A hierarchical Bayesian approach to assess learning and guessing strategies in reinforcement learning

Schaaf, J.V.; Jepma, M.; Visser, I.; Huizenga, H.M.

DOI
10.1016/j.jmp.2019.102276

Publication date
2019

Document Version
Final published version

Published in
Journal of Mathematical Psychology

License
Article 25fa Dutch Copyright Act

Link to publication

Citation for published version (APA):
Schaaf, J. V., Jepma, M., Visser, I., & Huizenga, H. M. (2019). A hierarchical Bayesian approach to assess learning and guessing strategies in reinforcement learning. *Journal of Mathematical Psychology*, *93*, [102276]. https://doi.org/10.1016/j.jmp.2019.102276

Journal of Mathematical Psychology 93 (2019) 102276

Contents lists available at ScienceDirect

Journal of Mathematical Psychology

journal homepage: www.elsevier.com/locate/jmp

A hierarchical Bayesian approach to assess learning and guessing strategies in reinforcement learning[☆]

Jessica Vera Schaaf[a,*], Marieke Jepma[a], Ingmar Visser[a,b,c], Hilde Maria Huizenga[a,b,c]

[a] *Department of Psychology, University of Amsterdam, Nieuwe Achtergracht 129-B, 1018 WS Amsterdam, The Netherlands*
[b] *Yield, Research Institute for Child Development and Education, Nieuwe Achtergracht 129-B, 1018 WS Amsterdam, The Netherlands*
[c] *ABC, Amsterdam Brain and Cognition Centre, Nieuwe Achtergracht 129-B, 1018 WS Amsterdam, The Netherlands*

ARTICLE INFO

Article history:
Received 25 January 2019
Received in revised form 3 July 2019
Accepted 6 August 2019
Available online xxxx

Keywords:
Reinforcement learning
Hierarchical modeling
Bayesian model fitting
JAGS
Strategy
Guessing

ABSTRACT

In two-armed bandit tasks participants learn which stimulus in a stimulus pair is associated with the highest value. In typical reinforcement learning studies, participants are presented with several pairs in a random order; frequently applied analyses assume each pair is learned in a similar way. When tasks become more difficult, however, participants may learn some stimulus pairs while they fail to learn other pairs, that is, they simply guess for a subset of pairs. We put forward the Reinforcement Learning/Guessing (RLGuess) model — enabling researchers to model this learning and guessing process. We implemented the model in a Bayesian hierarchical framework. Simulations showed that the RLGuess model outperforms a standard reinforcement learning model when participants guess: Fit is enhanced and parameter estimates become unbiased. An empirical application illustrates the merits of the RLGuess model.

1. Introduction

In reinforcement learning agents learn, by trial and error, which actions to take in which states to maximize the total amount of reward they receive (see e.g., Sutton & Barto, 2018). A simplified version of the reinforcement learning problem is often studied using n-armed bandit tasks. For example, in two-armed bandit tasks participants learn which stimulus in a stimulus pair is associated with the highest value. In general, participants are presented with multiple stimulus pairs in a randomized order (e.g., Frank, Seeberger, & O'Reilly, 2004; Kim, Shimojo, & O'Doherty, 2006; Pessiglione, Seymour, Flandin, Dolan, & Frith, 2006). In the analysis of such data, it is typically assumed that each stimulus pair is learned in a similar way; accordingly, computational models of this learning process typically apply the same learning algorithm to each stimulus. However, if multiple pairs have to be learned in parallel, the difficulty of the task increases (Collins & Frank, 2012), which may result in learning for some pairs, and in guessing for others. In this paper we propose a model for such a combined learning and guessing process.

Typically responses from participants that guessed are excluded from the analyses by removing all data from participants that fail to reach the minimum learning criterion (see e.g., Decker, Otto, Daw, & Hartley, 2016; Doll, Jacobs, Sanfey, & Frank, 2009; Eppinger, Mock, & Kray, 2009; Hämmerer, Li, Müller, & Lindenberger, 2011; Niv et al., 2015). This is problematic as this leads to data loss and therefore loss of power (Cohen, 1988). More importantly, if data are removed, this provides an incomplete description of behavior in reinforcement learning tasks. Yet, the alternative approach in which guessing responses are included is also not recommended as it may induce bias on the parameters governing the learning process.

To address this, we propose the Reinforcement Learning/Guessing (RLGuess) model — enabling researchers to model that participants learn some stimulus pairs while they guess at others. Our first goal is to compare the model fit of the RLGuess model to a standard reinforcement learning model when the data contain guessing responses and to test the parameter recovery capabilities of the model. Our second goal is to examine whether bias is induced on the parameters estimates in standard reinforcement learning models when participants guess at some stimulus pairs (i.e., when the model is misspecified).

The structure of the paper is as follows. First we briefly discuss the fundamentals of reinforcement learning models that are currently applied and outline why guessing responses are problematic in these models. Then we introduce the RLGuess model, and present a parameter recovery study to assess the performance of the RLGuess model and to examine the effects of model misspecification on the parameter estimates. We also apply the RLGuess model to existing data from a reinforcement

[☆] Funding: HMH, JVS and MJ were supported by the Dutch National Science Foundation, NWO, The Netherlands, (VICI 453-12-005).

[*] Corresponding author.
E-mail address: j.v.schaaf@uva.nl (J.V. Schaaf).

https://doi.org/10.1016/j.jmp.2019.102276
0022-2496/© 2019 Elsevier Inc. All rights reserved.

learning task in which multiple stimulus pairs had to be learned in parallel.

2. Reinforcement learning models

Commonly applied reinforcement learning models originate in the Rescorla–Wagner Model (Rescorla & Wagner, 1972; Sutton & Barto, 2018). In these models each person makes a series of binary choices across trials $t = \{1, 2, \ldots, T\}$. At each trial t the value estimate Q of the chosen option is updated via the following rule:

$$Q(t+1) = Q(t) + \eta\delta(t) \tag{1}$$

with

$$\delta(t) = R(t) - Q(t) \tag{2}$$

where Q is the value estimate and η indicates the learning rate. The prediction error δ is computed by subtracting the current value estimate from the obtained reward R. People thus update the value estimate by scaling the prediction error with the learning rate and then adding this to the estimated value at the previous trial. Learning rates close to 1 indicate that a person makes fast adaptations based on prediction errors and learning rates closer to 0 indicate slow adaptation. The value estimates of both options are used to determine the probability to choose either option. This probability is often computed via the following softmax decision rule (Luce, 1959):

$$P\left(c(t) = 1\right) = \frac{1}{1 + \exp(-\beta\left[Q_1(t) - Q_2(t)\right])} \tag{3}$$

where $P(c(t) = 1)$ is the probability to choose the first option at trial t, Q_1 is the value estimate of the first option and Q_2 is the value estimate of the second option. The parameter β is the inverse temperature, a parameter that indicates to what extent a person's choice is guided by the difference in value estimates.

3. RLGuess model

The reinforcement learning model presented in Section 2 contains two parameters to be estimated — learning rate and inverse temperature. Both parameters are fixed across stimulus pairs. In this way the model cannot account for participants that learned some stimulus pairs and guessed at others. To overcome this problem without excluding data, we augmented the reinforcement learning model with a strategy variable. Strategy z is either 0, indicating that the participant guessed, or 1, indicating that the participant learned. The strategy for each participant and each stimulus pair s is determined by stimulus-specific (s) learning state probability π_s, a proportion between 0 and 1. The higher the learning state probability of a stimulus pair, the more participants tend to learn that pair. A stimulus-specific learning state probability is modeled to capture that some stimuli might be more easily learned than others, for example because they are more salient (O'Doherty, 2004; Schutte, Slagter, Collins, Frank, & Kenemans, 2017) or require less working memory because they are more familiar (Stern, Sherman, Kirchhoff, & Hasselmo, 2001).

In a binary choice paradigm, when a participant guesses, all choices are made with probability .5; when a participant learns, each choice c is determined by the decision rule presented in Eq. (3). In this way responses originating from learning and guessing are separated and the learning and choice parameters are only estimated when a participant learns. Hereby the RLGuess model purifies the interpretation of these parameters while still providing a complete description of behavior in reinforcement learning tasks.

3.1. Model implementation

We implemented hierarchical extensions of the standard reinforcement learning (RL) model and the RLGuess model in R-3.4.3 (R Development Core Team & R Core Team, 2017). We implemented the RLGuess model with stimulus-varying learning state probabilities (RLGuess$_{\text{vary}}$) and the standard RL model (RL$_{\text{fix}}$). In order to assess the effect of stimulus-specific parameters we also implemented the RLGuess model with a fixed learning state probability across stimuli (RLGuess$_{\text{fix}}$) and an RL model with stimulus-specific inverse temperatures (RL$_{\text{vary}}$).

In a hierarchical approach individuals are assumed to be nested within a group and therefore the individual-level parameters are drawn from a group-level distribution. We chose to estimate parameters hierarchically because this improves accuracy of the individual-level parameter estimates (Efron & Morris, 1977; Lee & Webb, 2005; Shiffrin, Lee, Kim, & Wagenmakers, 2008) and therefore more sound conclusions can be drawn. In addition, we used a Bayesian framework because it yields the possibility to quantify uncertainty in the parameter estimates (Wagenmakers, Morey, & Lee, 2016).

3.1.1. Graphical model

A graphical model (Lee & Wagenmakers, 2013) of the RLGuess$_{\text{vary}}$ model is depicted in Fig. 1. In this figure, square nodes represent discrete variables and round nodes represent continuous variables. Nodes with a single border are stochastic whereas a double border indicates deterministic variables. Blank nodes indicate unobserved, that is latent, variables whereas shaded nodes indicate observed variables. Furthermore, arrows capture dependencies between nodes and encompassing plates depict independent replications of model structures.

3.1.2. Prior distributions

In the analysis, we assigned an uninformative beta prior distribution to the group-level mean of learning state probability. For the RLGuess$_{\text{vary}}$ model we sampled stimulus-specific (s) values from this distribution, $\pi_s \sim Beta(1, 1)$ (see Fig. 1), and for the RLGuess$_{\text{fix}}$ model we sampled one value, $\pi \sim Beta\,(1, 1)$. To obtain a stimulus-specific (s) strategy z per participant (p) the learning state probability was inserted into an individual Bernoulli distribution; for the RLGuess$_{\text{vary}}$ model, $z_{p,s} \sim Bernoulli(\pi_s)$, and for the RLGuess$_{\text{fix}}$ model, $z_{p,s} \sim Bernoulli(\pi)$.

We assigned beta prior distributions to the individual-level learning rate, $\eta_p \sim Beta(\mu_\eta\lambda_\eta, (1-\mu_\eta)\lambda_\eta)$, and inverse temperature, $\beta'_p \sim Beta(\mu_{\beta'}\lambda_{\beta'}, (1-\mu_{\beta'})\lambda_{\beta'})$ (Steingroever, Wetzels, & Wagenmakers, 2014). Only for the RL$_{\text{vary}}$ model the inverse temperature was made stimulus-specific, $\beta'_{p,s} \sim Beta(\mu_{\beta'}\lambda_{\beta'}, (1-\mu_{\beta'})\lambda_{\beta'})$. To estimate the learning rate and inverse temperature hierarchically, we replaced the rate and shape parameters in the beta distribution with a group-level mean and group-level precision. We assigned uniform prior distributions to the group-level means μ_η and $\mu_{\beta'}$, $U\,(.001, .999)$, as well as to the log-transformed group-level precisions $\log(\lambda_\eta)$ and $\log(\lambda_{\beta'})$, $U(\log{(2)}\,, \log{(600)})$. We set these prior distributions such that no strict restrictions to the range of individual differences were made. As the range of the inverse temperature is assumed to be [0,50] (Gershman, 2016), not [0,1] as the underlying beta distribution suggests, the following transformation was performed to the individual-level parameters: In the RLGuess$_{\text{vary}}$, RLGuess$_{\text{fix}}$ and RL$_{\text{fix}}$ model, $\beta_p = 50 * \beta'_p$, and in the RL$_{\text{vary}}$ model, $\beta_{p,s} = 50 * \beta'_{p,s}$ (Steingroever et al., 2014).

Learning State Probability

$\pi_s \sim \text{Beta}(1,1)$

Strategy

$z_{p,s} \sim \text{Bernoulli}(\pi_s)$

Choices

$$c_{p,s,t} \sim \begin{cases} \text{Bernoulli}(.5), & if\ z_{p,s} = 0 \\ \text{Bernoulli}(p_{p,s,t}), & if\ z_{p,s} = 1 \end{cases}$$

Prediction Error

$\delta_{p,s,t} \leftarrow R(c_{p,s,t}) - Q_{p,s,t}(c_{p,s,t})$

Value Estimate

$Q_{p,s,t+1}(c_{p,s,t}) \leftarrow Q_{p,s,t}(c_{p,s,t}) + \eta_p \delta_{p,s,t}$

Learning Rate

$\eta_p \sim \text{Beta}(\mu_\eta \lambda_\eta, (1 - \mu_\eta)\lambda_\eta)$

$\mu_\eta \sim \text{U}(.001, .999)$

$\log(\lambda_\eta) \sim \text{U}(\log(2), \log(600))$

Response Probability

$$p_{p,s,t} \leftarrow \frac{1}{1 + e^{-\beta_p (Q_{p,s,t}(1) - Q_{p,s,t}(2))}}$$

Inverse Temperature

$\beta_p \leftarrow 50 * \beta'_p$

$\beta'_p \sim \text{Beta}(\mu_{\beta'} \lambda_{\beta'}, (1 - \mu_{\beta'})\lambda_{\beta'})$

$\mu_{\beta'} \sim \text{U}(.001, .999)$

$\log(\lambda_{\beta'}) \sim \text{U}(\log(2), \log(600))$

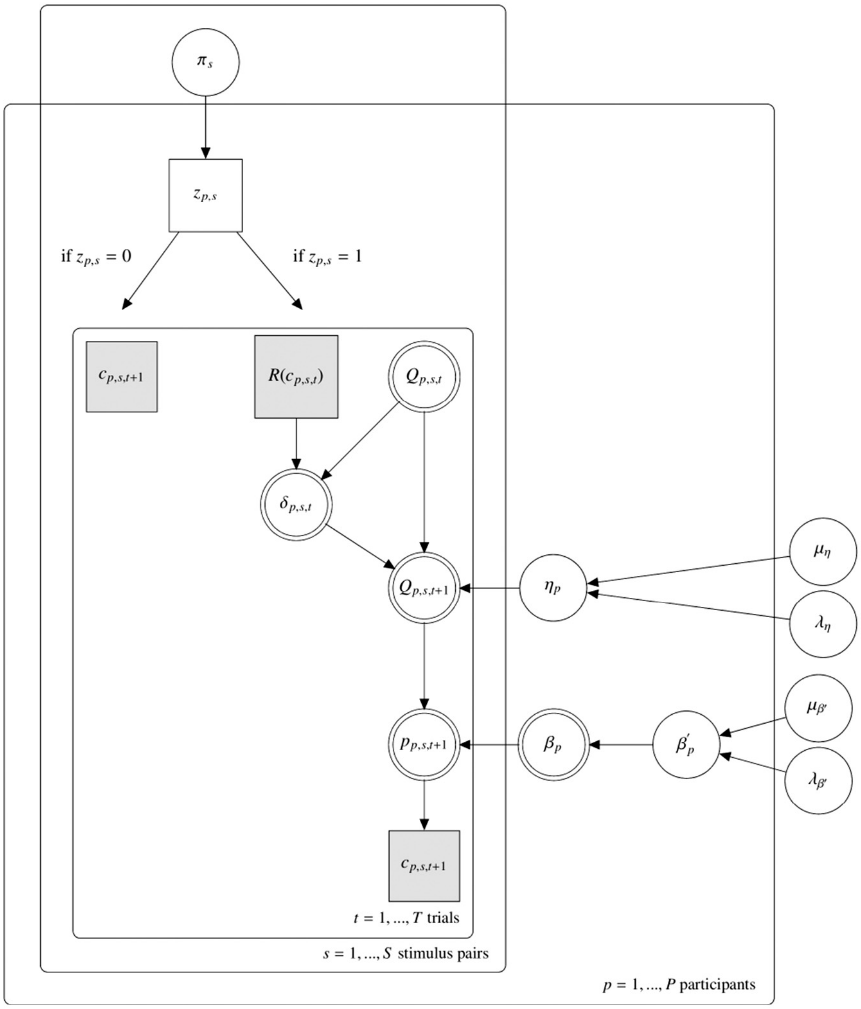

Fig. 1. Graphical representation of RLGuess$_{\text{vary}}$ model for choice $c_{p,t,s}$ of participant $p = \{1, 2, \ldots, P\}$ across trials $t = \{1, 2, \ldots, T\}$ of stimulus pair $s = \{1, 2, \ldots, S\}$. Obtained reward R is either 0 (i.e., negative outcome) or 1 (i.e., positive outcome).

3.2. Parameter estimation

We estimated the parameters of the RLGuess models and the RL models in JAGS (Plummer, 2003) by means of the R2jags package (Su & Yajima, 2015). JAGS uses Markov chain Monte Carlo (MCMC) sampling (e.g., Gamerman & Lopes, 2006; Gilks, Richardson, & Spiegelhalter, 1996) to obtain direct samples from the posterior distribution. As this posterior distribution cannot always be obtained analytically, MCMC sampling is used to characterize the distribution without knowing all of the distribution's mathematical properties (van Ravenzwaaij, Cassey, & Brown, 2018). Sampling chains are constructed that cover the entire posterior distribution. The narrower the distribution, the more certain one can be of the point estimate given by the average of the sampling chains (Lee & Wagenmakers, 2013). We initialized 3 sampling chains with 10,000 iterations each; from these 10,000 iterations half was removed as burn-in to minimize the influence of the chosen starting values. Furthermore, every 10th iteration was used to remove autocorrelation (thinning). Consequently, 3 x 500 = 1500 representative samples were obtained per parameter. Convergence of sampling chains was investigated using the R-hat statistic (Gelman & Rubin, 1992), a statistic that compares the variance between and within sampling chains; we interpreted values above 1.1 as convergence problems. When we encountered convergence problems, we reran the replication with 20,000 iterations, 10,000 samples removed as burn-in and a thinning factor of 20. The code for simulation, model implementation, model fit and analysis are provided on https://osf.io/uk684/. To illustrate the workings of MCMC sampling, an example of the learning curve of a simulated participant accompanied with the returned MCMC chains for the model parameters is presented in Fig. 1 of the Supplementary Materials.

4. Simulations

A simulation study was performed to compare model fit between the four models, to assess the parameter recovery capabilities of the RLGuess$_{\text{vary}}$ model, and to investigate parameter bias resulting from model misspecification. To do so, we simulated choices and rewards for 4 different stimuli with 24 trials each (total of 96 trials) for 38 participants in six simulation conditions. Hereafter we fit the four models to the simulated data sets thus obtained. We did this 100 times (replications).

To assess which of the four models recovered the simulated data sets best, we compared the Deviance Information Criterion (DIC; Spiegelhalter, Best, Carlin, & Van Der Linde, 2002). In this model comparison approach the deviance of the model – with lower values indicating better fit – is traded off against the number of free parameters.

To assess parameter recovery, point estimates of the group-level means of the model parameters were determined by averaging the 1500 posterior samples of that group-level mean. These group-level means were averaged over the 100 replications. In addition, for both the group-level and individual-level parameters, we computed the number of times the true parameter value lay within the 95% highest-density interval of the estimated posterior distribution of that parameter, and averaged across the 100 replications to determine the accuracy of the parameter estimates. We formally tested whether the difference between the true and estimated learning and choice parameters differed between the four models by means of Bayesian paired *t*-tests (Bååth, 2014; Kruschke, 2013). Finally, we calculated the proportion of sampled strategies (z) and rounded to integers (i.e., all proportions < 0.5 were rounded to 0, and > 0.5 to 1). We then determined the percentage of correctly classified strategies by averaging the proportion of rounded strategies that matched the simulated strategy across the 100 replications.

4.1. Simulation conditions

We simulated data in six conditions. In all conditions the group-level mean of learning rate was set to .280 and the group-level mean of inverse temperature to 6.6. First, a condition in which on average 80% of the participants learn a particular stimulus but each stimulus had a different probability to be learned (i.e., data generated given the RLGuess$_{\text{vary}}$ model); to accomplish this, we set the probability to adopt a learning strategy to $\pi = \{.65, .75, .85, .95\}$ for the four stimulus pairs. We used on average 80% congruent feedback (i.e., in 80% of the cases positive feedback following the most favorable choice and negative feedback following the least favorable choice; and in 20% of the cases negative feedback following the most favorable choice and positive feedback following the least favorable choice). This percentage is commonly used in reinforcement learning studies (e.g., Eppinger et al., 2009; Hauser, Iannaccone, Walitza, Brandeis, & Brem, 2015; van den Bos, Güroğlu, Van Den Bulk, Rombouts, & Crone, 2009). As both learning and guessing are simulated in this condition, with varying learning state probabilities across pairs, and the reward probabilities of both options (i.e., 80% for the most favorable option and 20% for the least favorable one) are dissimilar, we call this condition the Mixed (dissimilar/vary) Condition.

Second, we also simulated data in a condition in which again on average 80% of the participants learned each stimulus and each stimulus had a different probability to be learned (i.e., data generated given the RLGuess$_{\text{vary}}$ model), but it was harder to differentiate between learning and guessing responses. We accomplished this by lowering the percentage of congruent feedback from 80% to 60% in this Mixed (similar/vary) Condition. Because the difference between the percentages of rewards for both response options is smaller (i.e., 60% for the most favorable option and 40% for the least favorable one), response patterns that arise from learning are more similar to guessing responses than in the preceding condition.

Third, we simulated data in which on average 80% of the participants learn a particular stimulus but this probability was fixed across pairs (i.e., data generated given the RLGuess$_{\text{fix}}$ model); we accomplished this by setting the probability to learn each pair to $\pi = .8$. Again 80% congruent feedback was used. This is called the Mixed (dissimilar/fix) Condition.

Fourth, in the Mixed (similar/fix) Condition, on average 80% of the participants learn a particular stimulus and the probability was fixed across pairs (i.e., data generated given RLGuess$_{\text{fix}}$ model), but now with 60% congruent feedback.

Fifth, in the Learning (vary) Condition, all participants learn all stimulus pairs (i.e., $\pi = 1$). We varied the inverse temperature parameter across stimulus pairs (i.e., data generated given RL$_{\text{vary}}$ model). Again the percentage of congruent feedback was 80%.

In the final Learning (fix) Condition, we simulated data with the standard RL$_{\text{fix}}$ model in which all participants learn all pairs, a fixed inverse temperature across pairs and 80% congruent feedback.

We did not include Learning Conditions with 60% congruent feedback because we were mainly interested in the effect of feedback congruency on the strategy recovery capabilities of the RLGuess model when the data contain guessing responses. In each condition 100 replications were ran which resulted in a total of 6 x 100 = 600 simulated data sets.

4.2. Results

4.2.1. Model validation

Model selection by means of the DIC showed that in general the data generating model fitted the data best, although not in every replication (see Table 1).

4.2.2. Parameter estimates

We compared the true and estimated values of learning state probability π, learning rate η and inverse temperature β. A more thorough summary of the simulation results is provided in Table 1 of the Supplementary Materials. The posterior distributions of the group-level means of learning state probability, learning rate and inverse temperature are depicted in Fig. 2 and the parameter recovery capabilities of the four models are provided in Table 2. Both are further discussed below.

RLGuess$_{\text{vary}}$. The RLGuess$_{\text{vary}}$ model recovered all parameters adequately in all conditions, except in the Learning Conditions. In both Learning Conditions, the 95% highest-density interval did not contain the true learning state probability (0.0%) because this value was fixed at the bound of the beta prior distribution (i.e., 1). Furthermore, inspection of the distributions in Fig. 2 shows that the inverse temperature was slightly underestimated in the Learning (vary) Condition.

RLGuess$_{\text{fix}}$. The same applies to the RLGuess$_{\text{fix}}$ model: All parameters were adequately recovered in all conditions, except the learning state probabilities in both Learning Conditions and the inverse temperature in the Learning (vary) Condition.

RL$_{\text{vary}}$. The RL$_{\text{vary}}$ model inadequately recovered parameters in all conditions, except in the Learning (vary) Condition. Parameter estimation bias was most pronounced in the Learning (fix) Condition: Learning rate was underestimated. Inspection of the distributions in Fig. 2 shows that the estimated learning rates were generally lower than the true value. Estimated inverse temperatures on the other hand were generally higher than the true value and seemed to depend on the percentage of congruent feedback. Additional simulations in which we used 70% and 95% congruent feedback verified this pattern (see Fig. A.1): The higher the percentage of congruent feedback, the larger the overestimation of the inverse temperature.

RL$_{\text{fix}}$. The RL$_{\text{fix}}$ model inadequately recovered parameters in all conditions, except in the Learning (fix) Condition. In all other conditions inverse temperatures were underestimated. Inspection of the distributions in Fig. 2 shows the opposite pattern for the RL$_{\text{fix}}$ model compared to the RL$_{\text{vary}}$ model: Learning rates were generally higher than the true value and seemed to depend on the percentage of congruent feedback whereas estimated inverse temperatures were generally lower than the true value. Again, this pattern was verified in additional simulations (see Fig. A.1).

Taken together, these simulation results thus indicate that the RLGuess model outperforms standard reinforcement learning models when participants guess: Fit is enhanced and parameters are unbiased. Furthermore, model misspecification results in biased estimates of both learning rate and inverse temperature. In a standard model with fixed inverse temperature across pairs, learning rate is overestimated and inverse temperature underestimated. In a model with varying inverse temperature learning rate is underestimated and inverse temperature overestimated.

Table 1
Proportion of simulated data sets for which each model had the lowest DIC value per condition.

		Data generating model					
		$\text{RLGuess}_{\text{vary}}$		$\text{RLGuess}_{\text{fix}}$		RL_{vary}	RL_{fix}
		Mixed (dissimilar/vary)	Mixed (similar/vary)	Mixed (dissimilar/fix)	Mixed (similar/fix)	Learning (vary)	Learning (fix)
Data recovering model	$\text{RLGuess}_{\text{vary}}$	**56**	**74**	45[a]	37[a]	0	16[a]
	$\text{RLGuess}_{\text{fix}}$	44[a]	26[a]	**55**	**63**	0	28[a]
	RL_{vary}	0	0	0	0	**100**	0
	RL_{fix}	0	0	0	0	0	**56**

Note. In bold the model for which most data sets yielded the lowest DIC value.

[a] In case the best fitting model was not the data generating model, the difference between the DIC value of those models was often <10, that is 40/44, 23/26, 39/45, 29/37 and 40/44 for the five columns in Table 1 respectively.

Table 2
Percentage of cases where the models correctly classified the strategy (z) and where the 95% highest-density interval contained the true group-level mean of learning state probability (π), learning rate (η) and inverse temperature (β).

			Data generating model					
			$\text{RLGuess}_{\text{vary}}$		$\text{RLGuess}_{\text{fix}}$		RL_{vary}	RL_{fix}
			Mixed (dissimilar/vary)	Mixed (similar/vary)	Mixed (dissimilar/fix)	Mixed (similar/fix)	Learning (vary)	Learning (fix)
	Parameter							
Data recovering model	$\text{RLGuess}_{\text{vary}}$	z	94.3%	90.8%	94.2%	89.8%	86.4%	99.7%
		π[a]	94%	93%	96%	97%	0%	0%
		η	93%	94%	93%	94%	91%	96%
		β	94%	94%	94%	96%	89%	98%
	$\text{RLGuess}_{\text{fix}}$	z	93.9%	90.1%	94.3%	90.0%	86.8%	99.9%
		π	93%	97%	96%	96%	0%	0%
		η	93%	96%	92%	96%	90%	95%
		β	92%	94%	94%	96%	88%	98%
	RL_{vary}	η	60%	79%	70%	70%	90%	24%
		β	76%	84%	80%	86%	94%	30%
	RL_{fix}	η	88%	91%	88%	91%	79%	96%
		β	31%	32%	42%	45%	9%	98%

[a] For the $\text{RLGuess}_{\text{vary}}$ model the percentage of intervals containing the group-level mean of learning state probability was determined by averaging all samples over the four stimulus pairs, then determining the 95% highest-density interval and whether the true group-level mean fell within this interval, and finally averaging over the 100 replications.

5. Application to real data

5.1. Data

The four models were fit to reinforcement learning data collected by Kramer (2017). A total of 38 participants performed on a reinforcement learning task in which the correct spelling of a pseudo word needed to be learned from feedback. The pseudo word pairs were homophones (i.e., they sound the same; in Dutch). In this task (see Verburg, Snellings, Zeguers, & Huizenga, 2018) four different stimulus pairs (see Table 3) were learned in parallel with 24 trials each. Participants either gained nothing (0) or gained +10 points (see Fig. 3 for an example trial). On average, the percentage of congruent feedback was 65%, that is, in 65% of the cases positive feedback after the most favorable choice and negative feedback after the least favorable choice; and in 35% of the cases negative feedback after the most favorable choice and positive feedback after the least favorable choice. The data contained 0.7% missing values as a result of late responses; these responses were omitted from the analysis. The data are available at https://osf.io/uk684/.

5.2. Results

The $\text{RLGuess}_{\text{vary}}$ model (DIC = 3819.45) described the data better than the $\text{RLGuess}_{\text{fix}}$ model (DIC = 3829.83). It also fitted better than the RL_{fix} model (DIC = 3890.16) and the RL_{vary} model (DIC = 3938.04), suggesting that participants guessed at some stimulus pairs and that some pairs were more easily learned than others. This was supported by Bayesian t-tests on the estimated learning state probabilities (π): All probabilities differed from each other. On average each stimulus pair was learned by 78% to 94% of the participants (see Table 3).

Apart from general patterns in participants' choice behavior – formalized by the group-level means of learning state probability ($\mu_\pi = .87$), learning rate ($\mu_\eta = .23$) and inverse temperature ($\mu_\beta = 6.1$) – the RLGuess model is able to identify individual differences in the learning and guessing process. To illustrate the information that can be obtained about individual participants, the observed and predicted choices of four participants are shown in Fig. 4.

Both the $\text{RLGuess}_{\text{vary}}$ model and the $\text{RLGuess}_{\text{fix}}$ model indicate that participants 106 and 203 learned all stimulus pairs, even though learning was less clear for participant 203. The models suggest that participant 115 learned the last three pairs whereas (s)he most likely guessed at the first pair. Lastly, the models indicate that participant 120 guessed at the first and the third pair and learned the other two pairs.

If a participant learns all pairs (PP 106 and 203), all four models predict roughly the same choice pattern for that participant; when this learning strategy is clear (PP 106) the estimates of that participants' learning rate and inverse temperature are also very similar for the four models. However, when participants seem to guess at some pairs (PP 115 and 120), the predictive accuracy of the RL_{fix} and RL_{vary} model decreases, especially for guessed pairs, compared to the model predictions of the $\text{RLGuess}_{\text{vary}}$ and

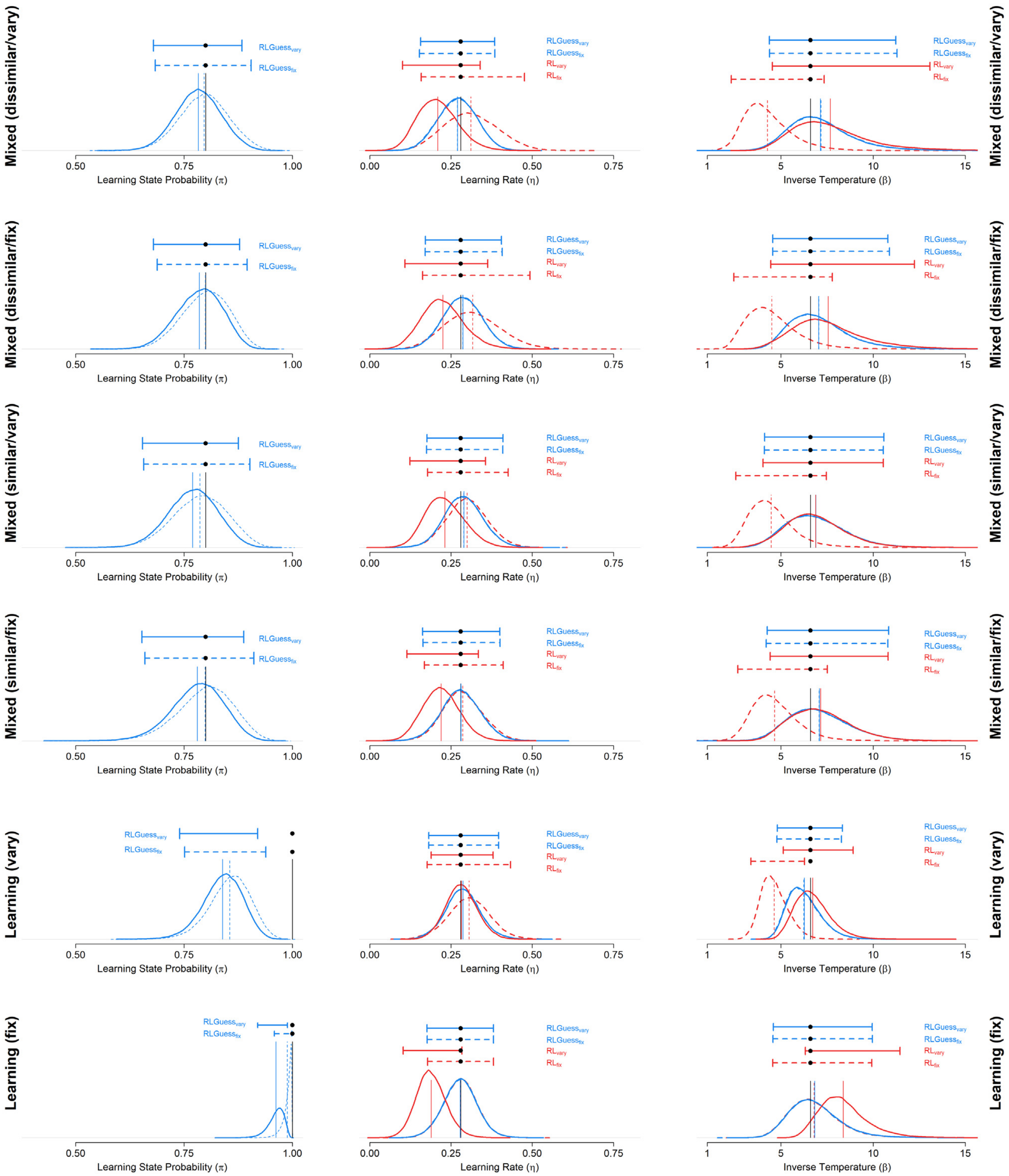

Fig. 2. The posterior distributions of the group-level means of learning state probability (π; left column), learning rate (η; middle column) and inverse temperature (β; right column) in the six simulated data conditions (six rows). In each panel the blue solid curve represents the posterior distribution of the group-level mean by the RLGuess$_{vary}$ model, the striped blue curve by the RLGuess$_{fix}$ model, the solid red curve by the RL$_{vary}$ model, and the striped red curve by the RL$_{fix}$ model. Vertical solid lines represent the true (black) and estimated (RLGuess$_{vary}$: blue, solid; RLGuess$_{fix}$: blue, striped; RL$_{vary}$: red, solid; RL$_{fix}$: red, striped) means of that distribution. Horizontal line segments on top of each panel indicate the 95% highest-density interval of the posterior distributions estimated by the four models; the black dot inside the interval indicates the true mean.

RLGuess$_{fix}$ model. Also the estimates of the learning and choice parameters (for PP 115 both learning rate and inverse temperature and for PP 120 mainly inverse temperature) of the RLGuess models and the RL models deviate.

6. Discussion

In this paper we proposed the RLGuess model — a reinforcement learning model augmented with a strategy variable,

Table 3
The means of the posterior distributions of the learning state probabilities (π), learning rates (η) and inverse temperatures (β) of the same participants displayed in Fig. 4 estimated by the four models.

		Model									
		RLGuess$_{\text{vary}}$				RLGuess$_{\text{fix}}$	RL$_{\text{vary}}$				RL$_{\text{fix}}$
	Stimulus Pair	*Kreip-Krijp*	*Preil-Prijl*	*Spreik-Sprijk*	*Strein-Strijn*		*Kreip-Krijp*	*Preil-Prijl*	*Spreik-Sprijk*	*Strein-Strijn*	
	Parameter										
	π[a]	.78	.86	.90	.94	.89	–				–
PP											
106	η	.18				.18	.17				.18
	β	8.12				8.28	10.7	5.56	10.6	9.56	8.04
203	η	.54				.51	.35				.48
	β	1.91				1.85	3.34	4.71	2.76	3.31	1.75
115	η	.46				.42	.11				.40
	β	1.98				2.06	3.76	7.42	6.15	3.74	1.96
120	η	.29				.29	.29				.26
	β	9.60				9.65	.92	12.2	1.63	8.54	2.99

[a] This learning state probability can be interpreted as the proportion of learning (per stimulus pair).

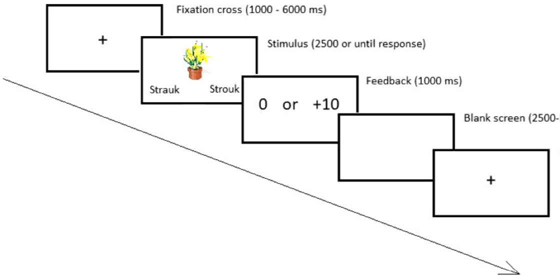

Fig. 3. Example trial of the reinforcement learning task administered by Kramer (2017).

enabling researchers to model that participants learn some stimulus pairs while they guess at others. In simulations we showed that, when the data contain guessing responses, the RLGuess model fits data better than standard reinforcement learning models and adequately recovers the learning and choice parameters. We also demonstrated the implications of using a standard reinforcement learning model when participants guess. In a standard model with fixed inverse temperature across pairs, their learning rate is overestimated and their inverse temperature underestimated, suggesting that participants make faster adaptations based on prediction errors and focus less on differences between the values of options than they actually do. In a model with varying inverse temperatures across pairs, their learning rate is underestimated and their inverse temperature overestimated, suggesting slower adaptations and more focus on differences between values. Therefore we argue that standard reinforcement learning models without considering guessing should only be applied when there is good reason to believe that guessing does not occur.

Other modeling approaches have previously been adopted to reduce the impact of choices unrelated to the learning process. Some models take into account lapses in attention by adding a "lapse rate" parameter to the softmax rule (see e.g., Economides, Kurth-Nelson, Lübbert, Guitart-Masip, & Dolan, 2015). Similarly, other models allow for occasional random choices by using an epsilon-greedy decision rule (Sutton & Barto, 2018; see e.g., Daw, O'Doherty, Dayan, Seymour, & Dolan, 2006; Speekenbrink & Konstantinidis, 2015). However, the lapse parameter and the epsilon in these previous approaches are not stimulus-specific, as is the case in the RLGuess model.

Applications

The RLGuess model could be used to clarify differences in choice behavior in various domains. In the developmental field, differences in the learning and guessing process could be related to different stages of development, both in child and adolescent samples (e.g., van den Bos et al., 2009; Verburg et al., 2018) as well as in samples consisting of seniors (e.g., Frank & Kong, 2008; Simon, Howard, & Howard, 2010). For example, the probability of guessing responses might decrease during childhood and adolescence while it may increase again in seniors. In the clinical field, clinical groups such as Parkinson patients (Frank et al., 2004) could be compared to their healthy counterparts. More broadly, the RLGuess model could be used to test the effect of experimental manipulations such as set size (Collins & Frank, 2012), feedback valence (Eppinger & Kray, 2011; Palminteri et al., 2012; Palminteri, Khamassi, Joffily, & Coricelli, 2015), feedback validity (Eppinger, Kray, Mock, & Mecklinger, 2008; Nieuwenhuis et al., 2002), or arousal (Lighthall, Gorlick, Schoeke, Frank, & Mather, 2013; Raio, Hartley, Orederu, Li, & Phelps, 2017) on the learning and guessing process.

Besides, modeling learning and guessing separately can strengthen functional magnetic resonance imaging (fMRI) results by removing guessing responses from the main analysis. Traditionally, prediction errors are correlated with blood-oxygen level dependent (BOLD) responses in the brain (e.g., O'Doherty et al., 2004; Pessiglione et al., 2006; van den Bos, Cohen, Kahnt, & Crone, 2012). When participants guess, however, choices are made randomly, they either do not compute prediction errors or do not use them to update their value estimates as assumed in reinforcement learning models. If these responses are included this adds noise to the main analysis and thus makes it more difficult to find prediction error related activity.

Future Directions/Extensions

In the RLGuess model, strategies are fixed across all trials of a stimulus pair. In other words, we assume that a participant either learns a stimulus from the first trial onwards or guesses. There is no room for switching between the two strategies during the task. It might be, however, that participants start off by guessing, but move on to a learning strategy once they have learned one of the other stimulus pairs, and, for example, working memory capacity is available (Collins & Frank, 2012). Such a process can be incorporated in the RLGuess model by modeling the onset of learning (Gallistel, Fairhurst, & Balsam, 2004) or by including a dynamic (see Busemeyer & Stout, 2002) learning

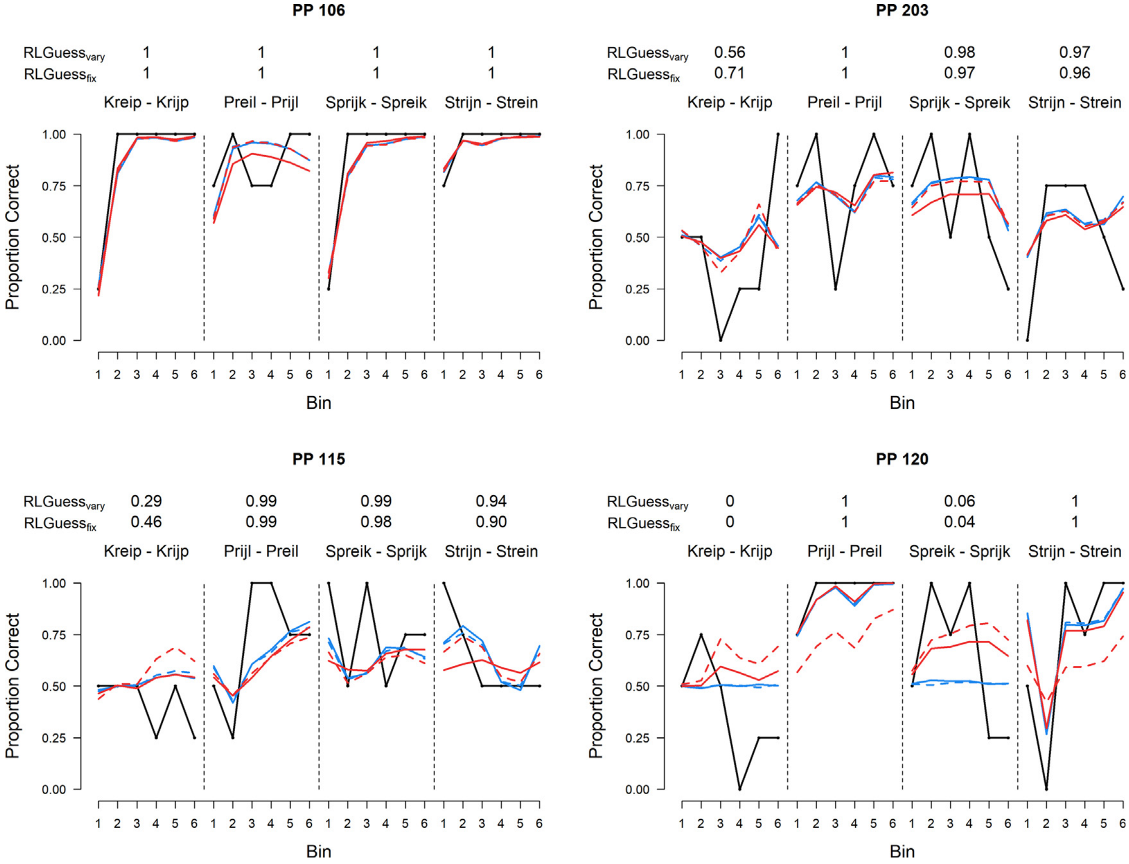

Fig. 4. Observed (black solid line) and predicted learning curve by RLGuess$_{vary}$ (blue solid), RLGuess$_{fix}$ (blue striped), RL$_{vary}$ (red solid) and RL$_{fix}$ model (red striped) of participants 106, 203, 115 and 120. On the *y*-axis the proportion of correct responses (i.e., choices for the option that yielded the highest reward); on the *x*-axis for all four stimulus pairs the 24 trials divided into 6 bins of 4 trials (96 trials in total). Note that we ordered the data by stimulus pair; in the experiment pairs were presented in a randomized order. Each vertical dotted line represents a new pair. The possible spellings of the pseudo words are presented above each pair; the first pseudo word represents the correct spelling. Above these spellings the portion of sampled strategies by the RLGuess$_{vary}$ (top) and RLGuess$_{fix}$ (bottom) model are denoted in which 0 = Guessing and 1 = Learning.

state probability. Second, the value updating mechanism used in standard reinforcement learning models assumes a monotonic learning process. A more flexible learning and guessing process could be incorporated by determining the probability of both strategies at each choice of a stimulus pair (Lee, Zhang, Munro, & Steyvers, 2011). Third, we saw in the empirical application that when participants first choose one of the options and during the task switch to the other option (see PP120 stimulus 1 in Fig. 4), these responses are classified as guessing. One could model these sudden changes in choice behavior by incorporating uncertainty about the unchosen option in the model; for example, by adding an "uncertainty bonus" to the softmax decision rule (Daw et al., 2006; Speekenbrink & Konstantinidis, 2015). Most likely, this improves model fit but also requires more free parameters.

Another possible extension is to estimate a learning rate for each stimulus pair separately. This would be meaningful if, for example, stimulus pairs differ in the percentage of congruent feedback and therefore prediction errors are more informative for some of the pairs, those with high feedback congruency, than for other pairs, those with low feedback congruency (e.g., Decker, Lourenco, Doll, & Hartley, 2015; Doll et al., 2009; Hämmerer et al., 2011). One could also decide to update not only the value estimate of the chosen response option, but also of the unchosen one. This adjustment would be suitable when, for example, deterministic feedback is used; in that case feedback also provides information on the unchosen option (e.g., Peters, Braams, Raijmakers, Koolschijn, & Crone, 2014; van der Schaaf, Warmerdam, Crone, & Cools, 2011; Van Leijenhorst, Crone, & Bunge, 2006). Other possible extensions are the inclusion of different types of learning strategies (e.g., Bartlema, Lee, Wetzels, & Vanpaemel, 2014), separate learning rates for positive and negative prediction errors (Daw, Kakade, & Dayan, 2002; Frank, Doll, Oas-Terpstra, & Moreno, 2009; Frank, Moustafa, Haughey, Curran, & Hutchison, 2007; Gershman, 2015; Niv, Edlund, Dayan, & O'Doherty, 2012) or inclusion of the propensity to switch between options independent of rewards (Christakou et al., 2013; Gershman, 2016; Gershman, Pesaran, & Daw, 2009).

7. Conclusion

To conclude, our results suggest guessing cannot be ignored in reinforcement learning tasks. Therefore, we put forward a simple and easy-to-apply model that can accurately describe a reinforcement learning process while considering participants might guess.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

CRediT authorship contribution statement

Jessica Vera Schaaf: Conceptualization, Formal Analysis, Validation, Visualization, Writing - original draft. **Marieke Jepma:**

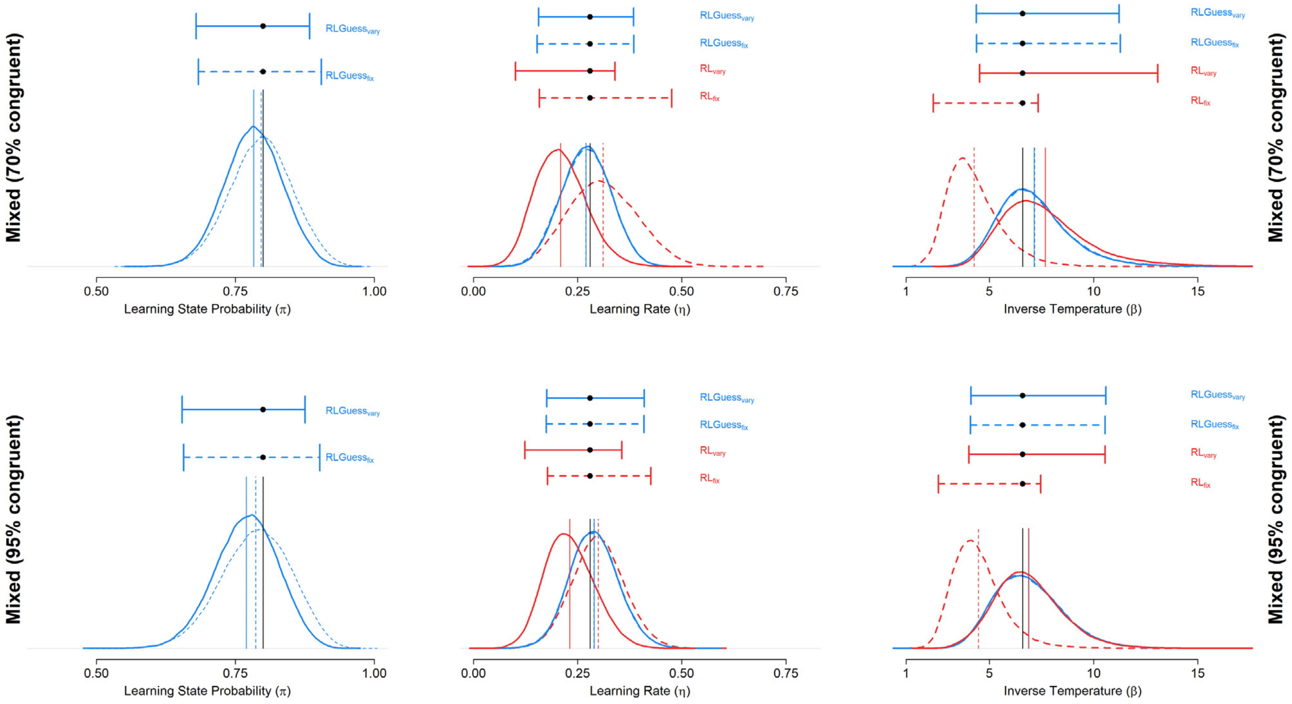

Fig. A.1. The posterior distributions of the group-level means of learning state probability (left column), learning rate (middle column) and inverse temperature (right column) in additional simulations where 80% of the participants learned each stimulus, and 70% (top row) and 95% (bottom row) congruent feedback (i.e., positive feedback following the most favorable choice and negative feedback following the least favorable choice) was used. In each panel the posterior distributions of the group-level means estimated by the RLGuess$_{\text{vary}}$ (blue solid), RLGuess$_{\text{fix}}$ (blue striped), RL$_{\text{vary}}$ (red solid) and RL$_{\text{fix}}$ (red striped) are depicted. The vertical solid lines represent the true (black) and estimated (RLGuess$_{\text{vary}}$: blue solid; RLGuess$_{\text{fix}}$: blue striped, RL$_{\text{vary}}$: red solid; RL$_{\text{fix}}$: red striped) means of that distribution. The horizontal line segment on top of each panel indicates the 95% highest-density interval; the black dot inside the interval indicates the true mean.

Conceptualization, Validation, Writing - review & editing, Supervision. **Ingmar Visser:** Conceptualization, Validation, Writing - review & editing, Supervision. **Hilde Maria Huizenga:** Conceptualization, Funding acquisition, Validation, Writing - review & editing, Supervision.

Acknowledgments

We would like to thank Johnny van Doorn and Maarten Speekenbrink for their critical comments on a preliminary version of the manuscript. We also thank Anne-Wil Kramer who shared her data with us.

Appendix. Results additional simulations

See Fig. A.1.

Appendix B. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jmp.2019.102276.

References

Bartlema, A., Lee, M., Wetzels, R., & Vanpaemel, W. (2014). A Bayesian hierarchical mixture approach to individual differences: Case studies in selective attention and representation in category learning. *Journal of Mathematical Psychology*, *59*, 132–150. http://dx.doi.org/10.1016/J.JMP.2013.12.002.

Bååth, R. (2014). Bayesian First aid. Retrieved November 2, 2018, from http://www.sumsar.net/blog/2014/01/bayesian-first-aid/.

Busemeyer, J. R., & Stout, J. C. (2002). A contribution of cognitive decision models to clinical assessment: Decomposing performance on the Bechara Gambling Task. *Psychological Assessment*, *14*(3), 253.

Christakou, A., Gershman, S. J., Niv, Y., Simmons, A., Brammer, M., & Rubia, K. (2013). Neural and psychological maturation of decision-making in adolescence and young adulthood. *Journal of Cognitive Neuroscience*, *25*(11), 1807–1823. http://dx.doi.org/10.1162/jocn_a_00447.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates.

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. http://dx.doi.org/10.1111/j.1460-9568.2011.07980.x.

Daw, N. D., Kakade, S., & Dayan, P. (2002). Opponent interactions between serotonin and dopamine. *Neural Networks*, *15*(4), 603–616. http://dx.doi.org/10.1016/S0893-6080(02)00052-7.

Daw, N. D., O'Doherty, J. P., Dayan, P., Seymour, B., & Dolan, R. J. (2006). Cortical substrates for exploratory decisions in humans. *Nature*, *441*(7095), 876–879. http://dx.doi.org/10.1038/nature04766.

Decker, J. H., Lourenco, F. S., Doll, B. B., & Hartley, C. A. (2015). Experiential reward learning outweighs instruction prior to adulthood. *Cognitive, Affective and Behavioral Neuroscience*, *15*(2), 310–320. http://dx.doi.org/10.3758/s13415-014-0332-5.

Decker, J. H., Otto, A. R., Daw, N. D., & Hartley, C. A. (2016). From creatures of habit to goal-directed learners: Tracking the developmental emergence of model-based reinforcement learning. *Psychological Science*, *27*(6), 848–858. http://dx.doi.org/10.1177/0956797616639301.

Doll, B. B., Jacobs, W. J., Sanfey, A. G., & Frank, M. J. (2009). Instructional control of reinforcement learning: A behavioral and neurocomputational investigation. *Brain Research*, *1299*, 74–94. http://dx.doi.org/10.1016/J.BRAINRES.2009.07.007.

Economides, M., Kurth-Nelson, Z., Lübbert, A., Guitart-Masip, M., & Dolan, R. J. (2015). Model-based reasoning in humans becomes automatic with training. *PLoS Computational Biology*, *11*(9), e1004463. http://dx.doi.org/10.1371/journal.pcbi.1004463.

Efron, B., & Morris, C. (1977). Stein's paradox in statistics. *Scientific American*, *236*(5), 119–127. http://dx.doi.org/10.1038/scientificamerican0577-119.

Eppinger, B., & Kray, J. (2011). To choose or to avoid: Age differences in learning from positive and negative feedback. *Journal of Cognitive Neuroscience*, *23*(1), 41–52.

Eppinger, B., Kray, J., Mock, B., & Mecklinger, A. (2008). Better or worse than expected? Aging, learning, and the ERN. *Neuropsychologia*, *46*(2), 521–539.

Eppinger, B., Mock, B., & Kray, J. (2009). Developmental differences in learning and error processing: Evidence from ERPs. *Psychophysiology*, *46*(5), 1043–1053. http://dx.doi.org/10.1111/j.1469-8986.2009.00838.x.

Frank, M. J., Doll, B. B., Oas-Terpstra, J., & Moreno, F. (2009). Prefrontal and striatal dopaminergic genes predict individual differences in exploration and exploitation. *Nature Neuroscience*, *12*(8), 1062–1068. http://dx.doi.org/10.1038/nn.2342.

Frank, M. J., & Kong, L. (2008). Learning to avoid in older age. *Psychology and Aging*, *23*(2), 392–398. http://dx.doi.org/10.1037/0882-7974.23.2.392.

Frank, M. J., Moustafa, A. A., Haughey, H. M., Curran, T., & Hutchison, K. E. (2007). Genetic triple dissociation reveals multiple roles for dopamine in reinforcement learning. *Proceedings of the National Academy of Sciences*, *104*(41), 16311–16316. http://dx.doi.org/10.1073/pnas.0706111104.

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

Gallistel, C. R., Fairhurst, S., & Balsam, P. (2004). The learning curve: Implications of a quantitative analysis. *Proceedings of the National Academy of Sciences*, *101*(36), 13124–13131. http://dx.doi.org/10.1073/pnas.0404965101.

Gamerman, D., & Lopes, H. F. (2006). *Markov chain Monte Carlo: Stochastic simulation for Bayesian inference*. Chapman and Hall/CRC Press, http://dx.doi.org/10.1002/1521-3773(20010316)40:6{\T1\textless}9823::AID-ANIE9823{\T1\textgreater}3.3.CO;2-C.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, *7*(4), 457–472. http://dx.doi.org/10.1214/ss/1177011136.

Gershman, S. J. (2015). Do learning rates adapt to the distribution of rewards? *Psychonomic Bulletin and Review*, *22*(5), 1320–1327. http://dx.doi.org/10.3758/s13423-014-0790-3.

Gershman, S. J. (2016). Empirical priors for reinforcement learning models. *Journal of Mathematical Psychology*, *71*, 1–6. http://dx.doi.org/10.1016/j.jmp.2016.01.006.

Gershman, S. J., Pesaran, B., & Daw, N. D. (2009). Human reinforcement learning subdivides structured action spaces by learning effector-specific values. *Journal of Neuroscience*, *29*(43), 13524–13531. http://dx.doi.org/10.1523/JNEUROSCI.2469-09.2009.

Gilks, W. R., Richardson, S., & Spiegelhalter, D. J. (1996). Introducing Markov chain Monte Carlo. In *Markov chain Monte Carlo in practice*. CRC Press.

Hämmerer, D., Li, S.-C., Müller, V., & Lindenberger, U. (2011). Life span differences in electrophysiological correlates of monitoring gains and losses during probabilistic reinforcement learning. *Journal of Cognitive Neuroscience*, *23*(3), 579–592. http://dx.doi.org/10.1162/jocn.2010.21475.

Hauser, T. U., Iannaccone, R., Walitza, S., Brandeis, D., & Brem, S. (2015). Cognitive flexibility in adolescence: Neural and behavioral mechanisms of reward prediction error processing in adaptive decision making during development. *Neuroimage*, *104*, 347–354. http://dx.doi.org/10.1016/J.NEUROIMAGE.2014.09.018.

Kim, H., Shimojo, S., & O'Doherty, J. P. (2006). Is avoiding an aversive outcome rewarding? Neural substrates of avoidance learning in the human brain. *PLoS Biology*, *4*(8), e233. http://dx.doi.org/10.1371/journal.pbio.0040233.

Kramer, A. (2017). *Information or motivation: An FMRI investigation into the effects of positive versus negative feedback*. University of Amsterdam.

Kruschke, J. K. (2013). Bayesian estimation supersedes the t test, 142(2), 573–603. http://dx.doi.org/10.1037/a0029146.

Lee, M. D., & Wagenmakers, E. J. (2013). *Bayesian cognitive modeling: A practical course*. Cambridge: Cambridge University Press, http://dx.doi.org/10.1017/CBO9781139087759.

Lee, M. D., & Webb, M. R. (2005). Modeling individual differences in cognition. *Psychonomic Bulletin and Review*, *12*(4), 605–621. http://dx.doi.org/10.3758/BF03196751.

Lee, M. D., Zhang, S., Munro, M., & Steyvers, M. (2011). Psychological models of human and optimal performance in bandit problems. *Cognitive Systems Research*, *12*(2), 164–174. http://dx.doi.org/10.1016/J.COGSYS.2010.07.007.

Lighthall, N. R., Gorlick, M. A., Schoeke, A., Frank, M. J., & Mather, M. (2013). Stress modulates reinforcement learning in younger and older adults. *Psychology and Aging*, *28*(1), 35.

Luce, R. D. (1959). Individual choice behavior. *Econometrica*, http://dx.doi.org/10.2307/1911299.

Nieuwenhuis, S., Ridderinkhof, K. R., Talsma, D., Coles, M. G., Holroyd, C. B., Kok, A., et al. (2002). A computational account of altered error processing in older age: Dopamine and the error-related negativity. *Cognitive, Affective, & Behavioral Neuroscience*, *2*(1), 19–36.

Niv, Y., Daniel, R., Geana, A., Gershman, S. J., Leong, Y. C., Radulescu, A., et al. (2015). Reinforcement learning in multidimensional environments relies on attention mechanisms. *Journal of Neuroscience*, *35*(21), 8145–8157. http://dx.doi.org/10.1523/JNEUROSCI.2978-14.2015.

Niv, Y., Edlund, J. A., Dayan, P., & O'Doherty, J. P. (2012). Neural prediction errors reveal a risk-sensitive reinforcement-learning process in the human brain. *Journal of Neuroscience*, *32*(2), 551–562. http://dx.doi.org/10.1523/JNEUROSCI.5498-10.2012.

O'Doherty, J. P. (2004). Reward representations and reward-related learning in the human brain: insights from neuroimaging. *Current Opinion in Neurobiology*, *14*(6), 769–776.

O'Doherty, J., Dayan, P., Schultz, J., Deichmann, R., Friston, K., & Dolan, R. J. (2004). Dissociable roles of ventral and dorsal striatum in instrumental conditioning. *Science*, *304*, 452–454.

Palminteri, S., Justo, D., Jauffret, C., Pavlicek, B., Dauta, A., Delmaire, C., . . ., & Pessiglione, M. (2012). Critical roles for anterior insula and dorsal striatum in punishment-based avoidance learning. *Neuron*, *76*(5), 998–1009.

Palminteri, S., Khamassi, M., Joffily, M., & Coricelli, G. (2015). Contextual modulation of value signals in reward and punishment learning. *Nature Communications*, *6*(8096).

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. http://dx.doi.org/10.1038/nature05051.

Peters, S., Braams, B. R., Raijmakers, M. E. J., Koolschijn, P. C. M. P., & Crone, E. A. (2014). The neural coding of feedback learning across child and adolescent development. *Journal of Cognitive Neuroscience*, *26*(8), 1705–1720. http://dx.doi.org/10.1162/jocn_a_00594.

Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In *Proceedings of the 3rd international workshop on distributed statistical computing*. http://dx.doi.org/10.1.1.13.3406.

R Development Core Team, & R Core Team (2017). *R: A Language and Environment for Statistical Computing*. http://dx.doi.org/10.1016/j.jssas.2015.06.002.

Raio, C. M., Hartley, C. A., Orederu, T. A., Li, J., & Phelps, E. A. (2017). Stress attenuates the flexible updating of aversive value. *Proceedings of the National Academy of Sciences*, *114*(42), 11241–11246.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black, & W. F. Prokasy (Eds.), *Classical conditioning II: Current research and theory* (pp. 64–99). New York: Appleton-Century-Crofts.

Schutte, I., Slagter, H. A., Collins, A. G., Frank, M. J., & Kenemans, J. L. (2017). Stimulus discriminability may bias value-based probabilistic learning. *PLoS One*, *12*(5), e0176205.

Shiffrin, R. M., Lee, M. D., Kim, W., & Wagenmakers, E. J. (2008). A survey of model evaluation approaches with a tutorial on hierarchical Bayesian methods. *Cognitive Science*, *32*, 1248. http://dx.doi.org/10.1080/03640210802414826.

Simon, J., Howard, J., & Howard, D. (2010). Adult age differences in learning from positive and negative probabilistic feedback. *Neuropsychology*, *24*(4), 534–541. http://dx.doi.org/10.1037/a0018652.

Speekenbrink, M., & Konstantinidis, E. (2015). Uncertainty and exploration in a restless bandit problem. *Topics in Cognitive Science*, *7*, 351–367. http://dx.doi.org/10.1111/tops.12145.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Van Der Linde, A. (2002). Bayesian Measures of model complexity and fit. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, *64*(4), 583–616. http://dx.doi.org/10.1111/1467-9868.00353.

Steingroever, H., Wetzels, R., & Wagenmakers, E. J. (2014). Absolute performance of reinforcement-learning models for the Iowa Gambling Task. *Decision*, *3*, 115–131. http://dx.doi.org/10.1037/dec0000005.

Stern, C. E., Sherman, S. J., Kirchhoff, B. A., & Hasselmo, M. E. (2001). Medial temporal and prefrontal contributions to working memory tasks with novel and familiar stimuli. *Hippocampus*, *11*(4), 337–346.

Su, Y.-S., & Yajima, M. (2015). *R2jags: Using R to run JAGS. R packages*. http://cran.r-project.org/package=R2jags.

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction*. http://dx.doi.org/10.1016/S1364-6613(99)01331-5.

van den Bos, W., Cohen, M. X., Kahnt, T., & Crone, E. A. (2012). Striatum–medial prefrontal cortex connectivity predicts developmental changes in reinforcement learning. *Cerebral Cortex*, *22*(6), 1247–1255. http://dx.doi.org/10.1093/cercor/bhr198.

van den Bos, W., Güroğlu, B., Van Den Bulk, B. G., Rombouts, S. A. R., & Crone, E. A. (2009). Better than expected or as bad as you thought? The neurocognitive development of probabilistic feedback processing. *Frontiers in Human Neuroscience*, *3*(52), http://dx.doi.org/10.3389/neuro.09.052.2009.

van der Schaaf, M. E., Warmerdam, E., Crone, E. A., & Cools, R. (2011). Distinct linear and non-linear trajectories of reward and punishment reversal learning during development: Relevance for dopamine's role in adolescent decision making. *Developmental Cognitive Neuroscience*, *1*(4), 578–590. http://dx.doi.org/10.1016/J.DCN.2011.06.007.

Van Leijenhorst, L., Crone, E. A., & Bunge, S. A. (2006). Neural correlates of developmental differences in risk estimation and feedback processing. *Neuropsychologia*, *44*, 2158–2170. http://dx.doi.org/10.1016/j.neuropsychologia.2006.02.002.

van Ravenzwaaij, D., Cassey, P., & Brown, S. D. (2018). A simple introduction to Markov chain Monte–Carlo sampling. *Psychonomic Bulletin and Review*, http://dx.doi.org/10.3758/s13423-016-1015-8.

Verburg, M., Snellings, P., Zeguers, M. H. T., & Huizenga, H. M. (2018). Positive-blank versus negative-blank feedback learning in children and adults. *Quarterly Journal of Experimental Psychology*, 1–11. http://dx.doi.org/10.1177/1747021818769038.

Wagenmakers, E. J., Morey, R. D., & Lee, M. D. (2016). Bayesian Benefits for the pragmatic researcher. *Current Directions in Psychological Science*, *25*, 168–176. http://dx.doi.org/10.1177/0963721416643289.