Appendix to 'Quantifying the effect of inter-trial dependence on perceptual decisions'

Ingo Fründ, Felix A. Wichmann, Jakob H. Macke

A1 Literature survey

As mentioned in the introduction, most researchers in the perceptual sciences are aware of the potential contamination of behavioural data by internal factors, at least during informal discussions. In stark contrast, we felt that rather few published papers actually discussed or corrected for internal factors. To confirm—or disconfirm—our impression, we attempted to obtain a rough estimate of the prevalence of serial dependency discussions in the relevant literature. To this end we searched the 2011 volume of a high-impact neuroscience journal (Nature Neuroscience) as well as in an established specialist journal for visual psychophysics (Journal of Vision).

We scanned one volume of Nature Neuroscience by searching pubmed (http://pubmed.com) for the search string ("Nature Neuroscience"[Journal] AND 2011[dp] AND (psychophysic*[TIAB] OR behavior*[TIAB] OR behaviour*[TIAB])). From the resulting articles we removed those that

1. referred to behavioral data in their abstract but did not analyze or record them in the study.
2. analyzed invertebrates (e.g. Drosophila or C. elegans) or used non-psychophysical methods such as food or water intake.

Application of the two criteria to the pubmed search resulted in a total of 14 articles from the 2011 volume of Nature Neuroscience. In addition we dropped one article that referred to psychophysics in the abstract but only reanalyzed average behavior from another study, resulting in a total of 13 articles from Nature Neuroscience.

We scanned one volume of The Journal of Vision using the journal's search function (http://www.journalofvision.org/search). We searched for articles that contained ("psychophysic*" OR "behavior*" OR "behaviour") in their "Title/Abstract" field and 2011 in their "Year" field. From the search results we manually omitted reviews and a number of papers that were published at the end of 2010 but were erroneously included in the

search results. In addition, we excluded one article about zebrafish and two that referred to psychophysics in their abstract but actually did functional magnetic resonance imaging without recording behavioral responses. Finally, we excluded one theoretical study that reanalyzed discrimination thresholds from another study. This left us with 41 articles from the Journal of Vision.

We carefully read the abstracts and methods sections of 54 articles in total and went through the entire manuscript to detect potentially misplaced methods descriptions (e.g. a brief description of the method at the beginning of the results section or in the discussion section).

Each paper was judged based on four criteria:

1. Did the article refer to sequential dependencies in the abstract?
2. Did the study perform any measures to avoid potential artifacts from inter-trial dependencies or did the article report checking for inter-trial dependencies in the methods section?
3. If no measures against inter-trial dependencies were taken, we assume that the analysis tacitly treated trials as independent realizations of a random variable. Was this explicitly mentioned in the article? We believe that this is the minimum level of awareness that could be expected for inter-trial dependencies.
4. Finally, we asked if measures to avoid potential artifacts from inter-trial dependencies were described in the supplemental material, in case such supplementary material existed. We only considered supplemental material if the main text referred to it in the context of behavioral data.

From the thirteen article in Nature Neuroscience, only a single one (Jaramillo & Zador, 2011) made the independence assumption explicit in their methods section. No article met the other three criteria. In the Journal of Vision five out of the relevant 41 articles published in 2011 referred explicitly or implicitly to inter-trial dependencies in their abstract, and two of these articles dealt with inter-trial dependencies in their methods section. However, one of these two articles—with two of the current authors as co-authors (IF and FAW)—was predominantly a simulation study with the focus on correcting the size of the confidence intervals of estimated parameters resulting from fitting a stationary observer model to non-stationary data (Fründ, Haenel, & Wichmann, 2011). The other article tried to avoid artifacts resulting from previous error trials by excluding trials from the analysis if they immediately followed an error trial (Meier, Flister, & Reinagel, 2011). Four articles in the Journal of Vision explicitly mentioned the assumption of independent trials in their methods section. Thus less than 10% of the relevant articles in the Journal of Vision discuss inter-trial dependencies, less than 5% do something about them. None of the Nature Neuroscience articles mention or discuss inter-trial dependencies and less than 8% mention the independence assumption.

A2 Illustration of psychometric function model

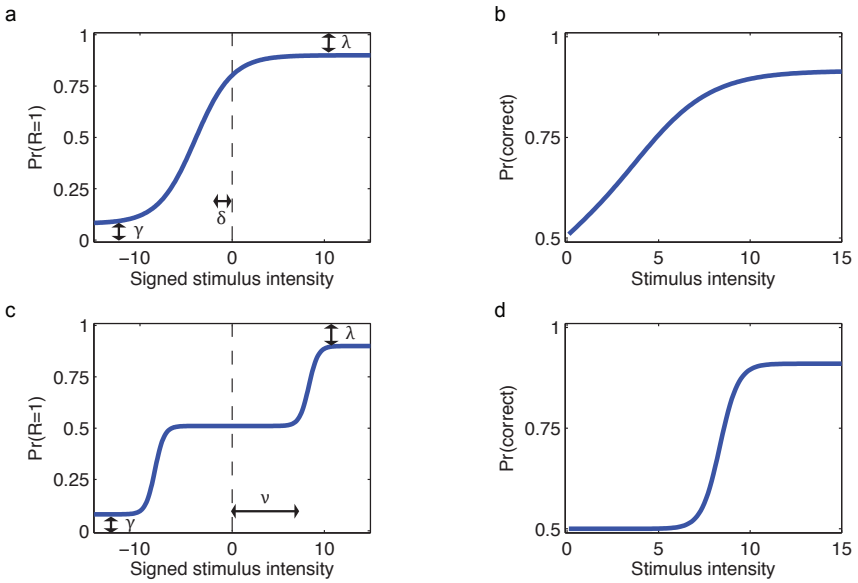

Figure A1: **Illustration of the psychometric function model and threshold. a)** Probability of response $r_t = 1$ as a function of the signed stimulus intensity in our model (without history dependence), with $\gamma = 0.08$, $\lambda = 0.1$, $\alpha = 0.5$, $\delta = 2$. **b)** Probability of correct response as a function of stimulus intensity, calculated from curve in panel a. **c)** Probability of response $r_t = 1$ as a function of the signed stimulus intensity in our model (without history dependence) and with sensory threshold, with $\gamma = 0.08$, $\lambda = 0.1$, $\alpha = 2$, $\delta = .1$, $\kappa = 4$, $\nu = 8$. **d)** Probability of correct response as a function of stimulus intensity, calculated from curve in panel c.

We are interested in modelling the effect of experimental history on perceptual decisions. Therefore, our psychometric function model relates the probability of a particular response to external covariates, and not (as is often done in psychophysics) the probability of a correct response. It is straightforward to convert our "left/right" psychometric function to a "correct/incorrect" one (see Figure A1 a, b). However, the resulting "correct/incorrect" psychometric curve would be steepest when the argument of the sigmoid $g(x)$ is 0, which (unless there is a left/right bias) occurs when the stimulus intensity is 0. This is in contrast with many psychophysical experiments which find that observers are at chance level for all stimuli which have an intensity that is less than some threshold ν, and therefore require a psychometric function that is flat at low intensities.

We explicitly modelled such a sensory threshold using an input non-linearity. We chose a "soft threshold" for this

$$u_\nu(x) = \frac{1}{\kappa}\log(1+\exp(x-\nu)^\kappa) - \frac{1}{\kappa}\log(1+\exp(-x-\nu)^\kappa). \quad (1)$$

For $\nu = 0$, we define $u_0(x) = x$. The effect of this function is that stimuli between $-\nu$ and ν are set to a value close to 0, while stimuli outside this interval are set to either $x-\nu$ (for positive x) or $x+\nu$ (for negative x). The value of κ defines the "softness" of the threshold. For $\kappa \to \infty$, we have

$$u_\nu(x) = \begin{cases} x+\nu & x < -\nu \\ x-\nu & x > \nu \\ 0 & \text{otherwise} \end{cases}$$

For finite κ, the transition between the three cases is smooth and the function remains differentiable at $-\nu$ and ν. We fixed $\kappa = 4$ which provided a good compromise between achieving values close to 0 between $-\nu$ and ν but keeping the function u_ν relatively smooth. We optimized the value of ν during the EM-optimization. Optimization of ν was done using Newton's procedure, keeping all other parameters fixed. This optimization step was introduced between the "expectation" step and the "maximization" step of the EM algorithm.

In the main text, we used a 2-AFC paradigm to describe our statistical framework. However, it can also easily be applied to detection (or yes/no) experiments like in our audio data-set. In this case, we drop the parameter describing stimulus identity and the encoding non-linearity u_ν (i.e. set $\nu = 0$), and use $r = 1$ to denote trials on which the observer indicated presence of the target (or responded with 'yes').

A3 Fitting the history-dependent psychometric function model to data

We want to fit a modified logistic regression model which also allows for 'performance asymptotes'. When modelling left/right responses, the asymptotes correspond to the probabilities that a subject would 'blindly' press left or right, without looking at (listening to) the stimulus. One way of incorporating these performance asymptotes is to define a latent variable which indicates when leftward or rightward 'guesses' did occur, and then to fit the model using expectation maximization (EM) algorithms.

We define r_t to be the binary response of the subject on trial t, x_t to be the 'effective' stimulus, i.e. a concatenation of the offset, the stimulus, and the history features on trial t, and ω their relative weights[1]. In addition, we define the (latent) variable $l_t \in \{0, 1, 2\}$.

[1]Note that we use ω in a slightly different way here than we do in the main text to keep the notation uncluttered.

If l_t is 0, we say that the subject guessed to a left response ($P(r_t = 1|l_t = 0) = 0$), if l_t is 1, it guessed a right response ($P(r_t = 1|l_t = 1) = 1$), and if l_t is 2, the subject actually looked at the stimulus, and responded "right" with probability $P(r_t = 1|l_t = 2) = g(\omega^\top x_t)$. We define the corresponding probabilities over l by $P(l = 0) = p_0 = \gamma$, $P(l = 1) = p_1 = \lambda$ and $P(l = 2) = p_2$. In addition, we define priors over both ω, $\pi_\omega(\omega)$ and over p, $\pi_p(p)$.

In the E-step of the EM-algorithm, we have to calculate the posterior probabilities of l_t given our observed data and our current estimate of parameters, $q_t(l) = P(l_t = l|x_t, r_t, \omega, p_{l_t})$. We get that

$$q_t(l) = \frac{P(r_t|l, x_t, \omega)P(l|x_t, p)}{\sum_{l'=0}^{2} P(r_t|l', x_t, \omega)P(l'|x_t, p)},$$

where each $P(r_t|l_t, x_t, \omega)$ is either 0, 1, or $g(\omega^\top x_t)$, and each $P(l_t|x_t, p)$ is one of the three p.

In the M-step, we have to find the parameter values that maximize the expected joint log-likelihood, where the expectation is over the possible values of lapse-variable, using the probabilities calculated above. Hence, we have to maximize

$$L(\omega, p) = \log(\pi(\omega)) + \sum_t \sum_{l_t=0}^{2} q_t(l_t) \log\left(P\left(y_t|x_t, l_t, \omega\right)\right) \tag{2}$$

$$+ \log(\pi(p)) + \sum_t \sum_{l_t=0}^{2} q_t(l_t) \log\left(P\left(l_t|x_t, \omega\right)\right) \tag{3}$$

$$= L_\omega(\omega) + L_p(p) \tag{4}$$

Here, π denotes the prior density of the parameter (see below). We simplify equation (2)

$$L_\omega(\omega) = \log \pi(\omega) + \sum_t q_t(2) \log P\left(\tilde{r}_t|\omega^\top x_t\right) + \text{const}$$

to find that it is very similar to the 'usual' cost function of logistic regression—the only difference is that each entry is now multiplied by $q_t(2)$, i.e. the probability that on a particular trial, the subject is not guessing. Thus, we can update ω using the standard iteratively reweighted least squares algorithm for logistic regression (e.g. (Dobson & Barnett, 2008)). We used independent normal distributions with mean 0 and precision 0.1 as priors for all elements of ω.

The update for p is closed-form, and does not require any numerical optimization. If we ignore the prior on p for the moment, we get

$$p_l = \frac{\sum_t q_t(l)}{\sum_{l'} \sum_t q_t(l')}$$

for $l \in \{0, 1, 2\}$. If we use a Dirichlet-prior on p with parameters α_D, we get

$$p_l = \frac{\alpha_D(l) - 1 + \sum_t q_t(l)}{\sum_{l'} \left(\alpha_D(l') - 1 + \sum_t q_t(l')\right)}.$$

For the current study we set $\alpha_D(l) = 1$ for all l. It is easy to enforce symmetry of the left and right lapses, by replacing their values by their average.

A4 Prediction-performance of different models as function of stimulus intensity for example observer

We show the prediction performance for each of the three models (quantified as prediction accuracy per block) as a function of stimulus intensity (see Figure A2). We note that, for the majority of blocks, the full model and the stimulus only model have identical performance (although they often did not yield the same predictions on a single stimulus level).

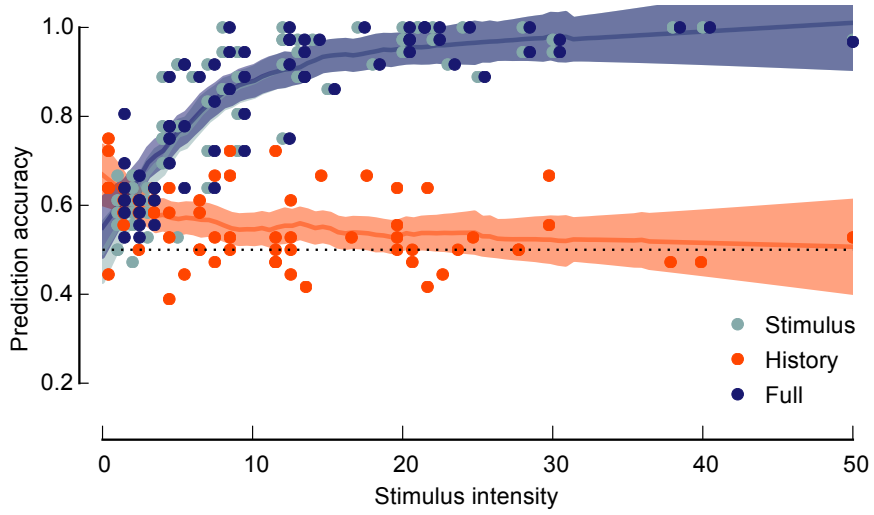

Figure A2: **Prediction performance as function of stimulus intensity for example observer.**

A5 Results on simulated data

To verify that our approach correctly identifies the presence or absence of history dependence, we simulated synthetic data which was matched in its statistial properties to the experimental data, but for which we new the ground-truth parameters which generated the data (see Figures A3, A4, A5, A6).

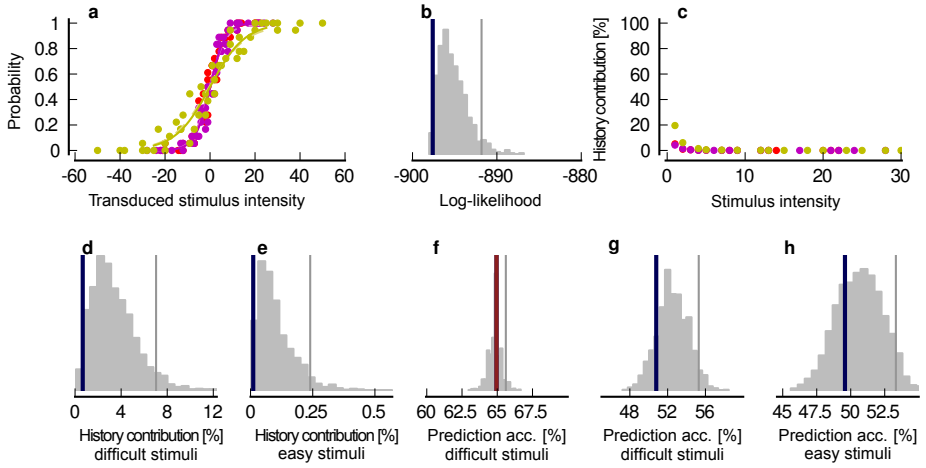

Figure A3: **Model fit to simulated data without history dependence.** Data was simulated using best-fitting parameters for observer pk in the main text, but by subsequently setting all history-couplings to 0. The number of trials N in the simulated data was matched to the experimental data for pk. Same format as Figure 1 in the main text, i.e. **a)** Psychometric function colours correspond to different experimental conditions. **b)** Log-likelihood of the full model (blue line). The grey histogram is the distribution on permuted data (grey), and vertical grey line marks its 95th percentile, the star marks statistical significance. **c)** Percentage of variance of decision variable explained as a function of stimulus intensity. **d)** Percentage of variance of decision variable explained by history on difficult trials. **e)** Same as d) but for easy trials. **f)** Prediction performance (percentage correct) of full model (including stimulus and history terms, blue line) in predicting observers' responses on difficult stimuli, and comparison with stimulus-only model (green line). **g)** Prediction performance of model with only history dependence and no stimulus dependence (blue line), and comparison with stimulus-only model (green line). **h)** Prediction performance of history-only model on easy stimuli (blue line).

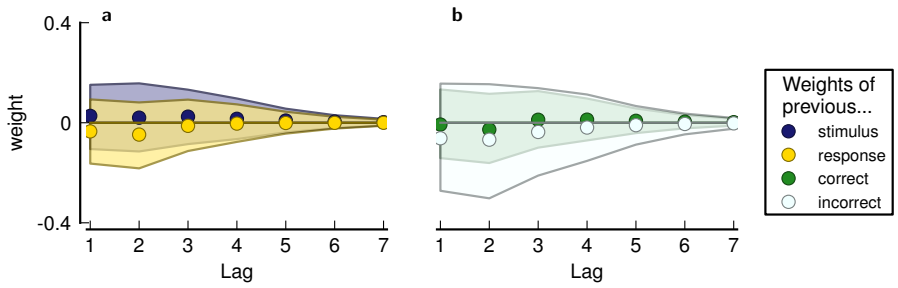

Figure A4: **History kernels recovered on simulated data without history dependence. a)** Weights assigned to previous stimuli and responses (coloured dots) and bootstrap confidence intervals (shaded regions). The lines at zero mark the true kernels that were used to generate the data. **b)** Weights assigned to previous correct and incorrect responses.

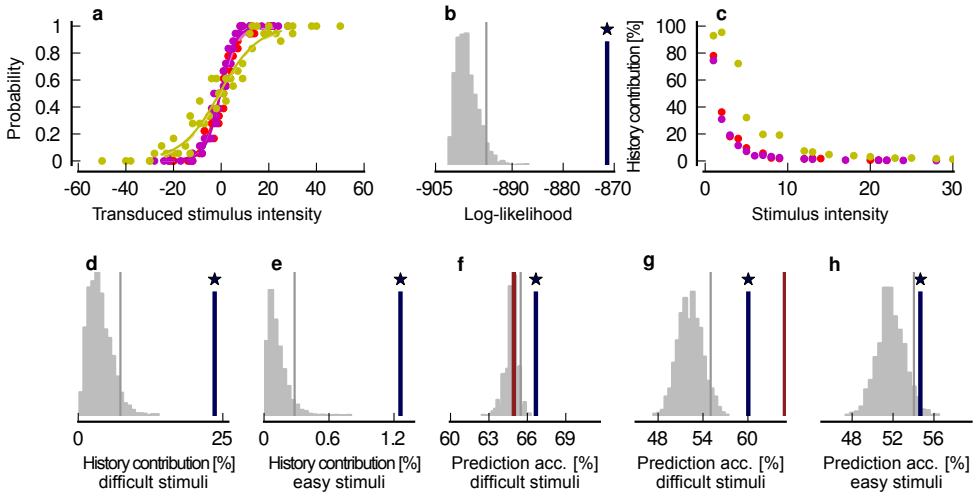

Figure A5: **Model fit to simulated data with known history dependence.** Data was simulated using best-fitting parameters for observer pk in the main text. **a-h)** Same format as Supplementary figure A3

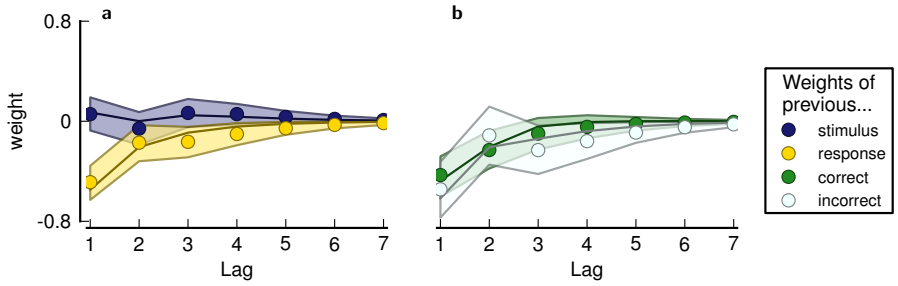

Figure A6: **History kernels recovered on simulated data with known history dependence.** Data was simulated using best-fitting parameters for observer pk in the main text. **a)** Weights assigned to previous stimuli and responses (coloured dots) and bootstrap confidence intervals (shaded regions). The lines mark the true kernels that were used to generate the data. **b)** Weights assigned to previous correct and incorrect responses.

A6 Connection between change in slope and history dependence in a simplified setting

Here we analytically connect the history-induced variance of the decision variable with the change in slope that one would obtain if history-dependence is falsely ignored. To simplify the analysis, we assume that $\gamma = \lambda = \delta' = \nu = 0$, i.e. the observer does not show any stimulus and history-independent lapses, has no left/right bias, and we do not need to consider the input non-linearity.

We can then write the psychometric function as

$$\Psi(\tilde{s}_t, h_t) = g\left(\alpha\tilde{s}_t + \sum_{k=1}^{K}\omega_k h_{t,k}\right) = g\left(\alpha\tilde{s}_t + \delta_t\right). \tag{5}$$

We further assume that the history features are scaled such that δ_t has mean 0 and variance σ^2. For this observer, a classical psychometric function would be fit to the stimulus-averaged probabilities $P(r = 1|\tilde{s}) = \mathbf{E}_\delta(g(\alpha\tilde{s} + \delta))$, where the average is over all history-dependent biases δ that were observed for stimulus $\tilde{s}$. We will first investigate how history dependency affects the probability of correct response for a stimulus with (signed) intensity $\tilde{s} > 0$, i.e. a correct response corresponds to $r = 1$ (the other case will follow by symmetry): For inputs $x > 0$, the sigmoid non-linearity $g(x)$ is concave, and therefore (by Jensens inequality (Cover & Thomas, 2006)) we get that

$$P(r = 1|\tilde{s}) = \mathbf{E}_\delta(g(\alpha\tilde{s} + \delta)) \leq g(\alpha\tilde{s} + \mathbf{E}_\delta(\delta)) = g(\alpha\tilde{s}). \tag{6}$$

Thus, averaging over different histories leads to a probability of a correct response $r = 1$ which is lower than the observer would have in the absence of history dependency. Importantly, our model gives us access to the underlying slope-parameter α, and therefore lets us correct the psychometric function for this (potential) performance- drop due to history dependency.

We approximate the logistic nonlinearity g by a rescaled Gaussian cumulative distribution function $g(x) \approx \Phi\left(x\sqrt{\pi/8}\right)$(see (Bishop, 2006) for details). We additionally assume that δ_t is approximately Gaussian (which is the case for weak history dependence $\sigma \ll 1$) to obtain

$$\begin{aligned} P(r = 1|\tilde{s}) &= \mathbf{E}_\delta g(\alpha\tilde{s} + \delta) && (7)\\ &\approx \mathbf{E}_\delta \Phi\left(\sqrt{\frac{\pi}{8}}(\alpha\tilde{s} + \delta)\right) && (8)\\ &= P\left(Y < \sqrt{\frac{\pi}{8}}(\alpha\tilde{s} + \delta)\right) && (9) \end{aligned}$$

where $Y \sim \mathcal{N}(0, 1)$ and Y is independent of δ. Therefore, $Y - \delta\sqrt{\frac{\pi}{8}}$ has variance $1 + \sigma^2\frac{\pi}{8}$

and thus

$$P(r=1|\tilde{s}) = P\left(\frac{Y-\delta\sqrt{\frac{\pi}{8}}}{\sqrt{1+\sigma^2\frac{\pi}{8}}} < \frac{\sqrt{\frac{\pi}{8}}\alpha\tilde{s}}{\sqrt{1+\sigma^2\frac{\pi}{8}}}\right) \tag{10}$$

$$= \Phi\left(\sqrt{\frac{\pi}{8}}\frac{\alpha\tilde{s}}{\sqrt{1+\sigma^2\frac{\pi}{8}}}\right) \tag{11}$$

$$= g\left(\frac{\alpha}{\sqrt{1+\sigma^2\frac{\pi}{8}}}\tilde{s}\right) \tag{12}$$

Thus, the slope-parameter of the psychometric function changes from α to $\alpha/\sqrt{1+\sigma^2\frac{\pi}{8}}$, i.e. it is 'rescaled' by division through $\sqrt{1+\sigma^2\frac{\pi}{8}}$. for weak history dependency. Thus, for weak history dependency and this simplified setting, there is a direct and simple relationship that tells us how the variability of the history-dependent bias reduces the slope of the psychometric function (or, vice versa, how correcting for this history dependency leads to a steeper slope.) The formula $1/\sqrt{1+\sigma^2\frac{\pi}{8}}$ gives us the factor by which the slope of the psychometric function at the inflection point needs to be multiplied to account for the effect of history-dependency. Furthermore, we note that for small σ, $1/\sqrt{1+\sigma^2\frac{\pi}{8}} \approx 1-\frac{\pi}{16}\sigma^2$.

A7 Detailed results for further observers

We show detailed results for two further observers. The first observer (kp) participated in the plaid-masking experiments and showed strong history dependence– for this observer, experimental history was a better predictor of perceptual choices than the presented stimulus (see Figures A7, A8, A9)). The second observer (gbh) is a very experienced psychophysical observer which participated in the discrimination experiment by Jäkel and Wichmann(Jäkel & Wichmann, 2006). History dependence for this observer was comparatively weak, yet statistically significant (see Figures A10, A11, A12).

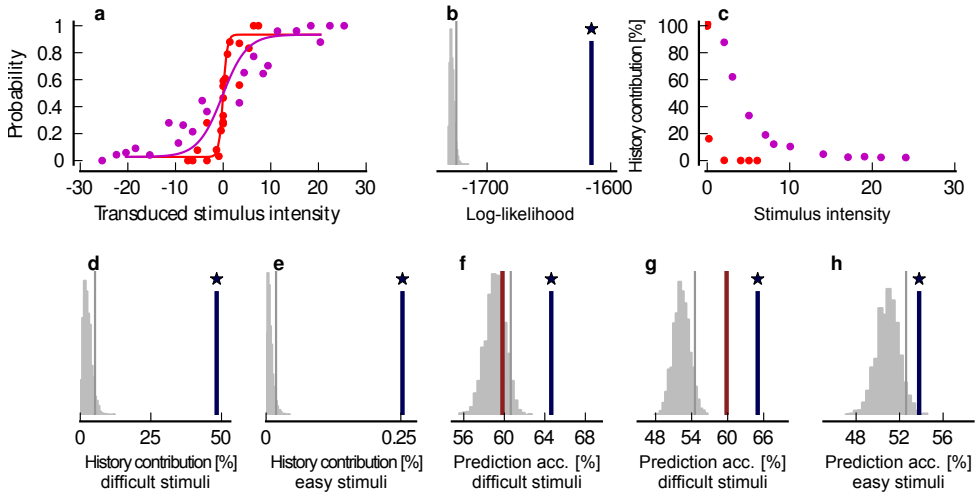

Figure A7: **History effects for observer kp. a-h)** Labels are as for Supplementary Figure A3. This observer has a strong effect of history.

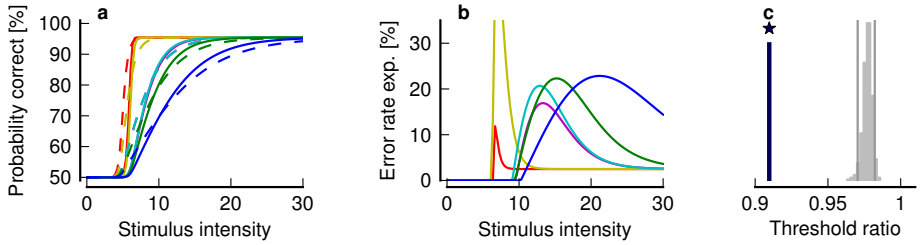

Figure A8: **Effects of history on the psychometric function for observer kp**. Labels are as for figure 2 in the main text. The psychometric function for this observer is altered by history effects, and there was a significant change in threshold. **a)** Psychometric functions indicating frequency of correct responses as a function of stimulus intensity. Dashed lines mark fits of a model without history terms. Colours correspond to different experimental conditions. **b)** Percentage of behavioural errors attributable to history, i.e. normalized difference between error rates predicted psychometric functions with or without history couplings. **c)** Ratio of 85% performance thresholds (blue line) between full model and conventional model, and null-distribution (grey histogram).

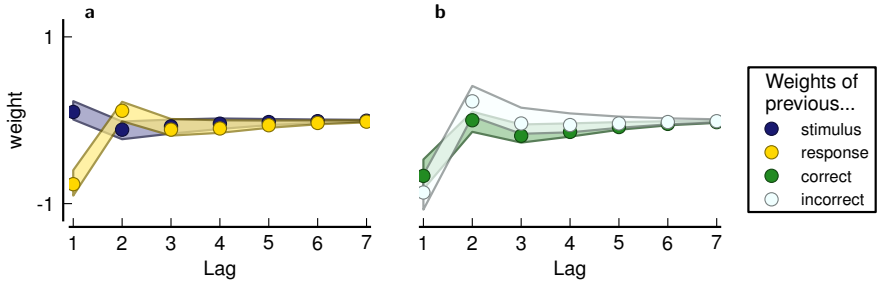

Figure A9: **History kernels for observer kp. a)** Weights for preceding stimuli and responses (dots), and 95% bootstrap confidence regions (shaded). **b)** Weights for preceding correct and incorrect responses (dots), and 95% bootstrap confidence regions (shaded). In accordance with part a), the effects of incorrect responses on previous trials is the same as the effect of correct responses.

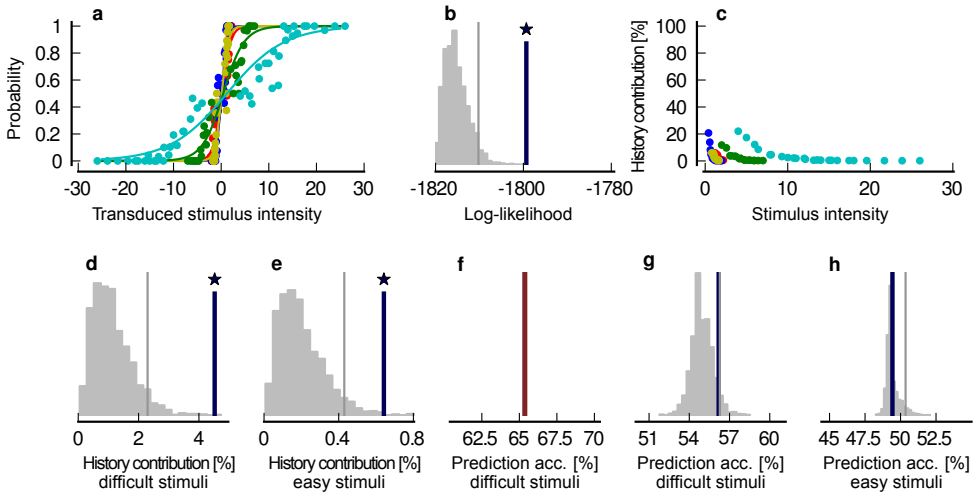

Figure A10: **History effects for observer gbh. a-h)** Labels are as for Supplementary figure A3. This observer has a weak history effect.

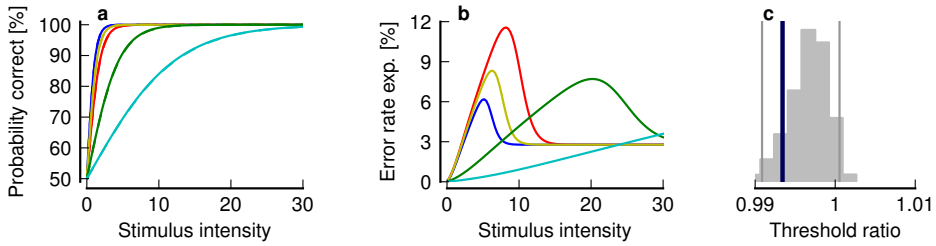

Figure A11: **Effects of history on the psychometric function for observer gbh. a-c)** Labels are as for Supplementary Figure A8.

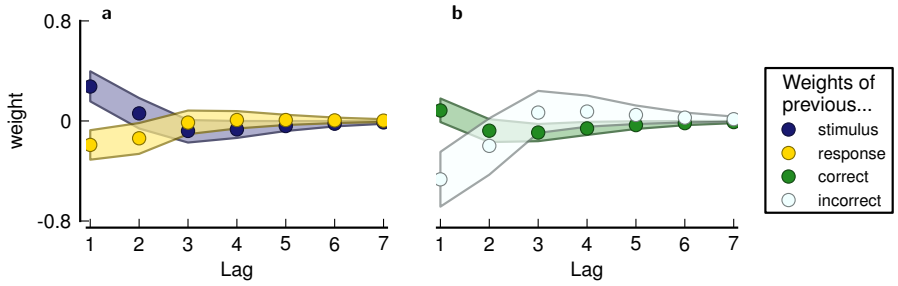

Figure A12: **History kernels for observer gbh. a)** Weights for preceding stimuli and responses (dots), and 95% bootstrap confidence regions (shaded). Previous stimuli and previous responses have significantly different directions of effects on the observers responses: Whenever stimulus and response differ (i.e. an incorrect response), the stimulus and response effects of this trial add up, when stimulus and response match (i.e. a correct response), the stimulus and response effects of this trial cancel partly. **b)** Weights for preceding correct and incorrect responses (dots), and 95% bootstrap confidence regions (shaded). As already expected from part a), the effect of incorrect responses on previous trials is larger than the effect of correct responses.

A8 Response and stimulus kernels for all observers

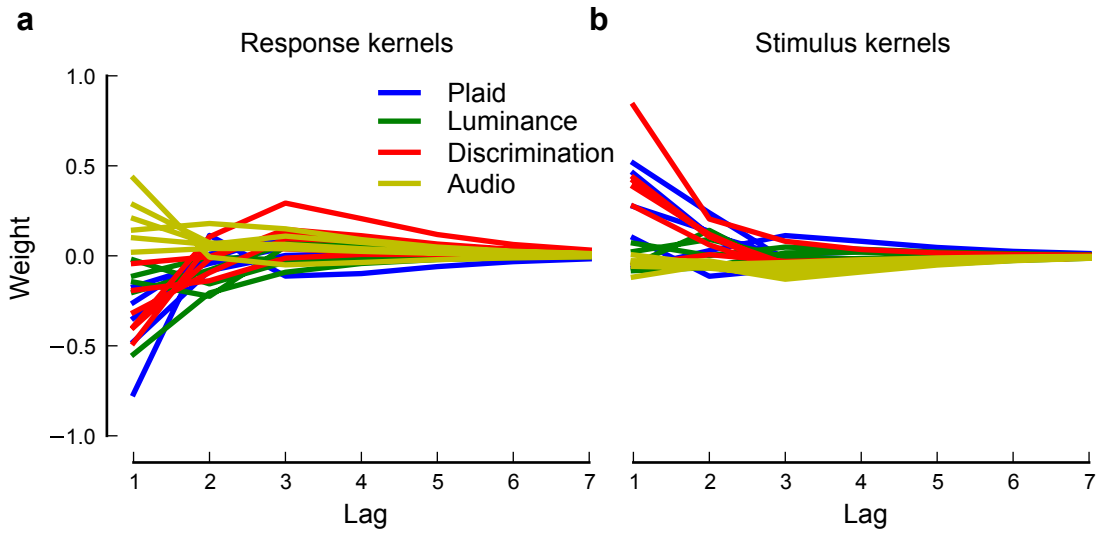

Figure A13: **History kernels for all observers a)** Response kernels for all observers. **b)** Stimulus kernels for all observers. The experimental design is coded by the color of the lines as in figure 4 in the main text

References

Bishop, C. M. (2006). *Pattern recognition and machine learning*. New York: Springer.

Cover, T. M., & Thomas, J. A. (2006). *Elements of Information Theory* (2nd ed.). Hoboken, New Jersey: Wiley.

Dobson, A. J., & Barnett, A. G. (2008). *An Introduction to Generalized Linear Models* (3rd ed.). Boca Raton: Chapman & Hall.

Fründ, I., Haenel, N. V., & Wichmann, F. A. (2011). Inference for psychometric functions in the presence of nonstationary behavior. *Journal of Vision*, *11*(6), 1-19.

Jäkel, F., & Wichmann, F. A. (2006). Spatial four-alternative forced-choice method is the preferred psychophysical method for naïve observers. *Journal of Vision*, *6*(11), 1307–1322.

Jaramillo, S., & Zador, A. M. (2011). The auditory cortex mediates the perceptual effects of acoustic temporal expectation. *Nature Neuroscience*, *14*(2).

Meier, P., Flister, E., & Reinagel, P. (2011). Collinear features impair visual detection by rats. *Journal of Vision*, *11*(3), 1-16.