Commentary on "Robust modeling in cognitive science", Michael D. Lee, Amy Criss, Berna Devezer, Christopher Donkin, Alexander Etz, Fábio P. Leite, Dora Matzke, Jeffrey N. Rouder, Jennifer S. Trueblood, Corey N. White, and Joachim Vandekerckhove

What do the rules for the wrong game tell us about how to play the right game?

Andrew Heathcote, University of Tasmania, andrew.heathcote@utas.ed.au

Psychological science has rightly become worried about questionable practices in experimental research, with a range of recent suggestions being made about remedies for this "replication crisis". To avoid similar problems in psychological-process modelling, Lee et al. (2019) propose ingenious adaptions of these remedies along with insightful new suggestions. Although in the main applauding these developments I question whether some of the lessons drawn from the replication crisis are applicable, particularly with respect to the confirmatory vs. exploratory dichotomy given the intrinsically explanatory nature of most psychological-process models.

In reading Lee et al. (2019) I find myself half and half (c.f., Newell, 1973)[1]. Half of me is quite content, impressed not only with their many insightful comments and suggestions but also that, at every turn which might stray into the unfortunate excesses that have sometimes plagued reactions to the "replication crisis", the authors pull back from the brink and offer sensible caveats. In this I suppose I should not be surprised, because this balanced assessment reflects a collective wisdom borne of experience garnered from their many excellent contributions. Yet half of me is distressed because I am confused as to how psychological modelling can draw lessons from the replication crisis and its many purported remedies. This is because, for my distressed half, the crisis is the child of a fundamentally flawed enterprise, phenomenon-driven research that tries to play 20 questions with nature, and so I wonder how solutions to an approach that should be abandoned in any case can be relevant to the approach that should replace it?

I hasten to add I do not mean to belittle experimental research and to unduly elevate theory; as the Nobel Laurette Robert Milikan observed, science walks forward on two feet, theory and experimentation. Rather, I think psychology has got into trouble by paying attention to experiments motivated by clever words about seductively interesting and plausible dichotomies that have tenuous links to cumulative and quantitative theory development (or worse still never claimed to in the first place!). Although this trap can beset all of the sciences, I think the problem is particularly acute for psychology because the objects of our study bring to bear a new "task virtual machine" (Dennis, 2005) in each paradigm we subject them to. It should be no surprise, then, that to achieve a science of humans adequate in power and complexity we need commensurately flexible and general quantitative models, models with parameters that can be first tuned to provide an adequate descriptive account of the (often various) ways that participants process information in a range of paradigms. It is only then, in my opinion, that we can use such models to understand the psychological processes that underpin observations about the brain and behaviour in the broad.

My content half applauds Lee et al.'s many useful contributions. For example, post-registration has the potential to save much duplicated and wasted effort. The point is well taken that Bayes Factors are valid as measure of the relative likelihood of the data under different models, but that without a prior we cannot speak of evidence for the model itself. Bookend models are a great idea to address the difficulties around judging absolute fit. In the same vein I have also often found them useful in order to judge whether model parameterizations selected with an eye to simplicity (and which may therefore misfit some aspects of the data) at least come from a model with more complex parameterizations that fit well (e.g., Rae et al., 2014). However, I part ways a little when Lee et al. say: "the only difference between statistical analysis and psychological-process modeling lies in the emphasis that psychological models place on substantive interpretation". By statistical analysis I take it they mean something like the general(ized) linear model, which by design can fit any systematic data pattern, with testing of model adequacy largely limited to the characterization of measurement error. Substantive interpretation is certainly one difference, but for me a more major difference is that process models make particular predictions about data. These predictions are often further restricted by the substantive interpretation of their parameters that disallows ventures into particular regions of parameter space (i.e., selective influence assumptions), or in some cases make such forays

[1]I acknowledge borrowing in my commentary from the style and substance of Alan Newell's wonderful "You can't play 20 questions with nature and win". I dedicate this commentary to the memory of Doug Mewhort, who brought Newell's commentary to my attention, and whose work and teaching inspired in me a life-long interest in psychological modelling. Doug leavened this interest with a dollop of caution borne of another idea he made me aware of, that participants should not be viewed as always the same, but instead as deploying highly flexible virtual machines that adapt to different tasks. Thanks also to Dora Matzke for discussions related to this commentary.

informative of an expanded theory. In short, such process models are predictive and so applications of them are intrinsically both explanatory and confirmatory. That is, although they can also be used in partially exploratory ways, the explanatory nature of process models means that they are not subject to the confirmatory-exploratory dichotomy that underlies preregistration in quite the same way as are purely statistical models.

I think there are important caveats to be drawn from the fact that the traversal of the garden of forking paths by a process model can be considerably more restricted than the meanderings of their statistical cousins. For one, a process model can be totally excluded from the Eden of adequately describing the data, a penalty their statistical counterparts can rarely suffer, and certainly not with the theoretically informative sequela that can attend such occurrences for process models. For models whose main contribution is making a mapping between manifest and latent variables (e.g., MPTs and Signal Detection Theory in non-ROC designs) the restriction in their predictions can be fairly minimal beyond selective influence assumptions. However, it is often much more than that, even in quite general frameworks like evidence-accumulation modelling, and clearly this is a characteristic that is highly valued (witness the considerable furore when Jones & Dzhafarov, 2014, suggested this was not the case, with replies by Heathcote, Brown, & Wagenmakers, 2015, and Smith, Ratcliff & McKoon, 2014). I would argue that in essence every new successful application of a psychological-process model that possesses such predictive restrictions constitutes passing a generalization test, which as Lee et al. note is a powerful model selection criterion. I believe passing such tests supports inferences in favour of a model whether or not the details of how it did so are pre-registered, and it is why such models are at the heart of a cumulative psychological science.

A second caveat is related to how goodness-of-fit is properly judged, and it amplifies Lee et al.'s perspicuous differentiation between core and ancillary assumptions. Some parameters almost always need to be adapted accommodate the way participants configure themselves to particular task goals, and often to accommodate other aspects of the task design. For example, in evidence-accumulation models participants can have sometimes idiosyncratic stimulus and response biases, and differences in rate variability between item classes may be necessitated by the stimuli that happen to have been used in a given experiment. Such adaptations are hard to anticipate in a pre-registration and making them should rarely constitute grounds to declare subsequent inferences entirely or even partially exploratory. Instead these ancillary adaptations leverage the model's fundamentally explanatory nature to provide a coherent and principled account of task specific differences. Importantly, adaptation is not to be discouraged; if not done well the parameter differences that typically bear on core issues for a particular investigation may not support valid inferences because the model does not provide an accurate distillation of the data.

Indeed, even when parameters directly bearing on the core issues must be adapted in unanticipated ways, I do not think this necessarily makes the analysis entirely exploratory. As an example, consider Rae et al.'s (2014) analysis, which sought to overturn the long-held assumption in evidence-accumulation modelling that an emphasis on speed vs. accuracy selectively influences only the amount of evidence required to trigger a choice. Suppose they had pre-registered the hypothesis that there was also an effect on the mean rate of evidence accumulation (as they found). Would their analysis then be exploratory if, in order to provide an accurate description of their data, they needed not only a difference in mean rates but also rate standard deviations? Should they forgo this addition to avoid that perception? I am sure Lee et al. would say no, a they sensibly emphasize that pre-registration is no substitute for good, theoretically motivated, judgement.

These considerations take me to the heart of my anxiety that, for psychological-process models it may be considerably harder than for statistical models to follow Lee et al.'s adjuration that "It is critical … that exploratory evidence should not be misinterpreted as confirmatory evidence". No doubt we must be mindful of extra flexibility that opens the possibility of pernicious practices, but I think it is also important to acknowledge that characteristics like predictiveness and meaningfulness considerably reduce the available scope for nefarious practices in ways that are not properly captured by the solutions that have been proffered for statistical models. Perhaps my worried half is also a little paranoid in thinking that those prone to more rigid thinking and games of 20 questions might seek to put the enterprise of modelling within a straitjacket that would slow progress in building unified understandings in psychology and the neurosciences. Indeed, rather than being reassured by Lee et al.'s analogy between preregistration and a dissertation research plan, my experience of the latter is that it rarely works beyond the first experiment for any but the most pedestrian varieties of translational research, and so frequently wastes a good deal of everyone's time. I very much hope Lee et al.'s many good ideas, and the ensuring debate about them, will lead to improved modeling practices, but I also hope that everyone can be mindful that for academics, almost more than anything else, time is precious, and so any scheme to prescribe modeling research practices needs to be mindful of the compliance (and hence opportunity) cost.

Of course, it is easy to be a critic without putting one's own ideas on the line, so in closing I offer two ways that I think scarce research time can be spent to the betterment of psychological-process modelling. The first is to provide a comprehensive account of the data as possible without a priori judgements about what aspects are most important, at least in the initial stage of fitting a model. No doubt particular characteristics, often summarised in a clever plot (e.g., a Conditional-Accuracy Function, Hübner, Steinhauserv& Lehle, 2010), are theoretically telling. However, models should be first fit to all of the data using sufficient statistics (e.g., through maximum likelihood or Bayes) and then the particular characteristics assessed. Directly maximizing fit to partial aspects risks a skewed assessment. Second, although we must continue to take seriously Newell's (1973) call to eschew 20 questions and "find some way to put it all together" by building integrative models that address broad ranges of phenomena (something that a reviewer rightly commented there is still too little of), I don't think that needs to entail a Tolkienesqe search for the "one true model to rule them all". It seems to me that accepting all models are wrong, but some are useful (Box, 1979) also entails acceptance of diversity by sometimes fitting several different models (see Newell, 1990, for a similar view). Although no doubt requiring extra effort, the payoff is that conclusions supported by all models are then to some degree robust to model uncertainty and point the way to an understanding in terms of mechanisms shared among the models (e.g., Walsh, Gunzelmann & Van Dongen, 2017). A further payoff is a potential to remedy the "Toothbrush Problem" discussed in Walter Mischel's 2008 presidential APS Observer column and encapsulated in a quote from an anonymous wit: "Psychologists treat other peoples' theories like toothbrushes — no self-respecting person wants to use anyone else's.". Maybe by embracing model uncertainty we can, as Alan Newell hoped, more quickly home in on the essential structure of the mind by cooperatively working on larger theoretical wholes than we now do.

References

Box, G. E. P. (1979). Robustness in the strategy of scientific model building, in Launer, R. L.; Wilkinson, G. N. (eds.), *Robustness in Statistics*, Academic Press, pp. 201–236.

Dennis, S. (2005). A memory-based theory of verbal cognition. *Cognitive Science*, *29*(2), 145–193.

Heathcote, A., Wagenmakers, E. J., & Brown, S. D. (2014). The falsifiability of actual decision-making models. *Psychological Review*, *121*, 676–678.

Hübner, R., Steinhauser, M., & Lehle, C. (2010). A dual-stage two-phase model of selective attention. *Psychological Review*, *117*, 759– 784.

Jones, M., & Dzhafarov, E. N. (2014). Unfalsifiability and mutual translatability of major modelling schemes for choice reaction time. *Psychological Review*, 121, 1–32.

Newell, A. (1973). You can't play 20 questions with nature and win: Projective comments on the papers of this symposium. W. G. Chase (ed.), *Visual Information Processing*, New York, Academic Press.

Newell, A. (1990). *Unified Theories of Cognition*. Harvard University Press.

Rae, B., Heathcote, A., Donkin, C., Averell, L. & Brown, S. (2014). The Hare and the Tortoise: Emphasizing speed can change the evidence used to make decisions. *Journal of Experimental Psychology: Learning, Memory & Cognition*, 40, 1226-1243.

Smith, P. L., Ratcliff, R., & McKoon, G. (2014). The diffusion model is not a deterministic growth model: Comment on Jones and Dzhafarov (2013). *Psychological Review*, *121*, 679-688.

Walsh, M. M., Gunzelmann, G., & Van Dongen, H. P. (2017). Computational cognitive modeling of the temporal dynamics of fatigue from sleep loss. *Psychonomic Bulletin & Review*, 24, 1785-1807.