

Quantifying the effect of inter-trial dependence on perceptual decisions

Ingo Fründ¹, Felix A. Wichmann², Jakob H. Macke^{3,*}

1 Bernstein Center for Computational Neuroscience, Technical University Berlin, Germany, Center for Vision Research, York University, Toronto, ON, Canada

2 Neural Information Processing Group, Eberhard Karls Universität, Max Planck Institute for Intelligent Systems, Bernstein Center for Computational Neuroscience, Tübingen, Germany

3 Max Planck Institute for Biological Cybernetics, Bernstein Center for Computational Neuroscience, Werner Reichardt Centre for Integrative Neuroscience, Tübingen, Germany, Gatsby Computational Neuroscience Unit, University College London, UK

* E-mail: jakob.macke@wsii.uni-tuebingen.de

Abstract

In the perceptual sciences experimenters study the causal mechanisms of perceptual systems by probing observers with carefully constructed stimuli. It has long been known, however, that perceptual decisions are not only determined by the stimulus, but also by internal factors. Internal factors could lead to a statistical influence of previous stimuli and responses on the current trial, resulting in *serial dependencies* which complicate the causal inference between stimulus and response. However, the majority of studies do not take serial dependencies into account, and it has been unclear how strongly they influence perceptual decisions. We hypothesize that one reason for this neglect is that there has been no reliable tool to quantify them and to correct for their effects. Here we develop a statistical method to detect, estimate and correct for serial dependencies in behavioural data. We show that even trained psychophysical observers suffer from strong history dependence. A substantial fraction of the decision-variance on difficult stimuli was independent of the stimulus but dependent on experimental history. We discuss the strong dependence of perceptual decisions on internal factors and its implications for correct data interpretation.

Introduction

In the perceptual sciences one of the central goals is to infer the causal mechanisms of perceptual systems by probing human observers or animals with carefully constructed or selected stimuli (Blackwell, 1952; Green & Swets, 1966). In practice, a tacit assumption underlying this approach is that the only systematic causal determinant of the perceptual decision is the presented stimulus on an individual trial, i.e. that each response is not influenced by previous responses or stimuli. This would imply that, given the stimulus, different trials of an experiment are statistically independent. While this assumption is undoubtedly convenient, it may nonetheless not be appropriate. Internal factors can lead to a statistical influence of previous stimuli and responses on the current trial (Green, 1964). In the current context internal factors are any factors that may influence an observer's decision other than the stimulus—internal dynamics, attentional and motivational state, adaptation and learning, as well as the stimulus or response history in the form of inter-trial dependencies.

In principle, the existence of such inter-trial dependencies has been recognized long ago (Verplanck, Collier, & Cotton, 1952; Senders & Sowards, 1952; Howarth & Bulmer, 1956; Green, 1964), and numerous studies have reported correlations between successive trials in experiments with human observers (Mori, 1998; Maljkovic & Nakayama, 1994; Lages & Treisman, 2010; Bode et al., 2012). Multiple studies have quantified correlations between responses (Howarth & Bulmer, 1956; Mori & Ward, 1995; Mori, 1998; Lages & Treisman, 1998, 2010; Maloney, Dal Martello, Sahm, & Spillmann, 2005) or response times (Peeke & Stone, 1972; Maljkovic & Nakayama, 1994; Otto & Mamassian, 2012; Wagenmakers, Farrell, &

Ratcliff, 2004). Some studies have suggested theoretical models for interdependencies between successive trials in particular tasks (Lockhead & King, 1983; Stewart, Brown, & Chater, 2002; Maljkovic & Martini, 2005; Stewart, Brown, & Chater, 2005; Martini, 2010; Raviv, Ahissar, & Loewenstein, 2012), and tested signatures of model assumptions (Baird, Green, & Luce, 1980; Green, Luce, & Duncan, 1977; Green, Luce, & Smith, 1980; Treisman & Williams, 1984). While there is some evidence suggesting that the effect of sequential dependencies on human psychometric functions is small (Verplanck & Blough, 1958), studies with animals found that taking inter-trial dependencies into account was of critical importance (Lau & Glimcher, 2005; Busse et al., 2011).

This substantial body of literature shows that the existence of non-stimulus effects on perceptual decisions is well-established in principle. However, they are commonly ignored in neuroscientific and psychophysical research: We analyzed the 2011 volume of a high-impact neuroscience journal (*Nature Neuroscience*) as well as of an established specialist journal for visual psychophysics (*Journal of Vision*), and found that only 1 out of 54 “candidate”-articles (Meier, Flister, & Reinagel, 2011) checked and corrected for inter-trial dependence (for details, see Appendix A1). The vast majority of articles did not explicitly state that they assume trials to be independent, but analyzed their data as if they were. Thus, despite considerable evidence for inter-trial dependencies in experiments that were designed to explicitly study them, these results are rarely applied in practice. This is potentially problematic, as non-stimulus determinants could both lead to spurious correlations between behaviour and measurements of neural activity as well as to (downward) biases in estimates of psychophysical performance.

We hypothesize that the widespread neglect to consider this effect is caused by three factors: First, there has been no generally applicable method for quantifying the effect of experimental history on perceptual decisions in psychophysical tasks. Second, there has been no systematic, direct quantification of the *magnitude* of trial by trial, non-stimulus determinants of perceptual decisions in the context of psychophysics with trained observers. Therefore, it is unknown what percentage of variance in typical psychophysical data is caused by the stimulus, and what percentage can be attributed to task-irrelevant experimental history. Third, while some early studies reported to find a weak or no effect of inter-trial dependence on psychophysical thresholds (Verplanck & Blough, 1958; Senders & Sowards, 1952) there have been no generally applicable and practical methods for quantifying and correcting such biases on a wide range of psychophysical data. Thus, it is still an open question how strong and problematic serial dependencies are across typical psychophysical tasks.

Our focus here is on a unified statistical description of the different types of history effects in binary responses in single-stimulus or two-alternative forced-choice designs. Using a range of different psychophysical data-sets from these paradigms, we will show that our method finds systematic dependencies between trials. Yet, typical measures of psychophysical performance, such as the psychometric function’s slope and threshold, are hardly influenced by these dependencies. By performing mathematical analysis of a simplified model, we describe why classical psychometric measures are robust to sequential dependencies.

While the kind of data considered here (binary responses) is common in psychophysics and widely used to measure perceptual thresholds with psychometric functions (e.g. (Wichmann & Hill, 2001)), there are also alternative psychophysical paradigms (e.g. those based on ternary responses or rating tasks) which are not compatible with this modelling framework. We do not address the question of sequential dependence in reaction times in visual search tasks, which have been characterized e.g. by (Maljkovic & Martini, 2005; Martini, 2010). We also exclusively investigate inter-trial sequential dependencies in behavioural tasks in which the experimenter assumes there to be none. This is in contrast to experiments investigating other internal factors violating the assumption of independent trials, such as adaptation (Chopin & Mamassian, 2012) or learning (Sugrue, Corrado, & Newsome, 2004; G. S. Corrado, Sugrue, Seung, & Newsome, 2005; O’Doherty, Hampton, & Kim, 2007; Lau & Glimcher, 2005). In such settings the experimenter is explicitly interested in behavioural changes over time and therefore aims to induce and characterize these changes.

Methods

A statistical model for capturing serial dependencies in psychophysical data

To study the causal influence of recent experimental history on the perceptual decisions of observers, we need a statistical model which can capture the influence of both the stimulus and previous trials on the observed response. The psychometric function is a commonly used model for psychophysical data which usually relates the probability of a correct response to the presented stimulus intensity (Treutwein & Strasburger, 1999; Wichmann & Hill, 2001; Kuss, Jäkel, & Wichmann, 2005). To model the effect of previous trials on the response, we modify this common formulation: We relate the probability of a particular behavioural response r_t to the presented stimulus intensity \tilde{s}_t . We used “signed” stimulus intensities $\tilde{s}_t := s_t z_t$ here, which consist of the product of the absolute intensity of the stimulus s_t and an identity factor z_t which codes when or where the target was presented. For example, in the 2-alternative forced-choice (2-AFC) task (Jäkel & Wichmann, 2006) considered below, we use the stimulus identity $z_t = 1$ to indicate that the second of two presented stimuli contained a luminance increment (‘target’), and set $r_t = 1$ if the observer also chose the second interval, ($z_t = -1$ or $r_t = -1$ otherwise, see below and Appendix A2 for details). Choice models in psychophysics usually have a bias-term δ which captures a stimulus independent tendency of observers to choose a particular response. To model sequential dependencies, we simply assume that δ is not constant, but may shift dependent on experimental history (Treisman & Williams, 1984). This is in accordance with a large number of experimental findings (Hock, Kelso, & Schöner, 1993; Lages & Treisman, 1998, 2010; Lages & M., 2010) and previous modeling attempts (Busse et al., 2011; G. S. Corrado et al., 2005; Raviv et al., 2012; Bode et al., 2012; Goldfarb, Wong-Lin, Schwemmer, Leonard, & Holmes, 2012; Green et al., 1977, 1980; Lockhead & King, 1983; Ward, 1979). Concretely, we assume that the bias-term δ can be written as a linear combination of ‘history features’, i.e. summary statistics of the events on preceding trials (Busse et al., 2011; G. S. Corrado et al., 2005)¹:

$$\delta(\mathbf{h}_t) = \delta' + \sum_{k=1}^K \omega_k h_{kt} =: \delta' + \delta_{\text{hist}}(\mathbf{h}_t). \quad (1)$$

Here, the vector h_{kt} can be taken to be any feature of the recent history which might potentially influence behavioural responses. We say that a data set exhibits history dependence, if given the current stimulus, the current response is statistically dependent on previous stimuli and previous responses, that is $P(r_t|\tilde{s}_t, \mathbf{h}_t) \neq P(r_t|\tilde{s}_t)$.

In our analyses below, we set \mathbf{h}_{kt} to be a concatenation of the last seven responses and stimulus-identities, that is $h_t = (r_{t-1}, \dots, r_{t-7}, z_{t-1}, \dots, z_{t-7})$, a vector of dimensionality $K = 14$.

The influence of history will then be modelled as a weighted sum of these history features, i.e. the history-couplings ω_k in equation (1) indicate how much the respective response/stimulus influences the current response. For example, $\omega_1 > 0$ indicates that the observer tended to repeat the previous response, and $\omega_1 < 0$ that there was a tendency to switch responses. Our this model captures co-variations between the observer’s responses and previous responses or stimuli. These co-variations could increase the variance of the resulting responses in a block², but could also lead to a decrease in variance or even leave it unchanged.

In our model, the probability of choice $r_t = 1$ (which denotes a right-ward or second-interval choice

¹Letting the bias-term vary dynamically from trial to trial complicates the interpretation of the model in terms of a signal and a decision criterion as defined in signal-detection theory. We here restrict ourselves to finding a statistical description of history effects, and we do not attempt to disentangle whether they influence the criterion or the signal in the decision process.

²Here, a block refers to a sequence of trials with constant (unsigned) stimulus intensity. Thus, trials within a block differ only with respect to the position of the stimulus (e.g. left/right in a spatial 2-AFC) and their history features.

in the context of 2-AFC) is given by

$$P(r_t = 1 | \tilde{s}_t, \mathbf{h}_t) = \gamma + (1 - \gamma - \lambda)g \left(\delta' + \sum_{k=1}^K \omega_k h_{kt} + \alpha u_\nu(\tilde{s}_t) \right). \quad (2)$$

Here, $\tilde{s}_t = z_t s_t$ is the signed stimulus intensity, λ and γ describe the probabilities of stimulus-independent responses to the right (γ) or to the left (λ) and $g(x)$ is a sigmoid function. In the following, we chose the logistic function $g(x) = 1/(1 + \exp(-x))$, and δ' and α are the offset and slope of the stimulus dependent part of the psychometric function. We note that an alternative view of the model is to assume that observers implicitly combine the stimulus with a trial-specific Bayesian prior assumption about whether the next target is in the first or second stimulus, and that this prior probability $P(z_t | \mathbf{h}_t)$ depends on the recent stimulus history (Wilder, Jones, & Mozer, 2010; Yu & Cohen, 2008).

In experiments with multiple experimental conditions, we allowed the slope α to be different across conditions, but assumed that the history couplings ω_k are constant across conditions. As our model describes the probability of particular responses (not the probability of the response being correct), we need to introduce a sensory threshold ν which accounts for the fact that observers perform at chance level whenever the stimulus has an intensity less than some sensory threshold ν . We use the non-linear threshold function u_ν which maps all stimuli with an intensity of less than ν to 0 (for details see Appendix A2 and Figure A1).

This model also has a modified intercept term δ' that can capture potential within trial biases that are unrelated to the experimental history (see for example (Ulrich & Vorberg, 2009) and (Garcia-Perez & Alcalá-Quintana, 2011) for more detailed treatments of these effects). In some cases, these within-trial biases are also associated by differences in the observer’s sensitivity. In that case, it would be possible to apply the model separately to trials of each presentation order, which is equivalent to the approach advocated in (Garcia-Perez & Alcalá-Quintana, 2011). Our formulation could also be used to include additional covariates in \mathbf{h}_t which describe the current trial, and which thus could be used to capture more complex within-trial effects, but this is not pursued here.

We also note that alternative parameterizations could be used to model the effect of history on the *slope* of the psychometric function. To model effects on slope, one could include features which are proportional to s_t . For example, by including a feature of the form $s_t z_{t-1}$, one could make the slope dependent on the position of the target on the previous trial.

It has been suggested that random guesses made by ‘undecided’ observers on difficult trials in forced-choice tasks could be a source of bias in psychophysical data (Garcia-Perez & Alcalá-Quintana, 2010). We did not explicitly model such indecision responses. However, if observers did make guesses which are temporally correlated, this could be captured by our history features, whereas a bias resulting from temporally uncorrelated guesses would affect our intercept term.

All parameters of the model were estimated from data using log-likelihood maximization (see below and Appendix A3 for details). In synthetic data sets (which satisfied our modelling assumptions) the model correctly identified the presence or absence and the magnitude of history dependence (see Appendix Figures A2-A5).

Parametrization of model and model-fitting

We modelled the influence of the previous seven responses and stimulus-identities. To avoid having to fit 14 parameters, we only considered history features that could be described by a linear filtering process

$$h_{kt} = \sum_{t'=1}^7 b_{kt'} y_{t-t'},$$

where $y \in \{z, r\}$, and the $b_{kt'} = \eta_k^{t'-1}$ denote three exponentially decaying filter kernels with decay constants $\eta_k \in \{0, 1/2, 1/4\}$ which are sensitive to fluctuations at different time scales. Each of these

filters was applied to the responses sequence and to the stimulus sequence. As these basis functions are strongly correlated, they would result in strongly correlated history features, which can result in numerical problems when trying to identify their parameters from data. We therefore orthogonalized the basis functions with respect to each other, i.e. ensured that they are of unit-norm and mutually orthogonal (Paninski, Pillow, & Simoncelli, 2004), resulting in new basis functions b'_k . Our 14-dimensional history feature \mathbf{h}_t thus lives in a 6-dimensional subspace, with its first 3 components given by the projections of the previous stimulus identities onto the basis functions and the last 3 components given by the projections of the previous responses. While our algorithm identifies the coefficients of these basis functions, we report the effective history-filters which can be reconstructed by multiplying the basis functions with their matching coefficients. We find the parameters $\alpha, \delta', \gamma, \lambda$, the sensory threshold ν and the history weights ω by maximizing the log-likelihood of the data under the response probabilities predicted by the model, $L = \sum_t \log P(r_t | \tilde{s}_t, \mathbf{h}_t)$. This likelihood can have multiple local maxima, and there are multiple constraints on parameters (e.g. $0 \leq \lambda \leq 1 - \gamma$) which renders naive gradient-based approaches problematic. We therefore used the Expectation Maximization (EM) algorithm (Bishop, 2006), an iterative algorithm which is guaranteed to find a (local) optimum for mixture models (see Appendix A3 for details). While this algorithm can not guarantee convergence to a global optimum, we have found empirically that, by using a modest number of restarts combined with heuristic starting-values, the algorithm typically converged to parameters that explained our data well, and were close to the true parameters on simulations with known ground truth.

Performance measures, correcting for history effects and statistical tests

Performance of the models was quantified using the log-likelihood of the data under the model, $L = \sum_t \log(P(r_t | h_t, \tilde{s}_t))$, across all trials indexed by t . On each trial, the choice is influenced by the effect of the stimulus $\delta_{\text{stim}}(s_t) = a \cdot u_\nu(\tilde{s}_t)$ as well as the effect of the history, $\delta_{\text{hist}}(h_t) := \delta(\mathbf{h}_t)$ (equation (2)). We thus quantified the relative influence of the history as the ratio between the variance of the history-influence and the sum of the variances, $\text{HistCont} = \frac{\text{Var}_t(\delta_{\text{hist}}(\mathbf{h}_t))}{\text{Var}_t(\delta_{\text{hist}}(\mathbf{h}_t)) + \text{Var}_t(\delta_{\text{stim}}(\tilde{s}_t))} 100\%$, where Var_t indicates that the variance is determined across all trials t . This measure quantifies to what extent fluctuations in the internal decision variable can be attributed to history, and we therefore refer to this measure as ‘‘history contribution to variance in the decision variable’’. This measure quantifies the *relative* contribution of the stimulus and the experimental history to fluctuations of the decision variable, but does not model additional noise in the decision process. Thus, if the stimulus intensity is 0, the internal variance of \tilde{s} will be zero as well and the history contribution to variance in the decision variable will be 100%. We excluded blocks with performance $< 55\%$.

To provide an *absolute* measure of the influence of history dependence on behavioural choice, we computed the accuracy of different models in predicting behaviour: On every trial t , the model provides two probabilities $P(r_t = 1)$ and $P(r_t = -1)$. In order to quantify how well a model predicted behavioural choice, we said that the model predicted $r_t = 1$ whenever $P(r_t = 1) > P(r_t = -1)$, i.e. in which $P(r_t = 1) > 0.5$, and said that the model predicted $r_t = -1$ otherwise. The percentage of correct predictions was calculated by counting the number of correct predictions. For the ‘history only’ model, the parameter α was set to 0. To correct the psychometric function for the effect of errors attributable to serial dependencies, we first fitted the full model (equation 2) to data, and then extracted a psychometric function by setting the history-couplings ω to 0. This model was then compared to a conventional psychometric function (equation 1). Confidence intervals for history-kernels were determined using a bootstrap procedure. After the values of the history features on each trial had been calculated, 2000 bootstrap data-sets of the same size were sampled from the data with replacement (Efron & Tibshirani, 1993). Confidence intervals were defined as the 2.5th and 97.5th percentiles of this distribution. As the full model has 6 more parameters than the history-free model, bootstrap samples can not be used to evaluate the statistical significance of history features (as, on each bootstrap set, the full model would have a higher

likelihood than the history-free one). We therefore used a permutation test: We permuted the sequence of trials randomly such that the history features could contain no information about the response, but that the association between stimulus and response was left intact. Thus, the permuted data-sets yielded an approximation of the distribution of likelihoods that would be expected under the null-hypothesis of no serial dependence. We fit the model to 2000 permutations of the original data-sets, and compared performance measures and other parameters to the 95th percentile of this permutation distribution. In cases in which the effect of history dependence is nonlinear, our model will capture a linear approximation to this nonlinear system, and will detect the presence of history dependence provided that its linear term is not negligible. Of course, weak history dependence on small data-sets might not reach the level of statistical significance and might therefore not be detected by the statistical tests described here.

Details of psychophysical data

All psychophysical experiments analyzed in this study were conducted in accordance with the regulations of the relevant institution (Max Planck Institute Tübingen and TU Berlin). We analysed data from four different experiments in which one of the co-authors was actively involved in (FAW); all data were considered “clean” by conventional analyses, i.e. showed no obvious signs of learning, fatigue, or equipment failure as determined by criteria outlined elsewhere (Wichmann & Hill, 2001).

Five observers had to detect luminance increments in different fields of Adelson’s checkerboard illusion (Maertens & Wichmann, 2012) (for a description of the paradigm (Maertens & Wichmann, 2013)). Each observer performed between 2 448 and 3 204 temporal 2-AFC trials and received only overall feedback after blocks of 36 trials, i.e. the average performance over 36 trials. Each 36-trial block contained 18 trials for each of the 2-AFC stimulus orderings, in a randomly shuffled trial order. As our permutation test applied to this data would also have a fixed number of alternatives in the null-distribution, it will also work correctly if blocks are constructed by shuffling rather than by sampling from a binomial distribution. In total, 14 256 trials from this experiment were analysed.

Five observers participated in a temporal 2-AFC plaid masking experiment (Wiebel & Wichmann, 2007); observers received auditory trial-by-trial feedback about whether or not their response was correct. After practice sessions, the observers performed between 3 255 and 4 100 trials. In these trials, the signal stimulus was randomly assigned to one of the two intervals, independently of all other trials. In total, 18 305 trials from this experiment were analysed.

Six observers performed a single-interval (yes-no) auditory tone in noise detection experiment. Each observer performed between 25 to 33 sessions. The initial 7 to 12 sessions were used for training and only the remaining 18 to 22 sessions were analysed. Data were collected in blocks of 50 to 60 trials. If a particular trial contained a signal or not was independent of all other trials. Observers received single-trial feedback for the first 10 trials of each block, and we therefore discarded these trials. For each observer, between 20 610 and 22 800 trials were analysed resulting in a total of 127 430 trials (Schönfelder & Wichmann, 2011).

Finally, we analysed data from a study by Jäkel and Wichmann (Jäkel & Wichmann, 2006). From this study, we analysed data from six observers, five naive at the beginning of the experiment and one highly trained observer. For each observer, we analysed between 2 640 and 5 665 spatial 2-AFC trials, i.e. a total of 27 060 trials. The signal was randomly presented either left or right of the fixation cross. The assignment of the signal to the left or right position did not depend on signal positions in previous trials.

Results

History dependence in low-level visual psychophysical task

First, we show an example of a human observer whose decisions are influenced by the recent experimental history. We analysed data from a two-interval forced-choice experiment in which a human observer had to decide as to whether a luminance increment appeared in the first or the second of two temporally separated presentations of Adelson’s checker board (Maertens & Wichmann, 2012)(see Methods for details). Performance in the task showed a clear and clean dependence on stimulus intensity (Fig. 1 a), as intended by the experimenter. Nevertheless, our model which accounts for history dependence provided a much better explanation of the data than a conventional model which assumes independent trials: the log-likelihood of our model ($l_{\text{full}} = -911.7$) was substantially larger than the one of the independent-trial model ($l_0 = -941.2$, corrected for difference in parameters using Akaike’s information criterion, AIC), and a permutation test against random trial sequences revealed that this performance-benefit was significant at level $p < 0.0005$ (Fig. 1 b, see section “Performance measures, correcting for history effects and statistical tests” for details, all p-values in the following are derived from this test unless stated otherwise).

To show that the effect of previous trials was not only statistically significant, but determined behaviour noticeably, we calculated how much of the trial-by-trial variability in the observer’s decision variable could be explained by history (see Methods for details). This quantity (history contribution to variance in the decision variable) would be 0% if trials were independent, and 100% if observers were exclusively influenced by the experimental history (It would also be trivially be 100% in tasks or conditions in which there is no signal). The influence of history decreases monotonically with stimulus intensity (Fig. 1 c). On difficult stimuli (for which the performance of the observer was between 55% and 75% correct) 32% of the variance of the observer’s decision variable depended on previous trials, rather than the current stimulus ($p < 0.0005$, Fig. 1 d). As expected, behaviour was primarily explained by the stimulus on easy trials (for which performance $> 75\%$ correct), on which 98% of the variance of the observer’s decision variable was explained by the stimulus, and only 1.6% of the variance of the observer’s decision variable was still history dependent ($p < 0.0005$, Fig. 1 e).

To elucidate the ability of the stimulus and experimental history to predict behaviour, we quantified on how many trials the different models correctly predicted the subjects’ decision: The prediction performance of the history-only model (i.e. for which the stimulus weight is set to 0) decreased with stimulus intensity (Appendix A6). For difficult stimuli, the full model (i.e. stimulus and history) predicted behaviour with 64% accuracy (Fig. 1 f), whereas the stimulus only achieved 63% and the history alone achieved 60% ($p < 0.0005$, chance level is 50%, Fig. 1 g). Our analysis shows that, on difficult stimuli which are of relevance for measuring psychophysical performance, perceptual decisions of this observer are almost as strongly influenced by internal decision variable-fluctuations induced by experimental history as they are by the experimental stimulus. Nevertheless, the fact that the prediction power was far away from 100% implies that much of the variability in subjects’ responses could neither be accounted for by the stimulus or experimental history.

Even on easy stimuli prediction performance of the history-only model was statistically significant, at 54% ($p < 0.03$, Fig. 1 h).

The additional benefit of the ‘full’ model over the stimulus only model might seem surprisingly small. This might be interpreted as indicating that the stimulus and the history signal were very correlated. However, this was not the case: For this observer, less than 0.05% of the stimulus component could be explained from the history component (maximum across observers was 0.11%). To better understand this effect and to assess whether our results about history contribution to variance in the decision variable were consistent with these measures of prediction performance, we investigated a simplified model in which the stimulus, the history-influence and the internal noise were approximated by independent Gaussian random variables. The variances of stimulus and history were chosen to be consistent with our results on

history contribution to variance in the decision variable, the variance of the internal noise distribution as adapted to match the observers performance on difficult trials. In this model, the predictive power of the history-only model was 59%, of the stimulus-only model 63%, and of the combined model 65%, which is in agreement with our empirical findings. On average across observers, this simplified model explained prediction accuracies with an error of 3%. Therefore, the small gain of the combined model relative to the other two models is consistent with our other findings, and in particular does not imply that the history is correlated with the current stimulus.

Effect of history dependence on psychometric functions

If observers are influenced by task-irrelevant information in previous trials, then this could lead to sub-optimal performance in the task. Hence, estimates of the capabilities of sensory processing systems could be negatively biased by the dependence of perceptual decisions on internal states. Our model allows us to provide a statistical description of psychophysical data which separates the effects of experimental history and the stimulus on the decision and, thus, it leads to a 'decontaminated' estimate of the psychometric function. We emphasize that this 'discounting' of history-induced errors yields a model-based estimate of psychophysical performance, which could be wrong if the modelling assumptions were inappropriate. Our model gives two generally different psychometric functions for the two response intervals (Ulrich & Vorberg, 2009; Garcia-Perez & Alcala-Quintana, 2011). We report the average of these two psychometric functions here because both psychometric functions were typically similar and our interest is on trial-by-trial effects rather than biases that might emerge within an individual trial.

After discounting errors induced by history, the predicted probabilities of correct responses differed from a conventional psychometric function (Fig. 2 a-b), i.e. some of the observers' errors could be attributed to the influence of task-irrelevant features. We quantified the impact of the history on the psychometric function by comparing the stimulus level at which our model or the independent-trial model predicted a performance of 85% correct. Accounting for history dependence yielded discrimination thresholds that were reduced to 96% of the original value, i.e. a reduction of at most 4% ($p < 0.0005$, average across conditions 97.2%, Fig. 2 c). Thus, the fact that conventional analyses can not discount the errors induced by dependence on previous trials leads to a slight underestimation of the stimulus sensitivity of this observer.

On the one hand, experimental history is clearly influences behaviour, but on the other hand, it only seems to have a weak effect on psychophysical thresholds. To better understand this apparent paradox, we considered a second simplified, analytically tractable model (see Appendix A6 for details). We note that, if the psychometric function were perfectly linear, even strong history-dependence would not lead to *any* degradation in performance, as stimulus-independent lapses in one direction and in the other direction would cancel perfectly. However, the psychometric function is a nonlinear model. We found that a history-induced *standard deviation* of $\sigma = \text{Std}(\delta_{\text{hist}}(h_t))$ leads to a quadratic reduction in the slope of the psychometric function, i.e. one which is proportional to the history-induced *variance* σ^2 . Thus, for typical history dependence ($\sigma < 1$), the reduction in the slope is even smaller (as $\sigma^2 < \sigma$). Furthermore, the proportionality factor of this relationship is also small, namely $\pi/16 \approx 0.2$. For example, for a standard deviation of $\sigma = 0.3$ one would only expect a reduction of the slope by $0.09 \times \pi/16 \approx 1.18\%$.

This relationship thus explains why even substantial history dependence only has a weak effect on the slope of the psychometric function (and hence thresholds) (Verplanck & Blough, 1958). Conversely, including history dependence in the model should lead to a steeper psychometric function, and thus lower thresholds. For our example observer (for which $\sigma = 0.52$), this simplified analysis predicted a reduction of threshold by 5.3%, which slightly over-estimates the empirically measured reduction of 4%. Across observers, the threshold changes predicted by this simplified model were correlated with those of the full model ($c=0.73$) and did not differ in their mean level (paired samples t -test, $t(21) = 0.57$, $p = 0.57$).

Serial dependencies across multiple observers and paradigms

To demonstrate that the results above are not an idiosyncrasy of this particular observer or task, we analysed data from 22 human observers collected in 4 different experimental paradigms, resulting in a substantial data-set of 187 051 trials in total. In all cases, observers were engaged in low-level psychophysical experiments and the data were considered “clean” by means of conventional analyses (Wichmann & Hill, 2001). However, stimulus material and task varied widely across the most commonly used experimental paradigms, including visual two alternative forced-choice tasks with the alternatives separated spatially and temporally and an auditory single interval task. In experiments with multiple experimental conditions, we conservatively assumed the structure of history dependence to be fixed across conditions. Thus, if observers have condition-dependent history dependence, our results would underestimate the true magnitude of history dependence.

We found significant history-dependence ($p < 0.05$) in 19 out of 22 observers, and modelling the non-stimulus determinants of the behavioural choices led to an average increase in log-likelihood of 0.009 ± 0.0020 per trial (s.e.m. across observers, Fig. 3 a, see Appendix A7-A12 for detailed results of two further observers). In other words, a data-set of 500 trials would be (on average) 77 times more likely under our model than under a conventional model assuming independent trials. Significant history dependence was found in all four experimental paradigms we investigated (see Fig. 3 a), and its strength varied considerably across observers.

On average, $13.7\% \pm 2.4\%$ (s.e.m.) of variance of the decision variable on difficult stimuli was determined by the experimental history (Fig. 3 b), and not by the presented stimulus, with values for individual observers as high as 48.2%. Experimental history was a meaningful predictor of behavioural choices on individual trials. On average, the model based on previous trials predicted $56.5\% \pm 1.0\%$ (s.e.m., chance level 50%, significant for 16 out of 22 observers) of responses on difficult stimuli correctly, compared to $64.8\% \pm 0.5\%$ for the stimulus and $65.5\% \pm 0.4\%$ for the combined model (Fig. 3 c). Across observers, the prediction performance of the full model was significantly better than for a model that only contained stimulus terms ($p < 10^{-3}$, permutation test of paired differences). For one observer, the history was in fact a better predictor of the behavioural choice than the presented stimulus.

We emphasize that the prediction accuracy of the history model was on par with the accuracy levels reported in many decoding studies which predict behavioural choice from functional imaging measurements (e.g. (Soon, Brass, Heinze, & Haynes, 2008))— thus, while the absolute numbers seem low, they could potentially be big enough to lead to confounds in the analysis of imaging data (Lages & Jaworska, 2012). As expected, performance on easy stimuli was largely driven by the stimulus, and previous trials explained only $1.4\% \pm 0.3\%$ of the variance of the decision variable, and predicted choices at $52.7\% \pm 0.5\%$ correct. We also note that the variance across observers is substantial, i.e. that there are large inter-personal differences in how strongly observers are affected by experimental history.

Across observers, the thresholds that were estimated by the model with history terms were $97.0\% \pm 0.5\%$ of those obtained without history (significant for 17 out of 22 observers). In other words, the observers sensitivity was on average 3.0% higher if errors caused by a reliance on task irrelevant information in previous trials were discounted. Thus, history dependence leads to a small, yet systematic underestimation of perceptual sensitivity. Nevertheless, in most analysis questions, such small differences in estimated thresholds will be negligible. The upper asymptotic performance was not systematically closer to 100% if history was modelled (reduction in asymptotic error rate by $0.2\% \pm 0.2\%$).

Finally, we also tested our modelling assumption that history-dependence would primarily lead to *shifts* in the psychometric functions, and not changes in the slope. We fit four conventional psychometric functions separately to trials that followed a left-response and trials that followed a right-response, as well as trials that followed a left stimulus and trials that followed a right stimulus. Differences in horizontal shifts were on average 2.3 times larger than differences in slopes (after normalizing with the estimates’ standard errors). In 35 out of 44 cases, the effect of the previous trial on the horizontal position of the psychometric function was larger than the effect on the slope of the psychometric function ($p = 0.0001$

binomial test).

What properties of the recent experimental history were predictive of behaviour? To answer this question, we visualized the filtering-kernels which capture how previous responses and previous stimulus identities (i.e. target locations) influence the decision on the current trial. The exemplary observer from the luminance-experiment was mostly sensitive to responses on the most recent trial, and showed a tendency to avoid previous choices, as evidenced by the predominantly negative filter-weights (Fig. 4 a). In contrast, the previous stimulus identities did not have a significant effect on decisions. This finding is in accordance with the fact that this observer did not receive trial-by-trial feedback about the identity of the preceding stimulus. Figure 4 b displays response kernels for all observers, and shows that the strongest influence was from the most recent trial (55.6% of the history contribution to variance in the decision variable was explained by the previous trial, all kernels are shown in appendix A8). Nevertheless, there was a significant influence of longer trial-lags, which contributed a log-likelihood gain of 0.0013 ± 0.004 (s.e.m.) per trial across all datasets (corrected using AIC, $p = 0.00001$ binomial test).

The combination of the filter weights associated with the previous stimulus or response can be viewed as the behavioural strategies employed by different observers (Figure 4 c). Negative filter weights associated with the previous response indicate a tendency to switch, and negative filter weights for the previous stimulus identity a tendency to avoid the previous target. The top half of the plot contains a continuum of ‘win stay-lose switch’ strategies, i.e. signify a tendency to repeat a response if it was successful and to otherwise alternate the response. Interactions between these effects also give rise to more interesting strategies: For example, along the upper part of the negative diagonal, the weights for previous stimulus and the previous response have same magnitude but opposite sign— thus, these effects cancel after correct responses (on which the previous stimulus equals the previous response) but they lead to history effects after incorrect responses. Observers tend to switch their responses after errors—a strategy that could be called ‘lose-switch’. A similar reasoning can be applied to the other diagonals as well.

The observers in the luminance experiment consistently associated negative weights to previous responses (average -1.13 ± 0.38 , s.e.m.), reflecting a tendency to switch responses from trial to trial. Weights associated with previous stimuli were heterogeneous and on average a factor of 3.85 ± 1.98 times smaller than the response weights. In contrast, most observers from two masking experiments (plaid mask in the first experiment and a sine grating mask of the same spatial frequency and orientation in the second experiment) are near the upper left diagonal of this plot, reflecting the fact that these observers primarily changed their response criterion after errors, which is likely to be a consequence of the fact that they received trial-by-trial feedback. In contrast, subjects in a yes-no audio experiment had weak weights associated with the previous stimulus (average weight -0.11) and stronger weights associated with previous responses (average weight 0.33). Thus, these observers showed slow fluctuations of their decision variable, which manifests itself in a tendency to repeat their previous responses. While the overall influence for longer trial-lags was substantially weaker, a clear clustering of weights according to experimental paradigm was still evident (Fig. 4 d).

We note that these filtering-kernels are not equivalent to cross-correlations between current and previous responses. For example, a direct effect which is confined to the previous trial would lead to cross-correlations even at (in theory) infinite time-lag, with an exponentially decaying strength. Similarly, and in contrast to correlations, these filter-kernels allow us to quantify whether the response of the subject is dependent on previous responses or previous stimulus-identities. For example, for our example observer, the cross-correlation between previous stimulus-identities and the current response is non-zero (as previous stimulus identities are correlated with previous responses are correlated with the current response), despite the fact that our analysis reveals that their *direct* influence is negligible. Thus, with our method, we can disentangle serial dependence on previous responses from a dependence on previous stimuli or feedback.

Discussion

Our results demonstrate that there exist significant and systematic causal determinants of perceptual decisions over and above the stimulus presented to the organism. Great care must thus be exercised when attempting to infer mechanisms from behavioural data, unless the unwanted, non-stimulus dependent internal factors are taken into consideration. In particular, we showed that the predictive power of experimental history was similar to values reported in studies predicting behavioural choices from functional imaging measurements (Soon et al., 2008; Lages & Jaworska, 2012). After correcting for the effect of history-dependence, observers were slightly more sensitive to the stimulus (i.e. had lower thresholds) than determined by conventional analyses. Our experiments were conducted with experienced observers and in tightly controlled settings. In addition, our model assumes the effect of history to be linear. If the true history dependence in the data is nonlinear, our model will only capture the linear kernel of the nonlinear system. It is therefore likely that we have explored the lower bound of serial dependence in behavioural studies. In addition, these effects might well be more pronounced for untrained observers, patients, animals or in other experimental designs, e.g. those employing adaptive procedures (Treutwein, 1995). Therefore, it might be unwarranted to extrapolate our findings or previous reports to other experimental settings.

The existence of serial dependencies has long been known (Verplanck et al., 1952; Senders & Sowards, 1952; Howarth & Bulmer, 1956). Although these dependencies have a modest influence on aggregated measures of performance (see also (Senders & Sowards, 1952; Verplanck & Blough, 1958)), trial-by-trial effects are consistent in individual trials. Our results show that, even for low-level perceptual tasks, perceptual decision making can not be modelled as being based on a feed-forward system which passively responds to external stimuli. As observers were strongly influenced by their previous responses—and not by previous stimuli—adaptation is unlikely to be an explanation for the observed inter-trial dependence found here: If the response patterns could be explained by adaptation, we would expect observers to avoid the response associated with previous stimuli. This should result in negative history weights associated with previous stimuli. Yet, for most observers, history weights associated with previous stimuli are very close to zero. Similarly, the fact that history dependence was also observed in experiments without trial-by-trial feedback makes it unlikely that either post-error dynamics (Goldfarb et al., 2012) or models of reinforcement learning (Dayan & Niv, 2008) could explain our results. In contrast, a more likely explanation for our results would be that human observers combine information from the stimulus with their prior beliefs about where they expect the next target. If they have an incorrect model of what constitutes a random sequence of events (Bar-Hillel & Wagenaar, 1991), and combine their single-trial expectations with stimulus information in a Bayesian fashion (Körding & Wolpert, 2004), then this would lead to precisely the type of sequential dependencies observed here.

Neglecting sequential dependencies can be problematic for four reasons: First, as we showed above, inter-trial dependencies can lead to small but systematic biases in characterizations of the performance limits of a sensory system. In addition, these biases might be bigger for behaving animals or other experimental paradigms. Explicitly analyzing sequential dependencies can help identifying these biases if they occur. Second, by analysing dependencies across trials, researchers might obtain interesting information from psychophysical data. For example, Chopin and Mamassian (Chopin & Mamassian, 2012) analysed sequential dependencies to draw conclusions about adaptive processing of orientation which would be inaccessible with averaged data. Third, statistical techniques for system identification have gained popularity in the perceptual sciences over the last years (Ahumada & Lovell, 1971; Kienzle, Franz, Schölkopf, & Wichmann, 2009; Macke & Wichmann, 2010). Therefore, it is conceivable that the application of system identification techniques to data contaminated by serial dependencies may recover wrong features or may show a disappointingly low correlation between inferred features and behaviour. Fourth—and arguably most importantly—many studies relating behavioural choices to measurements of neural activity or their correlates are dependent on the assumption that the perceptual choice is determined by the current stimulus (Nienborg & Cumming, 2009; Soon et al., 2008; Fründ, Busch,

Schadow, Körner, & Herrmann, 2007; Busch, Dubois, & VanRullen, 2009; Tong & Pratte, 2012), and could be led astray if internal states are a strong determinant of these choices. Similarly, the fact that behavioural choices can be predicted from neural measurements could be a consequence of sequential dependencies combined with a dependence of neural activity on previous choices (Lages & Jaworska, 2012).

Previous studies have reported history kernels for human reaction times (Maljkovic & Nakayama, 1994; Maljkovic & Martini, 2005; Martini, 2010), binary responses in animals (Lau & Glimcher, 2005; G. S. Corrado et al., 2005; Busse et al., 2011), and human categorization judgments (Stewart et al., 2005) that share qualitative similarities with the ones presented here. Furthermore, if effects in earlier studies had been expressed as kernels, these might also have supported this perspective (Senders & Sowards, 1952; Treisman & Williams, 1984). Nevertheless, given that these kernels were estimated in very different experiments or even using different measures, they are likely to be caused by different underlying mechanisms. Kernels estimated in studies of reaction times typically indicate that repetition of the same response becomes faster (Maljkovic & Nakayama, 1994; Maljkovic & Martini, 2005; Martini, 2010). Kernels derived describing an animal’s reward expectation reveal a tendency to expect reward in the same position (Lau & Glimcher, 2005; G. S. Corrado et al., 2005), or a combination of reward expectancy and sensory processes (Busse et al., 2011). In contrast, we note that studies of sequential dependence in “purely” perceptual processing—with two stimuli that are themselves difficult to discriminate—have typically not derived response kernels (Senders & Sowards, 1952; Verplanck et al., 1952). In addition, the classical studies of these effects (Senders & Sowards, 1952; Verplanck et al., 1952) used long sequences of stimuli with the same intensity and no intervening blank stimuli in single interval designs, a design that has been rarely used since the introduction of signal detection theory (Green & Swets, 1966). Some authors have explained sequential dependencies by means of a temporally varying criterion (Treisman & Williams, 1984; Lages & Treisman, 1998). We find consistent sequential dependencies in many different experimental designs, including forced choice designs, and these dependencies have a systematic—albeit small—effect on performance. Although sequential dependencies in different experiments have very different meanings and mechanisms, the statistical framework that we present can describe all of the history effects that refer to binary responses. Thus, our framework allows for a unified quantification and comparison of history effects even if they have very different psychological interpretations.

Previous studies typically found that the experimental history mainly shifted psychometric functions horizontally (Hock et al., 1993; Lages & Treisman, 1998, 2010; Lages & M., 2010). Therefore, our model resembles previous models that also described sequential dependencies as a form of trial by trial response bias rather than variations in an observer’s sensitivity to the stimulus (Busse et al., 2011; G. S. Corrado et al., 2005; Raviv et al., 2012; Bode et al., 2012; Goldfarb et al., 2012; Green et al., 1977, 1980; Lockhead & King, 1983; Ward, 1979). (Note that we here refer to psychometric functions which relate the stimulus intensity to the probability of a given binary response, and *not* to the probability of a *correct* response). Yet, it is possible that history has an influence on the slope of the psychometric function as well (see (Lages & M., 2010) for some indication of this). Our model is not able to capture these kinds of sequential dependencies and thus, the history dependence reported here should be treated as a lower bound.

The statistical methodology we presented makes it possible to quantify the strength of inter-trial dependence and to correct psychometric functions or other estimates of behavioural performance for the systematic influence of previous trials. Our framework makes it possible to track fluctuations of the internal decision variables on *individual* choices (G. Corrado & Doya, 2007; O’Doherty et al., 2007; Lau & Glimcher, 2005), and thus has the potential to reveal a rich source of information which had previously been buried by trial averaging. Given that time-series of behavioural observations are ubiquitous in neuroscience and related fields, our methods will be applicable to a wide range of experimental or clinical paradigms which measure human or animal performance. Combined with methods for single-trial analyses for neurophysiological recordings (Churchland, Yu, Sahani, & Shenoy, 2007), they thus have the potential to contribute to a more realistic understanding of perceptual decision making.

Acknowledgments

We thank M. Sahani for his support and important discussions during early stages of the project, F. Jäkel, M. Maertens, V. Schönfelder and C. Wiebel for sharing their data with us, M. Lages, D. Laming, S. Liebe, H. Nienborg and E. Simoncelli for fruitful discussions as well as S. Liebe, M. Maertens, T. Otto and V. Schönfelder for comments on the manuscript. This work was supported by the German Research Foundation (Sachbeihilfe FR 2854/1-1 to IF and FAW), by the Gatsby Charitable Foundation (JHM), an EU Marie Curie Fellowship (to JHM and M. Sahani) and the German Federal Ministry of Education and Research (BMBF) through the Bernstein Computational Neuroscience Programs Berlin (FKZ: 01GQ0414) and Tübingen (FKZ: 01GQ1002).

Supplementary material

Appendices A1 to A8 can be found in the Supplementary Information file.

References

- Ahumada, A. J., & Lovell, J. (1971). Stimulus Features in Signal Detection. *The Journal of the Acoustical Society of America*, *49*(6 (Part 2)), 1751–1756.
- Baird, J. C., Green, D. M., & Luce, R. D. (1980). Variability and Sequential Effects in Cross-Modality Matching in Area and Loudness. *Journal of Experimental Psychology: Human Perception and Performance*, *6*(2), 277–289.
- Bar-Hillel, M., & Wagenaar, W. A. (1991). The Perception of Randomness. *Advances in Applied Mathematics*, *12*, 428–454.
- Bishop, C. M. (2006). *Pattern recognition and machine learning*. New York: Springer.
- Blackwell, H. R. (1952). Studies of psychophysical methods for measuring visual thresholds. *Journal of the Optical Society of America*, *42*, 606–616.
- Bode, S., Sewell, D. K., Lilburn, S., Forte, J. D., Smith, P. L., & Stahl, J. (2012). Predicting Perceptual Biases from Early Brain Activity. *Journal of Neuroscience*, *32*(36), 12488–12498.
- Busch, N. A., Dubois, J., & VanRullen, R. (2009). The phase of ongoing EEG oscillations predicts visual perception. *Journal of Neuroscience*, *29*(24), 7869–7876.
- Busse, L., Ayaz, A., Dhruv, N. T., Katzner, S., Saleem, A. B., Schölvinck, M. L., ... Carandini, M. (2011). The Detection of Visual Contrast in the Behaving Mouse. *Journal of Neuroscience*, *31*(31), 11351–11361.
- Chopin, A., & Mamassian, P. (2012). Predictive properties of adaptation. *Current Biology*, *22*(7), 622–626.
- Churchland, M. M., Yu, B. M., Sahani, M., & Shenoy, K. (2007). Techniques for extracting single-trial activity patterns from large-scale neural recordings. *Current Opinion in Neurobiology*, *17*(5), 609–618.
- Corrado, G., & Doya, K. (2007). Understanding Neural Coding through the Model-Based Analysis of Decision Making. *Journal of Neuroscience*, *27*(31), 8178–8180.
- Corrado, G. S., Sugrue, L. P., Seung, H. S., & Newsome, W. T. (2005). Linear-Nonlinear-Poisson Models of Primate Choice Dynamics. *Journal of the Experimental Analysis of Behavior*, *84*(3), 581–617.
- Dayan, P., & Niv, Y. (2008). Reinforcement learning: The Good, The Bad and The Ugly. *Current Opinion in Neurobiology*, *18*, 185–196.
- Efron, B., & Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. Boca Raton: Chapman & Hall.
- Fründ, I., Busch, N., Schadow, J., Körner, U., & Herrmann, C. (2007). From perception to action: phase-locked gamma oscillations correlate with reaction times in a speeded response task. *BMC Neuroscience*, *8*(1), 27.

- Garcia-Perez, M., & Alcala-Quintana, R. (2010). The difference model with guessing explains interval bias in two-alternative forced-choice detection procedures. *Journal of Sensory Studies*, *25*(6), 876-898.
- Garcia-Perez, M., & Alcala-Quintana, R. (2011). Improving the estimation of psychometric functions in 2afc discrimination tasks. *Frontiers in Psychology*, *2*(96).
- Goldfarb, S., Wong-Lin, K., Schwemmer, M., Leonard, N. E., & Holmes, P. (2012). Can post-error dynamics explain sequential reaction time patterns? *Frontiers in Psychology*, *3*(213), 1-16.
- Green, D. M. (1964). Consistency of auditory detection judgments. *Psychological Review*, *71*(5), 392-407.
- Green, D. M., Luce, R. D., & Duncan, J. E. (1977). Variability and sequential effects in magnitude production and estimation of auditory intensity. *Perception & Psychophysics*, *22*(5), 450-456.
- Green, D. M., Luce, R. D., & Smith, A. F. (1980). Individual magnitude estimates for various distributions of signal intensity. *Perception & Psychophysics*, *27*(6), 483-488.
- Green, D. M., & Swets, J. A. (1966). *Signal Detection Theory and Psychophysics*. New York: Wiley.
- Hock, H. S., Kelso, J. A. S., & Schöner, G. (1993). Bistability and hysteresis in the organization of apparent motion patterns. *Journal of Experimental Psychology: Human Perception and Performance*, *19*(1), 63-80.
- Howarth, C. I., & Bulmer, M. G. (1956). Non-random sequences in visual threshold experiments. *Quarterly Journal of Experimental Psychology*, *8*(4), 163-171.
- 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.
- Kienzle, W., Franz, M., Schölkopf, B., & Wichmann, F. (2009). Center-surround patterns emerge as optimal predictors for human saccade targets. *Journal of Vision*, *9*(5), 7:1-15.
- Körding, K. P., & Wolpert, D. M. (2004, Jan). Bayesian integration in sensorimotor learning. *Nature*, *427*(6971), 244-7. doi: 10.1038/nature02169
- Kuss, M., Jäkel, F., & Wichmann, F. A. (2005). Bayesian inference for psychometric functions. *Journal of Vision*, *5*, 478-492.
- Lages, M., & Jaworska, K. (2012). How predictable are "spontaneous decisions" and "hidden intentions"? Comparing classification results based on previous responses with multivariate pattern analysis of fMRI BOLD signals. *Frontiers in Psychology*, *6*(56).
- Lages, M., & M., T. (2010). Sensory integration across modalities: how kinaesthesia integrates with vision in visual orientation discrimination. *Seeing and Perceiving*, *23*(5), 435-462.
- Lages, M., & Treisman, M. (1998). Spatial Frequency Discrimination: Visual Long-term Memory or Criterion Setting? *Vision Research*, *38*(4), 557-572.
- Lages, M., & Treisman, M. (2010). A Criterion Setting Theory of Discrimination Learning that Accounts for Anisotropies and Context Effects. *Seeing and Perceiving*, *23*, 401-434.
- Lau, B., & Glimcher, P. W. (2005). Dynamic Response-By-Response Models Of Matching Behavior in Rhesus Monkeys. *Journal of the Experimental Analysis of Behavior*, *84*, 555-579.
- Lockhead, G. R., & King, M. C. (1983). A Memory Model of Sequential Effects in Scaling Tasks. *Journal of Experimental Psychology: Human Perception and Performance*, *9*(3), 461-473.
- Macke, J. H., & Wichmann, F. A. (2010). Estimating predictive stimulus features from psychophysical data: The decision image technique applied to human faces. *Journal of Vision*, *10*(5), 22:1-24.
- Maertens, M., & Wichmann, F. A. (2012). Color and light: lightness and brightness: On the relationship between luminance increment thresholds and apparent brightness. In *Vision Sciences Society* (p. 66).
- Maertens, M., & Wichmann, F. A. (2013). When luminance increment thresholds depend on apparent lightness. *Journal of Vision*, *13*(6).
- Maljkovic, V., & Martini, P. (2005). Implicit short-term memory and event frequency effects in visual search. *Vision Research*, *45*, 2831-2846.

- Maljkovic, V., & Nakayama, K. (1994). Priming of pop-out: I. role of features. *Memory & Cognition*, *22*(6), 657–672.
- Maloney, L. T., Dal Martello, M. F., Sahn, C., & Spillmann, L. (2005). Past trials influence perception of ambiguous motion quartets through pattern completion. *Proceedings of the National Academy of Science, USA*, *102*(8), 3164–3169.
- Martini, P. (2010). System identification in Priming of Pop-Out. *Vision Research*, *50*, 2110–2115.
- Meier, P., Flister, E., & Reinagel, P. (2011). Collinear features impair visual detection by rats. *Journal of Vision*, *11*(3), 1–16.
- Mori, S. (1998). Effects of Stimulus Information and Number of Stimuli on Sequential Dependencies in Absolute Identification. *Canadian Journal of Experimental Psychology*, *52*(2), 72–83.
- Mori, S., & Ward, L. M. (1995). Pure feedback effects in absolute identification. *Perception & Psychophysics*, *57*(7), 1065–1079.
- Nienborg, H., & Cumming, B. G. (2009). Decision-related activity in sensory neurons reflects more than a neuron's causal effect. *nature*, *459*, 89–92.
- O'Doherty, J. P., Hampton, A., & Kim, H. (2007). Model-based fMRI and its application to reward learning and decision making. *Annals of the New York Academy of Sciences*, *1104*, 35–53.
- Otto, T. U., & Mamassian, P. (2012). Noise and Correlations in Parallel Perceptual Decision Making. *Current Biology*, *1*. doi: 10.1016/j.cub.2012.05.031
- Paninski, L., Pillow, J. W., & Simoncelli, E. P. (2004). Maximum likelihood estimation of a stochastic integrate-and-fire neural encoding model. *Neural Computation*, *16*, 2533–2561.
- Peeke, S. C., & Stone, G. C. (1972). Sequential effects in two- and four-choice tasks. *Journal of Experimental Psychology*, *92*(1), 111–116.
- Raviv, O., Ahissar, M., & Loewenstein, Y. (2012). How Recent History Affects Perception: The Normative Approach and Its Heuristic Approximation. *Plos Computational Biology*, *8*(10), e1002731.
- Schönfelder, V. H., & Wichmann, F. A. (2011). Extracting Auditory Cues in Tone-in-Noise detection with a Sparse Feature Selection Algorithm. In *BC11 : Computational Neuroscience & Neurotechnology Bernstein Conference & Neurex Annual Meeting*. doi: 10.3389/conf.fncom.2011.53.00074
- Senders, V., & Sowards, A. (1952). Analysis of Response Sequences in the Setting of a psychophysical Experiment. *The American Journal of Psychology*, *65*(3), 358–374.
- Soon, C. S., Brass, M., Heinze, H.-J., & Haynes, J.-D. (2008). Unconscious determinants of free decisions in the human brain. *Nature Neuroscience*, *11*, 543–545.
- Stewart, N., Brown, G. D. A., & Chater, N. (2002). Sequence Effects in Categorization of Simple Perceptual Stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*(1), 3–11.
- Stewart, N., Brown, G. D. A., & Chater, N. (2005). Absolute Identification by Relative Judgment. *Psychological Review*, *112*(4), 881–911.
- Sugrue, L. P., Corrado, G. S., & Newsome, W. T. (2004). Matching Behavior and the Representation of Value in the Parietal Cortex. *science*, *304*(5678), 1782–1787.
- Tong, F., & Pratte, M. S. (2012). Decoding Patterns of Human Brain Activity. *Annual Review of Psychology*, *63*, 483–509.
- Treisman, M., & Williams, T. C. (1984). A Theory of Criterion Setting With an Application to Sequential Dependencies. *Psychological Review*, *91*(1), 68–111.
- Treutwein, B. (1995). Adaptive Psychophysical Procedures. *Vision Research*, *35*(17), 2503–2522.
- Treutwein, B., & Strasburger, H. (1999). Fitting the psychometric function. *Perception & Psychophysics*, *61*(1), 87–106.
- Ulrich, R., & Vorberg, D. (2009). Estimating the difference limen in 2afc tasks: Pitfalls and improved estimators. *Attention, Perception & Psychophysics*, *71*, 1219–1227.
- Verplanck, W. S., & Blough, D. S. (1958). Randomized Stimuli and the Non-Independence of Successive Responses at the Visual Threshold. *Journal of General Psychology*, *59*, 263–272.

- Verplanck, W. S., Collier, G. H., & Cotton, J. W. (1952). Non-independence of successive response in measurements of the visual threshold. *Journal of Experimental Psychology*, *44*, 273–282.
- Wagenmakers, E.-J., Farrell, S., & Ratcliff, R. (2004). Estimation and interpretation of $1/f^\alpha$ noise in human cognition. *Psychonomic Bulletin & Review*, *11*(4), 579–615.
- Ward, L. M. (1979). Stimulus information and sequential dependencies in magnitude estimation and cross-modality matching. *Journal of Experimental Psychology: Human Perception and Performance*, *5*(3), 444–459.
- Wichmann, F. A., & Hill, N. J. (2001). The psychometric function: I. Fitting, sampling, and goodness of fit. *Perception & Psychophysics*, *63*(8), 1293–1313.
- Wiebel, C., & Wichmann, F. A. (2007). Oblique-and plaid-masking re-visited. In K. F. Wender, S. Mecklenbräuker, G. D. Rey, & T. Wehr (Eds.), *Experimentelle Psychologie. Beiträge zur 49. Tagung experimentell arbeitender Psychologen (TEAP)* (p. 153). Pabst Science.
- Wilder, M., Jones, M., & Mozer, M. C. (2010). Sequential effects reflect parallel learning of multiple environmental regularities. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, & C. A. (Eds.), *Advances in neural information processing systems 22* (p. 2053–2061). La Jolla, Ca: NIPS Foundation.
- Yu, A. J., & Cohen, J. D. (2008). Sequential effects: Superstition or rational behavior? In D. Koller, D. Schuurmans, Y. Bengio, & L. Bottou (Eds.), *Advances in neural information processing systems 21* (pp. 1873–1880).

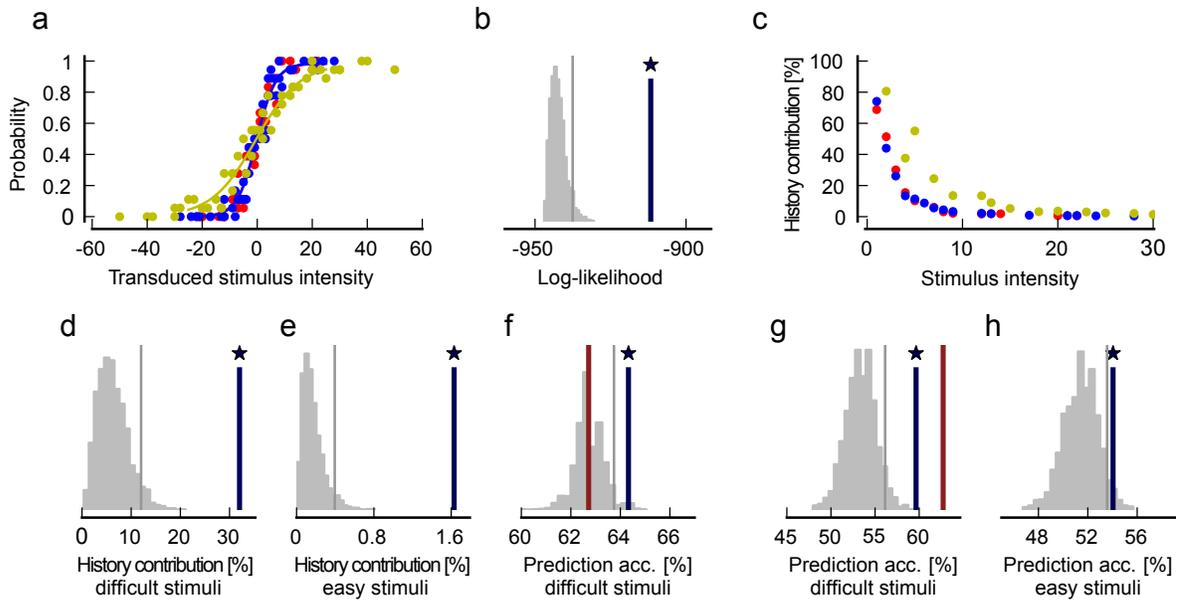


Figure 1. Data from example observer in luminance-experiment. **a)** Psychometric function and proportion of ‘second interval’ responses as a function of transduced stimulus intensity for observer pk, colours correspond to different experimental conditions. **b)** Log-likelihood of the full model (dark blue line). Here and in all panels, the grey histogram is the distribution on permuted data (grey), and vertical grey line marks its 95th percentile, the star marks statistical significance. **c)** History contribution to variance in the decision variable as a function of stimulus intensity. **d)** History contribution to variance in the decision variable 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, dark blue line) in predicting observers’ responses on difficult stimuli, and comparison with stimulus-only model (red line). **g)** Prediction performance of model with only history dependence and no stimulus dependence (dark blue line), and comparison with stimulus-only model (red line). **h)** Prediction performance of history-only model on easy stimuli (dark blue line).

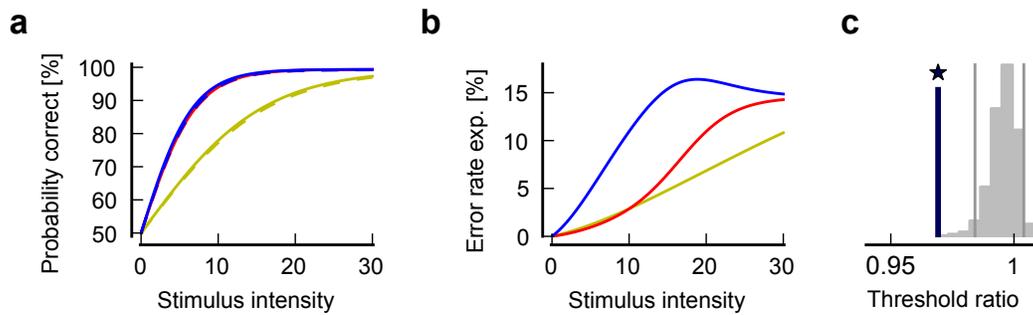


Figure 2. Effect of history dependence on the psychometric function for the example observer. **a)** Psychometric functions indicating frequency of correct responses as a function of stimulus intensity. Dashed lines mark fits of a model without history terms. Colours correspond to different experimental 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). Thresholds are lower for the full model with history terms.

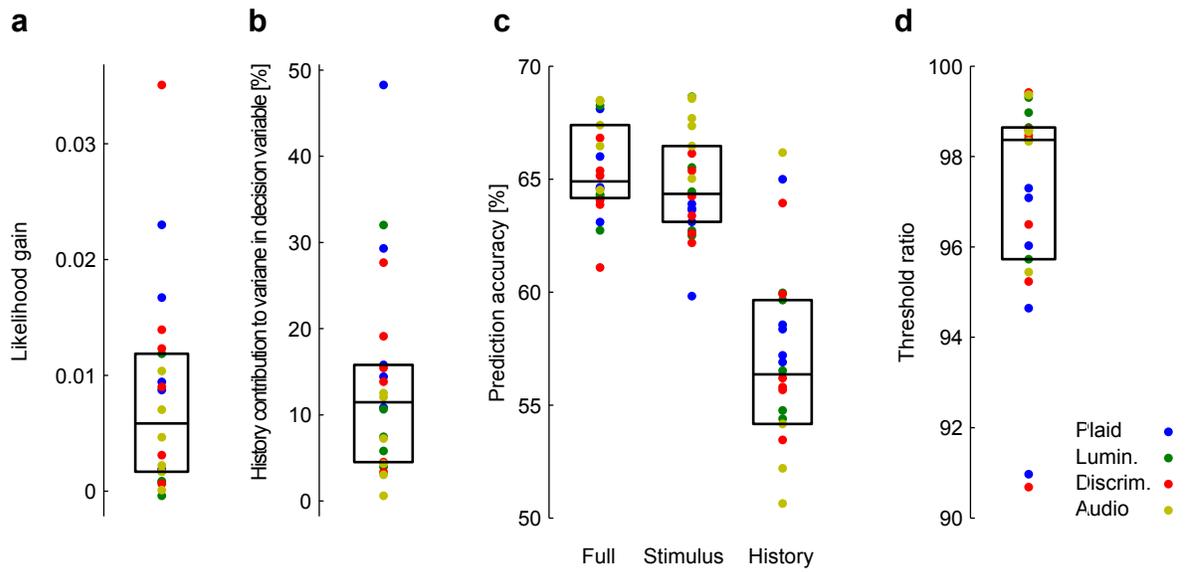


Figure 3. Summary results across observers. **a)** Likelihood change due to history: coloured dots mark the increase in likelihood with respect to a model with no history terms, corrected for different numbers of parameters using Akaike's criterion. Black box marks median and quartiles of distribution. **b)** history contribution to variance in the decision variable on difficult trials: coloured dots mark the fraction of total history contribution to variance in the decision variable on difficult trials. **c)** Prediction of behavioural responses from full model using both history and stimulus (left), the stimulus only (center) or the experimental history only (right). **d)** Ratio between 85%-performance thresholds estimated with the full model and with a model without history terms. Thresholds were consistently underestimated if history dependence was not taken into account.

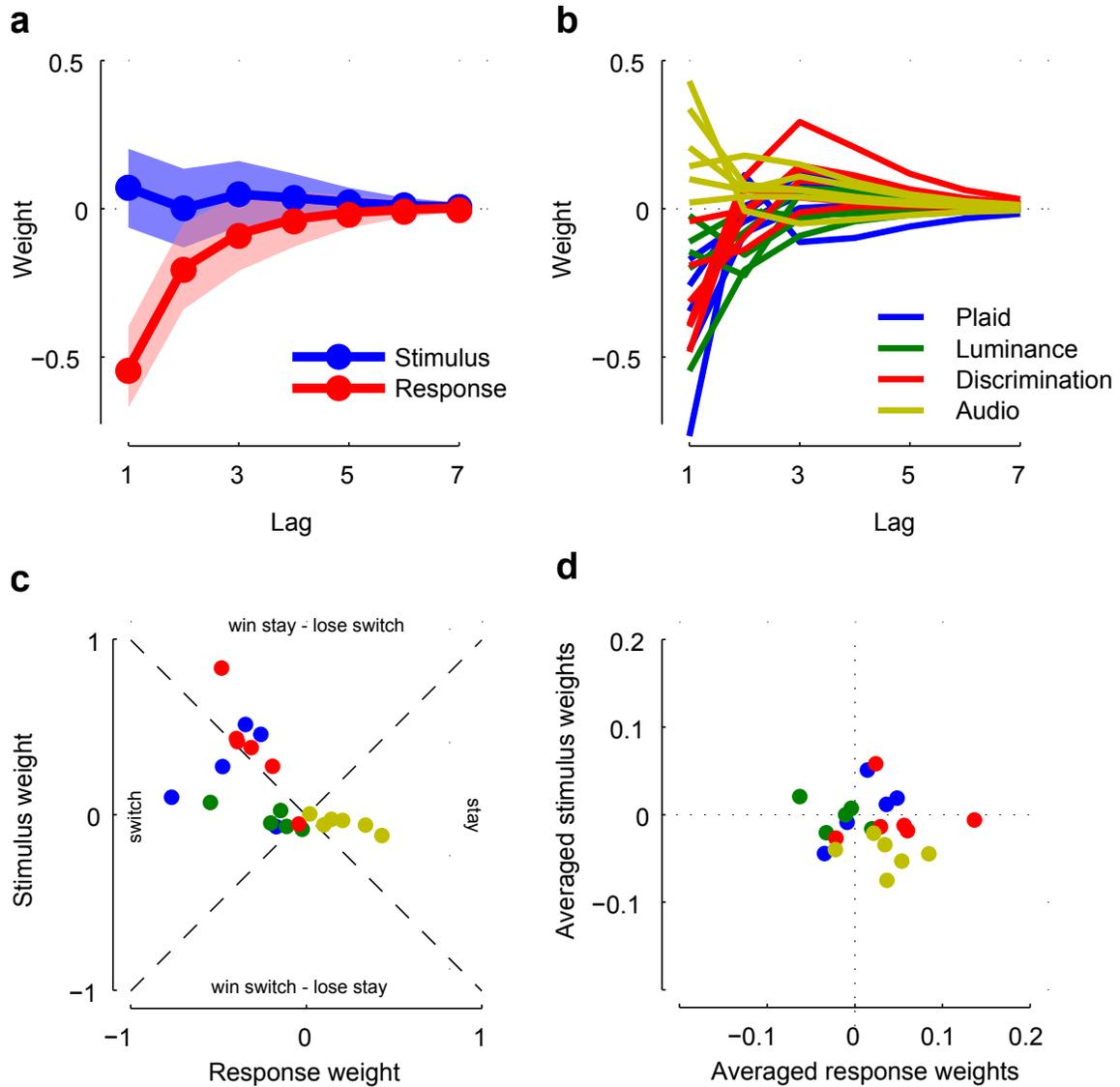


Figure 4. History couplings. **a)** Weights assigned to previous stimuli and responses for observer pk. Shaded regions mark 95% bootstrap confidence regions. **b)** Response kernels for all observers. Colour codes for the different experiments. Each position in this coordinate system corresponds to a behavioural strategy. For example, a measurement being in the upper quadrant shows that the observers had a tendency to repeat a response after a successful trial and to alternate otherwise. **c)** Weights assigned to previous response and previous stimulus identity across observers. **d)** Average of weights assigned to stimuli and responses more than one trial back.