All models are wrong, some are useful, but are they reproducible?

Commentary on Lee et al. (2019)

Michael David Wilson[1], Russell James Boag[2], Luke Strickland[3].

1 Future of Work Institute, Curtin University, Perth WA, Australia

2 University of Amsterdam, Amsterdam, Netherlands

3 University of Western Australia, Crawley WA, Australia

Author Note

Address correspondence to Michael David Wilson, Curtin University, 78 Murray Street, Perth, Western Australia 6000 (michael.d.wilson@curtin.edu.au).

Abstract

Lee et al. (2019) make several practical recommendations for replicable and useful cognitive modeling. They also point out that the ultimate test of the usefulness of a cognitive model is its ability to solve practical problems. Solution-oriented modeling requires engaging practitioners who understand the relevant applied domain but may lack extensive modeling expertise. In this commentary, we argue that for cognitive modeling to reach practitioners there is a pressing need to move beyond providing the bare minimum information required for reproducibility, and instead aim for an improved standard of transparency and reproducibility in cognitive modeling research. We discuss several mechanisms by which reproducible research can foster engagement with applied practitioners. Notably, reproducible materials provide a starting point for practitioners to experiment with cognitive models and evaluate whether they are suitable for their domain of expertise. This is essential because solving complex problems requires exploring a range of modeling approaches, and there may not be time to implement each possible approach from the ground up. Several specific recommendations for best practice are provided, including the application of containerization technologies. We also note the broader benefits of adopting gold standard reproducible practices within the field.

Lee et al. (2019) provide a number of practical recommendations for robust cognitive modeling. At several points they touch upon *reproducibility*[1] — the extent to which researchers make openly available their experimental data, code, descriptions of the software dependencies required to execute the code, and provide clear user documentation. Lee et al. propose a minimum standard for reproducibility: "to provide accessible modeling details that allow a competent person in the field to reproduce the results" (p. 6). Here we argue that there is a pressing need to move beyond this minimum standard of reproducibility, towards a gold standard, to facilitate the uptake of cognitive modeling in applied fields. By the gold standard of reproducibility, we refer to the practice of providing a complete and automated analytical pipeline that includes all materials to reproduce the results of a given study, accompanied by high quality documentation (Peng, 2011). Reproducible practices are increasingly supported by emerging technologies, such as dynamic document generation tools (e.g., R Markdown), version control and code/data sharing platforms (e.g., Github, Open Science Framework), and containerization technology (e.g., Docker, Singularity).

Lee et al. state that "ultimately, the test of the usefulness of a theory or model is whether it works in practical applications" (p8). Testing the applied utility of cognitive models requires engaging practitioners familiar with the problems that "solution-oriented modeling" attempts to solve. However, applied practitioners often do not have extensive modeling expertise, and as such there are barriers to engagement and communication with modeling experts. In this commentary, we argue that reproducible and open research practices can substantially enhance the adoption and understanding of cognitive models in the applied community. We focus on a

[1] This definition contrasts with *replicability*, the extent to which findings can be repeated in new experiments when there is no *a priori* reason to expect a different outcome.

specific applied field, human factors: a multi-disciplinary domain that focuses on the application of psychological principles to the engineering and design of workplace systems. However, our arguments also hold for a range of other fields of applied psychology.

In human factors research and practice, cognitive theory is frequently applied to model human performance in simulated task environments, and routinely translated to inform real-world decisions made by practitioners and system designers. Like many fields of psychology, human factors research often relies on flexible verbal theories, particularly when the research involves synthesizing data with anecdotal reports (e.g., accident analyses; expert interviews). However, human factors researchers are often interested in latent cognitive processes that require a model to identify, particularly in the context of simulated task environments. In addition, cognitive models have great potential utility in practice, for instance by providing a means to predict behavior when human in-the-loop testing is not feasible. As such, human factors can greatly benefit from the adoption of cognitive modeling (Byrne & Pew, 2009). There are also reciprocal benefits to modelers. Human factors paradigms provide excellent testbeds for evaluating model generalizability, can lead to novel theoretical insights, and inspire future model development. For example, recent evidence accumulation modeling of performance in a cognitively demanding air traffic control task has inspired the development of a detailed theory of how attentional capacity relates to evidence accumulation, with potentially broad applications (Boag, Strickland, Loft, & Heathcote, 2019).

The most straightforward and significant reason that reproducibility benefits applied engagement is that high-quality reproducible materials (e.g., well documented model code) provide a starting point for practitioners to experiment with models independently (i.e., without the need to procure outside help or expertise). Implementing a cognitive model from

mathematical descriptions alone can demand a massive amount of time and expertise, which practitioners do not typically possess or have easy access to. It is essential for applied practitioners to be able to experiment with a range of models to determine which is most appropriate to the problem at hand before dedicating the time and resources required to develop expertise and refine methodologies. Practitioners often face complex issues that could potentially benefit from a range of modeling approaches, and as such, being required to implement models by hand does not permit adequately exploring the solution space. Reproducible examples also provide a clear vignette of the required data structures to apply a model. This is critical for practitioners faced with complex data sets (e.g., from simulation software) which typically demand deliberation and effort to shape into the structure required for modeling.

One potential issue for practitioners hoping to quickly assess a range of cognitive models is that different model code can have dramatically different, and potentially conflicting, sets of dependencies, making it difficult to configure and debug environments. A related problem is that updates and changes to software dependencies can break previously working code (e.g., deprecated features in an R package), colloquially referred to as ‘code rot’. These threats to reproducibility can be mitigated through containerization - a class of technologies that enable encapsulating code in a ‘container’ that specifies an exact operating environment (for a comprehensive review see, Boettiger, 2005). We believe the initial time investment required to implement containerization is justified by the accessibility it affords modelers and practitioners.

Ideally, reproducible models should be implemented in generalized modeling frameworks that flexibly apply to a range of experimental designs. This would not necessarily benefit reproducibility, but would make it more straightforward to directly adapt models to the applied solution space and provide a clear method for assessing model generalizability. The best example

of such a framework may be ACT-R (Anderson & Lebiere, 1998), which has proven enormously useful and influential in the human factors literature (e.g., Laughery, Plott, Matessa, Archer, & Lebiere, 2012). We appreciate that in some cases, cognitive models cannot be easily generalized beyond the task in which they were developed. Even in these cases, reproducibility can facilitate engagement between modelers and practitioners. At the early stages of collaboration, reproducible examples could be adapted to address common questions from practitioners. For instance, one of the first questions we often receive when interacting with practitioners is "how many observations are required to fit this model?". Rather than relying on heuristics from modelers, practitioners could use reproducible recovery studies as an entry point to testing parameter recovery. Engaging practitioners in this type of model testing will promote substantive collaborations that are more likely to produce desired outcomes (e.g., by preventing situations where serious confounds emerge after significant investment of time and/or resources).

We acknowledge that the gold standard of reproducibility we have outlined here is a lofty ambition that may not always be feasible. For example, modelers may be constrained by limited expertise in the technologies that cultivate reproducible research, or be unable to make the time investment required to fully refine and document their code. Fortunately, this need not discourage modelers from making ongoing efforts to move towards the gold standard, because reproducibility is not an all-or-nothing proposition. Rather, there is a spectrum of possibilities between a study being unreproducible and meeting the gold standard (Peng, 2011). The closer modeling research falls to the gold standard on this spectrum, the more useful it will be to applied practitioners. Indeed, we would expect most reproducibility efforts to inevitably be an iterative and collaborative process in which incremental improvements in coding and documentation practices accumulate from successive implementations.

Reproducibility is not the only pathway to increasing adoption of cognitive modeling in practice, and of course there are broader and more substantive benefits to reproducibility within the field. However, moving forward, if we seriously seek to evaluate the robustness and quality of models by their practical value, then we must take the necessary steps to ensure our methods can realistically be applied by practitioners.

References

Anderson, J. R., & Lebiere, C. (1998). *The atomic components of thought*. Mahwah, N.J: Lawrence Erlbaum Associates.

Boag, R. J., Strickland, L., Loft, S., & Heathcote, A. (2019). Strategic attention and decision control support prospective memory in a complex dual-task environment. *Cognition*, *191*, 103974. https://doi.org/10.1016/j.cognition.2019.05.011

Boettiger, C. (2015). An introduction to Docker for reproducible research, with examples from the R environment. *ACM SIGOPS Operating Systems Review*, *49*(1), 71–79. https://doi.org/10.1145/2723872.2723882

Byrne, M. D., & Pew, R. W. (2009). A History and Primer of Human Performance Modeling. *Reviews of Human Factors and Ergonomics*, *5*(1), 225–263. https://doi.org/10.1518/155723409X448071

Laughery, K. R., Plott, B., Matessa, M., Archer, S., & Lebiere, C. (2012). Modeling Human Performance in Complex Systems. In G. Salvendy (Ed.), *Handbook of Human Factors and Ergonomics* (4th ed., pp. 931–961). https://doi.org/10.1002/9781118131350.ch32

Lee, M. D., Criss, A. H., Devezer, B., Donkin, C., Etz, A., Leite, F. P., … Vandekerckhove, J. (2019, January 10). Robust modeling in cognitive science. https://doi.org/10.31234/osf.io/dmfhk

Peng, R. D. (2011). Reproducible Research in Computational Science. *Science*, *334*(6060), 1226–1227. https://doi.org/10.1126/science.1213847