Computational Brain & Behavior
https://doi.org/10.1007/s42113-019-00041-2

ORIGINAL PAPER

On Testing and Developing Cognitive Models

Thomas J. Palmeri[1]

Abstract
The target article, "Robust Modeling in Cognitive Science," proposes a number of recommended practices in computational modeling in response to the growing "crisis of confidence" facing many scientific disciplines, including psychology and neuroscience. Those of us who do modeling, write about modeling, teach modeling, and mentor modelers worry deeply about best practices and any new suggestions for making modeling more transparent, trusted, and robust are welcome. Many of the recommendations seem uncontroversial. My commentary focuses on forms of preregistration and postregistration, which constitute three of the four key ideas highlighted as take-home recommendations at the conclusion of the target article. I have chosen to consider these recommendations by reflecting on my own past experiences developing new models and modeling approaches.

Keywords Computational modeling · Preregistration · Categorization · Decision making

I applaud this distinguished team of researchers for proposing ways to make cognitive modeling more transparent, trusted, and robust. My commentary considers how a few of these recommendations might have impacted my own past work had these recommendations been made years ago. Some of the recommendations regarding good practices in model fitting and model comparison seem uncontroversial and some are embodied in some way in modern textbook treatments of cognitive modeling (e.g., Farrell and Lewandowsky 2018). My comments focus on forms of preregistration and postregistration, which constitute three of the four key ideas highlighted as take-home recommendations at the conclusion of the article.

Some of my earliest cognitive modeling projects contrasted alternative models of category learning, such as ALCOVE (Kruschke 1992), the rational model (Anderson 1990), and the configural cue model (Gluck and Bower 1988), on their ability to predict errors made when learning different types of categorization problems (Nosofsky et al. 1994a)[1] and learning categories at different levels of abstraction (Palmeri 1999). I could well imagine work like this having been preregistered. These were well-established models, all of which could be implemented precisely following their descriptions in the literature, with a goal that was a straightforward comparison of model predictions, fitted and evaluated using well-established techniques, at least for their time,[2] with failures of certain models that were qualitative in nature, not merely quantitative by some particular metric. From what I recollect (after many years), the way the models were fitted and evaluated were loosely "preregistered" in that we did not deviate from the way similar models had been fitted and evaluated in previous work. Preregistration would have memorialized decisions made before the modeling work began.

While I remain to be entirely convinced that preregistration would have made our work stronger (e.g., Adam 2019), or would have facilitated the peer review process, I understand that one prime motivator for such recommendations is a lack of trust, manifest as a "crisis of confidence" (Pashler and Wagenmakers 2012). Would preregistration or even a registered modeling report have allayed the criticism about certain examples of modeling leveled by Roberts and Pashler (2000)? Perhaps, though I suspect not. As the authors of the target article note well, "preregistration is no substitute for good judgment." To the extent that preregistration—whether

[1] An early example of "adversarial" collaboration (Kahneman and Klein 2009) in cognitive modeling.

✉ Thomas J. Palmeri
thomas.j.palmeri@vanderbilt.edu

[1] Department of Psychology, Vanderbilt University, 111 21st Avenue South, 301 Wilson Hall, Nashville, TN 37240, USA

[2] We have certainly evolved over the years to using more robust modeling methods (Farrell and Lewandowsky 2018), from minimizing sum-squared-error (SSE) in these early publications, to minimizing chi-squared or maximizing likelihood, to using Bayesian estimation and model comparison when possible (e.g., Annis and Palmeri 2018, 2019).

Published online: 26 July 2019

required or rewarded—encourages a modeler to think critically about each step of the modeling process, this could help push modelers to engage in better (if not best) practices.

I wonder what goes into a preregistration and how deviations from preregistered model predictions are to be interpreted. Johansen and Palmeri (2002) had participants learn novel categories and tracked over the course of category learning how individual participants generalized their category knowledge to untrained test items in three experiments. Early in learning, participants generalized on the basis of single diagnostic dimensions, consistent with the use of simple categorization rules. Later in learning, participants generalized in a manner consistent with the use of similarity-based exemplar retrieval, attending to multiple stimulus dimensions.

The category structure used in experiment 1 was first used by Medin and Schaffer (1978) to contrast predictions of exemplar and prototype models. The category structures used in experiments 2 and 3 were constructed by me to contrast rule and exemplar models. I created dozens of different category structures over the course of many weeks, testing each on what kinds of generalization from trained items to test items would be predicted by a particular exemplar model, the generalized context model (GCM) (Medin and Schaffer 1978; Nosofsky 1986). I selected for experiment 2 and 3 category structures that were different from one another that predicted different generalization patterns depending on whether a participant was using a single-dimension rule versus exemplar similarity to categorize. "Ideally, a preregistered model could take the form of the precise predictions that are made by the model" (Lee et al., in press, p. 2)—I could well have preregistered these generalization predictions by the model, anticipating the rule generalizations early in learning and the exemplar generalizations later in learning.

But my predictions would have been wrong. After collecting the data, we saw that in experiments 2 and 3, some prominent generalization patterns had emerged over the course of category learning that I had not anticipated in my initial predictions generated using the GCM. We ended up fitting the data using ALCOVE, a connectionist implementation of the GCM that learns categories, and found that ALCOVE predicted at the end of learning the full distribution of generalization patterns, including the prominent ones I had not initially predicted using GCM. Maybe ALCOVE was too flexible? Maybe with the right choice of parameters, ALCOVE could account for any types of generalizations? No. We conducted tens of thousands of simulations of ALCOVE over a dense grid of possible parameter combinations (see also Pitt et al. 2006), which took several weeks on what were state-of-the-art workstations at the time, and found instead that ALCOVE predicted these prominent generalizations over most parameter combinations. My predictions were wrong. I erroneously assumed that the static GCM would make the same predictions as the learning ALCOVE model. If I had preregistered the wrong predictions with GCM, would that have weakened the explanatory power of ALCOVE? I hope not. ALCOVE predicts the prominent generalization patterns over a wide range of parameter values (as well as by optimizing parameters to fit the details of observed data) whether I preregistered those predictions beforehand or discovered those predictions after the data were collected. Preregistering might well have highlighted my inability to anticipate the surprises lurking in models with variability, nonlinearities, and parallelism that learn (Hintzman 1990), but would not have impacted whether or not a model truly predicts an observation in a (nearly) parameter-free manner.

Fitting an existing model and comparing existing models may well lend themselves to some form of preregistration since the models are already specified and the fitting and comparison approach can be selected, described, and justified. But it is unclear where and how preregistration comes in when developing a new model or a new modeling approach. I have been fortunate to work with some great collaborators on the development of new cognitive models, such as RULEX (Nosofsky et al. 1994b), EBRW (Nosofsky and Palmeri 1997; Palmeri 1997), the interactive race model (Boucher et al. 2007), and the gated accumulator model (Purcell et al. 2010). As the authors rightly note (Lee et al. 2019, p. 6), "model development is a creative activity that often proceeds in [an] incremental and exploratory fashion."

When developing EBRW, we wanted to create a model that could predict both errors and RTs during categorization and predict how those changed with learning and expertise. We were guided theoretically by the GCM, instance theory (Logan 1988), and accumulation of evidence models of decision making as our building blocks, but we did not have a fully complete blueprint of how those blocks might come together until we started to generate simulations and try to fit the model to data. When creating a model like the gated accumulator (Purcell et al. 2010), we were also creating a new approach to model-based cognitive neuroscience (Palmeri et al. 2017; Turner et al. 2017) that used the spike rates recorded from neurons in awake-behaving primates to drive an accumulation of evidence model to predict saccade decisions. How best to use the spiking data to drive model predictions, how to aggregate spikes and behavioral data across sessions, and how to both evaluate the fits of the model to behavioral data and evaluate the predictions of neural dynamics from accumulator dynamics were all discoveries that emerged over the course of a couple of years of model development and exploration. In all of the cases of developing new models and modeling approaches that I have been involved with, I cannot think of a time when it would have been most appropriate or sensible to preregister our theoretical plans.

In such cases, the authors of the target article instead propose postregistration documentation. Some elements of this documentation make eminent sense for any computational laboratory, like keeping detailed modeling records, using modern version control methods, maintaining onsite and offsite shared repositories, and establishing digital laboratory protocols (e.g., Noble 2009; Rouder et al. 2019). Sadly, too often, my digital records are a photo of the notes on the 8-foot white board in my lab using my iPhone and my lack of diligence in establishing and enforcing digital protocols has occasionally created challenges in finishing a project when someone leaves the lab.

But these recommendations go beyond mere digital hygiene to suggest that "in exploratory model development,"[3] the postregistration documentation should be "made public at the time of publication or even as the research is being done" (Lee et al., in press, p. 6). While I admit to having watched with morbid fascination the live webcam of H.M.'s famous brain being cryogenically sliced (Annese et al. 2014), is there really a use for a public digital record of all the undistilled hunches, the soul-crushing, time-sucking theoretical rabbit holes, and ideas that with hindsight were sheer lunacy that are all part of forging new theoretical ground? At least in the work I have been involved in, the most important lessons learned during model development—the models that fail—are either given a prominent place within the body of an article or are described in footnotes, appendices, and supplementary materials; these failed approaches seemed like good ideas to us, so they might be to someone else. Such failures also help explain and understand what works and justify why models might need certain complexity because simpler alternative may fail in important ways. Some of these theoretical journeys also make their way, albeit in distilled form, into the formal classes I teach and the informal mentoring I provide to folks working with me—and I suppose into this commentary as well. Raw thoughts are not raw data, whether when creating new models or designing new experiments.

Keeping good records of modeling steps (in programming a model, simulating a model, fitting a model to data, contrasting alternative models) is unquestionably important. There must be sufficient detail in any modeling article—whether in the body of the article itself, in an appendix, or, when article space is severely limited, in supplementary information—to allow any competent modeler to reproduce the model predictions. It is incumbent upon authors to be mindful about providing sufficient detail and upon editors and reviewers to demand such detail. But, in my view, those details will be carefully distilled from the records of modeling steps, not be a raw copy of a physical or digital laboratory notebook in a postregistration document.

I recognize that there is a danger in being perceived as railing[4] against robustness. Who would ever not want to be robust—"one might as well ask for acne" (Mook 1983). But the appropriate boundary conditions for when preregistration is appropriate, or even necessary, and what kinds of information might actually be useful for science, if anything, in a postregistration when developing new models needs very careful consideration by the field before becoming expected—or demanded—practice in computational modeling.

Funding Information TJP is supported by NSF grant SMA 1640681 and NEI grant R01 EY021833.

References

Adam, D. (2019). A solution to psychology's reproducibility problem just failed its first test. *Science*. Retrieved from https://www.sciencemag.org/news/2019/05/solutionpsychology-s-reproducibility-problem-just-failed-its-first-test. Accessed 23 May 2019.

Anderson, J. R. (1990). *The adaptive character of thought*. Hillsdale: Erlbaum.

Annese, J., Schenker-Ahmed, N. M., Bartsch, H., Maechler, P., Sheh, C., Thomas, N., ... & Klaming, R. (2014). Postmortem examination of patient H.M.'s brain based on histological sectioning and digital 3D reconstruction. *Nature Communications*, *5*, 3122.

Annis, J., & Palmeri, T.J. (2018). Bayesian statistical approaches to evaluating cognitive models. *Wiley Interdisciplinary Reviews in Cognitive Science*.

Annis, J., & Palmeri, T. J. (2019). Modeling memory dynamics in visual expertise. *Journal of Experimental Psychology: Learning, Memory, and Cognition*.

Boucher, L., Palmeri, T. J., Logan, G. D., & Schall, J. D. (2007). Inhibitory control in mind and brain: an interactive race model of countermanding saccades. *Psychological Review, 114*, 376–397.

Farrell, S., & Lewandowsky, S. (2018). *Computational modeling of cognition and behavior*. Cambridge University Press.

Gluck, M. A., & Bower, G. H. (1988). Evaluating an adaptive network model of human learning. *Journal of Memory & Language, 27*, 166–195.

Hintzman, D. L. (1990). Human learning and memory: connections and dissociations. *Annual Review of Psychology, 41*, 109–139.

Johansen, M. K., & Palmeri, T. J. (2002). Are there representational shifts during category learning? *Cognitive Psychology, 45*, 482–553.

Kahneman, D., & Klein, G. (2009). Conditions for intuitive expertise: a failure to disagree. *American Psychologist, 64*(6), 515–526.

Kruschke, J. K. (1992). ALCOVE: an exemplar-based connectionist model of category learning. *Psychological Review, 99*, 22–44.

Logan, G. D. (1988). Toward an instance theory of automatization. *Psychological Review, 95*, 492–527.

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review, 85*, 207–238.

[3] "Exploratory" is such an unfortunate word since it is so often hedged in science in ways that connote "merely exploratory." Creating a model that for the first time instantiates a new set of theoretical principles, or accounts for a new type of phenomenon, or establishes links between brain and behavior in a new way is a deeply exploratory process. Whereas fitting an existing model might take a few weeks for well-mentored member of a laboratory, creating a new model or modeling approach, at least in my experience, can take many months if not years of deep, scientific exploration by a team of collaborators.

[4] And my intent is not to rail.

Mook, D. G. (1983). In defense of external invalidity. *American Psychologist, 38*(4), 379–387.

Noble, W. S. (2009). A quick guide to organizing computational biology projects. *PLoS Computational Biology, 5*(7), e1000424.

Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of Experimental Psychology: General, 115*, 39–57.

Nosofsky, R. M., Gluck, M., Palmeri, T. J., McKinley, S. C., & Glauthier, P. (1994a). Comparing models of rule-based classification learning: a replication and extension of Shepard, Hovland, and Jenkins (1961). *Memory & Cognition, 22*, 352–369.

Nosofsky, R. M., Palmeri, T. J., & McKinley, S. C. (1994b). Rule-plus-exception model of classification learning. *Psychological Review, 101*, 53–79.

Nosofsky, R.M., & Palmeri, T.J.(1997). An exemplar-based random walk model of speeded classification. *Psychological Review, 104*, 266–300

Palmeri, T. J. (1997). Exemplar similarity and the development of automaticity. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23*, 324–354.

Palmeri, T. J. (1999). Learning hierarchically structured categories: a comparison of category learning models. *Psychonomic Bulletin & Review, 6*, 495–503.

Palmeri, T. J., Love, B. C., & Turner, B. M. (2017). Model-based cognitive neuroscience. *Journal of Mathematical Psychology, 76*, 59–64.

Pashler, H., & Wagenmakers, E.-J. (2012). Editors' introduction to the special section on replicability in psychological science: a crisis of confidence? *Perspectives on Psychological Science, 7*, 528–530.

Pitt, M. A., Kim, W., Navarro, D. J., & Myung, J. I. (2006). Global model analysis by parameter space partitioning. *Psychological Review, 113*, 57–83.

Purcell, B. A., Heitz, R. P., Cohen, J. Y., Schall, J. D., Logan, G. D., & Palmeri, T. J. (2010). Neurallyconstrained modeling of perceptual decision making. *Psychological Review, 117*, 1113–1143.

Roberts, S., & Pashler, H. (2000). How persuasive is a good fit? A comment on theory testing. *Psychological Review, 107*(2), 358–367.

Rouder, J., Haaf, J. M., & Snyder, H. K. (2019). Minimizing mistakes in psychological science. *Advances in Methods and Practices in Psychological Sciences, 2*(1), 3–11. https://doi.org/10.1177/2515245918801915.

Turner, B. M., Forstmann, B. U., Love, B., Palmeri, T. J., & Van Maanen, L. (2017). Approaches to analysis in model-based cognitive neuroscience. *Journal of Mathematical Psychology, 76*, 65–79.