



The Importance of Standards for Sharing of Computational Models and Data

Russell A Poldrack¹ · Franklin Feingold¹ · Michael J Frank² · Padraig Gleeson³ · Gilles de Hollander^{4,5} · Quentin J. M. Huys³ · Bradley C. Love³ · Christopher J. Markiewicz¹ · Rosalyn Moran⁶ · Petra Ritter⁷ · Timothy T. Rogers⁸ · Brandon M. Turner⁹ · Tal Yarkoni¹⁰ · Ming Zhan¹¹ · Jonathan D. Cohen¹²

Published online: 7 October 2019

© Society for Mathematical Psychology 2019

Abstract

The target article by Lee et al. (in review) highlights the ways in which ongoing concerns about research reproducibility extend to model-based approaches in cognitive science. Whereas Lee et al. focus primarily on the importance of research practices to improve model robustness, we propose that the transparent sharing of model specifications, including their inputs and outputs, is also essential to improving the reproducibility of model-based analyses. We outline an ongoing effort (within the context of the Brain Imaging Data Structure community) to develop standards for the sharing of the structure of computational models and their outputs.

Keywords Data sharing · Standards · Reproducibility · Computational models

The issues raised by Lee et al. (in review) regarding reproducibility and robustness in model-based approaches in cognitive science are well stated, and the proposed solutions would have important positive impacts on the field. In this commentary, we focus on a related, complementary issue that was not discussed in detail by Lee et al., yet that we believe to be equally

important for improving reproducibility: the transparent sharing of model specifications, including their inputs and outputs. One strong motivation for our interest in the specification of model structure, inputs, and outputs comes from the increasing prevalence of model-based analyses in neuroimaging and cognitive neuroscience (Frank 2015; Turner et al. 2017).

The problem of reproducing computational models from the descriptions included in most published studies is comparable with that of reproducing statistical analyses: often, many of the details (parameter values and/or procedures) needed to replicate the results are not explicitly described and are difficult to determine from the (often custom) code used to implement the model. This poses a problem not only for replicating results of models published by others but often even when trying to re-implement a model developed earlier in the same laboratory using a software and/or hardware environment that is no longer accessible.

There are several different levels at which a modeling result might be reproducible (Benureau and Rougier 2017), all of which have important implications for computational cognitive science. First, a researcher may wish to rerun the same code in exactly the same way as reported, in order to simply reproduce the results. Even this baseline level of reproducibility can be challenging across time within a single lab, as software interfaces and computational environments change. Second, a researcher may wish to rerun the same code while varying parameters or inputs (including different input datasets), in order to assess

✉ Russell A Poldrack
russpold@stanford.edu

¹ Stanford University, Stanford, USA

² Brown University, Providence, USA

³ University College London, London, UK

⁴ University of Zurich, Zurich, Switzerland

⁵ Spinoza Centre for Neuroimaging, Amsterdam, the Netherlands

⁶ King's College London, London, UK

⁷ Charité Universitätsmedizin Berlin & Berlin Institute of Health, Berlin, Germany

⁸ University of Wisconsin-Madison, Madison, USA

⁹ The Ohio State University, Columbus, USA

¹⁰ University of Texas at Austin, Austin, USA

¹¹ National Institute of Mental Health, NIH, Bethesda, USA

¹² Princeton University, Princeton, USA

robustness, check for identifiability of parameters, or assess the ability of the model to recover parameters from simulated data. This requires an additional level of clarity regarding the structure of the inputs, outputs, and parameters. Finally, full reproducibility requires the ability to rerun the same model using a different implementation, which also enables assessment of generalizability. The latter is necessary both to assess the reliance of any result on specific algorithmic or implementational choices, as well as to identify any potential software errors (that occur with increasing likelihood in any large software project, especially when software engineering best practices are not employed, as is often the case with scientific software). Achieving these ends requires a formal model specification, in terms of the equations, function definitions, and/or algorithms that precisely and clearly communicate the theory from which any particular software implementation follows (Cooper and Guest 2014), and that can be used to reproduce the published results (Fum et al. 2007).

Careful comparative analyses of different computational cognitive modeling approaches applied to the same data, which would be greatly facilitated by reproducible analysis methods, can even lead to important theoretical results. For example, Ratcliff and Childers (2015) performed a comparative analysis of different fitting routines for the diffusion decision model (DDM), a prominent model of behavior on speeded choice tasks. They showed that different fitting routines can lead to different parameter estimates. Similarly, Donkin et al. (2011) and Goldfarb et al. (2014) fitted both the DDM and a competing response time model, the linear ballistic accumulator (LBA) model to the same data sets. They showed that using different models, although conceptually similar, can lead to slightly different interpretations of response time patterns. Unfortunately, comparative studies such as these are currently very time-intensive, because different packages need to be installed and learned and these packages often have very different input/output specifications. An alternative approach would be to code the models themselves from scratch, but this is potentially even more time-intensive and prone to errors. A common standard for the input/output structure of such packages and, better still, a standard format for specification of the models themselves would make important modeling comparison studies such as those mentioned above considerably less time-intensive and much easier to reproduce on independent data sets.

The sharing of code is clearly necessary for reproducibility (Eglen et al. 2017; Gleeson et al. 2017), but is rarely sufficient, as nicely stated by Buckheit and Donoho (1995): “An article about computational science... is not the scholarship itself, it is merely advertising of the scholarship. The actual scholarship is the complete software development environment and the complete set of instructions which generated the figures.” Sharing of the execution environment is now possible using containerized computing environments (such as Docker or Singularity), but containerization alone only enables the first level of reproducibility outlined above. The higher levels of

reproducibility require a clear way to specify the structure of the model so that it can be modified and/or replicated in a different environment, and compared with others.

The need for standards

The availability of standards for the organization of data and models has the potential to greatly enhance the utility of shared models and results; without such standards, inordinate effort is required on the part of one researcher simply to understand the organization of the information shared by another. Lee et al. state the following regarding the sharing of modeling details:

Modelers should always endeavor to make their models available (Baumgaertner et al., 2018). The motivating goal of ensuring availability is to preserve the rights of others to reach independent conclusions about model-based inferences. A minimum standard, then, is to provide accessible modeling details that allow a competent person in the field to reproduce the results. This is likely to include mathematical and statistical description, an algorithm or pseudo-code, user documentation, and so on. Providing these details in a sufficiently precise form makes a model available, and means it is likely to be used and understood more broadly than by a specific researcher or a single lab.

While we concur with the goals described above, we believe that achieving them on a meaningful scale will only be possible through the development and use of community standards for the description of computational models. This is because of the growing diversity of ways in which models can be expressed, and the fact that different investigators often arrive at different interpretations of the degree of precision needed for full reproducibility. An instructive example comes from the sharing of data within the brain imaging community which suffered from similar problems of a proliferation of formats and practices with regard to precision of description and organization. This community has benefited substantially from the development of community standards, first of a common file format (NIfTI) supported by all major software packages, and more recently through the Brain Imaging Data Structure (BIDS) (Gorgolewski et al. 2016), that defines a schema for the organization and naming of data files and folders, as well as a framework for the specification of metadata that are important for data analysis at multiple levels. Whereas BIDS first arose as a standard for raw magnetic resonance imaging data, it has been extended by the community to a number of additional data types including magnetoencephalography (MEG) and electroencephalography (EEG).

In addition, ongoing work is extending the standard to derived data types, such as the outputs of neuroimaging analyses and the statistical models used to analyze these data.

Inspired by the increasing relevance of computational models in neuroimaging research and supported by the NIH BRAIN Initiative, a group of researchers (including the authors of this commentary) met in April 2019 to initiate the development of a standard framework for the description of computational models relevant to neuroscience and psychology. The breadth of modeling approaches represented by the attendees was intentionally wide, from cognitive models to dynamic neural models to spiking neural networks, such that while the inspiration came from neuroimaging research, the ultimate product is meant to be useful for a much broader range of modeling approaches (from single neurons to higher order cognition).

A first goal of the standards effort is to develop a scheme to organize the inputs to a computational model (such as behavioral or neuroscientific measurements) and the outputs of a modeling run. In many cases, this effort dovetailed with prior work in describing derived data types within BIDS, such as connectivity matrices that may be the result of structural or functional connectivity analysis, but which also may inform brain-network or statistical models. A further need was identified to standardize the representation of stimuli and training corpora used in empirical studies and/or computational simulations, which were left free-form in the original BIDS specification but would benefit from greater structure in the modeling context. Finally, several classes of derivatives specific to computational and statistical models were identified, including descriptions of parameter distributions, traces of internal model variables, and synthetic neuroimaging data.

A second (and arguably much more ambitious) goal of the effort was to develop a common framework for describing the structure of computational models themselves. The upside of a common specification for computational models is potentially enormous: researchers would be able to implement the same model in multiple environments; to easily inspect, evaluate, compare, and extend one another's models; and to express even relatively complex models using a common generative syntax. However, the barriers to developing such a standard are also considerable. One challenge is the need to balance expressivity against simplicity: it is probably impossible for a relatively compact specification to capture the full breadth of computational models used in all of neuroscience and psychology, so difficult choices will have to be made in balancing the desire for comprehensiveness with the goal of practicality. The BIDS community heavily emphasizes the Pareto Principle or 80/20 Rule—a recognition that, with regard to adoption, the sweet spot for a new standard is a specification that is as simple as possible while capturing as wide a range of common use cases as possible. We expect that the same philosophy is likely to be fruitful in the case of computational models.

One potentially promising approach to developing a standard for model specification is description in the form of a computational graph. This is an approach that is seeing increasing use in other domains (such as the data sciences and machine learning) and has already been adopted by at least two open-source model specification projects that address very different levels of analysis in neuroscience: NeuroML (Cannon et al. 2014), focused on spiking network models containing elements ranging from integrate and fire neurons to biophysically detailed cells, and PsyNeuLink (psyneurolink.org), focused on higher level models of system-level brain and cognitive function. Despite these differences in focus, both share the idea that computational elements (whether neurons, neural populations, entire brain areas, or even abstract cognitive functions) can be expressed as a graph, the nodes of which describe the function carried out by each computational element (as well as its parameters), and the edges of which describe the exchange of information among them (i.e., the sources of inputs and destinations of outputs for the functions of the nodes they connect).

The initial goal of this approach would be to provide a standard format that allows the description of a model to be imported into and executed by any run-time environment that supports the standard. A key requirement of this approach is the ability to express the functions of nodes in a standard form. A first step in this direction would be the definition of an ontology of the most commonly used functions—describing these in a canonical mathematical form (e.g., using a common vocabulary of function and parameter names)—while allowing references to existing function libraries (such as MATLAB, Numpy, or any publicly accessible repository) for functions not in the ontology. For example, a model that included neural network layers that linearly combine and apply a logistic transformation to their inputs would be able to reference the functions of nodes as “LinearCombination” and “Logistic,” using parameter names specified for these in the ontology. Specifications of functions not in the ontology (for example, the analytic expression for the distribution of reaction times and mean accuracy of a drift diffusion process) would be required to reference an existing library or a published form of the function (Navarro and Fuss 2009). Such a framework would have the ability to encode a very broad class of models, ranging from neural networks to Bayesian models to large-scale models of neural dynamics. This could facilitate exchange across the existing diversity of environments that focus on particular levels of analysis and/or theoretical approaches, such as NEURON (Carnevale and Hines 2006), Emergent (Aisa et al. 2008), Nengo (Bekolay et al. 2014), The Virtual Brain (Sanz Leon et al. 2013), and ACT-R (Anderson et al. 2004). During the recent BIDS workshop, the group developed an outline for such an approach that provides a promising starting point for a standard model description format that can help bridge between and integrate

existing computational modeling efforts in neuroscience and psychology.

Moving forward, our group plans to further develop the standard for computational model description as part of the BIDS ecosystem, though its utility far beyond brain imaging is a possible argument for its deployment as an independent standard (inspired by, but separate from, the BIDS initiative). We would also note that many of the issues raised here are not specific to BIDS and are relevant to any attempt to develop new standards for computational models. This process will undoubtedly involve many iterations over an expanding set of concrete use cases. Researchers interested in contributing to this effort can find more information at the BIDS Github page, where all discussion regarding the standards takes place (<https://github.com/bids-standard/bids-specification/issues/230>).

Funding Information Development of this article was supported by the BRAIN Initiative and National Institute of Mental Health (1R24MH114705). P.R. funding sources: H2020 VirtualBrainCloud 826421, Human Brain Project 785907 and ERC 683049; German Research Foundation CRC 1315, CRC 936 and RI 2073/6-1; Berlin Institute of Health & Foundation Charité, Johanna Quandt Excellence Initiative.

References

- Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., & Qin, Y. (2004). An integrated theory of the mind. *Psychological Review*, *111*, 1036–1060.
- Baumgaertner, B., Devezer, B., Buzbas, E.O., and Nardin, L.G. (2018). A model-centric analysis of openness, replication, and reproducibility.
- Bekolay, T., Bergstra, J., Hunsberger, E., Dewolf, T., Stewart, T. C., Rasmussen, D., Choo, X., Voelker, A. R., & Elias Smith, C. (2014). Nengo: A python tool for building large-scale functional brain models. *Frontiers in Neuroinformatics*, *7*, 48.
- Benureau, F. C. Y., & Rougier, N. P. (2017). Re-run, repeat, reproduce, reuse, replicate: Transforming code into scientific contributions. *Frontiers in Neuroinformatics*, *11*, 69.
- Buckheit, J.B., and Donoho, D.L. (1995). WaveLab and reproducible research. In *Wavelets and statistics*, A. Antoniadis, and G. Oppenheim, eds. (New York, NY: Springer New York), pp. 55–81.
- Cannon, R. C., Gleeson, P., Crook, S., Ganapathy, G., Marin, B., Piasini, E., & Silver, R. A. (2014). LEMS: A language for expressing complex biological models in concise and hierarchical form and its use in underpinning NeuroML 2. *Frontiers in Neuroinformatics*, *8*, 79.
- Carnevale, N.T., and Hines, M.L. (2006). *The NEURON book* (Cambridge University Press).
- Donkin, C., Brown, S., Heathcote, A., & Wagenmakers, E.-J. (2011). Diffusion versus linear ballistic accumulation: Different models but the same conclusions about psychological processes? *Psychonomic Bulletin & Review*, *18*, 61–69.
- Eglen, S. J., Marwick, B., Halchenko, Y. O., Hanke, M., Sufi, S., Gleeson, P., Silver, R. A., Davison, A. P., Lanyon, L., Abrams, M., Wachtler, T., Willshaw, D. J., Pouzat, C., & Poline, J. B. (2017). Toward standard practices for sharing computer code and programs in neuroscience. *Nature Neuroscience*, *20*, 770–773.
- Frank, M. J. (2015). *Linking across levels of computation in model-based cognitive neuroscience. An introduction to model-based cognitive neuroscience* (pp. 159–177).
- Fum, D., Del Missier, F., and Stocco, A. (2007). The cognitive modeling of human behavior: Why a model is (sometimes) better than 10,000 words.
- Gleeson, P., Davison, A. P., Silver, R. A., & Ascoli, G. A. (2017). A commitment to open source in neuroscience. *Neuron*, *96*, 964–965.
- Goldfarb, S., Leonard, N. E., Simen, P., Caicedo-Núñez, C. H., & Holmes, P. (2014). A comparative study of drift diffusion and linear ballistic accumulator models in a reward maximization perceptual choice task. *Frontiers in Neuroscience*, *8*, 148.
- Gorgolewski, K. J., Auer, T., Calhoun, V. D., Craddock, R. C., Das, S., Duff, E. P., Flandin, G., Ghosh, S. S., Glatard, T., Halchenko, Y. O., Handwerker, D. A., Hanke, M., Keator, D., Li, X., Michael, Z., Maumet, C., Nichols, B. N., Nichols, T. E., Pellman, J., Poline, J. B., Rokem, A., Schaefer, G., Sochat, V., Triplett, W., Turner, J. A., Varoquaux, G., & Poldrack, R. A. (2016). The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments. *Scientific Data*, *3*, 160044.
- Navarro, D. J., & Fuss, I. G. (2009). Fast and accurate calculations for first-passage times in Wiener diffusion models. *Journal of Mathematical Psychology*, *53*, 222–230.
- Ratcliff, R., & Childers, R. (2015). Individual differences and fitting methods for the two-choice diffusion model of decision making. *Decision*, *2*, 237–279.
- Sanz Leon, P., Knock, S. A., Woodman, M. M., Domide, L., Mersmann, J., McIntosh, A. R., & Jirsa, V. (2013). The Virtual Brain: A simulator of primate brain network dynamics. *Frontiers in Neuroinformatics*, *7*, 10.
- Turner, B. M., Forstmann, B. U., Love, B. C., Palmeri, T. J., & Van Maanen, L. (2017). Approaches to analysis in model-based cognitive neuroscience. *Journal of Mathematical Psychology*, *76*, 65–79.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.