Test-retest reliability of canonical reinforcement learning models

Laura Weidinger (lauraweidinger@outlook.com)
Center for Adaptive Rationality, Max-Planck-Institute for Human Development, 14195 Berlin, Germany

Andrea Gradassi (a.gradassi@uva.nl)
University of Amsterdam, Department of Psychology,1018 WS Amsterdam, The Netherlands

Lucas Molleman (l.s.molleman@uva.nl)
University of Amsterdam, Department of Psychology, 1018 WS Amsterdam, The Netherlands

Wouter van den Bos (w.vandenbos@uva.nl)
University of Amsterdam, Department of Psychology, 1018WS Amsterdam, The Netherlands
Center for Adaptive Rationality & Max Planck UCL Centre for Computational Psychiatry and Ageing, Max-Planck-Institute for Human Development,14195 Berlin, Germany

Abstract:
Reinforcement learning (RL) paradigms are commonly used in Cognitive Science research on human learning. These paradigms are often used in combination with computational models to estimate individual differences in learning parameters. Recently, it has been proposed that such parameter estimates can be used to better understand psychiatric conditions (Montague, Dolan, Friston, & Dayan, 2012). However, to be used as such, it is essential that the test-retest reliability of these paradigms and computational models is established. The present study seeks to close this gap by investigating the test-retest reliability of standard RL models in the context of two canonical paradigms: a probabilistic RL task with gain and loss feedback and a reversal learning task (Cools, Clark, Owen, & Robbins, 2002; Frank, Seeberger, & O'reilly, 2004). This study obtained test results from n=150 participants for each task via the online testing platform Amazon Mechanical Turk with a between-test interval of five weeks. Several standard versions of Rescorla Wagner models are fitted to the choice data in R to study the test-retest reliability of resulting parameter estimates. Test-retest reliability is studied in regard to behavioral measures and model parameters.

Keywords: reinforcement learning; computational modelling; test-retest reliability

Introduction

Psychology is experiencing a replicability crisis. One of the potential causes underlying this problem may be lacking test-retest reliability of canonical test methods (Leppink & Pérez-Fuster, 2017). Cognitive Science and Cognitive Neuroscience increasingly use reinforcement learning tasks for computational modelling of choices to infer underlying patterns of learning. Concerns have been raised about the robustness of such computational modelling results, with test-retest reliability problems shown in relation to oversimplified, or overparameterized, computational models (Collins & Frank, 2012), misalignment of the model and experiment design (Spektor & Kellen, 2018), and assumptions about dynamic versus fixed parameters (Nassar & Gold, 2013). Some of these problems could be prevented by first establishing the model identifiability and recovery (Palminteri, Wyart, & Koechlin, 2017). However, the participants or task design may also play a role in performance on test re-test reliability (e.g. due to strong path dependency in dynamic tasks). No prior research reported an empirical test of reliability of parameter estimates across different sets of computational models. The present study seeks to close this gap by testing the test-retest reliability of two canonical RL paradigms: a probabilistic RL task with gain and loss feedback (Pessiglione, Seymour, Flandin, Dolan, & Frith, 2006) and a reversal learning task (Cools et al., 2002).

Methods

Participants

Participants located in the United States completed each task via the online testing platform Amazon Mechanical Turk with a between-test interval of five weeks. Participants were allowed to take part in each task once and all participants included during T1 were invited to re-take the task five weeks later. The probabilistic gain/loss (PGL) task was completed by (n=142) participants during T1 and (n=93) during T2. The reversal learning (RVL) task was completed by (n=154) during T1 and (n=64) during T2. Behavioral analysis and computational modelling included participants whose performance met inclusion criteria during both T1 and T2, (n=69, m/f: 44/25, age=35(11)) in the PGL task and (n=47, m/f: 23,23, age=39(12)) in the RVL task (i.e. 'returners'). Exclusion criteria included failing to provide a valid MTurk ID, timing out on >20% of trials, and comments after completing the task that indicated misunderstanding the task. Participants were excluded when overall accuracy dropped <50% (PGL task) or below 55% (RVL task) or when they chose stimuli with the same laterality >20 (PGL) or >30 (RVL) times in a row.

Tasks

Two canonical decision-making tasks were tested: a probabilistic RL task with gain and loss feedback (PGL, Pessiglione et al., 2006) and a reversal learning task (Schlagenhauf et al., 2014; Cools et al., 2002). All aspects of task design were kept as in the cited studies except for slightly adapted stimuli. Choice stimuli in the RVL task were geometric shapes and in the PGL task, images of everyday objects. Feedback stimuli were an image of a \$1 bill with the headline “Gain”, an image of a crossed out \$1 bill with the headline “Loss” and an image of a neutral grey box with the headline “Neutral”.

Behavioral analysis

Three behavioral measures were analyzed: accuracy, win stay and lose shift. Accuracy is defined as the number of times the stimulus with a higher reward probability was chosen, divided by all trials. Win stay is the number of repeated choices in trials following positive feedback divided by all trials following positive feedback. Lose shift is the number of shift responses following negative feedback divided by all trials following negative feedback. Timed-out trials as well as trials with 50-50 reward probability (only occurred in RVL task) were excluded. Test-retest reliability of each behavior was studied using Pearson’s correlation and the Intra-class correlation coefficient (ICC (3,k)) over all returners between T1 and T2. ICC(3,k) scores were interpreted following (Koo & Li, 2016), with r<0.5 indicating ‘poor’, $.5 < \mathrm{r} < .75$ ‘moderate’, $.75 < \mathrm{r} < .9$ ‘good’, and $\mathrm{r} > .9$ ‘excellent’ reliability.

Computational modelling

Reinforcement learning algorithms were fitted to participants’ choice behavior to infer underlying parameter values (Sutton & Barto, 1998). Specifically, different adaptations of the Rescorla-Wagner model were used (Rescorla & Wagner, 1972). In this model, choices result from a trial-by-trial calculation of anticipated outcomes (V) of a choice (c), weighed by prediction errors (δ) and the learning rate (α).

$$V_{\mathrm{c},t+1} = V_{\mathrm{c},t} + \alpha\delta_{\mathrm{V_{c,t}}}$$

The prediction error constitutes the trial-by-trial mismatch between an anticipated outcome and the observed outcome.

$$\delta_{\mathrm{V_{c,t}}} = R_t - V_{\mathrm{c},t}$$

A modification was applied using two learning rates to differentiate between learning from positive and negative prediction error . A second modification was to add a parameter to weigh the extent to which participants use feedback to infer value-updates about the unchosen stimulus. Three variations of this ‘double-updating’ parameter were tested together with one and two learning rates. First, $\kappa = 0$ to model the absence of such inference, second, $\kappa = 1$ to model full updating, assuming anticorrelated reward for the two stimuli, and third, κ as free parameter $0 < \kappa < 1$ for individually weighted updating of the unchosen stimulus trial by trial. The third modificsation was to add a free parameter γ moderating the decay of the learning rate(s) over the course of the task, tested with one and two learning rates. All models were fitted in the programming language *R*. In total, 8 models were fitted to choices in each task for T1 and T2, resulting in 32 model fittings in total. A softmax function was used to generate a trial-by-trial probability of the observed choice behavior, given the modelled value estimates and accounting for decision noise in the free parameter (θ).

$$\mathrm{p(choice)} = \frac{\exp(\theta * V_{\mathrm{choice}})}{\sum \exp\,(\theta * V_{\mathrm{choice}}) + \exp(\theta * V_{\mathrm{\sim choice}})}$$

Free parameters were initialized at random values $(0 < \alpha, \kappa, \gamma < 1)$ and $(0 < \theta < 10)$ for each participant and constrained to these parameter boundaries except the decay parameter, which was constrained $(0 < \gamma < 4)$. All models were fitted using a general-purpose optimization algorithm based on the Nelder-Mead method (Nelder & Mead, 1965). Each model was fitted to each participant with 20 random initial parameter values to avoid getting stuck in local minima. The best fitting parameter estimates as indicated by lowest AIC value were stored for each subject.

Results

Behavioral results

Participants learn in both tasks

In the PGL task, included participants in T1 (n=119) achieved a mean hit rate of 69.25% (SD=10.41%) and during T2 (n=78) of 72.33%(9.49%). Returners (n=69) achieved hit rates of 70.44%(10.36%) during T1 and 72.63%(8.55%) during T2, with no significant difference in accuracy between T1 and T2, $t(68,2) = 1.59, p = .116$. In the RVL task, included participants at T1 (n=92) achieved a hit rate of 62.13%(10.58%) and those at T2 (n=52) 63.09%(10.23%). IQ scores and n-back scores were highly correlated in both tasks between T1 and T2, RVL IQ: $\mathrm{r} = .61, \mathrm{p} < .001$, n-back: $\mathrm{r} = .65, \mathrm{p} < .001$ and PGL IQ: $\mathrm{r} = .72, \mathrm{p} < .001$, n-back: $\mathrm{r} = .37, \mathrm{p} < .01$.

Returners (n=47) achieved a mean hit rate of 65.72%(9.41%) during T1 and of 57.03%(8.11%) during T2, both significantly above chance, T1: $\mathrm{t} = .63, \mathrm{p} < .001$, T2: $\mathrm{t} = .55, \mathrm{p} < .001$, however with a significant drop in accuracy between T1 and T2, $\mathrm{t}(47,2) = 6.79, \mathrm{p} < .001$ (**Fig. 1 a-b**).

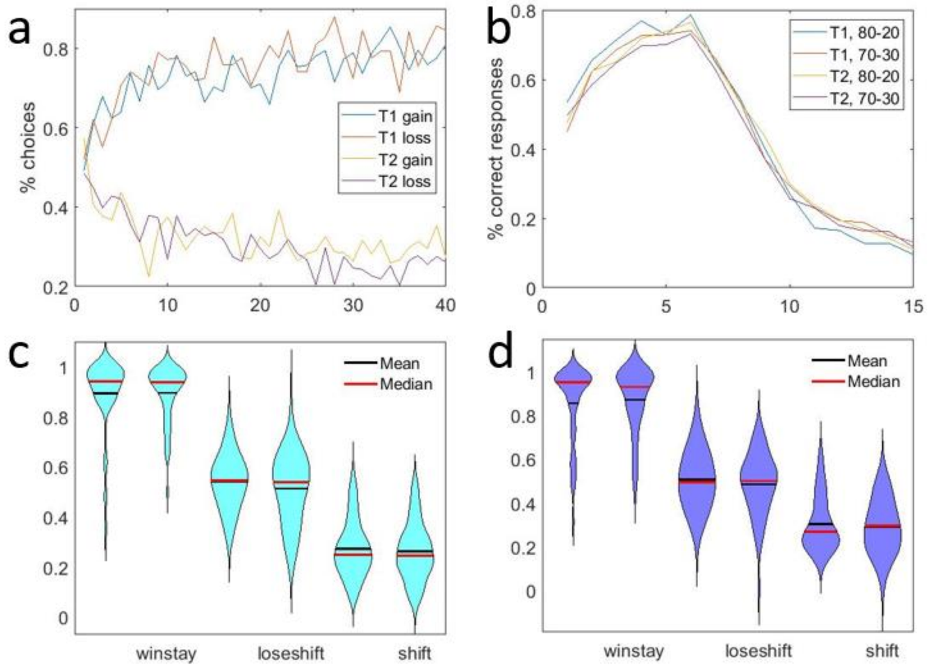

Figure 1: Choices (a-b) and behavioral measures T1 and T2 (c-d). PGL task (a,c), RVL (b,d).

Accuracy is more reliable in RVL than PGL task

In the PGL task, Pearson's correlation coefficient of accuracy between T1 and T2 was significant but small, $\mathrm{r}(69) = .28, \mathrm{p} = .019$. The ICC was significant but 'poor', $\mathrm{ICC}(3,\mathrm{k}) = .43, \mathrm{p} = .01$. In the RVL task, the correlation of accuracy between T1 and T2 was significant, $\mathrm{r}(47) = .5, \mathrm{p} < .001,$ and the ICC was significant and 'moderate', $\mathrm{ICC}(3,\mathrm{k}) = .67, \mathrm{p} < .001$.

Win stay and lose shift are reliable in both tasks

In the PGL task, Pearson's correlation coefficient was significant for win stay: $r(69) = .53, \mathrm{p} < .001$ and lose shift: $r(69) = .45, \mathrm{p} < .001$. This was mirrored by 'moderate' ICC scores, win stay: $\mathrm{ICC}(3,\mathrm{k}) = .67, \mathrm{p} < .001$, lose shift: $\mathrm{ICC}(3,\mathrm{k}) = .61, \mathrm{p} < .001$. Equally, in the RVL task, these behaviors were correlated between T1 and T2, win stay: $r(47) = .62, \mathrm{p} < .001$, lose shift: $r(47) = .55, \mathrm{p} < .001$ and corresponding ICC values were 'good', win stay: $\mathrm{ICC}(3,\mathrm{k}) = .76, \mathrm{p} < .001$, and 'moderate', lose shift: $\mathrm{ICC}(3,\mathrm{k}) = .7, \mathrm{p} < .001$.

Computational modelling results

In the PGL task, best model fit during T1 and T2 was achieved by the model with two learning rates and no double updating $(\mathrm{T1{:}\,AIC} = 139.98, \mathrm{T2{:}\,AIC} = 141.32)$. In the RVL task, best model fit during T1 and T2 was achieved by the model with two learning rates and individually weighted double-updating, $(\mathrm{T1{:}\,AIC} = 211.23, \mathrm{T2{:}\,AIC} = 212.51)$. Adding a decay parameter did not improve model fit in either task. In the PGL task, no estimated parameters by the best-fitting model exhibited a significant correlation between T1 and T2, although a trend emerged for θ, $r(69) = .23, \mathrm{p} = .057$. The ICC was significant but 'poor' for θ, $\mathrm{ICC}(3,\mathrm{k}) = .37, \mathrm{p} = .029$. No other model yielded more than one significantly correlated parameter estimate between T1 and T2.

In the RVL task, estimated values by the best fitting model were correlated for α_{neg}, $r(47) = .37, \mathrm{p} = .01$, mirrored by a 'moderate' ICC, $\mathrm{ICC}(3,\mathrm{k}) = .54, \mathrm{p} = .005$. In all other models with two learning rates, estimates for α_{neg} were significantly correlated between T1 and T2 as well. No model yielded more than two significantly correlated parameter estimates between T1 and T2. In both tasks and for both T1 and T2 the learning rates related to positive and negative prediction error were correlated with win stay and lose shift behavior respectively (with two exceptions, **Table 1**).

Table 1: Correlation of behavior and learning rates.

Task		**T1**	**T2**
PGL	Win stay, α_{pos}	.49**	.31**
	Lose shift, α_{neg}	.15	.05
RVL	Win stay, α_{pos}	.65**	.49**
	Lose shift, α_{neg}	.52**	.55**

*** p<.05, ** p<.001**

Discussion

Most behavioral measures, such as win stay and lose shift, showed substantial individual differences (**Fig. 1c-d**) and moderate to good test re-test reliability in both tasks over a time span of five weeks. This suggests both tasks capture robust individual differences in learning. Furthermore, the correlation between win stay and lose shift behaviour and learning rates from positive and negative prediction errors respectively suggests these RL models capture crucial behavioural phenotypes. However, in most models, including the best fitting model, only one parameter was correlated between T1 and T2. For the PGL task, no specific pattern emerged, whereas for the RVL task, the negative learning rate (α_{neg}) appears to be a crucial and robust factor in determining individual differences.

Our results suggest more work is needed to ensure reliable parameter estimates from reinforcement learning models. Considering experimental paradigms, one advantage of RVL over PGL may be that it requires more steady learning, potentially leading to a better fit of learning models. Although more dynamic learning tasks introduce more variability in behavior, it is possible that in this case they result in more robust parameter estimates. Second, we have not explored all possible RL models. Future work will compare a larger model space, including Bayesian models and additional parameters (e.g. 'stickiness'). Lastly, work is planned to compare model performance between fitting procedures (Log Likehood vs Hierarchical Bayesian). For instance, Bayesian approaches like *Stan* produce parameter estimates for each participant as probability distributions instead of point-estimates, which helps mitigate parameter-identifiability problems.

Conclusion

Test-retest reliability of research methods is critical for generating robust findings and making inferences about individual differences. We found behavioral measures of canonical reinforcement learning paradigms show moderate to good reliability between test sessions with a five-week interval. However, the parameters estimated through standard computational RL models did not (yet) show such robust results. Our results urge caution when interpreting estimated parameter values as individual differences in latent processes underlying learning. Further work is needed to investigate ways of improving test-retest reliability of parameter estimates.

References

Collins, A. G. E., & Frank, M. J. (2012). How much of reinforcement learning is working memory, not reinforcement learning? A behavioral, computational, and neurogenetic analysis. *European Journal of Neuroscience*, *35*(7), 1024–1035. https://doi.org/10.1111/j.1460-9568.2011.07980.x

Cools, R., Clark, L., Owen, A. M., & Robbins, T. W. (2002). Defining the neural mechanisms of probabilistic reversal learning using event-related functional magnetic resonance imaging. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *22*(11), 4563–4567. https://doi.org/20026435

Frank, M. J., Seeberger, L. C., & O'reilly, R. C. (2004). By carrot or by stick: cognitive reinforcement learning in parkinsonism. *Science (New York, N.Y.)*, *306*(5703), 1940–1943. https://doi.org/10.1126/science.1102941

Koo, T. K., & Li, M. Y. (2016). A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *Journal of Chiropractic Medicine*, *15*(2), 155. https://doi.org/10.1016/J.JCM.2016.02.012

Leppink, J., & Pérez-Fuster, P. (2017). We need more replication research – A case for test-retest reliability. *Perspectives on Medical Education*, *6*(3), 158–164. https://doi.org/10.1007/s40037-017-0347-z

Montague, P. R., Dolan, R. J., Friston, K. J., & Dayan, P. (2012). Computational psychiatry. *Trends in Cognitive Sciences*, *16*(1), 72–80. https://doi.org/10.1016/J.TICS.2011.11.018

Nassar, M. R., & Gold, J. I. (2013). A Healthy Fear of the Unknown: Perspectives on the Interpretation of Parameter Fits from Computational Models in Neuroscience. *PLoS Computational Biology*, *9*(4), e1003015. https://doi.org/10.1371/journal.pcbi.1003015

Nelder, J., & Mead, R. (1965). A simplex method for function minimization. *The Computer Journal*. Retrieved from https://academic.oup.com/comjnl/article-abstract/7/4/308/354237

Palminteri, S., Wyart, V., & Koechlin, E. (2017). The Importance of Falsification in Computational Cognitive Modeling. *Trends in Cognitive Sciences*, *21*(6), 425–433. https://doi.org/10.1016/J.TICS.2017.03.011

Pessiglione, M., Seymour, B., Flandin, G., Dolan, R. J., & Frith, C. D. (2006). Dopamine-dependent prediction errors underpin reward-seeking behaviour in humans. *Nature*, *442*(7106), 1042–1045. https://doi.org/10.1038/nature05051

Rescorla, R., & Wagner, A. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. *Classical Conditioning II: Current Research and Theory*, (2), 64–99.

Schlagenhauf, F., Huys, Q. J. M., Deserno, L., Rapp, M. A., Beck, A., Heinze, H.-J., … Heinz, A. (2014). Striatal dysfunction during reversal learning in unmedicated schizophrenia patients. *NeuroImage*, *89*, 171–180. https://doi.org/10.1016/J.NEUROIMAGE.2013.11.034

Spektor, M. S., & Kellen, D. (2018). The relative merit of empirical priors in non-identifiable and sloppy models: Applications to models of learning and decision-making. *Psychonomic Bulletin & Review*, 1–22. https://doi.org/10.3758/s13423-018-1446-5

Sutton, R. S., & Barto, A. G. (1998). Reinforcement Learning: An Introduction. *MIT Press, Cambridge, MA, A Bradford Book*. https://doi.org/10.1109/TNN.1998.712192