

Assessing the Potential of Computational Modeling in Clinical Science

Peter F. Hitchcock

Department of Psychology
Drexel University
pfh26@drexel.edu

Angela Radulescu

Department of Psychology
Princeton University
angelar@princeton.edu

Yael Niv

Department of Psychology
Princeton University
yael@princeton.edu

Chris R. Sims

Department of Psychology
Drexel University
crs346@drexel.edu

Abstract

There has been much recent interest in using reinforcement learning (RL) model parameters as outcome measures in clinical science. A prerequisite to developing an outcome measure that might co-vary with a clinical variable of interest (such as an experimental manipulation, intervention, or diagnostic status) is first showing that the measure is stable within the same subject, absent any change in the clinical variable. Yet researchers often neglect to establish test-retest reliability. This is especially a problem with behavioral measures derived from laboratory tasks, as these often have abysmal test-retest reliability. Computational models of behavior may offer a solution. Specifically, model-based analyses should yield measures with lower measurement error than simple summaries of raw behavior. Hence model-based measures should have higher test-retest reliability than behavioral measures. Here, we show, in two datasets, that a pair of RL model parameters derived from modeling a trial-and-error learning task indeed show much higher test-retest reliability than a pair of raw behavioral summaries from the same task. We also find that the reliabilities of the model parameters tend to improve with time on task, suggesting that parameter estimation improves with time. Our results attest to the potential of computational modeling in clinical science.

Keywords: computational psychiatry, psychometrics, test-retest reliability, individual differences, trial-by-trial modeling

Introduction

There has been much recent interest in translating reinforcement learning (RL) tasks into assays for use in clinical science (Hitchcock, Radulescu, Niv, & Sims, 2017; Maia & Frank, 2011). Clinical science studies typically seek to test whether some parameter or set of parameters (such as computational modeling parameters, or statistical summaries of raw behavior) co-vary with an independent variable. The independent variable might be an experimental manipulation, an intervention, or a difference in diagnostic status. For example, a large analysis of an RL task found depression and anhedonia (independent variables) were both associated with lower values of a reward sensitivity parameter (an outcome measure) (Huys, Pizzagalli, Bogdan, & Dayan, 2013).

To show that a potential outcome measure co-varies in a meaningful way with an independent variable, it is first necessary to establish that the measure is stable within-subject in the absence of change in the independent variable. That is, the measure must show high test-retest reliability. If the test-retest reliability of a candidate outcome measure is low, it is unlikely that it cleanly samples some latent construct, such as a cognitive or learning process, that is stable within individuals. A low-reliability measure is unlikely to have much long-term utility in clinical science, as it is likely to be too noisy to meaningfully and reliably co-vary with independent variables of interest (Rodebaugh et al., 2016).

Despite the importance of establishing the test-retest reliability of potential outcome measures, researchers often neglect to do this. Neglect is especially common for behavioral measures in laboratory tasks, even though such measures often have abysmal test-retest reliability (Lilienfeld, 2014). An (in)famous example is the dot probe task (Rodebaugh et al., 2017). The task was used for decades, in dozens of clinical studies, before it was shown to have close to 0 test-retest reliability (Price et al., 2015; Schmukle, 2005). The measure's lack of stability likely explains continued, widespread replication failures in studies employing it. More generally, the low reliability of measures from many laboratory paradigms poses a serious threat to progress in clinical science (see Rodebaugh et al., 2016 for discussion).

Traditionally, the measures derived from laboratory paradigms have been statistical summaries of raw behavior. For example, the outcome measure typically used in the dot probe task is average reaction time. Yet it has long been known that these measures tend to be highly variable and imprecise (Mischel, 1968). In addition to the fact that models can expose variables that are latent in behavior (e.g., learning rate), one source of excitement about applying RL modeling in clinical science comes from the idea that modeling behavior trial-by-trial will allow for the creation of outcome variables with lower measurement error (Hitchcock, 2017; Huys, Maia, & Frank, 2016).

One consequence of decreased measurement error should be more stability of measures at test and retest. However, no empirical study (of which we are aware) has compared head-to-head the test-retest reliability of RL model parameters with measures summarizing raw behavior. Thus, we compared the test-retest reliability of model parameters and behavioral summary measures derived from the Dimensions Task (Leong, Radulescu, et al., 2017; Niv et al., 2015; Radulescu et al., 2016), a multidimensional bandit with high selective attention demands. Radulescu et al., 2016 (study 2) found large differences between older adults and younger adults on two measures—one behavioral (accuracy; $g = .94$) and the other a computational modeling parameter (*decay*; older adults median = .52, younger adults median = .42)—suggesting the task has promise as a sensitive measure of individual differences.

Methods

We compared the test-retest reliability of two behavioral and model parameter measures derived from the Dimensions Task in two datasets. Datasets are from Niv et al. (2015; hereafter **D1**) and Radulescu et al. (2016, study 2; hereafter **D2**).

Specifications. The datasets differed in total trials as well as the number of trials that comprised a game. (In each “game” one dimension of the bandits determined reinforcement probability; reinforcement contingencies reset and the participant had to learn contingencies anew in each game; see Niv et al., 2015.) In D1 ($N = 22$), subjects played 500 trials (number of trials per game was drawn from a Uniform(15,25) distribution, for a total of $M=22.27$, $SD=1.45$ games per subject). In D2 ($N = 54$), subjects played ~1400 trials ($M=46.43$, $SD=5.41$ games; subjects stopped playing after exactly 40 minutes; all games were 30 trials).

Measures. Following Radulescu et al. (2016), the behavioral measures were accuracy (trials with a correct response/total trials) and number of games learned (a game was defined as learned if the participant selected the most rewarding bandit on 6 consecutive trials). The model parameters came from an RL model with decay of weights of unchosen stimuli, and were d (decay rate) and $\beta*\eta$ (the product of inverse temperature and the learning rate). We used $\beta*\eta$ as a single parameter because we have previously found in these data that β and η are highly correlated (Hitchcock et al., 2017), consistent with known identifiability issues between inverse temperature and learning rate parameters in RL models (e.g., Schönberg et al., 2007). When we used $\beta*\eta$, the four measures were only modestly correlated in both datasets (Figure 1a). For more details on the Dimensions Task, including the computational model (known as *feature RL + decay*) and its free parameters, see Niv et al. (2015). See Hitchcock et al. (2017) and Radulescu et al. (2016) regarding the clinical potential of the task.

Test-retest reliability. Test-retest reliability of behavioral and computational modeling measures was assessed via ‘A-1’ intraclass (ICC) correlation coefficients (McGraw & Wong, 1996). ICCs were calculated on parameter fits and behavioral statistics by splitting the data into first (test) and second (retest) halves, and then calculating the ICC for each measure across these halves. Each half consisted of a set of games. Of note, because reward contingencies were reset in each game, each game (and hence also each half) was independent. Also of note, the “halving” was approximate because of the task’s structure into games; specifically, the halfway split was made at the first game change after half the total trials had elapsed. A high ICC (that is, a higher correlation between first and second half scores for a given measure) reflects high within-subject stability in the measure.

Results and Discussion

In support of the premise that RL modeling can yield more precise, stable measures of individual differences than summaries of raw behavior (see also, Hitchcock, 2017), model parameters (d and $\beta*\eta$) showed higher test-retest reliability than measures summarizing raw behavior (accuracy and number of games learned) in both datasets (Figure 1b). Specifically, whereas the ICCs of the behavioral measures in both datasets were nearly 0, consistent with ICCs of other laboratory behavioral measures such as measures from the dot-probe task (Price et al., 2015; Schmukle, 2005), the ICCs of the model parameters were much higher. Notably, in D2, the parameter with the highest test-retest reliability (d) outperformed the behavioral measure with the highest test-retest reliability (accuracy) by a factor of more than 4 (.68 versus .16).

Of note, the test-retest reliability of both model parameters was significantly higher in D2 than D1. Two differences in the task specifications could have led to this difference: (1) each game had more trials in D2 (30/game compared to an average of 20/game in D1); (2) many more games were played in D2 (~46/participant compared to ~22/participant in D1). Each of these changes could conceivably have enhanced reliability in D2. To better understand the difference responsible for the increase, we computed in D2 test-retest reliabilities for pairs of game subsets of varying length (range: 10-18), with the games in each subset drawn randomly and without replacement from the first and second halves of the task (i.e., 10-18 games per half from the first and second half). For example, the two vectors of game draws from the first and the second halves of the task for a given participant in a subset in which n games were drawn might look like: vector 1: {1, 2, 5, ..., 21}; vector 2: {24, 25, 28, ..., 47}. Since this procedure of randomly drawing a

at a single visit (Epstein, 1979; Lilienfeld, 2014). These early results suggest the possibility that, as computational psychiatry progresses and modeling approaches are refined, model parameters may eventually achieve the holy grail of psychometrics: high reliability *and* high validity.

Acknowledgements

This work was supported by ARO grant W911NF-14-1-0101 (YN), NIMH grant R01MH098861 (YN & AR), and NSF research grant DRL-1560829 (CRS).

References

- Epstein, S. (1979). The stability of behavior: I. On predicting most of the people much of the time. *Journal of Personality and Social Psychology*, 37(7), 1097-1126.
- Hitchcock, P.F. (2017). Computational Modeling and Reform in Clinical Science. [preprint; osf.io/mvxfk] *Open Science Framework*.
- Hitchcock, P.F., Radulescu, A., Niv, Y., Sims, C.R. (2017). Translating a Reinforcement Learning Task into a Computational Psychiatry Assay: Challenges and Strategies. In *Proceedings of the 39th Annual Meeting of the Cognitive Science Society*.
- Huys, Q. J., Pizzagalli, D. A., Bogdan, R., & Dayan, P. (2013). Mapping anhedonia onto reinforcement learning: a behavioural meta-analysis. *Biology of mood & anxiety disorders*, 3(12), 1-16.
- Huys, Q. J., Maia, T. V., & Frank, M. J. (2016). Computational psychiatry as a bridge from neuroscience to clinical applications. *Nat. Neuro.*, 19(3), 404-413.
- Leong, Y. C., Radulescu, A., Daniel, R., DeWoskin, V., & Niv, Y. (2017). Dynamic Interaction between Reinforcement Learning and Attention in Multidimensional Environments. *Neuron*, 93(2), 451-463.
- Lilienfeld, S. O. (2014). The Research Domain Criteria (RDoC): an analysis of methodological and conceptual challenges (pp. 13-14). *Beh. Res. and Ther.*, 62, 129-139.
- Maia, T. V., & Frank, M. J. (2011). From reinforcement learning models to psychiatric and neurological disorders. *Nature Neuroscience*, 14(2), 154-162.
- McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological methods*, 1(1), 30-46
- Mischel, W. (1968). *Personality and Assessment*. New York: Wiley.
- Niv, Y., Daniel, R., Geana, A., Gershman, S. J., Leong, Y. C., Radulescu, A., & Wilson, R. C. (2015). Reinforcement learning in multidimensional environments relies on attention mechanisms. *J. of Neurosci.*, 35(21), 8145-8157.
- Podsakoff, P. M., MacKenzie, S. B., Lee, J. Y., & Podsakoff, N. P. (2003). Common method biases in behavioral research: a critical review of the literature and recommended remedies. *Journal of applied psychology*, 88(5), 879-903.
- Price, R. B., Kuckertz, J. M., Siegle, G. J., Ladouceur, C. D.,... & Amir, N. (2015). Empirical recommendations for improving the stability of the dot-probe task in clinical research. *Psych. Assessment*, 27(2), 365-376.
- Radulescu, A., Daniel, R., & Niv, Y. (2016). The effects of aging on the interaction between reinforcement learning and attention. *Psychology and Aging*, 31(7), 747-757.
- Rodebaugh, T. L., Scullin, R. B., Langer, J. K., Dixon, D. J., Huppert, J. D., Bernstein, A., Zvielli, A., & Lenze, E. J. (2016). Unreliability as a threat to understanding psychopathology: The cautionary tale of attentional bias. *Journal of Abnormal Psychology*, 125(6), 840-851.
- Schmukle, S. C. (2005). Unreliability of the dot probe task. *European Journal of Personality*, 19(7), 595-605.
- Schönberg, T., Daw, N. D., Joel, D., & O'Doherty, J. P. (2007). Reinforcement learning signals in the human striatum distinguish learners from nonlearners during reward-based decision making. *Journal of Neuroscience*, 27(47), 12860-12867.