at other universities can also delay the start of a research project. Federal code explicitly allows a university to "rely upon the review of another qualified IRB... [to avoid] duplication of effort" (CFR §46.114), and some IRBs are content to have only the lead researcher proceed through her own campus review. Other Boards insist that all participating investigators clear their own campus IRBs. With respect to overseas research, solo or with foreign collaborators, although federal policy recognizes and makes allowances for international variability in ethics regulation (CFR §46.101(h)), some US IRBs require review by a foreign government or research setting or by the foreign colleague's university's IRB, not considering that not all universities or states, worldwide, have IRBs. Multiple review processes can make coordinated review for a jointly written proposal difficult. Add to that different Boards' interpretations of what the code requires, and one has a classic instance of organizational coordination gone awry.

In Sum

On many campuses political (and other social) scientists doing field research are faced with educating IRB members and administrative staff about the ways in which their methods differ from the experimental studies performed in hospitals and laboratories. Understanding the federal regulations can put researchers on more solid footing in pointing to permitted research practices that their local Boards may not recognize. And knowing IRB-speak can enable clearer communications between researchers and Board members and staff. Though challenging, educating staff as well as Board members potentially benefits all field researchers, graduate students in particular, some of whom have given up on field research due to IRB delays, often greater for research that does not fit the experimental model (van den Hoonard 2011).

IRB review is no guarantee that the ethical issues relevant to a particular research project will be raised. Indeed, one of our concerns is the extent to which IRB administrative processes are replacing research ethics conversations that might otherwise (and, in our view, should) be part of departmental curricula, research colloquia, and discussions with supervisors and colleagues. Moreover, significant ethical matters of particular concern to political science research are simply beyond the bounds of US IRB policy, including recognition of the ways in which current policy makes "studying up" (i.e., studying societal elites and other power holders) more difficult.

Change may still be possible. In July 2011, OHRP issued an Advanced Notice of Proposed Rulemaking, calling for comments on its proposed regulatory revisions. As of this writing, the Office has not yet announced an actual policy change (which would require its own comment period). OHRP has proposed revising several of the requirements discussed in this essay, including allowing researchers themselves to determine whether their research is "excused" (their suggested replacement for "exempt"). Because of IRB policies' impact, we call on political scientists to monitor this matter. Although much attention has, rightly, been focused on Congressional efforts to curtail National Science Foundation funding, as IRB policy affects *all* research engaging human participants, it deserves as much disciplinary attention.

References

Schrag, Zachary M. 2010. *Ethical Imperialism: Institutional Review Boards and the Social Sciences, 1965-2009.* Baltimore, MD: Johns Hopkins University Press.

Schwartz-Shea, Peregrine and Yanow, Dvora. 2014. Field Research and US Institutional Review Board Policy. Betty Glad Memorial Symposium, University of Utah (March 20-21). `http://poli-sci.utah.edu/2014-research-symposium.php`.

Stark, Laura. 2012. *Behind Closed Doors: IRBs and the Making of Ethical Research.* Chicago: University of Chicago Press.

US Code of Federal Regulations. 2009. Title 45, Public Welfare, Department of Health and Human Services, Part 46, Protection of human subjects. `http://www.hhs.gov/ohrp/humansubjects/guidance/45cfr46.html`.

van den Hoonaard, Will. C. 2011. *The Seduction of Ethics.* Toronto: University of Toronto Press.

Yanow, Dvora and Schwartz-Shea, Peregrine. 2008. "Reforming institutional review board policy." *PS: Political Science & Politics* 41(3): 484-94.

Andreoni, James. 1989. "Giving with Impure Altruism: Applications to Charity and Ricardian Equivalence." *The Journal of Political Economy* 97(6): 1447-58.

Building and Maintaining R Packages with `devtools` and `roxygen2`

Jacob Montgomery
Washington University in St. Louis
jacob.montgomery@wustl.edu

Ryan T. Moore
American University
rtm@american.edu

Political methodologists increasingly develop complex computer code for data processing, statistical analysis, and data visualization – code that is intended for eventual distribution to collaborators and readers, and for storage in replication archives.[1] This code can involve multiple functions stored in many files, which can be difficult for others to read, use, or modify. In many cases, even loading the various files containing the needed functions and datasets can be a time-consuming chore.

For researchers working in R (R Core Team 2014), creating a package is an attractive option for organizing and distributing complex code. A basic R package consists of a set of functions, documentation, and some metadata. Other components, such as datasets, demo, or compiled code may

[1] Supplementary materials, including the code needed to build our example R package, are available at `https://github.com/jmontgomery/squaresPack`.

also be included. Turning all of this into a formal R package makes it very easy to distribute it to other scholars either via the Comprehensive R Archiving Network (CRAN) or simply as a compressed folder. Package creation imposes standards that encourage both the proper organization and coherent documentation of functions and datasets. Packages also allow users to quickly load and use all the relevant files on their systems. Once installed, a package's functions, documentation, datasets, and demonstrations can be accessed using R commands such as `library()`, `help()`, `data()`, and `demo()`.

However, transforming R code into a package can be a difficult and tedious process requiring the generation and organization of files, metadata, and other information in a manner that conforms to R package standards. It can be particularly difficult for users less experienced with R's technical underpinnings. In this article, we demonstrate how to develop a simple R package involving only R code, its documentation, and the necessary metadata. In particular, we discuss two packages designed to streamline the package development process – `devtools` and `roxygen2` (Wickham 2013; Wickham, Danenberg, and Eugster 2011). We begin by describing the basic structure of an R package and alternative approaches to package development, maintenance, and distribution. We compare the steps required to manually manage files, directories, and metadata to a more streamlined process employing the `devtools` package. We conclude with a discussion of more advanced issues such as the inclusion of datasets and demo files.[2]

1. R Package Basics

R package development requires building a directory of files that include the R code, documentation, and two specific files containing required metadata.[3] In this section, we walk through the basic components of an R package. As a running example, we create an R package containing two functions, which are stored in separate files named `addSquares.R` and `subtractSquares.R`.

```
## Function 1: Sum of squares
addSquares <- function(x, y){
  return(list(square=(x^2 + y^2), x = x, y = y))
  }

## Function 2: Difference of squares
subtractSquares <- function(x, y){
  return(list(square=(x^2 - y^2), x = x, y = y))
  }
```

We build the package in the directory `~/Desktop/MyPackage/`, where a `*.R` file containing each function is stored. The R package resides in a single directory whose title matches the package name. The directory must contain two metadata files, called `DESCRIPTION` and `NAMESPACE`, and two subdirectories containing the relevant R code and documentation. Thus, our `squaresPack` package will be structured as follows.

```
squaresPack
├── DESCRIPTION
├── NAMESPACE
├── R
│   ├── addSquares.R
│   └── subtractSquares.R
└── man
    ├── addSquares.Rd
    └── subtractSquares.Rd
```

Populating the package directory consists of four basic steps. First, we store all R source code in the subdirectory `R`. A good standard is to include each R function as a separate file. Second, corresponding documentation should accompany all functions that users can call. This documentation, which explains the purpose of the function, its inputs, and the values of any output, is stored in the subdirectory `man`. For example, the file `addSquares.Rd` would appear as follows.

```
\name{addSquares}
\alias{addSquares}
\title{Adding squared values}
\usage{
 addSquares(x, y)
}
\arguments{
 \item{x}{A numeric object.}
 \item{y}{A numeric object with the same dimensionality as \code{x}.}
}
\value{
 A list with the elements
 \item{squares}{The sum of the squared values.}
 \item{x}{The first object input.}
 \item{y}{The second object input.}
}
\description{
 Finds the squared sum of numbers.
}
\note{
 This is a very simple function.
}
\examples{
myX <- c(20, 3); myY <- c(-2, 4.1)
addSquares(myX, myY)
}
\author{
  Jacob M. Montgomery
}
```

Third, the directory must contain a file named `DESCRIPTION` that documents the directory in a specific way. The `DESCRIPTION` file contains basic information including the package name, the formal title, the current version number, the date for the version release, and the name of the author and maintainer. Here we also specify any dependencies on other R packages and list the files in the `R` subdirectory.

```
Package: squaresPack
Title: Adding and subtracting squared values
Version: 0.1
```

[2]The code and examples below were written for Mac OS X (10.7 and 10.8) running `R` version 3.1.0 with `devtools` version 1.5. Some adjustment may be necessary for authors using Linux. R package creation using Windows machines is not recommended. A useful online tutorial for creating an R package in RStudio using `devtools` and `roxygen2` is currently available at: `https://www.youtube.com/watch?v=9PyQlbAEujY`

[3]The canonical source on package development for R is "Writing R Extensions" (R Core Team 2013).

```
Author: Jacob M. Montgomery and Ryan T. Moore
Maintainer: Ryan T. Moore <rtm@american.edu>
Description: Find sum and difference of squared values
Depends: R (>= 3.0.0)
License: GPL (> = 2)
Collate:
  'addSquares.R'
  'subtractSquares.R'
```

Finally, the NAMESPACE file is a list of commands that are run by R when the package is loaded to make the R functions, classes, and methods defined in the package "visible" to R and the user. As we discuss briefly below, details on class structures and methods can be declared here. For squaresPack, the NAMESPACE file tells R to allow the user to call our two functions.

```
export(addSquares)
export(subtractSquares)
```

2. Approaches to Package Development and Maintenance

Authors can create and update packages in several ways, arrayed on a continuum from "very manual" to "nearly automated." At the "very manual" end, the author starts by creating the directory structure, each of the required metadata files, and a documentation file for each function. A "semi-manual" approach initializes the package automatically, but then requires that maintainers update the metadata and create documentation files for new functions as they are added. We describe this latter approach to build readers' intuition for what happens behind the scenes in the "nearly automated" approach we detail in Section 2.2, and because this approach requires nothing beyond base R.

2.1. Semi-manual Package Maintenance

A "semi-manual" procedure automatically initializes the package, but may require substantial bookkeeping as development proceeds.

Package creation: After the author loads the required functions into her workspace, she provides `package.skeleton()` with the package name and a list of the functions to be included.

```
setwd("~/Desktop/MyPackage/") ## Set the working directory
source("addSquares.R") ## Load functions into workspace
source("subtractSquares.R")
package.skeleton(name = "squaresPack",
                 list = c("addSquares", "subtractSquares"))
```

This creates the package directory using the proper structure, generates blank documentation files with the appropriate file names, and includes a helpful 'Read-and-delete-me' file that describes a few of the next steps. After the package is created, the author edits the DESCRIPTION, NAMESPACE, and help files, and the package is ready to compile and submit to CRAN. To compile the package, check it for errors, and install it on the author's instance requires three steps (shown below for the Terminal prompt in the Mac OS),

```
R CMD build --resave-data=no squaresPack
R CMD check squaresPack
R CMD INSTALL squaresPack
```

Package maintenance and submission: Superficially, the process described above may not seem cumbersome. However, calling `package.skeleton()` again (after deciding to add a new function, for example) will overwrite the previously-created directory, so any changes to documentation or metadata will be lost. Thus, after the original call to `package.skeleton()`, the author should manually add new data, functions, methods, and metadata into the initial skeleton. Adding new arguments to an existing function requires editing associated help files separately. Thus, a minimal list of required steps for updating and distributing an R package via this method includes the steps shown below.[4]

1. Edit DESCRIPTION file
2. Change R code and/or data files.
3. Edit NAMESPACE file
4. Update man files
5. `R CMD build --resave-data=no pkg`
6. `R CMD check pkg`
7. `R CMD INSTALL pkg`
8. Build Windows version to ensure compliance by submitting to: `http://win-builder.r-project.org/`
9. Upload (via Terminal below, or use other FTP client):
   `> ftp cran.r-project.org`
   `> cd incoming`
   `> put pkg_0.1-1.tar.gz`
10. Email R-core team: `cran@r-project.org`

This approach comes with significant drawbacks. Most importantly, editing the package requires altering multiple files stored across subdirectories. If a new function is added, for instance, this requires updating the R subdirectory, the DESCRIPTION file and usually the NAMESPACE file. In more complicated programming tasks that involve class structures and the like, such bookkeeping tasks can become a significant burden. Moreover, the process of actually building, checking, and submitting a package can involve moving between multiple user directories, user interfaces, and software.

We have authored four R packages over the course of the last six years. To organize the manual updating steps, one of us created an 17-point checklist outlining the actions

[4] We omit some maintenance details such as updating the LICENSE file, the Changelog, and unit testing.

required each time a package is edited. We expect that most authors will welcome some automation. The packages `devtools` and `roxygen2` can simplify package maintenance and allow authors to focus more on improving the functionality and documentation of their package.

2.2. `devtools` and `roxygen2`

`devtools` streamlines several steps: it creates and updates appropriate documentation files, it eliminates the need to leave R to build and check the package from the terminal prompt, and it submits the package to `win-builder` and CRAN and emails the R-core team from within R itself. After the initial directory structure is created, the only files that are edited directly by the author are contained in the `R` directory (with one exception – the `DESCRIPTION` file should be reviewed before the package is released). This is possible because `devtools` automates the writing of the help files, the `NAMESPACE` file, and updating of the `DESCRIPTION` file relying on information placed directly in `*.R` files.

There are several advantages to developing code with `devtools`, but the main benefit is improved workflow. For instance, adding a new function using more manual methods requires creating the code in a `*.R` file stored in the `R` subdirectory, specifying the attendant documentation as a `*.Rd` file in the `man` subdirectory, and updating the `DESCRIPTION` and `NAMESPACE` files. In contrast, developing new functions with `devtools` requires only editing a single `*.R` file, wherein the function and its documentation are written simultaneously. `devtools` then updates the documentation (using the `roxygen2` package), and package metadata with no further attention.

Writing `*.R` files: Thus, one key advantage of using `devtools` is that the `*.R` files will themselves contain the information for generating help files and updating metadata files. Each function is accompanied by detailed comments that are parsed and used to update the other files. Below we show how to format the `addSquares.R` file to create the same help files and `NAMESPACE` file shown above.

```
#' Adding squared values
#'
#' Finds the sum of squared numbers.
#'
#' @param x A numeric object.
#' @param y A numeric object with the same dimensionality as \code{x
    }.
#'
#' @return A list with the elements
#' \item{squares}{The sum of the squared values.}
#' \item{x}{The first object input.}
#' \item{y}{The second object input.}
#' @author Jacob M. Montgomery
#' @note This is a very simple function.
#' @examples
#'
#' myX <- c(20, 3)
#' myY <- c(-2, 4.1)
#' addSquares(myX, myY)
#' @rdname addSquares
#' @export
addSquares<- function(x, y){
  return(list(square=(x^2 + y^2), x = x, y = y))
}
```

The text following the `#'` symbols is processed by R during package creation to make the `*.Rd` and `NAMESPACE` files. The `@param`, `@return`, `@author`, `@note`, `@examples`, and `@seealso` commands specify the corresponding blocks in the help file. The `@rdname` block overrides the default setting to specify the name of the associated help file, and `@export` instructs R to add the necessary commands to the `NAMESPACE` file. We now walk through the steps required to initialize and maintain a package with `devtools`.

Setting up the package: Creating an R package from these augmented `*.R` files is straightforward. First, we must create the basic directory structure using

```
setwd("~/Desktop/MyPackage/") ## Set the working directory
create("squaresPack")
```

Second, we edit the `DESCRIPTION` file to make sure it contains the correct version, package name, etc. The `create()` call produces a template file. The author will need to add some information to this template `DESCRIPTION` file,[5] such as

```
Author: Me
Maintainer: Me <me@myemail.edu>
```

`devtools` will automatically collate all R files contained in the various subdirectories. Third, place the relevant R scripts in the `R` directory. Finally, making sure that the working directory is correctly set, we can create and document the package using three commands:

```
current.code <- as.package("squaresPack")
load_all(current.code)
document(current.code)
```

The `as.package()` call loads the package and creates an object representation of the entire package in the user's workspace. The `load_all()` call loads all of the R files from the package into the user's workspace as if the package was already installed.[6] The `document()` command creates the required documentation files for each function and the package, as well as updates the `NAMESPACE` and `DESCRIPTION` files.

Sharing the package: Next, the author prepares the package for wider release from within R. To build the package, the author runs `build(current.code, path=getwd())`. The analogous `build_win()` command will upload the package to the `http://win-builder.r-project.org/` website. This builds the package in a Windows environment and emails the address of the maintainer in the `DESCRIPTION` file with results in about thirty minutes. Both of these compressed files can be uploaded onto websites, sent by email,

[5] The `DESCRIPTION` file should not contain any blank lines. If the template file contains any, these will either need to be deleted or filled in.

[6] The help files and demo files will only be available using after running `install(current.code)`, which is equivalent to `R CMD INSTALL` in the Terminal.

or stored in replication archives. Other users can download the package and install it locally.

The package can be submitted to CRAN without the need to leave R. We provide a minimal checklist for editing and submitting an existing R package using `devtools`:

1. Edit R code and/or data files
2. Run `as.package()`, `load_all()`, and `document()`
3. Check the code: `check(current.code)`
4. Make a Windows build: `build_win(current.code)`
5. Double-check the `DESCRIPTION` file
6. Submit the package to CRAN: `release(current.code, check=FALSE)`

The `check()` command is analogous to the `R CMD check` from the terminal, but it also (re)builds the package. Assuming that the package passes all of the required checks, it is ready for submission to CRAN. As a final precaution, we recommend taking a moment to visually inspect the `DESCRIPTION` file to ensure that it contains the correct email address for the maintainer and the correct release version. Finally, the `release()` command will submit the package via FTP and generate the required email. This email should come from the same address listed for the package maintainer in the `DESCRIPTION` file.

3. Extensions: Documentation, Data, Demos, and S4

Often, R packages include additional documentation, datasets, and commands that can be executed using the `demo()` function. These can illustrate the package's functionalities, replicate results from a published article, or illustrate a set of results for a collaborator. Some authors may also work in the S4 framework, which requires more documentation and some tricks for setting up the package to pass CRAN checks. A somewhat more developed R package might consist of a directory structured as follows.

```
squaresPack
├── DESCRIPTION
├── NAMESPACE
├── R
│   ├── addSquares.R
│   ├── subtractSquares.R
│   ├── exampleDataset.R
│   └── squaresPack-package.R
├── man
│   ├── squaresPack.Rd
│   ├── addSquares.Rd
│   ├── subtractSquares.Rd
│   └── exampleDataset.Rd
└── data
    └── exampleDataset.rda
└── demo
    ├── 00Index
    ├── addSquares.R
    └── subtractSquares.R
```

3.1. Package Documentation

One common feature of many packages is some simple documentation of the package itself. Using `devtools`, this requires the author to include a chunk of code in some file in the `R` subdirectory. In our example, we include a file called `squaresPack-package.R` containing the following code. (Note the use of the `@docType` designation and that no actual R code is associated with this documentation.)

```
#' squaresPack
#'
#' The squaresPack package performs simple arithmetic calculations.
#' @name squaresPack
#' @docType package
#' @author Ryan T. Moore: \email{rtm@@american.edu} and
#' Jacob M. Montgomery: \email{jacob.montgomery@@wustl.edu}
#' @examples
#'
#' \dontrun{
#' demo(addSquares)
#' demo(subtractSquares)
#' }
#'
NULL
```

3.2. Datasets

Another common feature of many `R` packages is the inclusion of datasets. Datasets are typically stored as `*.rda` objects and must be located in the `data` subdirectory. Documenting the dataset is similar to documenting the package itself. In our example, we created a separate `exampleDataset.R` file with the following content.

```
#' Example Dataset
#'
#' This line could include a brief description of the data
#'
#' The variables included in the dataset are:
#' \itemize{
#' \item\code{Variable1} A vector of random numbers
#' \item\code{Variable2} Another vector of random numbers
#' }
#'
#' @name exampleDataset
#' @docType data
NULL
```

3.3. Demo Files

The demo file provides examples for particular functions or the package as a whole. Demo files should contain a single R script that will be run when the user calls `demo(addSquares)` or `demo(subtractSquares)`. Since this command will also be run during the normal R check, authors may want to omit any extremely slow or time-consuming command.

As the directory structure above shows, the `demo` subdirectory must include an index file named `00Index` listing the included demo files, one per line, with a short description:

```
addSquares      Demo file for addSquares
subtractSquares Demo file for subtractSquares
```

Each demo file is a `*.R` file that ends with code that is run when the user types, for example, `demo(addSquares)`:

```
dx <- 1:2
dy <- 3:4
addSquares(dx, dy)
```

3.4. S4 Considerations

Finally, some authors may work in an S4 environment, which requires the specification of both class structures, generics, and class-specific methods.[7] In S4 development, every class, subclass, and method must have a help file. To handle this, one can include a list of 'aliases' in the help files. That is, one can make one help file for the class definition also work for some of the more trivial class methods that may not require their own documentation. To do this, one includes multiple class-specific methods in the `@alias` block. One can point multiple classes and methods to the same help file using the `@rdname` command.

4. Conclusion

We illustrate how the `devtools` package can aid package authors in package maintenance by automating several steps of the process. The package allows authors to focus on only editing `*.R` files since both documentation and metadata files are updated automatically. The package also automates several steps such as submission to CRAN via ftp.

While we believe that the `devtools` approach to creating and managing R packages offers several advantages, there are potential drawbacks. We routinely use other of Hadley Wickham's excellent packages, such as `reshape`, `plyr`, `lubridate`, and `ggplot2`. On one hand, each of them offers automation that greatly speeds up complex processes such as attractively displaying high-dimensional data. However, it can also take time to learn a new syntax for old tricks. Such frustrations may make package writers hesitant to give up full control from a more manual maintenance system. By making one's R code conform to the requirements of the `devtools` workflow, one may lose some degree of flexibility.

Yet, `devtools` makes it simpler to execute the required steps efficiently. It promises to smoothly integrate package development and checks, to cut out the need to switch between R and the command line, and to greatly reduce the number of files and directories that must be manually edited. Moreover, the latest release of the package contains many further refinements, such as building packages directly from GitHub repositories, creating vignettes, and creating "clean" environments for code development. While developing R packages in a manner consistent with `devtools` requires re-learning some techniques, we believe that it comes with significant advantages for speeding up development and reducing the frustration commonly associated with transforming a batch of code into a package.

References

Chambers, John M. *Software for Data Analysis: Programming with R*. New York, NY: Springer.

R Core Team. 2013. "Writing R Extensions." `http://cran.r-project.org/doc/manuals/R-exts.html`, Version 3.0.0.

R Core Team. 2014. *R: Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0.

Wickham, Hadley. 2013. *devtools: Tools to Make Developing R Code Easier*. R package version 1.2. `http://CRAN.R-project.org/package=devtools`.

Wickham, Hadley, Peter Danenberg, and Manuel Eugster. 2011. *roxygen2: In-source documentation for R*. R package version 2.2.2. `http://CRAN.R-project.org/package=roxygen2`.

[7] For additional details on S4 programming, see Chambers (2008).

Rice University
Department of Political Science
Herzstein Hall
6100 Main Street (P.O. Box 1892)
Houston, Texas 77251-1892

The Political Methodologist is the newsletter of the Political Methodology Section of the American Political Science Association. The support of the Department of Political Science at Rice University in helping to defray the editorial and production costs of the newsletter is gratefully acknowledged.

Subscriptions to *TPM* are free for members of the APSA's Methodology Section. Please contact APSA (202-483-2512) if you are interested in joining the section. Dues are $29.00 per year and include a free subscription to *Political Analysis*, the quarterly journal of the section.

Submissions to *TPM* are always welcome. Articles may be sent to any of the editors, by e-mail if possible. Alternatively, submissions can be made on diskette as plain ascii files sent to Justin Esarey, 108 Herzstein Hall, 6100 Main Street (P.O. Box 1892), Houston, TX 77251-1892. LaTeX format files are especially encouraged.

TPM was produced using LaTeX.

THE SOCIETY FOR
POLITICAL METHODOLOGY
ad indicia spectate