What should be the data sharing policy of cognitive science?

Mark A. Pitt and Yun Tang

Department of Psychology, Ohio State University

Author Note

Address correspondence to either author: Department of Psychology, 1835 Neil Avenue, Ohio State University, Columbus, OH, 43210. E-mail: pitt.2@osu.edu, tang.162@osu.edu

Running head: Data sharing in cognitive science

Keywords: data sharing, repository, open science

Abstract

There is a growing chorus of voices in the scientific community calling for greater openness in the sharing of raw data that leads to a publication. In this commentary, we discuss the merits of sharing, common concerns that are raised, and practical issues that arise in developing a sharing policy. We suggest that the cognitive science community discuss the topic and establish a data sharing policy.

What should be the data sharing policy of cognitive science?

Addyman and French (2012) call on modelers to make their models accessible to cognitive scientists by not only making the code for running the model available to all, but also providing an interface that enables those with minimal modeling experience to interact with and learn about the model. This request is one example of a growing trend championed by scientists and policy makers to make science more accessible and open. Publication of the manifesto also provides an opportunity for members of the cognitive science community to discuss how open our science should be. The purpose of this commentary is to extend the call of Addyman and French to data sharing. Many disciplines have strengthened their data-sharing policies in recent decades. How strong of a policy should cognitive science have? Below is a summary of some of the issues to consider in formulating an opinion.

What are the responsibilities of the scientist?

An answer to this question can help guide the development of a data-sharing policy in cognitive science. Currently the community expects scientists to share their discoveries through journal publications, which include the presentation of summary data, such as descriptive statistics and model performance metrics. Is it also reasonable to require researchers to share their raw data, by which we mean those values that were used in the statistical analyses or modeling exercises? Advocates of open science argue in the affirmative, noting that with the wide-spread availability of internet access, open-source software, and data repositories, there are no barriers for making raw data available to anyone who wants it. Many individuals practice open science already, providing publications and other materials (e.g., stimuli, data analysis code, experimentation software, and even data) on their professional websites.

Data sharing benefits science

There is a growing belief that data sharing is sufficiently important for scientific communication that it is an obligation of the scientist. This attitude is evident in calls for open science and in the policies of national scientific organizations and federal agencies. The National Research Council (NRC) issued a report on the topic (Committee on Responsibilities of Authorship in the Biological Sciences, National Research Council, 2003), in which they strongly advocate data sharing, arguing that it is a responsibility of the scientist that comes with publication of the article that grew out of the data. Except where concerns of commercialization or security are present, data should be freely available to members of the scientific community. Complete transparency is maximally beneficial for scientific progress because it provides the opportunity to verify, evaluate, and build on published work. In response to a recognized need for addressing challenges in this area, the NRC created the Board on Research Data and Information, which focuses on improving the management, policy, and use of digital data and information for science and society.

Consistent with these efforts, major federal funding agencies have been developing and revising policies to encourage data sharing. The National Science Foundation has enhanced its data archiving policy by requiring grant applicants to submit a Data Management Plan that describes how the proposed work will conform to policies on the dissemination and sharing of research results. The National Institutes of Health has a similar requirement of large-dollar grants, and a main funder in Britain, the Economic and Social Research Council, makes data sharing a condition of its research grants. Professional societies and journals are moving toward stricter sharing policies (e.g., American Society of Naturalists, American Economics Review), especially

when there are established repositories for a particular type of data. For example, *Science* requires microarray data be submitted to an established repository prior to publication.

What is prompting the move to greater sharing? There is clear evidence that sharing benefits science. Examples of this are numerous in genetics, where repositories of data have been available to researchers for a decade. One of the biggest success stories in this field has been the completion of the Human Genome Project (HGP). Technology and resources generated and shared by HGP have had a major impact on research across the life sciences and will benefit future generations of researchers. Piwowar, Vision, and Whitlock (2011) studied reuse of data sets in the Gene Expression Omnibus database and estimated that 2,711 data sets deposited in 2007 had made third-party contributions to more than 1,150 published articles by the end of 2010.

Data sharing benefits science in many ways. As illustrated in the preceding paragraph, it stimulates new research. Researchers design a study with a particular purpose in mind, and the data collected are used to answer that question. Sharing makes it possible for those same data to be repurposed to address questions not thought of by those who originally generated it. A perhaps more common form of repurposing is a secondary analysis that grows out of the original study, such as testing an additional hypothesis or providing further insight into a particular result. Such analyses can increase or decrease the scientific impact of the study, and so are important to perform. They are often the types of questions that can arise during the review of a manuscript, where curious reviewers might ask that additional analyses be included prior to publication. Once published, a reader has to query the author in the hope that the analysis was run. If the data are publically available, the author would not have to be bothered and the reader would be free to answer the question to her satisfaction as well as pursue others.

An example of each type of repurposing in cognitive science can be found in a recent study in decision making. Erev et al. (2010) collected data in three monetary gamble tasks, and made the data available to other researchers in the form of a modeling competition aimed at brainstorming on theory development and promoting quantitative modeling. More than a dozen teams conducted secondary analyses of the shared data, either by using existing models or by introducing a modeling framework new to the topic (e.g. Gonzalez & Dutt, 2011). After analyzing the shared data, some researchers (e.g. Abdellaoui, L'Haridon, & Paraschiv, 2011; Camilleri & Newell, 2011) designed follow-up experiments to test new hypotheses that grew out of the original study. These are clear instances in which data sharing accelerated multiple discoveries.

Another type of secondary analysis that data sharing greatly facilitates is meta-analysis (Vickers, 2006). After a paradigm has been used or phenomenon has been studied for a while, it can be desirable to integrate data across multiple experiments to evaluate the reliability of findings and obtain more accurate estimates of population parameters and effect sizes. Meta-analyses are a relatively straightforward means of providing such information (Schmidt, 1992). Studies involving data aggregation in this manner are not common in cognitive science, but would surely benefit the study of mind. They would be trivial to perform if the data were easily accessible.

Data sharing can also contribute to science education, especially training undergraduate and graduate students in courses in statistics and research methods (Whitlock, 2011). Students can be directed to published data sets on a topic of interest, and have the opportunity to learn to reproduce analyses in a publication. Researchers themselves could benefit from the availability of such data sets when learning new statistical and modeling techniques. This also extends to the

development of new statistical methodologies. Ready access to multiple data sets that share the same properties would aid developers in evaluating the robustness and applicability of new methods.

Piwowar et al. (2011) suggest that data reuse is sensible financially. Reuse decreases the cost of research for the whole community by reducing the cost of each experiment, whether measured in federal funding dollars or in time and lab resources (e.g., equipment, participants) required for replication. Although there are clearly occasions when it is wisest to initiate a study by first replicating a past result, this is not always the case. Reuse is especially attractive when expensive technologies such as imaging are required for replication (Yarkoni, Poldrack, Essen, & Wager, 2010).

What is the best way to share data? Some researchers place data on their personal websites, but this method of sharing suffers from two problems: permanency and discoverability. The researchers is tasked with ensuring links are always valid, which is likely a low priority, and broken links are not uncommon because of changes in an institution's network infrastructure or relocation due to employment. Who maintains the site after the researcher has retired? The use of a personal website also reduces the visibility of the data because it can be found only by visiting the researcher's site.

Archiving data in an online repository frees researchers from the hassles of maintenance while also increasing its discoverability (Vision, 2010). Repositories are libraries of data sets, and becoming the standard method of data storage across disciplines. Like libraries, their contents can be indexed along multiple dimensions so that users can not only locate the data set that led them to the repository, but also browse (un)related data sets. Repositories range from bare-bones versions in which the data are listed only with the accompanying manuscript, much

like a personal website (e.g., American Economics Review) to those with significant curation (e.g. ncbi.nlm.nih.gov/geo and datadryad.org), which have more advanced search capabilities and provide additional content that can enhance the user experience.

Two recent events are likely to accelerate sharing via repositories. Datacite's (datacite.org) digital object identifier (DOI) for data sets has become an ISO standard, opening the way for the adoption of a single convention for a persistent link to a data set. Thomson Reuters will soon launch the Data Citation Index, which will make it possible to track the reuse of data sets, thereby providing those who collected the data with appropriate recognition. It will also link articles listed in its Web of Knowledge directly to the corresponding data sets in accredited repositories (using the Datacite DOI).

Finally, it would make sense that a data repository in cognitive science would house cognitive models as well (e.g., cmr.osu.edu), as modeling is central to the science. To achieve the functionality that Addyman and French (2012) desire, the modeling section of a repository could be designed after www.runmycode.org, a new economics modeling web site that affords varying levels of interactivity, from merely depositing code to creating a web page on which simulations could be run.

In summary, data sharing has scientific, public policy, and pedagogical value. It is promoted by national research organizations, public and private, and some societies and journals because they recognize that greater openness leads to a better and a more cumulative science.

Concerns about sharing data

It might be easy to agree in principle to share data, but crafting a policy that ensures all parties involved are treated fairly requires care and sensitivity to the idiosyncrasies of the science. Any policy must protect the participants whose data will be shared and the investigators who

collected the data. Protection of human participants is comparatively much more straightforward given that national and institutional guidelines are well developed. Through IRBs, safeguards are put in place to ensure risks to participants are minimal and confidentiality and anonymity are maintained. One of the biggest worries is ensuring data are de-identified prior to sharing, a procedure we suspect most cognitive scientists already practice.

By collecting the data, the investigators possess some intellectual rights to the data. One of these should be having a say in the conditions under which the data are made available to others. The original researchers might have plans to perform follow-up analyses or reuse their data to address other theoretical questions. They should be given the opportunity to pursue these ideas without fear of competition from others. There could also be proprietary or contractual obligations that could restrict sharing.

How are the rights of the original investigators protected while at the same time providing others access to the data? One solution is to give researchers the option to embargo their data for a time period (e.g., six months to one year) after publication of the manuscript (Whitlock, 2011). This solution is a compromise that provides the researcher with time to complete or make significant headway on planned research, while also fulfilling the obligation to make the data available to the community in a timely manner. If the data are sensitive in any way, contracts that place restrictions on their use, as well as restricting the pool of qualified users, can be put in place.

A reluctance to share can stem from a deeper concern that sharing could harm one's career. By making data publicly available, the researcher loses control of how it will be used and could be the target of abuse (e.g., purposely hunting for errors or discrepancies). A means of addressing this concern must be part of a data sharing policy, as there is no place for bullying in

science. To minimize such unpleasant interactions, which are probably very rare, Vickers (2006) suggests that when data are reused, the original researcher be involved in the reuse in some way. This could include collaborating with those who want to reuse the data, reviewing the manuscript that emerges from the re-analysis, or being given the opportunity to write a reply should the re-analysis be published. Such a policy also minimizes the misuse of data. That said, validation is part of the scientific process. Scrutiny should be permitted so that science can be corrected when necessary (the reproducible research movement has a similar goal). This is especially important if the findings have significant societal consequences (Godlee, Smith, & Marcovitch, 2011).

Other concerns about sharing can be found in OHBM (2001) and Gøtzsche (2011). Although they vary in their relevance for cognitive science, like the above issues, none is insurmountable. Assuming data sharing is considered worthwhile by the community, a policy should be put in place with the recognition that it is a work in progress.

The pragmatics of sharing data

Unless data sharing is part of the culture (e.g., astrophysics), any policy without some means of enforcement will probably result in poor compliance (Guttmacher, Nabel, & Collins, 2009). Wicherts, Borsboom, Kats, and Molenaar (2006) requested data sets from the authors of 141 articles that appeared in three journals published by the American Psychological Association, which has an explicit policy that authors share their data with those who ask for it. Only 26% of the data sets were provided. Compliance is only somewhat better in disciplines in which data repositories exist and journals expect submission. Alshekh-Ali, Qureshi, Al-Mallah, and Ioannidis (2011; see also Savage & Vickers, 2009) surveyed data availability in 500 articles published in journals with the 50 highest impact factors in all of science, which differ widely in their policies on data sharing. Of those that had a sharing policy, only 41% of the papers fully

complied. Piwowar (2011) reports a similar level of compliance in archiving gene expression microarray data (45%), for which there are multiple centers for data storage and strict enforcement by some journals. Research in this field can be funded by large NIH grants, which require an explicit policy on sharing as a condition of funding. Although sharing was more likely from such studies, it was surprisingly low.

Why are sharing rates not higher? In a survey inquiring about the data-sharing attitudes and practices of geneticists, the most common reason for denying a request for data was the amount of work required (Campbell et al, 2002). Researchers are busy people, and locating even recently collected data, not to mention trying to reconstruct how the data are organized, can take too much time than one is willing to spend. This attitude is not unexpected because for most scientists, requests for data are sufficiently rare that data are not prepared for storage with sharing in mind. But as with new rules in federal grant and IRB applications, the process can become routine after one has implemented a system. To become a priority, sharing has to be valued. In genetics, those who re-use their own data or data from others are more likely to share, suggesting the value of sharing increases with experience (Piwowar, 2011).

Reluctance to share is not widespread in all disciplines, and changes in attitude have led some to revise policy. After the results of a survey asking ecologists and evolutionary biologists showed that 95% favor archiving their data, editors of multiple journals and publishers agreed to require all authors of accepted manuscripts to submit their data to an online database (Whitlock, McPeek, Rausher, Rieseberg, & Moore, 2010). Such a change no doubt arose from extended discussion and eventual agreement among most community members that the change was in the best interest of the discipline.

Consensus building is moving more slowly in one discipline closer to home. From 2000 to 2006, the Journal of Cognitive Neuroscience required all fMRI data to be submitted to fmridc.org as a condition of publication. In as few as four years after its creation, tangible benefits of the archive were appearing (van Horn, Grafton, Rockmore, & Gazzaniga, 2004). Openfmri.org is a new outlet for such data, and renewed calls are being made to increase sharing in the field (Visscher & Weissman, 2011; Yarkoni et al, 2010).

Money can have a profound effect on a sharing policy. Who is going to pay for the cost of building and maintaining a repository? Societies and publishers? Although federal funds might be available to create a repository, sustainability is the more costly. However easy and automated the submission and storage and retrieval processes, personnel will be needed to maintain the site and ensure best practices in data curation. Relatedly, what level of reuse, which might occur slowly, will be necessary to justify the investment? For these reasons, no-frills repositories (e.g., Dataverse) can be a sensible first step.

There are also concerns about the financial repercussions of implementing a data-sharing policy that veers too far from the norms of other journals in a field. Would such a policy threaten the existence or prestige of the journal because authors would avoid publishing in it? If circulation and citations drop, there could be pressure from the publisher to reverse the policy, as a journal's policies must reflect the opinions of the community it serves. Such concerns might seem worrisome enough for some to prefer to maintain the status quo. However, another recent example, from in the field of economics, shows change can be smooth and successful. In 2005, the *American Economics Review*, a leading journal in the field, implemented a strict data availability policy (www.aeaweb.org/aer/data.php) that requires submission of all data and analysis programs upon submission of a manuscript for review. Other economics journals

followed suit and adopted the same policy. That this change has been in place for so long suggests any negative impact has thus far been negligible.

Summary

Calls for more open data sharing policies are growing louder. They come not just from within the scientific community, where evidence continues to accumulate that sharing benefits science, but also from society and policy makers, who expect greater accountability and transparency by those whose research funding and salaries are supported by tax dollars. In this climate, what are our obligations? A healthy discussion of this issue should be followed by implementation of a policy.

References

Addyman, C. & French, R. (2012). Computational modeling in cognitive science: A manifesto for change. *Topics in Cognitive Science, 4(3),* 332-341. doi: 10.1111/j.1756-8765.2012.01206.x

Abdellaoui, M., L'Haridon, O., & Paraschiv, C. (2011). Experienced vs. described uncertainty: Do we need two prospect theory specifications? *Management Science, 57,* 1879-1895. doi:10.1287/mnsc.1110.1368

Alshekh-Ali, A. A., Qureshi, W., Al-Mallah, M. H., and Ioannidis, J. P. A. (2011). Public availability of published research data in high-impact journals. *PLoS ONE* 6(9):e24357. doi:10.1371/journal.pone.0024357

Camilleri, A. R., & Newell, B. R. (2011). When and why rare events are underweighted: A direct comparison of the sampling, partial feedback, full feedback and description choice paradigms. *Psychonomic Bulletin & Review, 18,* 377-384. doi:10.3758/s13423-010-0040-2

Campbell, E. G., Clarridge, B. R., Gokhale, M., Birenbaum, L., Hilgartner, S., Holtzman, N. A., & Blumenthal, D. (2002). Data withholding in academic genetics: Evidence from a national survey. *Journal of American Medical Association, 287(4),* 473-480. doi:10.1001/jama.287.4.473

Committee on Responsibilities of Authorship in the Biological Sciences, National Research Council. (2003). *Sharing Publication-Related Data and Materials: Responsibilities of Authorship in the Life Sciences.* Washington, D.C.: The National Academies Press.

Erev, I., Ert, E., Roth, A. E., Haruvy, E., Herzog, S. M., Hau, R. … Lebiere, C. (2010). A choice prediction competition: Choices from experience and from description. *Journal of Behavioral Decision Making, 23,* 15-47. doi:10.1002/bdm.683

Godlee, F., Smith, J., & Marcovitch, H. (2011). Wakefield's article linking MMR vaccine and autism was fraudulent. *BMJ,* 342:c7452. doi:10.1136/bmj.c7452

Gonzalez, C., & Dutt, V. (2011). Instance-based learning: Integrating sampling and repeated decisions from experience. *Psychological Review, 118,* 523-551. doi:10.1037/a0024558

Gøtzsche, P. C. (2011). Why we need easy access to all data from all clinical trials and how to accomplish it. *Trials,* 12(1):249. doi:10.1186/1745-6215-12-249

The Governing Council of the Organization for Human Brain Mapping (OHBM). (2001). Neuroimaging databases. *Science, 292(5522),* 1673-1676. doi:10.1126/science.1061041

Guttmacher, A. E., Nabel, E. G., & Collins, F. S. (2009). Why data-sharing policies matter. *Proceedings of the National Academy of Sciences, 106(40),* 16894-16894. doi:10.1073/pnas.0910378106

Koslow, S. H. (2000). Should the neuroscience community make a paradigm shift to sharing primary data? *Nature Neuroscience, 3(9),* 863-865. doi:10.1038/78760

Piwowar, H. (2011). Who shares? Who doesn't? Factors associated with openly archiving raw research data. *PLoS ONE, 6*(7):e18657. doi:10.1371/journal.pone.0018657

Piwowar, H. A., Vision, T. J., & Whitlock, M. C. (2011). Data archiving is a good investment. *Nature, 473 (7347),* 285-285. doi: 10.1038/473285a

Savage, C. J., & Vickers, A. J. (2009). Empirical study of data sharing by authors publishing in PLoS journals. *PLoS ONE* 4(9):e7078. doi:10.1371/journal.pone.0007078

Schmidt, F. L. (1992). What do data really mean? Research findings, meta-analysis, and cumulative knowledge in psychology. *American Psychologist, 47*(10), 1173-1181. doi:10.1037/0003-066X.47.10.1173

Van Horn, J. D., Grafton, S. T., Rockmore, D., & Gazzaniga, M. S. (2004). Sharing neuroimaging studies of human cognition. *Nature Neuroscience, 7(5),* 473-481. doi:10.1038/nn1231

Vickers, A. J. (2006). Whose data set is it anyway? Sharing raw data from randomized trials. *Trials* 7:15. doi:10.1186/1745-6215-7-15

Vision, T. J. (2010). Open Data and the Social Contract of Scientific Publishing. *BioScience, 60,* 330-331. doi:10.1525/bio.2010.60.5.2

Visscher, K.M. & Weissman, D.H. (2011). Would the field of cognitive neuroscience be advanced by sharing functional MRI data? *BMC Medicine*, 9:34. doi:10.1186/1741-7015-9-34

Whitlock, M. C., McPeek, M. A. Rausher, M. D. Rieseberg, L. & Moore, A. J. (2010). Data archiving. *American Naturalist. 175*, 145-146. doi:10.1086/650340

Whitlock, M. C. (2011). Data archiving in ecology and evolution: Best practices. *Trends in Ecology & Evolution, 26(2),* 61-65. doi:10.1016/j.tree.2010.11.006

Wicherts, J. M., Borsboom, D., Kats, J., & Molenaar, D. (2006). The poor availability of psychological research data for reanalysis. *American Psychologist, 61*, 726-728. doi:10.1037/0003-066X.61.7.726

Yarkoni, T., Poldrack, R. A., Essen, D. C. V., & Wager, T. D. (2010). Cognitive neuroscience 2.0: Building a cumulative science of human brain function. *Trends in Cognitive Sciences, 14(11),* 489-496. doi:10.1016/j.tics.2010.08.004