Computational Brain & Behavior (2019) 2:218–222
https://doi.org/10.1007/s42113-019-00056-9

ORIGINAL PAPER

The Really Risky Registered Modeling Report: Incentivizing Strong Tests and HONEST Modeling in Cognitive Science

Wolf Vanpaemel[1,2]

Published online: 2 October 2019
© Society for Mathematical Psychology 2019

Abstract
In the Registered Modeling Report format, authors specify the experimental design, models, and inference mechanism before they have seen the data and reviewers evaluate the level of detail and quality of the proposed plan. While useful, the Registered Modeling Report is limited in its ability to incentivize strong tests. I propose an extension to the Registered Modeling Report format, the Really Risky Registered Modeling Report, in which reviewers are required to evaluate whether a bad fit between the model predictions and the empirical data is plausible. The two crucial additions are that authors include prior predictions in the protocol and that reviewers set a data prior in Stage 1. Only protocols containing predictions that are not almost guaranteed to be confirmed when brought in contact with empirical data are eligible for in principle acceptance. Adopting the Really Risky Registered Modeling Report will lead to strong model tests and solid evidence.

Keywords Pre-registration · Theory testing · Model evaluation · Strong test · Data prior

Lee et al. (2019) have collected a convenient overview of good practices that will help making cognitive modeling more robust, credible, and interpretable. Their suggestions range from tried out but not sufficiently appreciated (e.g., blind and crowd-sourced analyses), over advocated but not yet applied (e.g., using utilities in model evaluation; Vanpaemel and Lee 2012a), to crisp and novel (e.g., a new journal format called Registered Modeling Report). It is no surprise that the proposals made in the last category are the least developed ones.

In particular, their proposal of a Registered Modeling Report provides an interesting starting point, but could be taken much further to guarantee a strong test of the models under consideration. In this commentary, I build upon their proposal to propose a Really Risky Registered Modeling Report, that is specifically designed with the goal of providing a strong test of how well the model approximates the empirical data. First, I outline the workflow and the intended benefits of the Registered Modeling Report. After arguing why it does not go far enough as a tool to incentivize strong tests, I provide a first skeleton of what a Really Risky Registered Modeling Report could look like, taking the lead from the seminal work of Roberts and Pashler (2000).

I am grateful for the stimulating questions and observations of two anonymous reviewers.

✉ Wolf Vanpaemel
wolf.vanpaemel@kuleuven.be

[1] University of Leuven, Leuven, Belgium

[2] Faculty of Psychology and Educational Sciences, University of Leuven, 3000, Leuven, Belgium

The Registered Modeling Report

The proposed Registered Modeling Report is a new journal format, inspired by the well-known Registered Report format that is gaining momentum in experimental psychology (see cos.io/rr and Chambers 2013). Its workflow is roughly as follows:

1. Before data are inspected, and ideally before data collection has been started, researchers write a preregistration protocol, in which they specify, and ideally justify, at least the following aspects:
   - the experimental design or the data collection mechanism (the playground)
   - the models they will test in this design (the players)
   - the formal criterion used to evaluate and compare the models when brought in contact with the empirical data (the rules)
   - ideally, the predictions made by the models in the design of interest

2. In Stage 1, this protocol is peer reviewed by reviewers who assess the level of detail of the protocol, as well as the reasonableness and quality of the design, the data collection, and the modeling. When the protocol is considered to be of sufficient detail and quality, and if the research is of sufficient interest for the readership of the journal, it is offered in principle acceptance (IPA), possibly after several rounds of revision.
3. The researchers collect the empirical data and test the models as planned. When deviations from the protocol are needed, these deviations are clearly disclosed and justified, and ideally, their impact is discussed (for example by reporting the results based on following through with the protocol as planned as well).
4. The manuscript, now updated with the modeling results, is reviewed again in Stage 2. If the protocol has been sufficiently adhered to, the manuscript is accepted, no matter what the outcome is.[1]

This format has several clear benefits. Preregistration—making important model building and model evaluation decisions independent of the observed data, and publicly documenting the decisions—has three main advantages. First, it reduces, or in some cases even eliminates, the risk that some of the modeling decisions are made to nudge the results in a favorable direction. Next, it helps in convincing readers that such nudging did not take place, and thus increases the readers' confidence in the results. Further, it pushes authors to carefully think through their experimental design and modeling strategy, which increases the quality and rigor of the research, and contrasts favorably with adopting a thoughtless "let's see how we go" attitude.

Registered Modeling Reports—having the preregistered protocol reviewed and offering IPA—come with several additional advantages, on top of the general advantages of preregistrations. First, it adds another layer of quality control to the preregistration process. Because the reviewers' feedback is no longer provided after the fact, it can, if useful, be implemented in the research rather than being buried in the limitations section of the final paper. Second, when evaluating the merit of the research, reviewers will focus on whether the research questions are important and the research design is sound, rather than being overwhelmed by the empirical and statistical results. Finally, the fact that publication is guaranteed not only helps in keeping the file drawer empty but it also assures that the time spent in collecting and modeling the data will be rewarded by the most important academic currency, lessening the need for merciless data and model torture.

[1]It is not clear what kind of deviations are tolerable for a guaranteed acceptance. This is currently an unresolved but important issue, not in the least because modeling is an inherently creative and iterative endeavor, making deviations from the protocol highly expected.

In their closing sentence, Lee et al. (2019) highlight a final advantage of Registered Modeling Reports, claiming that it "could incentivize the field to test models that make risky predictions, providing strong tests of theory and potentially rapid progress". By connecting risky predictions to strong tests, the authors seem to subscribe to a falsificationist perspective on theory testing (Popper 1959). The rationale is that if a model makes predictions that are almost guaranteed to be confirmed, there is not much to be learned from finding that these predictions are confirmed. A model can only be put to a real test if it makes risky predictions (Lakatos 1978; Meehl 1990; Roberts and Pashler 2000). If it does not make any risky predictions, it is immune to testing. Observing that the empirically observed data confirm the model predictions is completely unimpressive and should not increase the credence of the model. Only by surviving a strong test—a test that provides a genuine challenge to the model—meaningful evidence for the model is provided. Surviving a weak test—a test on which the model is almost impossible to fail—does not provide any evidence at all and corresponds to what Mayo (2019) calls a BENT case: Bad evidence. No test. To make real progress, we need HONEST cases: High opportunity for no evidence. Strong test.

Flexibility in model building and model evaluation drastically decreases the riskiness of the model predictions and hence the severity of the test and the informativeness, meaningfulness, and persuasiveness of the evidence. One commonly used type of flexibility in modeling can be called PARKing: predicting after the results are known. If researchers are given the freedom to tweak their models in the light of the observed data, so that the models can accommodate the observed data, the agreement between the model predictions (which are no longer genuine predictions, but rather postdictions) and the data is underwhelming and utterly unconvincing. We find ourselves in BENT, not HONEST territory. By limiting the modeler degrees of freedom, preregistration (be it unreviewed or as part of a Registered Modeling Report) goes some way in increasing the riskiness of the predictions and, thus, the severity of the test and the possibility of meaningful evidence. However, preregistration alone (both unreviewed and as implemented in the Registered Modeling Report format) does not go far enough. In its current form, the Registered Modeling Report does not guarantee risky predictions, strong tests, and solid evidence.

Because in the Registered Modeling Report format the flexibility to turn a bad prediction into a better one is denied or at least severely restricted, it becomes more likely that a prediction is risky, that the test is strong, and that the evidence is solid. However, without directly addressing the question Are the predictions risky?, it is impossible to answer the questions Is the experimental design a strong test

for the model? and, eventually, Is the evidence for the model solid? Because the suggested workflow for the Registered Modeling Report does not provide a head-on answer to the first question, it leaves the final two questions unanswered. As a result, the Registered Modeling Report fails to fully reach the goal of incentivizing the use of strong tests. To remedy this shortcoming, I propose a modified version of the Registered Modeling Report, which could be called the Really Risky Registered Modeling Reports.

The Really Risky Registered Modeling Report

There is a very simple and sure-fire way of having strong tests and solid evidence: asking reviewers to directly assess whether the model makes risky predictions. This is exactly what is implemented in the Really Risky Registered Modeling Report. Following Lakatos (1978) and Roberts and Pashler (2000, 2002), I consider a prediction to be risky if it is plausible that it will not be confirmed. From this perspective, the possibility of a bad fit between model predictions and empirical data is not sufficient to be risky. What matters is that a bad fit is plausible, and this is a key criterion in Stage 1 of the Really Risky Registered Modeling Report.

Somewhat counter-intuitively, to assess whether a model makes risky predictions, it is not terribly important to know what the model predicts. What matters most is what the model does not predict, and, in particular, whether these non-predictions contain plausible outcomes. This implies that making very precise predictions (i.e, being low in complexity) is no guarantee for making risky predictions: a model makes precise predictions if it rules out a lot of outcomes, but if the ruled out outcomes are not plausible, the predictions are not risky. Making novel predictions is also not a guarantee for being risky: if a model makes probabilistic predictions, the model can predict novel outcomes while at the same time not excluding any plausible outcomes at all. However, predictions that are novel and precise at the same time are quite likely to be risky. A model ruling out a lot of outcomes (because it makes precise predictions) but does not rule out the implausible outcomes (because it makes novel predictions) is likely to rule out plausible outcomes. Thus, predictions that are precise and that are unlikely to be made without the model are probably risky predictions.

Evaluating whether a test is strong, or equivalently, whether a bad fit is plausible, can be done before the data are being observed. It requires two ingredients: the predictions made by the model in a chosen experimental design, and a quantification of which outcomes are considered to be plausible in this design. The first ingredient can easily be obtained using the prior predictive distribution from Bayesian statistics (Box 1980), which provides a distribution of possible unobserved outcomes, based on the model (i.e., the likelihood and the parameter prior). The second ingredient, which is a novel concept, can be called a data prior.[2] It can be informed by previous empirical results and by competing theories and models, and, crucially, depends on the model under consideration. Vanpaemel (in press) provides a worked example focusing on category learning, demonstrating how a data prior can be specified, and how the riskiness of the predictions and, the severity of a test can be assessed. While determining a data prior is challenging, time consuming, and unconventional, it is impossible to assess the severity of a test, and thus the strength of the evidence without it.

At its core, the Really Risky Registered Modeling Report has two crucial additions with respect to the Registered Modeling Report. The first addition involves the content of the protocol. In the Really Risky Registered Modeling Report, authors specify the same elements as in a Registered Modeling Report. However, additionally, they include the model predictions for the selected design. In the Registered Modeling Report, the predictions made by the model for the experimental design at hand are welcome additional information. In the Really Risky Registered Modeling Report, in contrast, they are a crucial element. Optionally, the authors could include in the protocol what they consider a reasonable data prior by indicating which outcomes they think are plausible for their experimental design and model.

The second addition relates to what the task of the reviewers is during Stage 1. In the Really Risky Registered Modeling Report, the reviewers critically evaluate the same criteria as in the Registered Modeling Report. However, additionally, they are asked to specify their data prior, ideally providing a brief empirical or theoretical justification for their assessment. By putting the model predictions, the operationalization of what constitutes a good fit, and the data prior together, it is possible to assess the key question: is a bad fit plausible for the model and design under consideration? Only if this is the case, the model will be challenged by the data. The model is not practically guaranteed to survive contact with the data, so bringing model and data into contact will be informative. The proposed study is HONEST, and if the model passes the test, the evidence for the model is persuasive. If, in contrast, a bad fit is not plausible, the model is immune to testing in the selected design, and the study is BENT. Only an HONEST study is eligible for IPA.

The data priors as specified by the reviewers (and by the authors, if applicable) are included in the final paper,

[2] Contrary to what the name might suggest, a data prior is needed to assess riskiness of predictions even if one is working outside of the Bayesian framework.

with proper attribution. The final paper also contains a comparison of the empirically observed data against the model predictions. If the fit between model and data is bad, the data have provided evidence against the model. If the fit is good, the data have provided meaningful evidence in support of the model.

Discussion

In the Really Risky Registered Modeling Report, a key role is played by prior predictions, which are predictions in the real sense, made independent of the observed data. One way to make such predictions is provided by the prior predictive distribution from Bayesian statistics. Prior predictives have been used in cognitive modeling to measure the complexity of a model (Vanpaemel 2009) and as the basis of a prior predictive check, which can guide determining parameter priors and can help evaluating the a priori reasonableness of the model more broadly (Lee 2018; Lee and Vanpaemel 2018; Schad et al. 2019). In the Really Risky Registered Modeling Report, the prior predictive plays yet another role and is used to assess the severity of the test and the testability of the model, as well as to empirically evaluate the model.

Making useful or risky predictions, in the sense that empirical confirmation of these predictions provides solid evidence, will not be easy and will most likely require stepping away from using the much loved vague priors. The push away from vague priors should be welcomed as a positive side effect of the Really Risky Registered Modeling Report format, or of the use of prior predictives for model checking and model evaluation more generally. Especially in cognitive modeling, priors are as capable of expressing meaningful theory as is the likelihood (Vanpaemel 2010) and are sometimes even the only places where crucial theoretical assumptions can be captured (see Vanpaemel and Lee 2012b, for a worked example). Using the prior predictive forces modelers to think about what their parameters mean and to express relevant theorizing about these parameters in informative priors, and thus increases the theoretical commitment and empirical content of a model.

My proposal for the Really Risky Registered Modeling Report should be treated as an initial, rough effort, which leaves many practical questions unanswered. If it would ever be implemented, plenty of issues need to be resolved. Since theorists do not have the habit of giving the plausibility of outcomes a lot of thought, developing good practices or ideally principles to guide this thinking will be a tough nut to crack. Not only is it unclear what exact information should inform a data prior but there will also be a need to find out how a data prior could be optimally communicated. More fundamentally, the data prior will depend on the exact format in which the potentially high-dimensional prior predictive distribution is summarized. It will likely take several iterations to find out which format is best-suited for eliciting meaningful data priors. Further, given the inherent subjectivity of the data prior, it is expected that different reviewers might strongly disagree regarding the plausibility of the excluded outcomes, which might reflect their theoretical viewpoints and even personal biases. Finding adequate ways to deal with such inevitable disagreement will not be straightforward. Disagreement among reviewers is not unique to the Really Risky Registered Modeling Report. Reviewers of traditional Registered Reports are typically asked to evaluate the importance of the research questions and the soundness of the design, a task which is bound to give rise to conflicting assessments. Another issue shared with the traditional Registered Report format is what level of tolerance should be adopted for deviations from the registration protocol. Finally, the Really Risky Registered Modeling Report might prove not to be feasible, since it might turn out to be unreasonable to task already overtaxed reviewers with another challenging job. Setting a data prior requires intimate knowledge of the model and the experimental paradigm, which severely limits the potential pool of qualified reviewers.

Ultimately, installing a research practice of strong tests does not even depend on the willingness of an editorial board to implement the Really Risky Registered Modeling Report format. The task done by the reviewers—setting a data prior, and evaluating whether a bad fit is plausible—could be done using crowd sourcing. Once model predictions have been derived in a given experimental design, these could be made publicly available or privately shared with experts. Anybody with access to the predictions can assess whether the model rules out plausible outcomes for the given design and stimulus set and can share their assessment for the design and model under consideration on a public repository. These assessments are made part of the preregistered protocol and of the final paper (with proper attribution) and will help readers of the final paper to evaluate the severity of the test and the weight of the evidence.

The Registered Modeling Report is an interesting first step towards strong tests and solid evidence. However, better evidence and stronger tests can be provided by directly ensuring that a bad fit is plausible. This assessment can be done by reviewers in a Really Risky Registered Modeling Report format, for which I have outlined a rudimentary proposal. Basically, in this format, the reviewers are expected to explicitly answer a version of what Platt (1964) calls "The Question": Does the proposed experiment make a bad fit plausible? Only when the answer is yes, collecting data is not a waste of time and will provide real

scientific progress. When put to practice, no doubt several implementational details of the Really Risky Registered Modeling Report format will need fine tuning, but I hope that the skeleton I provided can serve as a bold and useful starting point for incentivizing strong tests and HONEST cognitive modeling.

References

Box, G.E. (1980). Sampling and Bayes' inference in scientific modelling and robustness. *Journal of the Royal Statistical Society: Series A (General)*, *143*(4), 383–404.

Chambers, C.D. (2013). Registered reports: A new publishing initiative at Cortex. *Cortex*, *49*(3), 609–610.

Lakatos, I. (1978). *The methodology of scientific research programmes*. Cambridge: Cambridge University Press.

Lee, M.D. (2018). Bayesian methods in cognitive modeling. *The Stevens' handbook of experimental psychology and cognitive neuroscience*, *5*, 37–84.

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

Lee, M.D., & Vanpaemel, W. (2018). Determining informative priors for cognitive models. *Psychonomic Bulletin and Review*, *25*(1), 114–127.

Mayo, D.G. (2019). *Statistical inference as severe testing*. Cambridge: Cambridge University Press.

Meehl, P.E. (1990). Appraising and amending theories: The strategy of Lakatosian defense and two principles that warrant it. *Psychological Inquiry*, *1*(2), 108–141.

Platt, J.R. (1964). Strong inference. *Science*, *146*, 347–353.

Popper, K.R. (1959). *The logic of scientific discovery*. London: Hutchinson.

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

Roberts, S., & Pashler, H. (2002). Reply to Rodgers and Rowe (2002). *Psychological Review*, *109*, 605–607.

Schad, D.J., Betancourt, M., Vasishth, S. (2019). Toward a principled Bayesian workflow in cognitive science. arXiv:1904.12765.

Vanpaemel, W. (2009). Measuring model complexity with the prior predictive. In Bengio, Y., Schuurmans, D., Lafferty, J., Williams, C.K.I., Culotta, A. (Eds.) *Advances in neural information processing systems*, (Vol. 2 pp. 1919–1927).

Vanpaemel, W. (2010). Prior sensitivity in theory testing: An apologia for the Bayes factor. *Journal of Mathematical Psychology*, *54*, 491–498.

Vanpaemel, W. (in press). Strong theory testing using the prior predictive and the data prior. *Psychological Review*.

Vanpaemel, W., & Lee, M.D. (2012a). The Bayesian evaluation of categorization models: Comment on Wills and Pothos. *Psychological Bulletin*, *138*(6), 1253–1258.

Vanpaemel, W., & Lee, M.D. (2012b). Using priors to formalize theory: Optimal attention and the generalized context model. *Psychonomic Bulletin and Review*, *19*, 1047–1056.

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