Supplementary Information for

"What can performance in the IEDS task tell us about attention shifting in clinical groups?"

James M. Yearsley[a]*, Sebastian B. Gaigg[a], Dermot M. Bowler[a], Melanie Ring[ab] & Corinna Haenschel[a]

a Department of Psychology,

City, University of London,

EC1V 0HB

b Department of Child and Adolescent Psychiatry,

University Hospital Carl Gustav Carus,

Technische Universität Dresden,

Dresden, Germany

Author Note

* james.yearsley@city.ac.uk

Data and code connected to this paper can be found on the Open Science Framework https://osf.io/cg2m3/

Supplementary Information

Appendix A: Description of the Model

The basic object in the model is a normalised vector $a(t)$ representing the belief that each of the four possible features (two shapes and two lines) is the correct cue. Each of the elements of a(t) can take values from 0 to 1. We will adopt a convention that $a_{1,2}$ are the relevant dimensions, $a_{3,4}$ are the irrelevant ones, and a_1 is the correct feature. On any given trial two stimuli will be presented, each made up of one line and one shape. The probability of selecting the stimuli made up of shape i and line j on trial t of a particular block is given by,

$$p(s_{ij}, t) = \frac{a_i^{d(t)}(t) + a_j^{d(t)}(t)}{\sum_k a_k^{d(t)}(t)}, \quad \text{(S1)}$$

where $d(t)$ is a decision consistency parameter, which decreases as $d(t) = d_0 e^{-\lambda(t-1)}$ where d_0 and λ are constants which we will set to $d_0 = 3$ and $\lambda = \frac{1}{20}$ in the main body of the paper[1]. The factor of $(t-1)$ in the exponent ensures the decision consistency parameter is d_0 at the first trial of each stage (ie $t = 1$.) The probability of choosing a particular stimulus is therefore given by a modification of Luce's choice rule, in a similar way to Bishara et al (2010), such that the decision consistency, i.e. the rate at which the participant picks the stimulus they think is most likely to be correct, decreases throughout a particular block. This encodes the idea that a participant unable to solve the puzzle after a few trials is likely to give up and lose focus, and allows the model to fail to complete a block if the learning rate is too low. Other forms for the decision consistency parameter are possible, for example a gaussian form might be used to encode the idea of an attention 'limit', but this would not change the basic intuition.

Note that in principle, d_0, λ and even the form of $d(t)$ could be allowed to vary, and model fitting could be employed to select the 'best' values. However we are working with limited amounts of data in this paper, and it is therefore important to restrain any tendency to let the number of free parameters multiply.

At the start of the experiment attention is initially equally split between the two shapes (the first two blocks only include one feature type.) When a new block begins the initial attention vector is set according to the final attention vector $a(t_f)$ from the previous block. For example, if the current stage is a reversal, the new attention vector a' related to the previous one a, by,

$$a'(1) = a_2(t_f), \quad a'_2(1) = a_1(t_f),$$
$$\text{and} \qquad a'_3(1) = a_3(t_f), \quad a'_4(1) = a_4(t_f). \qquad \text{(S2)}$$

For the C_D, CD, ID, and ED stages, we assume the introduction of a new feature, or arrangement, causes some attention to be switched to the irrelevant dimension. For the C_D and CD stages this happens according to,

$$a'_k(1) = S \times a_k(t_f) + (1 - S) \times \left(\frac{1}{4}\right), \qquad \text{for } k = 1,2,3,4 \qquad \text{(S3)}$$

where $0 \leq S \leq 1$ determines how much attention remains on the original dimension. For the rest of this paper we set $S = .95$. Very low values of S, seem hard to justify, and some limited investigation indicated that model performance is relatively insensitive to the exact value so long as it is < 1. This could be investigated more systematically in future work.

For the ID stage, the exemplars change. Information about which, eg, white line is correct is therefore lost, but we assume participants retain the knowledge that the, eg, white lines are a diagnostic dimension. We therefore assume that attention is split over feature type, so,

$$a_1'(1) = a_2'(1) = S\times\frac{(a_1(t_f) + a_2(t_f))}{2} + (1-S)\times\left(\frac{1}{4}\right)$$
$$a_3'(1) = a_4'(1) = S\times\frac{(a_3(t_f) + a_4(t_f))}{2} + (1-S)\times\left(\frac{1}{4}\right) \quad \text{(S4)}$$

This means that at the ID stage attention is mostly redistributed with a dimension.

Finally, for the ED stages, the exemplars change again, but now the relevant dimension switches. We assuming that, like the ID stage, participants mostly redistribute attention within a dimension, but now the relevant dimension has switched. We therefore implement this as,

$$a_1'(1) = a_2'(1) = S\times\frac{(a_3(t_f) + a_4(t_f))}{2} + (1-S)\times\left(\frac{1}{4}\right)$$
$$a_3'(1) = a_4'(1) = S\times\frac{(a_1(t_f) + a_2(t_f))}{2} + (1-S)\times\left(\frac{1}{4}\right) \quad \text{(S5)}$$

The final ingredient in the model is the rule for updating attention after a choice and feedback. In general for positive feedback the attention vector updates according to,

$$a(t+1) = a(t) + r\sigma(t) \quad \text{(S6)}$$

or

$$a(t+1) = a(t) + p\sigma(t) \quad \text{(S7)}$$

for negative feedback.

To specify σ we need to account for the fact that there are two possible sets of stimuli a participant could be presented with, since the correct relevant feature (1) could be paired with either of the two irrelevant features (3,4). If the correct feature is paired with (3) then,

$$\begin{aligned}\sigma_1(t) &= (1-f)\times a_2(t) + f\times a_4(t)\\ \sigma_2(t) &= -a_2(t)\\ \sigma_3(t) &= (1-f)\times a_4(t) + f\times a_2(t)\\ \sigma_4(t) &= -a_4(t)\end{aligned} \tag{S8}$$

or if the correct feature is paired with (4) then,

$$\begin{aligned}\sigma_1(t) &= (1-f)\times a_2(t) + f\times a_3(t)\\ \sigma_2(t) &= -a_2(t)\\ \sigma_3(t) &= -a_3(t)\\ \sigma_4(t) &= (1-f)\times a_3(t) + f\times a_2(t)\end{aligned} \tag{S9}$$

Here f is a parameter that controls how much attention is switched between dimensions, and is necessary because all feedback in the IEDS is ambiguous (excepting the SD and SDr stages), since a correct choice could have been the result of either of the two features being correct. A value of $f = 0$ essentially means no attention switching, and we will see this gives rise to poor performance at the ED stage. However a value of $f = 1$ is also problematic, since then attention is switched too readily, and this can also result in poor performance. For the SD and SDr stages we set $f = 0$, since these stages involve only a single dimension.

Let us try to give a brief overview of the logic behind these feedback functions σ. Suppose I see a trial where the choices are between feature 1 paired with feature 3, and feature 2 paired with feature 4. If I select the first option and am correct, I learn that features 2 and 4 are definitely incorrect, and I also learn that either feature 1 or feature 3 is correct. In other words,

information gained about which features are incorrect is unambiguous, whereas information gained about which features are correct is ambiguous.

To deal with the unambiguous information we simply remove a proportion of the attention from the features we learn to be incorrect. This attention moves to the features we learn could be correct, so that attention mostly moves within a dimension (eg from feature 2 to feature 1) but can also move across dimensions (eg from feature 2 to feature 3.) We have discussed the case where our choice is correct, but it is easy to see that an incorrect choice gives participants exactly the same information, so this does not change the feedback functions σ.

To summarise, our model is based on similar ideas to that of Bishara et al (2010) and is rather simple conceptually. The key object is an attention vector which determines the probabilities for picking each of the two stimuli. Attention is updated as a result of either positive or negative feedback, and also effectively changes from block to block as the rules for correct choices change. Given feedback, the attention associated with the two features which make up the selected stimuli change. Attention either increases or decreases depending on whether the guess was correct or an error, but there is also a tendency for attention to transfer between dimensions since all feedback is ambiguous. The model has three free parameters, r, p, f, and several which we assume are fixed but which could in principle vary given the right experimental manipulations (we briefly explore the effect of varying λ in the next section).

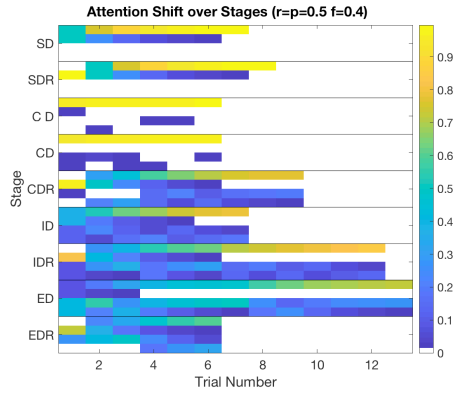
Crucially our model is designed to be simple enough that closed form expressions for the probability of choosing a given stimuli on a given trial can be derived (assuming we know the history of stimuli seen and choices made up to that point). This means we can very efficiently fit this model to experimental data, which we will see in a later section. Since Eq.(S1) is essentially a conditional probability of the choice on trial t given information about previous choices and

stimuli presentations, it would be possible to write down probabilities for any sequence of responses, given the information about which stimuli were presented. However, for ease we instead fit the model using a 'one step ahead' approach, similar to that in Bishara et al (2010), so we only need to use Eq.(S1) for the fits.

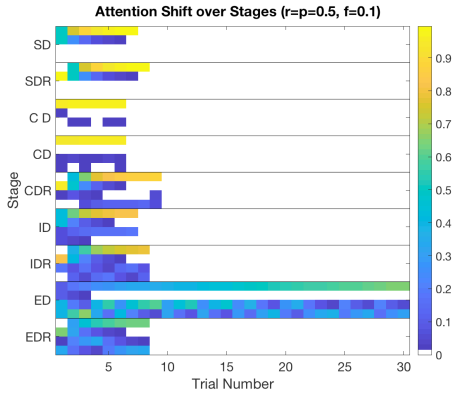
Appendix B: Additional Simulation Information

In this appendix we want to provide some more information from the basic simulations of the model. Firstly, we explore the way attention is transferred between dimensions during the task for the example parameter sets used to generate Figure 1 in the main text. In Figure S1 panels (a,b) we see a visualisation of an example attention vector for the 'control' and reduced attention switching parameter sets. For each trial we have plotted a 'heat map' of the values of the four components of the attention vector. Lighter colours represent larger values. For example looking at Figure S1 panel (a) we can see for the first stage, SD, attention is initially equally split between the first and second component, representing the two possible shapes shown in this block (the SD and SDr stages feature stimuli with only a single dimension). Over the course of the block attention gradually shifts from the incorrect to the correct dimension, indicated by the progressively brighter colour for this dimension. Similar patterns can be seen in the other blocks, with some initial attention split changing over the course of the trials, and the correct dimension eventually ending up with the largest attention weight.

In the case of reduced attention switching, Figure S1 panel (b) we clearly see how attention weight persists on the irrelevant dimensions through the course of the ED block, in other words, even after 30 trials the model is still paying attention to the irrelevant dimension.



(a) Attention vector over stages and blocks for a choice of $r = p = 0.5, f = 0.4$. Colour codes represent magnitude of each component of the attention vector.



(b) Attention vector over stages and blocks for a choice of $r = p = 0.5, f = 0.1$. Colour codes represent magnitude of each component of the attention vector.

Figure S1. The effect of changing the attention switching parameter, f, on performance for fixed $r = p = 0.5$. Here we are plotting the components of the attention vector $a(t)$ at each trial. This lets us examine how rapidly attention is switched from incorrect to correct features at the stages progress. The parameter sets produce similar behaviour until the ED stage (note the different scales on the x-axis), at which point the lower value of f hurts attention switching, and it takes many more trials for the model to switch attention to the correct feature. However once this switch happens, performance at the subsequent EDr stage is very similar for the two parameter sets.

Next, we want to explore the effect of changing the decision consistency parameters, in particular λ. This parameter can be thought of as capturing a participant's ability to sustain focus on the task. Recall, this parameter controls the way attention is mapped to choice probability as follows; the probability of selecting the stimuli made up of shape i and line j on trial t of a particular block is given by,

$$p(s_{ij}, t) = \frac{a_i^{d(t)}(t) + a_j^{d(t)}(t)}{\sum_k a_k^{d(t)}(t)}, \tag{S100}$$

where $d(t)$ is a decision consistency parameter, which decreases as $d(t) = d_0 e^{-\lambda(t-1)}$. In the main body of the paper we set $d_0 = 3$ and $\lambda = \frac{1}{20}$, but we want to briefly explore the impact of changing λ on model predictions. The reason for our interest is that there is evidence that ADHD, which commonly co-occurs with ASD, is associated with difficulties in sustained attention, so it may be the case that at least a sub-group of ASD participants (i.e., those with co-occurring ADHD) experience difficulties in sustained attention. It is therefore interesting to explore whether difficulties in sustaining attention, modelled through larger values of λ, produce behaviour which looks like diminished attention switching.

We can see the effect of varying λ in Figure S2. Starting with good overall rates of learning and attention shifting (red line), varying λ produces a very characteristic pattern of errors, with difficulties at the ED stage but also earlier stages, particularly CDr. This does not appear to match any of the data sets examined in the main text, although we have not fit the model with λ as a free parameter to any data, and this could be explored in future work.

Figure S2. Examining the effects of decrease in ability to sustain attentional. Comparing performance for r = p = 0.5, f =0.4, with $\lambda = \frac{1}{20}$ (red line) and $\lambda = \frac{1}{10}$ (blue line). Although reduced ability to sustain attention does seem to cause problems at the ED stage, it also reduces performance at earlier stages. Panel descriptions are as for Figure 1.

Appendix C: Details of the model fits for the City Data

Full data and code for these fits is available on the OSF https://osf.io/cg2m3/

The model allows us to write down an explicit probability for a participant choosing a particular stimulus on a given trial, given the set of choices and the past information about the trials and responses. We use a 'one step ahead' method, where we use the model to predict, for every trial, the probability of making the correct choice given the actual history of responses up to that point. This was then fit using MCMC methods in JAGS (Plummer et al, 2003). We took the priors for the three parameters to be uniform in the range [0, 1] and ran fits with five chains, 50,000 samples and a burn in of 5,000 samples. Chain convergence was assessed using the Rhat statistic, and all chains had good convergence behaviour by this metric.

We then extracted the mean of the posteriors for the three parameters, and used these in our analysis.

Appendix D: Details of the model fits for the Summary Data

Full data and code for these fits is available on the OSF.

We used an implementation of the Approximate Bayesian Computation - Partial Rejection Control (ABC-PRC) algorithm introduced by Sisson et al (2007, 2009) and previously employed by one of us to fit models of serial recall (Poirier et al, 2019). ABC-PRC provides a compromise between pure rejection sampling, which is simple but inefficient, and more

sophisticated algorithms like ABC Differential Evolution (Turner & Sederberg, 2012), which can be more efficient when parameters are correlated, but which are more complex.

ABC-PRC works by repeatedly sampling from a prior over the parameter space until it finds a set of parameters which generate a set of summary statistics sufficiently close to the data. When this happens, the algorithm stores these parameter values and moves on to the next particle in the generation. Once all particles in a generation have been associated with parameter sets, the algorithm gives each particle a weight depending on the prior, and then begins a new generation, sampling from the previous generation with probabilities given by the weights, and repeatedly perturbing around the previous parameter values until a set is found producing summary statistics even closer to the data. For full details, see Sisson et al. (2007; note also the errata, Sisson et al., 2009)

Under ABC-PRC, the posterior estimates for the parameters are just the fraction of particles in the final generation with that parameter value. Posterior predicted distributions of the summary statistics are also easily obtained. The important parameters for ABC-PRC are the number of particles (set to 10,000 for all fits reported here), the details of the priors (uniform of 0-1 in all cases), the proposal distributions, and the number of generations and tolerances for each generation. Setting the number of generations and the tolerances requires some trial and error. Lower tolerances will tend to result in a better match between model and data, but at some point the computational cost becomes prohibitive. Equally some of the mis-fitting between data and model is likely due to the presence of processes not captured by the model (eg loss of concentration or motivation), in which case there will be a lower limit to the tolerance which can be achieved.

Fits were performed on the Solon High Performance Computing cluster at City, University of London. For each particle we simulated an experiment with the same number of participants as reported in the relevant study. As well as the three parameters of interest, r, p, f , which we assumed represent the mean for the group, we assumed each participant in the group had their parameters drawn from a distribution with the group mean and a variance given by an additional set of parameters (λ_l, λ_f) for the learning rates (r, p) and attention switching (f). These parameters were given normal priors with means of 20 and SDs of 3. Posterior estimates for these parameters did not differ substantially from these priors and we conclude the model is not sensitive to the degree of heterogeneity in the groups.

To assess whether the parameter estimates for different fits were reliably different from each other, eg whether the ASD group had a lower best fitting attention switching parameter than the control group, we began by computing posterior difference distributions. This was done by pairing the particles for the TD and ASD fit, subtracting the final values of the ASD from the TD, and repeating for 100 different pairings. This gives a distribution of differences between the best fitting values, which are shown in Figure S3-5. Full size versions of these figure can be found on the OSF page, https://osf.io/cg2m3/ .

For the majority of these comparisons we can use the Savage-Dickey approximation (Wagenmakers et al 2010) to compute an approximate Bayes factor for the null hypothesis (the true difference is zero) against the alternative hypothesis. However, the Savage-Dickey approximation assumes the posterior is continuous around zero, which is questionable for some of the distributions. We therefore also compute the 95% Highest Density Intervals for the posterior difference, which ought to contain zero if the true difference is zero.

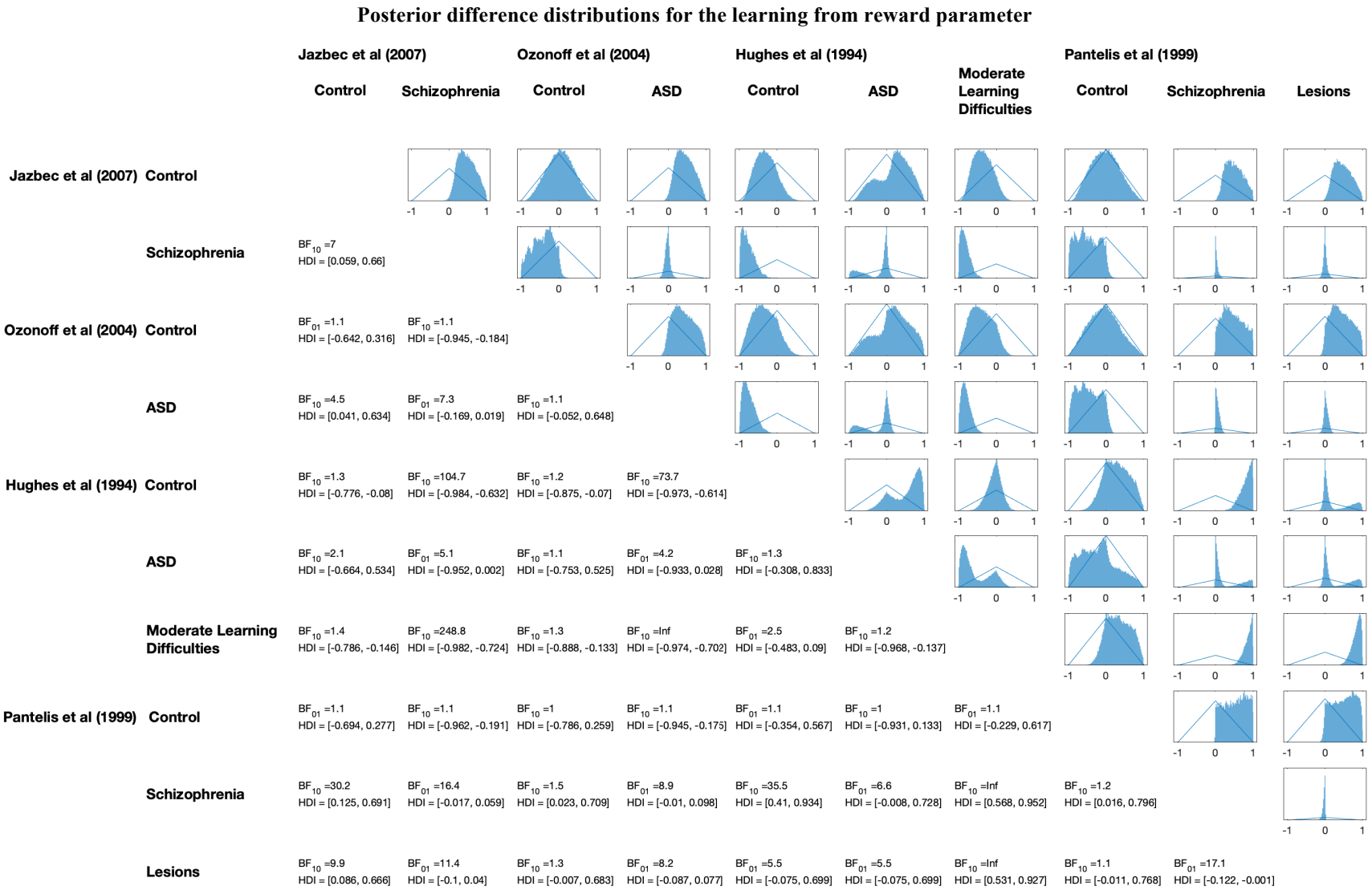


Figure S3: Posterior difference distributions for the learning from reward parameter, computed between fits for all data sets examined in this paper. The differences are computed for the plots as the fit for the row group minus the fit for the column group. The Bayes Factors and HDIs refer to the plot in the opposite position.

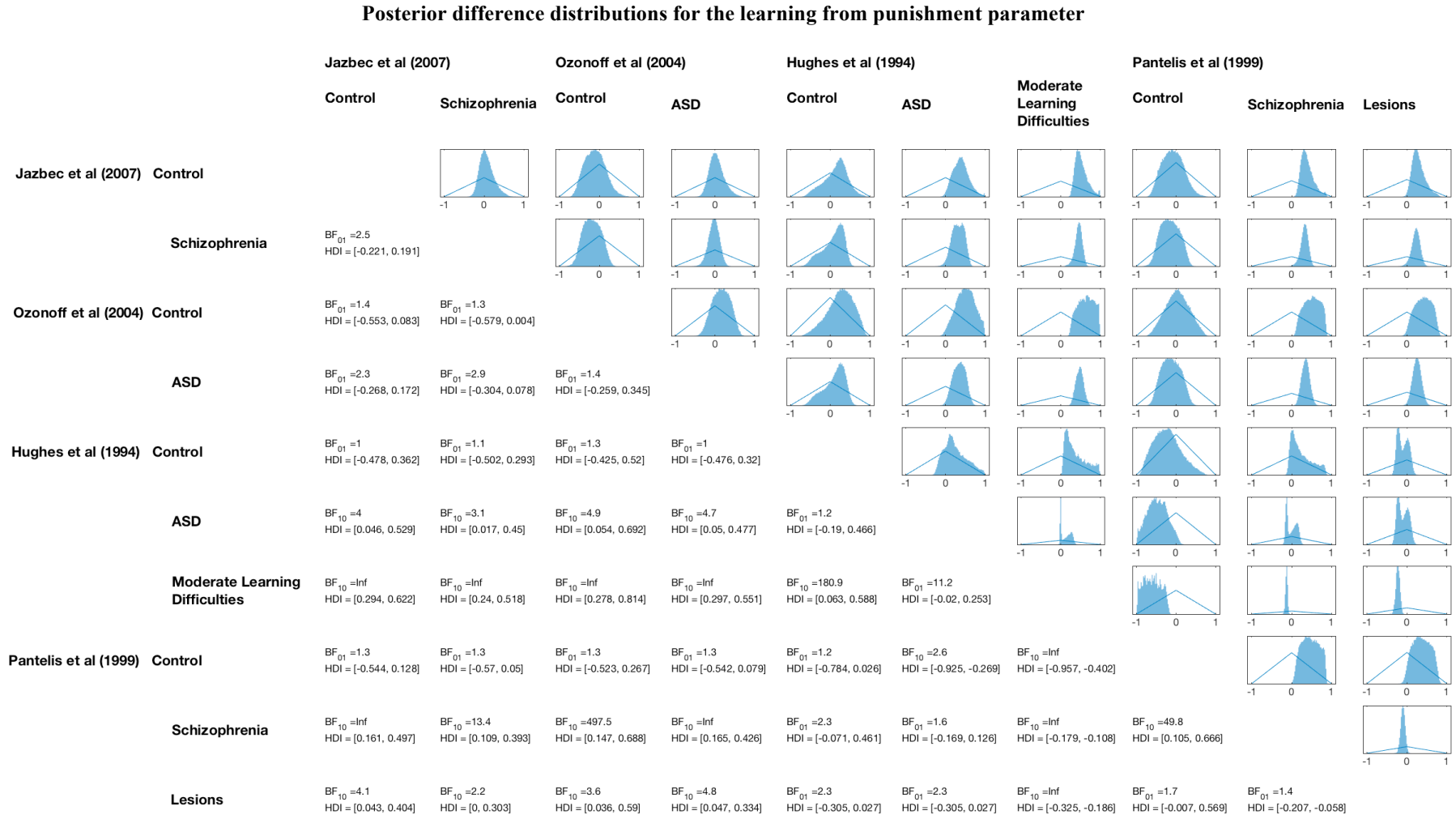


Figure S4. Posterior difference distributions for the learning from punishment parameter, computed between fits for all data sets examined in this paper. The differences are computed for the plots as the fit for the row group minus the fit for the column group. The Bayes Factors and HDIs refer to the plot in the opposite position.

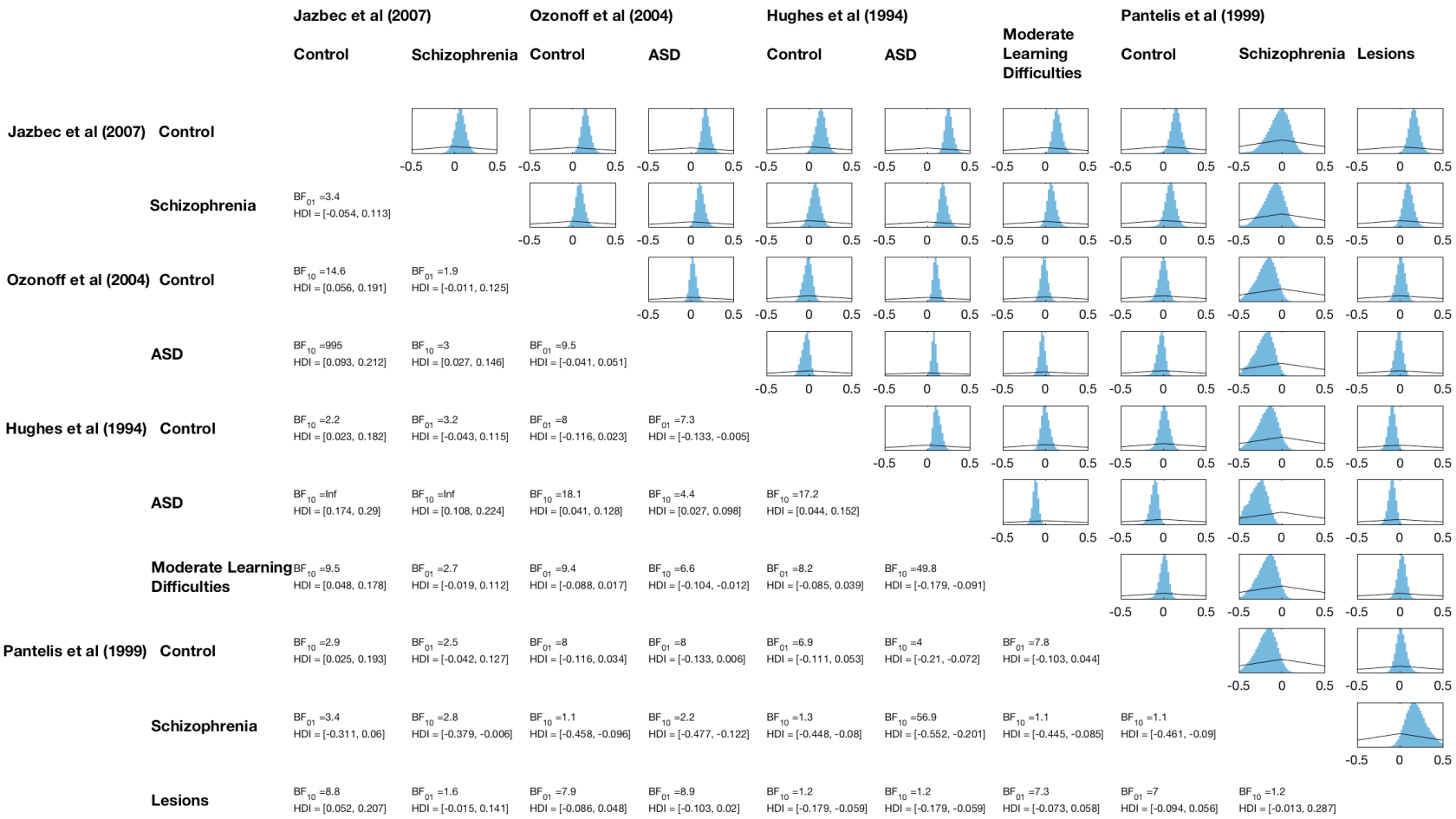


Figure S5. Posterior difference distributions for the attention switching parameter, computed between fits for all data sets examined in this paper. The differences are computed for the plots as the fit for the row group minus the fit for the column group. The Bayes Factors and HDIs refer to the plot in the opposite position. The possible range is [-1,1], only the central region is shown, for ease of viewing.

Footnotes

[1] Picking $d_0 = 3$ means that, at trial 1, if the ratio of attention for stimuli A over stimuli B is 2:1, the ratio of the probability of selecting A over B is around 9:1. The model is therefore quite sensitive early in each block to the stimuli with the larger attention. This provides a compromise between picking the stimuli with the higher attention, and probability matching, where the options would be chosen with probabilities given by the attention weights.

Picking $\lambda = \frac{1}{20}$ then means the model is probability matching by around trial 23, and by trial 50 if the ratio of attention for stimuli A over stimuli B is 9:1, the ratio of the probability of selecting A over B is around 2:1. This encodes the idea that, by the time a participant has answered almost 50 trials and failed to progress, it takes a very large difference in attention to produce a reliable response.

It is unlikely that the model fits depend strongly on these choices. We refit the data from Jazbec et al (2007) with $d_0 = 1$ and, while the best fit parameters were different, the comparison between the SZ and TD group yielded the same conclusions.

Additional References for the Supplementary Information

Plummer, M., et al. (2003). Jags: A program for analysis of bayesian graphical models using gibbs sampling. In Proceedings of the 3rd international workshop on distributed statistical computing (Vol. 124, p. 125).

Poirier, M., Yearsley, J.M., Saint-Aubin, J., Fortin, C., Gallant, G. and Guitard, D. (2019). Dissociating visuo-spatial and verbal working memory: It's all in the features. Memory and Cognition, 47(4), pp. 603–618. doi:10.3758/s13421-018-0882-9.

Sisson, S. A., Fan, F., & Tanaka, M. A. (2007). Sequential Monte Carlo without likelihoods. Proceedings of the National Academy of Sciences. 104, 6, 1760–1765. doi:https://doi.org/10.1073/pnas. 0607208104

Sisson, S.A., Fan, Y., & Tanaka, M.M. (2009). Correction for Sisson et al., Sequential Monte Carlo without likelihoods. Proceedings of the National Academy of Sciences of the United States of America, 106, 16889.

Turner, B. M., & Sederberg, P. B. (2012). Approximate Bayesian computation with differential evolution. Journal of Mathematical Psychology, 56, 375–385. doi:https://doi.org/10.1016/j.jmp.2012. 06.004

Wagenmakers, E-J, Lodewyckx, T, Kuriyal, H & Grasman, R (2010). Bayesian hypothesis testing for psychologists: A tutorial on the Savage-Dickey method. *Cognitive Psychology* 60, 158-189.