Stable Individual Differences in Occasion Setting

Steven Glautier and Ovidiu Brudan

Southampton University

7410 words 1/4/19

Author Note

Correspondence should be addressed to Steven Glautier, School of Psychology, University of Southampton, Southampton, SO17 1BJ, United Kingdom. E-mail: spg@soton.ac.uk, Tel: (+44) 023 8059 2589. Acknowledgements to student volunteer research assistant Jessica Richards for help with data collection.

Abstract

In the current investigation we classified participants as inhibitors or non-inhibitors depending on the extent to which they showed conditioned inhibition in a context that had been used for extinction of a conditioned response. This classification enabled us to predict participant responses in a second experiment which used a different design and a different experimental task. In the second experiment a feature-negative discrimination survived reversal training of the feature to a greater extent in the non-inhibitors than in the inhibitors and this result was supported by Bayesian analyses. We propose that the fundamental distinction between inhibitors and non-inhibitors is based on a tendency to utilise first-order (direct associations) or second-order (occasion-setting) strategies when faced with ambiguous information and that this classification is a stable individual differences attribute. (138 words)

Keywords: associative learning, inhibition, occasion-setting, response-recovery, feature-negative discrimination, reversal, individual differences

Stable Individual Differences in Occasion Setting

Introduction

Associative learning plays a crucial role in the survival of organisms by facilitating the acquisition of responses to stimuli which signal significant events. However, acquisition is only half of the story because in an ever-changing environment a response that was once appropriate may become redundant or even maladaptive if it continues to occur when the environmental conditions change. For example an animal may learn that a particular location is a good source of food but if it continues returning to that location after the source is exhausted it will waste energy that could be better spent foraging elsewhere. Extinction is one process by which an organism may adapt to changes in environmental conditions. Procedurally, extinction involves presenting a conditioned stimulus (CS) alone, without the unconditioned stimulus (US) it was previously paired with. During extinction the conditioned response (CR) produced by the CS is observed to decline in magnitude and probability until at some point extinction appears to be complete. Superficially extinction resembles unlearning and in one of the most widely cited theoretical models of associative learning, the Rescorla-Wagner model (Rescorla & Wagner, 1972), extinction is exactly that – the undoing, without leaving a trace, of a previously learned association.

However, it is well established that extinction cannot be understood as simple unlearning. Spontaneous recovery and renewal are among those phenomena which demonstrate that traces of the original learning survive extinction (e.g. Bouton, 1994). Spontaneous recovery refers to renewed CRs that occur when the CS is presented after a delay following extinction. Brooks and Bouton trained rats with a tone CS signalling delivery of food then presented the CS alone during an extinction phase. Responding clearly declined during extinction. Animals were then given test presentations of the CS either five hours or six days after extinction. Animals tested five hours after extinction showed a slight increase in responding to the CS compared to that seen at the end of extinction. In contrast, animals tested six days after extinction showed dramatically increased responding to the CS compared to that seen at the end of extinction. In fact

responding in the test was slightly higher than it was before extinction, a clear spontaneous recovery effect showing that extinction did not simply erase what had been learned during acquisition (Brooks & Bouton, 1993). Renewal refers to renewed CRs consequent to a contextual change after extinction. Bouton and King (Experiment 1) trained rats with a tone CS signalling electric shock in one context, context A: (A: $T+$ trials[1]), and then extinguished the CS in another context, context B: (B: $T-$ trials), before testing the CS back in context A:. Clear extinction effects were seen but responding returned when the CS was tested in context A:. These results were supported by further tests which showed that the renewal effect was not mediated by the excitatory properties of context A: (Bouton & King, 1983). Other rat studies, where testing was carried out in a novel context C: (an ABC design as opposed to an ABA design), confirmed that renewal effects do not depend on the excitatory properties of the test context (Bouton & Bolles, 1979).

Results such as these present important problems for learning theories and in what follows we describe two leading explanations for renewal to set the stage for the experiments to be presented below. The central question is, how we can understand the decline in responding that is seen during extinction when there is clear evidence that the original learning remains intact? We argue that two leading explanations for renewal, one based on context inhibition and one based on occasion-setting, are not mutually exclusive (e.g. Bouton & Nelson, 1994). The experiments reported below show that human participants may be categorised into one of two groups. In one group, inhibitors, renewal seems to be controlled by inhibitory associations involving the experimental context. In another group, non-inhibitors, we argue that renewal is controlled by an occasion-setting mechanism. The classification of participants into these two groups appears to be relatively stable, allowing predictions to be made across different experimental procedures, and indicates some practical implications. We leave

[1] The colon indicates that identifier A refers to a contextual cue. In contrast an undecorated identifier (e.g. T) refers to a discrete cue. When an experiment only involves a single US the '+' sign indicates a trial with the US and a '-' sign indicates a trial without the US.

consideration of these practical implications for the Discussion to focus here on two theoretical explanations for renewal.

In associative models learning is conceptualised as changes in the strength of associative links between mental representations of CSs and USs. We review here the operation of the Rescorla-Wagner model as a 'standard' model of associative learning (Rescorla & Wagner, 1972). Although the Rescorla-Wagner model was developed as a model of Pavlovian conditioning in animals its principles are sufficiently general to have been successfully imported into new domains. The Rescorla-Wagner model has been considered a viable candidate model in a variety of human learning tasks including predictive, causal, and Pavlovian learning (e.g. Chapman & Robbins, 1990; Dickinson, Shanks, & Evenden, 1984; Lachnit, 1988). In the Rescorla-Wagner model, in the case of simple excitatory learning, where a CS is repeatedly paired with a US, the association strength, V, increases towards an asymptote and is greater than zero. When $V > 0$ the CS is said to be excitatory and presentation of the CS activates the US representation – informally presentation of the CS leads to an expectation of the US through spreading activation. V can also be reduced when an expected US fails to occur, for example during extinction, and this may result in V becoming negative in which case the CS is said to be an inhibitor. When there is an inhibitory CS-US association presentation of the CS effectively suppresses expectation of the US. The Rescorla-Wagner model formally explains renewal effects though a mechanism known as 'protection-from-extinction' in which the extinction context develops inhibitory associative strength as detailed in the following paragraph.

$$\Delta V = \alpha\beta(\lambda - \Sigma V) \tag{1}$$

Equation 1 is the fundamental Rescorla-Wagner model learning equation. In Equation 1 ΔV is the change in the associative strength between the mental representation of a predictive stimulus (such as a tone CS) and the representation of the outcome (such as a shock US) that occurs on a single learning trial. ΔV is a function of two learning rate parameters, α for the CS and β for the US, and the parenthesised

error term. In the error term λ is the value of the US on that trial (usually modelled as 1 or 0 for the occurrence and non-occurrence of the US, respectively) and ΣV is the summed associative strength of all the predictors that are present on the trial. To see how inhibitory learning takes place during extinction consider the associative strength of cue D after an acquisition phase involving a series of $A\!:\!D+$ trials. Asymptotically $V_{A\!:} + V_D \rightarrow 1$ and ${}^{V_{A\!:}}\!/_{V_D} = {}^{\alpha_{A\!:}}\!/_{\alpha_D}$. Following acquisition there is an extinction phase involving $B\!:\!D-$ trials. At the start of extinction $V_{B\!:} = 0$ and $\Sigma V = V_{B\!:} + V_D > 0$. During extinction $\lambda = 0$ and since $\alpha > 0$ and $\beta > 0$ then $\Delta V < 0$. Since $V_{B\!:} = 0$ at the start of extinction $V_{B\!:}$ becomes negative during extinction and V_D declines. Learning (and responding) during extinction stops when $V_{B\!:} + V_D = 0$ and at this point $B\!:$ is inhibitory ($V_{B\!:} < 0$) while D remains excitatory ($V_D > 0$). The presence of inhibitory $B\!:$ is said to protect D from further extinction.

Protection from extinction effects have been demonstrated by presenting a discrete inhibitory CS in compound with a CS during extinction (e.g. Rescorla, 2003). However, there are few experiments which have found evidence for the context developing inhibition during extinction, as would be expected following the theoretical analysis above, calling into question the proposal that protection from extinction could explain the renewal effects that we are considering. For example, in the study mentioned above Bouton and King used a summation test to look for context inhibition (Bouton & King, 1983). A summation test involves presenting an excitatory cue in compound with a putative inhibitor (Rescorla, 1969) and in this case Bouton and King presented an excitatory CS in the extinction context but no inhibition was detected. More recently Polack, Laborda, and Miller (2012) reported evidence for extinction contexts becoming inhibitory in a summation test, and also in a retardation test in which learning is acquired more slowly in the presence of the inhibitor than in its absence (Rescorla, 1969). Polack et al. found clearest evidence for context inhibition during extinction when using short inter-trial-intervals and suggest that this variable may explain why there are few reports of context inhibition in the literature. Again using a summation test but this time using human subjects both Nelson, Sanjuan, Vadillo-Ruiz, Perez, and

Leon (2011) and Havermans, Keuker, Lataster, and Jansen (2005) found reduced conditioned suppression when an excitatory cue was presented in the extinction context indicating that the context was inhibitory. However, these results are ambiguous since Nelson et al. (2011) demonstrated that conditioned suppression was also reduced when the excitatory cue was presented in an associatively neutral context. Thus, any putative conditioned inhibitory effect of the extinction context added nothing to conditioned suppression produced by a novel cue-context combination. However, Glautier, Elgueta, and Nelson (2013) found clear evidence for context inhibition with appropriate controls using a predictive learning task (see Discussion, page 17) so it seems as though it may be premature to rule-out a role for protection from extinction in extinction and renewal effects.

Partly as a result of failures to confirm that extinction contexts become conditioned inhibitors alternatives to the protection from extinction explanation for renewal have been developed. Occasion setters are stimuli that can be shown to influence the expression of an association between a CS and a US without themselves being directly associated with the US. Instead they function by controlling an 'and-gate' which switches the CS-US association on or off (Bouton, 1994; Holland, 1992; Swartzentruber, 1995). An occasion-setting mechanism is illustrated in Figure 1 where it is contrasted with the Rescorla-Wagner model. Figure 1 shows the hypothesised associative structures that are formed after acquisition (left-hand side) and after extinction (right-hand side). CS_D was paired with the US during an acquisition phase to produce an $D \rightarrow US$ association (left) and then, in a new context, context B:, CS_D was extinguished by presentation without the US. According to the Rescorla-Wagner model, as explained above, this produces an inhibitory association between context B: and the US (bottom right). In contrast, in an occasion-setting model, it is assumed that context B: forms an inhibitory link with the $D \rightarrow US$ structure (top-right) so that when context B: is present the association is switched off and switched on otherwise. We term direct associations between CSs and the US 'first-order' and associations which operate on first-order associations 'second-order'. It should be apparent that

renewal of responding is predicted for the second-order model, as well as the first-order model, because when CS_D is presented outside of context B: the $D \rightarrow US$ association will be active.

The current investigation follows-up the work of Glautier et al. (2013). It was based on an analysis of renewal in which it was assumed that the two mechanisms outlined above could operate. During an extinction phase of a renewal experiment we assumed that participants could suppress responding by conditionalising an $D \rightarrow US$ association that had been learned during the acquisition phase on the experimental context or by learning that the context was inhibitory as illustrated in Figure 1. In Glautier et al., although there was a clear overall context inhibition effect, approximately 50% of participants showed some responding to a test cue when it was presented in the extinction context. It was hypothesised that extinction in participants who failed to suppress responding in the extinction context summation test would have had extinction performance controlled by second-order associations. This is because one of the defining features of an occasion-setter is that its occasion-setting function is specific to a particular CS-US relation (e.g. Holland, 1989, 1992). Thus, if B: has occasion set the $CS_D \rightarrow US$ relation then it should not affect another CS-US relation (e.g. $CS_G \rightarrow US$), unless that CS-US has also been occasion-set (Lamarre & Holland, 1987). In contrast, conditioned inhibition shows no such specificity since it operates on the US representation. Thus, if B: has become inhibitory it should suppress responding to any excitatory cue it is compounded with.

Of course a classification of participants based on a single test has little value unless there is some independent predictive value of that classification. Therefore we report below on two sequentially conducted experiments. Experiment 1 was a renewal experiment, based closely on Glautier et al. (2013), which provided an inhibition score for each participant. On the basis of these scores participants were classified as inhibitors or non-inhibitors. Then, in Experiment 2, participants underwent feature negative training, followed by reinforcement of the feature, and then a test to see if the feature negative discrimination was disrupted. Feature-negative training is a procedure

in which a cue is reinforced when it is presented alone and non-reinforced whenever it occurs in the presence of another cue (the feature negative). It was predicted that the feature negative discrimination would be maximally disrupted in the inhibitors, i.e. those predisposed to learn first order solutions. This would be consistent with previous reports which have shown that feature negative discriminations in which the feature is trained as an occasion-setter using serial presentation can be maintained after reinforced trials with the feature (e.g. Holland, 1984, 1992; Rescorla, 1987). The rationalisation of this prediction is illustrated in Figure 2. After training with $I+$ and $IJ-$ trials responding could be controlled by first-order associative structures (left-hand side, bottom) or by second-order structures (left-hand side, top). Following reinforcement of feature J the second-order associative structures formed by the non-inhibitors will have an excitatory input to the US representation from J (right-hand side, top). There is also an excitatory input to the US representation from I, but this association is gated closed by the presence of J. In contrast the inhibitors will have excitatory inputs to the US representation from both I and J (right-hand side, bottom). Consequently we predicted stronger responses in the IJ compound test for the inhibitors than for the non-inhibitors – the inhibitors will lose the original feature negative discrimination to a greater extent than the non-inhibitors.

Experiment 1

Participants took part in a computer-based predictive learning task during which they viewed a series of on-screen trials and were told that their task was to learn to predict the outcome of each trial on the basis of visual cues presented at the start of each trial. The trial outcomes were coloured flashes on the computer screen and the cues were visually distinctive objects, based on size, colour, distortion, and decoration variations of a 3D cube, which 'fell' on each trial from the top to the bottom of the computer screen. Towards the end of the trial, which lasted about 5s in total, the objects passed a 'sensor' located in the bottom part of the screen and the coloured flashes, when they occurred, were timed with and said to be triggered by the object

passing the sensor. Trials could take place in different contexts, each context being a visually distinctive 3D environment in which the falling objects were observed. On each trial participants had to press a key to indicate their expectation, with respect to the outcomes, before the objects passed the sensor. There were three response options available; key-R to predicted a red flash, key-G to predict a green flash, or no-key to predict no flash. Participants were instructed to make as many correct predictions as possible but minimise incorrect predictions. Further details of the task are given in Glautier et al. (2013) and an illustrative video can be found at https://tinyurl.com/y6v5unpj.

Method

Procedures were approved by the University of Southampton Research Governance Office and the School of Psychology's Ethics Committee.

Participants. Eighty participants took part. They were recruited by word of mouth and posted advertisement. Their mean age was 20.7 years (range 16-39) and they included 24 males. Participants were recruited in two distinct samples. The first sample of 28 was recruited, and tested, by JR from a local community in Wiltshire, UK. The second sample of 52 was recruited, and tested, by OB from the University of Southampton Highfield campus. Participants in the first sample were tested in various convenient community environments and paid £4 for their participation whilst participants in the second sample were tested in psychology research laboratories and given course credit for participation.

Apparatus. For the first sample the experiment was run on a laptop computer with a screen measuring 30.5 cm x 19.2 cm (W x H) running at 60hz. For the second sample the experiment was run on personal computers with screens measuring 41 cm x 26 cm (W x H) running at 75hz. In both cases the displays used 32 bit colour mode and pixel resolutions of 1440 x 900 and were controlled by computer programs written by the first author in Microsoft C# language using Microsoft XNA Game Studio Version 3.1 for rendering of the experimental scenario.

Design and procedure. Participants received a brief verbal introduction to the procedures and then signed a consent form before reading a more detailed on-screen description of the task. This description is provided in full as part of the illustrative video. Participants then had the opportunity to ask questions before the experimental procedure began. Table 1 shows the design for Experiment 1. The experiment contained acquisition, extinction, summation test, and recovery test phases. The table shows the trial types that were presented in each phase. The experimental context varied for different phases. Acquisition, extinction, and the recovery test took place in the three different contexts i.e. it was an ABC design. The screen backgrounds that served the roles of contexts A:, B:, and C: were selected randomly without replacement from four possibilities for each participant. Context B: was used for the summation test. The cue objects presented on each trial, serving the roles of D, E, F, and G, were selected randomly without replacement from 16 possibilities for each participant. The coloured flashes serving the roles of X and Y were selected randomly without replacement from two possibilities (red and green) for each participant. Outcome Z designates the no-flash outcome. The phases were divided into blocks with each block containing equal numbers of each trial type. Trial order was randomised independently for each participant within block so there could be no more than four repeats of a trial type in a single sequence. Responses made in the summation test were used to classify participants as either non-inhibitors or inhibitors. After completing Experiment 1 participants went onto Experiment 2.

Results

The results presented below were obtained from analyses undertaken using R, JAGS, and associated packages (Plummer, 2017; R Core Development Team, 2012). The R-code, JAGS model specifications, and raw data to check and reproduce the analyses reported below can be found at https://osf.io/xwp2d/. As noted above data was collected in two distinct samples. Preliminary analysis of data from the first sample (n=28) indicated support for the hypothesis under test and so was followed-up with

continued data collection in a second sample (n=52). Data from both samples (N=80) was combined for the analyses reported below. An earlier draft of this paper presents the data separately for both samples, all patterns are closely matched in both samples (Glautier & Brudan, 2018). Participants learned to respond appropriately to each cue during the acquisition phase and during the extinction phase. Responses to cues D and G were of focal interest and these are plotted in Figures 3a and 3b respectively. In Figure 3a it can be seen that x-responses were increasingly likely to be made to cue D over the course of acquisition – participants correctly predicted outcome X when cue D was present, and then, during extinction, the probability of x-responses to cue D declined.

Figure 3b shows that x-responses were also acquired to cue G during the acquisition phase and that responses to cue G were markedly suppressed after extinction, on average, during the summation test suggesting that the extinction context had become inhibitory for outcome X. However, 31 out of our 80 participants made at least one x-response to cue G during the two trials of the summation test, suggesting that there was less context inhibition for outcome X for these participants than for those who made no x-responses during the summation text. Accordingly those who made no x-responses during the summation test were classed as inhibitors for the purposes of Experiment 2, the remaining participants were classed as non-inhibitors. It was noted that extinction progressed more rapidly for the inhibitors than for the non-inhibitors. This was not anticipated but to examine the question of whether or not the probability of making an x-response during extinction differed for the non-inhibitors and inhibitors we considered a region of practical equivalence (ROPE) around the mean of posterior distribution for the inhibitors and a 95% credible interval around the mean of the posterior for the non-inhibitors (Kruschke, 2015). The ROPE was set at the mean probability of responding during extinction $\pm 0.1 \times \sigma_{pooled}$ corresponding to a small effect (Cohen, 1988). The boundaries of the ROPE $[0.17, 0.19]$ excluded the boundaries of the credible interval for the non-inhibitors $[0.27, 0.42]$ suggesting the observed difference is substantial.

Finally, for both groups, there was a clear recovery effect in block 12 when cue *D* was presented in a novel context i.e. an ABC recovery effect was observed.

Experiment 2

Experiment 2 was also a computer-based predictive learning task consisting of a series of trials during which participants tried to predict the outcome of each trial on the basis of visual cues presented at the start of each trial. However, instead of predicting coloured flashes of the computer screen, participants learned to predict payouts for cards dealt in a fictitious casino game on the basis of visually distinctive symbols and colours borne on each card. Each trial consisted of a 'hand' containing one or two cards being 'dealt' before the participants adjusted an on-screen indicator to judge the likelihood that the hand would be a winning hand. Ratings were made on an 11-point integer scale ($[0 \ldots 10]$) labelled '10 Win', '5=Win or Lose', and '0 Lose'. Participants made their judgements in their own time and were asked to make their judgements as accurate as possible, to reflect the true value of the cards in play. There were no actual payments made for the hands dealt in this task. Once the participants had made their rating the hand was 'turned' to reveal whether or not it was a winning hand. Feedback was given for 2s during which onscreen text appeared flashing 'Win!' or 'Lose!' with a brief auditory alerting stimulus after which the next trial began. Further details of the task are given in Glautier (2013, Experiment 1) and an illustrative video can be found at https://tinyurl.com/yb3te7oj.

Method

The participants and apparatus were the same as in Experiment 1.

Design and procedure. Experiment 2 followed Experiment 1 directly after participants had a brief verbal introduction, read a more detailed onscreen description, which is provided in full as part of the illustrative video, and after an opportunity to ask questions. Table 2 shows the design for Experiment 2. The experiment contained feature negative and feature reversal learning phases followed by a feature negative survival phase. The feature negative survival phase consisted of reminder and test

trials. We were interested primarily in the feature negative training ($I+$, $IJ-$ trials) but because feature negative discriminations can be solved on the basis of cue cardinality (one cue reinforced, two cues non-reinforced) we included a concurrent feature positive discrimination ($K-$, $KL+$ trials) to ensure that participants attended to the identity of the cues. The table shows the trial types presented in each phase. Phases were not differentiated by context changes. Characters I, J, K, L, and M represent different cards, the plus and minus signs indicate the outcome, win or lose, that occurred for that trial type. The cards serving the cue roles I, J, K, L, and M were chosen at random without replacement from 182 possibilities for each participant. The 182 possibilities were formed by combination of 14 different foreground symbols and 13 different background colours – each card was marked with a foreground symbol presented on a background colour. When two cards were presented at once (on the $IJ-$ and $KL+$ trials) they were presented symmetrically on either side of the vertical midline of the screen with a 3cm space in between the left and right card. The left/right location for cards in pairs was selected randomly on each trial; single cards were always presented on the vertical midline of the screen. As with Experiment 1 the phases were divided into blocks with equal numbers of each trial type within each block and trial order was randomised independently for each participant within block so there could be no more than four repeats of a trial type in a single sequence. The critical feature negative survival test trials involved presentation of the IJ cue compound and in this test we were interested in response differences between the participants classified as either non-inhibitors or inhibitors on the basis of their responses in the summation test of Experiment 1. Inhibitors were defined as those participants who completely suppressed responding to cue G when it was presented in context B: after extinction (blocks 10,11 of Experiment 1).

Results

Figures 4 and 5 show the progress of the feature negative and feature reversal phases respectively. Figures 4a and 4b show that participants learned to respond

appropriately to cues during the feature negative phase and the fact that responding was appropriate for the $I+/IJ-$ discrimination and for the $K-/KL+$ discrimination confirms that responses were based on cue identity rather than on cue cardinality. Figure 5 shows that learning was also successful during the feature reversal phase where cue J was trained as an outcome predictor. As the existence of group differences in the feature reversal phase could be expected (see General Discussion on 18) the data for the feature reversal was plotted trial by trial in case any averaging of the trials within block masked differences. However, as can be seen, there was no clear evidence for group differences in these stages of the experiment.

Figure 6 shows the critical data from the feature negative survival test phase. The inhibitors and non-inhibitors did not differ in their ratings for cue I during the reminder presentations, however their responses to the IJ test were different with higher ratings given by the inhibitors relative to the non-inhibitors. Two Bayesian analyses of the ratings in the IJ test were carried out. Bayesian analyses were chosen for two main reasons – due to their inherent suitability for sequential updating of parameter estimates over repeated runs of an experiment and to obtain full distributional information on the parameters of interest (Dienes, 2011; Kruschke, 2013; Wagenmakers, Lodewyckx, Kuriyal, & Grasman, 2010). Under conventional null-hypothesis-testing combining data across experiments requires special techniques to be applied to avoid inflated Type I error rate (Armitage, Berry, & Matthews, 2002) and the order in which the data comes in can influence the result. In contrast Bayesian approaches are explicitly based upon updating of estimates as each new piece of evidence comes in and the order of data arrival does not affect the final conclusions. In the first analysis a Bayesian t-test (Morey & Rouder, 2018) was used to assess whether or not there was a difference between inhibitors and non-inhibitors on responding to the test on IJ in the feature negative survival phase. The test used a non-informative Jeffreys prior on the variance and a Cauchy prior on the effect size with scale $\frac{\sqrt{2}}{2}$. The analysis produced a Bayes factor of 3.73 showing that the probability of the null hypothesis being true is 3.7 less likely given the data – substantial evidence in favour of the alternative hypothesis

(Jeffreys, 1961).

The second analysis was used to obtain posterior distribution estimates for ratings given by the inhibitors and non-inhibitors in the test on IJ in the feature negative survival phase. The analysis was based on examples given in Kruschke (2015) and Lee and Wagenmakers (2014). Figure 7 shows a graphical representation of the JAGS model that was used. The ratings of the non-inhibitors (x_i) were assumed to come from a Gaussian distribution with mean μ and standard deviation σ. The ratings of the inhibitors (y_i) were assumed to come from a Gaussian distribution with mean $\mu + \delta$ and standard deviation σ. The prior on μ was a Gaussian with mean μ_x equal to the observed mean from the non-inhibitor group and standard deviation $\sigma_\mu = 100 \times \sigma_{xy}$ where σ_{xy} was the observed standard deviation pooled across both groups. The prior on σ was a uniform distribution over the range $[\sigma_{xy} \times 1/100 \ldots \sigma_{xy} \times 100]$. The prior on δ was a Gaussian with mean equal to zero and standard deviation $\sigma_\delta = 100 \times \sqrt{2} \times \sigma_{xy}$. Using $\sqrt{2}$ is based on the fact that the variance of the difference between two normally distributed random variables x and y is $\sigma_x^2 + \sigma_y^2$ (Weisstein, 2017), therefore the standard deviation of the difference between two normally distributed random variables x and y, both of which have standard deviation σ_{xy}, is $\sqrt{2} \times \sigma_{xy}$. Given these priors a JAGS model was run with three chains, each with randomly generated initial values (within constraints), over 50000 iterations discarding the first 5000 'burn-in' samples. The chains were observed to converge and the posterior distributions presented in Figure 8 were constructed by pooling across chains.

The means of the posterior distributions shown in Figure 8 for the non-inhibitor and inhibitor group ratings were 1.08 (μ) and 2.54 ($\mu + \delta$) respectively and the common standard deviation (σ) estimated for both groups was 2.53. To further examine the question of whether or not the ratings differed for the non-inhibitors and inhibitors we can consider the region of practical equivalence (ROPE) around the mean of the posterior for the non-inhibitors and the 95% credible interval around the mean of the posterior for the inhibitors (Kruschke, 2015). The ROPE was set at $\mu \pm 0.1 \times \sigma_{xy}$ corresponding to a small effect (Cohen, 1988). The boundaries of the ROPE $[0.83, 1.33]$

excluded the boundaries of the credible interval for $\mu + \delta$, $[1.65, 3.44]$.

Figure 8 also shows the raw data and indicates that there are outliers in both groups. Outliers can shift the mean of an estimated distribution and the normal distribution, which assigns very small probabilities to extreme values, is susceptible to such influences. Therefore a follow-up analysis was done in which the x_i and y_i were assumed to come from t-distributions with parameters (μ, σ, and ν), and ($\mu + \delta$, σ, and ν) respectively. This is an approach to robust estimation and the heavier tails of the t-distribution, when $\nu < 30$, can accommodate extreme values with less of an impact on the estimated mean (Kruschke, 2015). However, this robust analysis did not produce an appreciable difference in the outcome. The boundaries of the ROPE under this robust analysis were $[0.64, 1.14]$ and the boundaries of the credible interval were $[1.61, 3.39]$. The posterior means for the non-inhibitors and inhibitors were 0.89 and 2.49, respectively, a close match to those from the 'standard' analysis and consistent with the the estimate of $\nu = 30.41$. If anything this robust analysis indicated a larger difference between the means than the standard analysis.

General discussion

In Experiment 1 we provided a clear demonstration of an ABC renewal effect – restoration of an extinguished response consequent to a change of context, Figure 3a. We also showed suppression of responding to cue G when it was presented in the extinction context, Figure 3b, suggesting that context inhibition had developed during extinction, a requirement of the protection from extinction account of renewal. A reasonable objection to the context inhibition claim is that a control condition was not included, the suppressed response to G in the extinction test may have occurred even if context B: had not been used for extinction. However, the procedures used in the current experiments matched closely those used in Glautier et al. (2013) where appropriate control contexts were used and the results obtained then and now were also closely matched. From Glautier et al. (2013), collapsing over the two experiments and experimental conditions (Single and Multiple context extinction), the mean rating for

the test cue equivalent to G in the current study (Figures 2 and 3 cue G from Glautier et al.) was 0.3 (SE=0.03, n=95) compared with 0.57 (SE=0.04, n=73) for the control conditions, again collapsing over the two experiments and conditions (No extinction and No extinction no $A \rightarrow X$). In the current investigation collapsing over both tests on cue G, as done in Glautier et al. (2013), the mean rating for cue G was 0.23 (SE=0.04, n=80).

However, again repeating the findings of Glautier et al. (2013), we noted that a substantial proportion of participants (31/80, approximately 40%) were classified as non-inhibitors on the basis that they showed at least one response to cue G in the two trials of the summation test. As outlined in the introduction we hypothesised that classifying participants as inhibitors and non-inhibitors, according to whether or not they responded in the summation test, would indicate the associative structures that were used to resolve the ambiguity in the predictive value of cue A during extinction. Importantly, we observed that the classification established in Experiment 1 enabled prediction of performance in the feature negative reversal survival test in Experiment 2. Specifically, a feature negative discrimination survived reversal training of the feature to a greater extent among the non-inhibitors than amongst the inhibitors (Figure 6), consistent with the logic of the proposed associative structures as described in the Introduction. Our conclusion that non-inhibitors and inhibitors differed on feature negative survival after reversal training of the feature was supported by Bayesian analyses. In the first analysis a Bayesian t-test provided evidence of a difference between inhibitors and non-inhibitors on responding in the feature negative reversal survival test, showing that the null hypothesis should be judged to be 3.7 less likely than the alternative given the data, implying substantial evidence in favour of the alternative hypothesis (Jeffreys, 1961). The second analysis was used to obtain full posterior distributions representing the responses on the feature negative reversal survival test. The means of the posterior distributions, shown in Figure 8, for the non-inhibitor and inhibitor group ratings were 1.08 and 2.54 respectively and the common standard deviation estimated for both groups was 2.53.

Although the group differences for the feature negative reversal survival test are clear group differences could also be expected during the feature reversal phase of Experiment 2 where the $J+$ trials effectively form a retardation test for inhibition – thus, if J was inhibitory for the inhibitors then development of responding to the $J+$ trials should lag behind that of the non-inhibitors. However, there was no evidence of a group difference in responding to cue J (Figure 5) and this aspect of the data remains puzzling. This failure to observe a group difference in this phase is somewhat ambiguous in the absence of comparison of learning between the putative inhibitor J and a novel stimulus.

Until this point we have considered that participants can be classified as those who tend to use first-order strategies and those who tend to use second-order strategies to resolve the ambiguity that arises during extinction and our results are consistent with the view that this categorisation remains stable over two different tasks. In the Introduction we outlined the Rescorla-Wagner model as a standard example of a first-order associative model and a second-order occasion-setting model and we derived the prediction that participants using first-order strategies would respond more strongly in the IJ compound test after reinforcement of feature J than participants using second-order strategies. However, whilst failure to observe abolition of feature negative discrimination performance after reinforcement of the feature has frequently been used in arguments to support the view that learning in both animals and humans involves second-order associative structures (Baeyens et al., 2004; Morell & Holland, 1993; Trask, Thrailkill, & Bouton, 2017) this observation has also been discussed in relation to stimulus configuration (Shanks, Charles, Darby, & Azmi, 1998; Williams, 1995). Therefore we now consider whether or not it would be better to think about the differences between our participants in terms of configural learning. Perhaps a distinction along such an axis fits the facts just as well as differences along an axis in which participants differ in their deployment of first and second-order strategies.

The investigations of Shanks et al. (1998) and Williams (1995) both contained experiments using predictive tasks with human participants in which a feature negative

discrimination survived feature reversal training and the authors identified conditions in which a configural associative model (Pearce, 1987, 1994) could explain survival. Focusing on Shanks et al.'s discussion of the Pearce model, note first that this model makes use of similar principles to those used in the Rescorla-Wagner model but that the associations that are learned are associations between mental representations of whole patterns (stimulus configurations) and the US. For example in the Rescorla-Wagner model during feature negative training ($I+/IJ-$ trials) an excitatory association between cue I and the US is formed alongside an inhibitory association between cue J and the US. In contrast, in Pearce's model an excitatory association between cue I and the US is formed alongside an inhibitory association between a configural cue IJ and the US. The effect of reinforced feature trials in the Rescorla-Wagner model has already been covered (page 9) and it clearly implies a strong response should then be seen when the IJ compound is presented for test. In Pearce's model the reversal training will result in an excitatory representation forming between J and the US, thus there are now three associations ($I \rightarrow US$, $IJ \dashv US$, and $J \rightarrow US$ where the arrow and stopped arrow indicate excitatory and inhibitory associations, respectively) which need to be taken into account to understand the predicted response in the IJ compound test. The response to the IJ compound would be determined by the state of the $IJ \dashv US$ association and by generalisation between the cues I, J, and the IJ configuration. Strong generalisation from J would cause loss of the feature negative discrimination but if there was little or no generalisation then the $J+$ trials would have little effect on the discrimination (Shanks et al., 1998).

Suppose now that our non-inhibitor group should be more properly labelled 'configural with limited generalisation' and that the inhibitor group should be labelled 'configural'. Presumably this classification could be used to explain the Experiment 2 results (c.f. Figure 6) where the $J+$ trials had less impact on the I/IJ discrimination among the non-inhibitors than among the inhibitors. But would a configural with limited generalisation group be predicted to show weak inhibition in the Experiment 1 summation test? The answer is clearly no. Recall that the non-inhibitor group

responded more to cue G in that test than the inhibition group. Analysis in terms of configural learning indicates that at the time of the summation test there would be two relevant associations, established during the acquisition and extinction phases, namely $A{:}G \rightarrow US$ and $B{:}D \dashv US$, and responding to the novel test $B{:}G$ depends on generalisation between $B{:}G$ and the configurations $A{:}G$ and $B{:}D$. $B{:}G$ shares one common element with $A{:}G$ and one common element with $B{:}D$ and a response is predicted if the absolute value of excitatory association exceeds the absolute value of inhibitory association i.e. a response should be observed if $(V_{A:G} - V_{B:D}) > 0$. In fact responding should increase with $S \times (V_{A:G} - V_{B:D})$ where S is given in Equation 2. Equation 2 defines the similarity between two stimuli X and Y where N_c, N_x, and N_y give the number of elements common to X and Y, the number of elements in X, and the number of elements in Y, respectively. The parameter $d = 2$ in standard configural models (Kinder & Lachnit, 2003; Pearce, 1994) and increasing d corresponds to reduced generalisation, by reducing S, hence a response is less likely if there is weaker generalisation i.e. in the configural with limited generalisation group (AKA non-inhibitor) than if there is stronger generalisation i.e. in the configural group (AKA inhibitor). We observed the opposite therefore a configural account of the group differences in Experiment 1 results is incompatible with a configural account of the group differences in Experiment 2.

$$S = \left(\frac{N_c}{\sqrt{N_x \times N_y}} \right)^d \tag{2}$$

We also considered an alternative configural approach to that provided by Pearce (1994). Discussions of the adequacy of the Rescorla-Wagner model soon lead to the observation that some discriminations e.g. negative patterning ($A+$, $B+$, and $AB-$ trials) cannot be 'solved' using the Rescorla-Wagner model as presented up until this point. Therefore, since there are many examples showing both animals and humans can solve such discriminations (e.g. Rescorla, 1972; Shanks et al., 1998; Young, Wasserman, & Johnson, 2000), a unique-cue modification of the Rescorla-Wagner model has been proposed (e.g. Rescorla, 1972, 1973). In this modification of Rescorla-Wagner unique

configural cues represent stimulus conjunctions e.g. it is assumed that a compound of cues A and B would be represented as ABc where c represents the conjunction of A and B. Suppose that our inhibitor group employ unique-cue configural strategies and therefore should be properly labelled 'configural' and the non-inhibitors do not and should be labelled 'elemental'. Taking the results of Experiment 1, we can explain the lower response to $B{:}G$ the among the configural (inhibitors) participants compared to the elemental (non-inhibitors) participants. This is partly the result of the fact that V_G would be lower for configural participants than for the elemental participants at the point of the summation test. This is because, among the configural participants, G would have been in competition with context $A{:}$ and a configural cue representing the conjunction of $A{:}$ and G during acquisition. In contrast, for the elemental participants, G would have been in competition with only the context $A{:}$ during acquisition. However, the putative configural (inhibitors) participants were observed to show larger responses than the elemental (non-inhibitors) participants in the test on IJ in Experiment 2 but if we really were comparing configural against elemental participants the test on IJ should show the opposite result. This is because, for the configural participants, the IJ compound would contain a unique-cue for the IJ conjunction and this unique-cue would have become inhibitory during the $IJ-$ trials, resulting in lower responses. Simulations, summaries of which can be found at https://osf.io/xwp2d/, were carried out to confirm these analyses.

Before we conclude by a consideration of some of the practical implications of the current findings a comment on one further aspect of the data is warranted. Although we can be reasonably confident that there exists a real difference between inhibitors and non-inhibitors the data shows that even amongst the inhibitors that the feature negative discrimination was not fully reversed by the $J+$ trials of the feature reversal phase – the response to the IJ compound was not strong (Figure 6). One possible explanation for this is that there were insufficient $J+$ trials during the reversal phase but this is rather unsatisfactory because responding to cue J was strong by the end of the phase (Figure 5). Another possibility is that inhibitors do not treat the IJ compound in a strictly

elemental additive fashion (e.g. as outlined in the replaced elements model of Wagner, Brandon, Mowrer, & Klein, 2001). Indeed studies of summation effects in animals and in humans indicate a complex picture in which responding to a compound is not a straightforward function of responding to the elements (e.g. Glautier, Redhead, Thorwart, & Lachnit, 2010; Pearce, Redhead, & George, 2002). Clearly there are other possibilities and directions for further study but the simple prediction derived in the Introduction, based on Figure 2, that responding in the IJ test would be greater for the inhibitors than the non-inhibitors was supported by our results.

We conclude with a brief review of some of the more practical implications of the current findings. First, there is a possible link between individual differences in conditioned inhibition, as discussed and studied above, and the general concept of inhibition which has been linked to a range of disorders including addiction, attention deficit hyperactivity disorder, and personality disorders, and the personality trait of impulsivity (Dawe & Loxton, 2004; He, Cassaday, Bonardi, & Bibby, 2013; Robbins, Gillan, Smith, de Wit, & Ersche, 2012). Second, there are possible implications for cue exposure treatments for anxiety and addiction, among other disorders. Cue exposure treatments have a clear and specific rationale in terms of extinction but the fact that there appear to be different underlying mechanisms through which extinction can be achieved suggests different ways to improve cue exposure outcomes. Taking links between conditioned and other types of inhibition first, it is already clear that it is an oversimplification to consider inhibition as a unitary construct. There are at least three recognised types of inhibition namely motor, cognitive, and attentional each of which is measured using different tasks and for which different neural loci of control have been identified (e.g. Eagle et al., 2008; Nigg, 2000). Despite the fact that conditioned inhibition has been widely studied in the learning literature it is not clear where it sits in overall taxonomy of inhibition. There are few examples in which it has been studied alongside other measures of inhibition (e.g. Stroop task, Stop-Signal Reaction Time) or in relation to disorders in which inhibition is impaired. Among the studies where conditioned inhibition has been examined there have been reports of weaker conditioned

inhibition linked to schizotypy, schizophrenia, and personality disorders (He, Cassaday, Howard, Khalifa, & Bonardi, 2011; He, Cassaday, Park, & Bonardi, 2012; Migo et al., 2006) but no evidence of a link between conditioned inhibition and Tourette's Syndrome (Heym, Kantini, Checkley, & Cassaday, 2014) nor between conditioned inhibition and behavioural inhibition as measured by the BIS component of the BIS-BAS questionaire (Carver & White, 1994; He et al., 2013). It is difficult to draw strong conclusions given the number of studies of this kind that currently exist but one lesson from the current research is that a finer grained analysis may be useful. For example, a standard conditioned inhibition test may simply sort participants according to preferred strategy (first or second order) rather than measuring conditioned inhibition strength *per se* complicating analysis of the relationship between conditioned inhibition and other types of inhibition.

The identification of individual differences in the mechanisms underlying behavioural extinction also has some implications for attempts to improve cue-exposure treatments for addiction and other disorders. Cue-exposure therapies have been highly successful in treatments for phobias and obsessional compulsive disorders (e.g. Choy, Fyer, & Lipsitz, 2007) but less so in the case of addictions (e.g. Conklin & Tiffany, 2002; Dawe, Rees, Mattick, Sitharthan, & Heather, 2002; Drummond & Glautier, 1994; Kavanagh et al., 2006). Differences in cue-exposure effectiveness for addiction and other disorders could arise because of differences between the drug-based and e.g. fear-based conditioning or because of differences in extinction mechanisms between addiction and e.g. anxiety disorder populations. With respect to population differences it is well established that high levels of impulsivity are linked with addiction (e.g. Dawe & Loxton, 2004) but given the uncertain nature of the relationship between impulsivity and conditioned inhibition it is not clear that a connection could be made between impulsivity and cue-exposure effectiveness. Nevertheless, as demonstrated in the current research, there are individual differences in extinction mechanisms which suggest lines of investigation to improve cue-exposure treatment across the board but which may be particularly valuable in addiction. Multiple-context cue-exposure therapies may improve

treatment outcomes (Shiban, Pauli, & Mühlberger, 2013; Shiban, Schelhorn, Pauli, & Muehlberger, 2015) and this effect could be mediated either through a reduction in protection from extinction or by increasing the number of stimulus elements that could exert occasion-setting control (Glautier et al., 2013). However, some suggestions for improving cue-exposure therapy seem more likely to be effective for inhibitors than for non-inhibitors. If extinction is based on a Rescorla-Wagner like process then increasing prediction error during extinction, by presenting a compound of multiple excitatory cues, should deepen extinction and this effect has been observed in some studies (e.g. Leung, Reeks, & Westbrook, 2012; Thomas & Ayres, 2004) but not in others (e.g. Griffiths, Holmes, & Westbrook, 2017; Holmes, Griffiths, & Westbrook, 2014). The results of the studies reported herein, showing stable individual differences in the extent to which extinction is based on conditioned inhibition, suggests that cue-exposure therapy could be tailored towards these individual differences. Multiple-context cue-exposure therapies is one way to address this and extinction based on excitatory cue compounds should produce the greatest benefits for those identified as inhibitors. In addition, although there was no strong *a priori* reason to expect differences between inhibitors and non-inhibitors during extinction we did see that inhibitors may appear to be fully extinguished when responding may be masked by context inhibition (Figure 3a), a result which would be of interest to follow-up in future studies.

References

Armitage, P., Berry, G., & Matthews, J. N. S. (2002). *Statistical methods in medical research*. Malden, MA: Blackwell.

Baeyens, F., Vervliet, B., Vansteenwegen, D., Beckers, T., Hermans, D., & Eelen, P. (2004). Simultaneous and sequential feature negative discriminations: Elemental learning and occasion setting in human pavlovian conditioning. *Learning and Motivation*, *35*, 136–166. doi:10.1016/s0023-9690(03)00058-4

Bouton, M. E. (1994). Conditioning, remembering, and forgetting. *Journal of Experimental Psychology: Animal Behavior Processes*, *20*, 219–231. doi:10.1037//0097-7403.20.3.219

Bouton, M. E. & Bolles, R. C. (1979). Contextual control of the extinction of conditioned fear. *Learning and Motivation*, *10*, 445–466. doi:10.1016/0023-9690(79)90057-2

Bouton, M. E. & King, D. A. (1983). Contextual control of the extinction of conditioned fear: Tests for the associative value of the context. *Journal of Experimental Psychology-Animal Behavior Processes*, *9*, 248–265. doi:10.1037//0097-7403.9.3.248

Bouton, M. E. & Nelson, J. B. (1994). Context-specificity of target versus feature inhibition in a feature-negative discrimination. *Journal of Experimental Psychology-Animal Behavior Processes*, *20*, 51–65. doi:10.1037/0097-7403.20.1.51

Brooks, D. C. & Bouton, M. E. (1993). A retrieval cue for extinction attenuates spontaneous-recovery. *Journal of Experimental Psychology-Animal Behavior Processes*, *19*, 77–89. doi:10.1037/0097-7403.19.1.77

Carver, C. S. & White, T. L. (1994). Behavioural inhibition, behavioural activaton, and affective responses to impending reward and punishment: The BIS/BAS scales. *Journal of Personality and Social Psychology*, *67*, 319–333. doi:10.1037/0022-3514.67.2.319

Chapman, G. B. & Robbins, S. J. (1990). Cue interaction in human contingency judgement. *Memory & Cognition*, *18*, 537–545. doi:10.3758/bf03198486

Choy, Y., Fyer, A. J., & Lipsitz, J. D. (2007). Treatment of specific phobia in adults. *Clinical Psychology Review*, *27*, 266–286. doi:10.1016/j.cpr.2006.10.002

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. New Jersey: Lawrence Erlbaum Associates. doi:10.4324/9780203771587

Conklin, C. A. & Tiffany, S. T. (2002). Applying extinction research and theory to cue-exposure addiction treatments. *Addiction*, *97*, 155–167. doi:10.1046/j.1360-0443.2002.00014.x

Dawe, S. & Loxton, N. J. (2004). The role of impulsivity in the development of substance use and eating disorders. *Neuroscience and Biobehavioral Reviews*, *28*, 343–351. doi:10.1016/j.neubiorev.2004.03.007

Dawe, S., Rees, V. W., Mattick, R., Sitharthan, T., & Heather, N. (2002). Efficacy of moderation-oriented cue exposure for problem drinkers: A randomized controlled trial. *Journal of Consulting and Clinical Psychology*, *70*, 1045–1050. doi:10.1037//0022-006x.70.4.1045

Dickinson, A., Shanks, D., & Evenden, J. (1984). Judgment of act-outcome contingency - the role of selective attribution. *Quarterly Journal of Experimental Psychology Section A-Human Experimental Psychology*, *36*, 29–50. doi:10.1080/14640748408401502

Dienes, Z. (2011). Bayesian versus orthodox statistics: Which side are you on? *Perspectives on Psychological Science*, *6*, 274–290. doi:10.1177/1745691611406920

Drummond, D. C. & Glautier, S. (1994). A controlled trial of cue exposure treatment in alcohol dependence. *Journal of Consulting and Clinical Psychology*, *62*, 809–817. doi:10.1037//0022-006x.62.4.809

Eagle, D. M., Baunez, C., Hutcheson, D. M., Lehmann, O., Shah, A. P., & Robbins, T. W. (2008). Stop-signal reaction-time task performance: Role of prefrontal cortex and subthalamic nucleus. *Cerebral Cortex*, *18*, 178–188. doi:10.1093/cercor/bhm044

Glautier, S. (2013). Revisiting the learning curve (once again). *Frontiers in Psychology*, *4*. doi:10.3389/fpsyg.2013.00982

Glautier, S., Elgueta, T., & Nelson, J. B. (2013). Extinction produces context inhibition and multiple-context extinction reduces response recovery in human predictive learning. *Learning & Behavior*, *41*, 341–352. doi:10.3758/s13420-013-0109-7

Glautier, S., Redhead, E., Thorwart, A., & Lachnit, H. (2010). Reduced summation with common features in causal judgments. *Experimental Psychology*, *57*, 252–259. doi:Doi10.1027/1618-3169/A000030

Glautier, S. & Brudan, O. (2018). Preprint GlautierAndBrudan2018 Stable individual differences in occasion setting. doi:10.31219/osf.io/trwbq

Griffiths, O., Holmes, N., & Westbrook, R. F. (2017). Compound stimulus presentation does not deepen extinction in human causal learning. *Frontiers in Psychology*, *8*. doi:10.3389/fpsyg.2017.00120

Havermans, R. C., Keuker, J., Lataster, T., & Jansen, A. (2005). Contextual control of extinguished conditioned performance in humans. *Learning and Motivation*, *36*, 1–19. doi:10.1016/j.lmot.2004.09.002

He, Z., Cassaday, H. J., Howard, R. C., Khalifa, N., & Bonardi, C. (2011). Impaired Pavlovian conditioned inhibition in offenders with personality disorders. *Quarterly Journal of Experimental Psychology*, *64*, 2334–2351. doi:10.1080/17470218.2011.616933

He, Z., Cassaday, H. J., Park, S. B. G., & Bonardi, C. (2012). When to hold that thought: An experimental study showing reduced inhibition of pre-trained associations in schizophrenia. *PLOS ONE*, *7*. doi:10.1371/journal.pone.0042175

He, Z., Cassaday, H., Bonardi, C., & Bibby, P. (2013). Do personality traits predict individual differences in excitatory and inhibitory learning? *Frontiers in Psychology*, *4*, 245. doi:10.3389/fpsyg.2013.00245

Heym, N., Kantini, E., Checkley, H. L. R., & Cassaday, H. J. (2014). Tourette-like behaviors in the normal population are associated with hyperactive/impulsive ADHD-like behaviors but do not relate to deficits in conditioned inhibition or response inhibition. *Frontiers in Psychology*, *5*. doi:10.3389/fpsyg.2014.00946

Holland, P. C. (1984). Differential-effects of reinforcement of an inhibitory feature after serial and simultaneous feature negative discrimination-training. *Journal of Experimental Psychology-Animal Behavior Processes*, *10*, 461–475. doi:10.1037//0097-7403.10.4.461

Holland, P. C. (1989). Transfer of negative occasion setting and conditioned inhibition across conditioned and unconditioned stimuli. *Journal of Experimental Psychology-Animal Behavior Processes*, *15*, 311–328. doi:10.1037/0097-7403.15.4.311

Holland, P. C. (1992). Occasion setting in pavlovian conditioning. *Psychology of Learning and Motivation-Advances in Research and Theory*, *28*, 69–125. doi:10.1016/s0079-7421(08)60488-0

Holmes, N. M., Griffiths, O., & Westbrook, R. F. (2014). The influence of partner cues on the extinction of causal judgments in people. *Learning & Behavior*, *42*, 289–303. doi:10.3758/s13420-014-0146-x

Jeffreys, H. (1961). *Theory of Probability* (3rd Edition). Oxford: Oxford University Press.

Kavanagh, D. J., Sitharthan, G., Young, R. M., Sitharthan, T., Saunders, J. B., Shockley, N., & Giannopoulos, V. (2006). Addition of cue exposure to cognitive-behaviour therapy for alcohol misuse: A randomized trial with dysphoric drinkers. *Addiction*, *101*, 1106–1116. doi:10.1111/j.1360-0443.2006.01488.x

Kinder, A. & Lachnit, H. (2003). Similarity and discrimination in human pavlovian conditioning. *Psychophysiology*, *40*, 226–234.

Kruschke, J. K. (2013). Bayesian estimation supersedes the t-test. *Journal of Experimental Psychology-General*, *142*, 573–603. doi:10.1037/a0029146

Kruschke, J. K. (2015). *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan* (2nd edition). London: Elsevier/Academic Press.

Lachnit, H. (1988). Convergent validation of information-processing constructs with Pavlovian methodology. *Journal of Experimental Psychology-Human Perception and Performance*, *14*, 143–152. doi:10.1037/0096-1523.14.1.143

Lamarre, J. & Holland, P. C. (1987). Transfer of inhibition after serial feature negative discrimination training. *18*, 319–342. doi:10.1016/0023-9690(87)90001-4

Lee, M. D. & Wagenmakers, E. J. (2014). *Bayesian cognitive modeling: A practical course*. Cambridge University Press.

Leung, H. T., Reeks, L. M., & Westbrook, R. F. (2012). Two ways to deepen extinction and the difference between them. *Journal of Experimental Psychology-Animal Behavior Processes*, *38*, 394–406. doi:10.1037/a0030201

Migo, E. M., Corbett, K., Graham, J., Smith, S., Tate, S., Moran, P. M., & Cassaday, H. J. (2006). A novel test of conditioned inhibition correlates with personality measures of schizotypy and reward sensitivity. *Behavioural Brain Research*, *168*, 299–306. doi:10.1016/j.bbr.2005.11.021

Morell, J. R. & Holland, P. C. (1993). Summation and transfer of negative occasion setting. *Animal Learning & Behavior*, *21*, 145–153. doi:10.3758/bf03213394

Morey, R. D. & Rouder, J. N. (2018). BayesFactor: Computation of Bayes factors for common designs. Retrieved from https://CRAN.R-project.org/package=BayesFactor

Nelson, J. B., Sanjuan, M. D., Vadillo-Ruiz, S., Perez, J., & Leon, S. P. (2011). Experimental renewal in human participants. *Journal of Experimental Psychology-Animal Behavior Processes*, *37*, 58–70. doi:10.1037/a0020519

Nigg, J. T. (2000). On inhibition/disinhibition in developmental psychopathology: Views from cognitive and personality psychology and a working inhibition taxonomy. *Psychological bulletin*, *126*, 220. doi:10.1037//0033-2909.126.2.220

Pearce, J. M. (1987). A model of stimulus generalisation for Pavlovian conditioning. *Psychological Review*, *94*, 61–73.

Pearce, J. M. (1994). Similarity and discrimination - a selective review and a connectionist model. *Psychological Review*, *101*, 587–607. doi:10.1037//0033-295x.101.4.587

Pearce, J. M., Redhead, E. S., & George, D. N. (2002). Summation in autoshaping is affected by the similarity of the visual stimuli to the stimulation they replace.

Journal of Experimental Psychology-Animal Behavior Processes, *28*, 175–189. doi:10.1037/0097-7403.28.2.175

Plummer, M. (2017). JAGS: Just another Gibbs sampler. Retrieved from https://sourceforge.net/projects/mcmc-jags/files/

Polack, C., Laborda, M., & Miller, R. (2012). Extinction context as a conditioned inhibitor. *Learning & Behavior*, 24–33. doi:10.3758/s13420-011-0039-1

R Core Development Team. (2012). R: A language and environment for statistical computing. Computer Program. Vienna, Austria: R Foundation for Statistical Computing.

Rescorla, R. A. (1969). Pavlovian conditioned inhibition. *Psychological Bulletin*, *72*, 77–94. doi:10.1037/h0027760

Rescorla, R. A. (1972). "configural" conditioning in discrete-trial bar pressing. *Journal of Comparative and Physiological Psychology*, *79*, 307–317.

Rescorla, R. A. (1973). Evidence for a "unique stimulus" account of configural conditioning. *Journal of Comparative and Physiological Psychology*, *85*, 331–338.

Rescorla, R. A. (1987). Facilitation and inhibition. *Journal of Experimental Psychology-Animal Behavior Processes*, *13*, 250–259. doi:10.1037/0097-7403.13.3.250

Rescorla, R. A. (2003). Protection from extinction. *Learning & Behavior*, *31*, 124–132. doi:10.3758/bf03195975

Rescorla, R. A. & Wagner, A. R. (1972). A theory of Pavlovian conditioning: variations in the effectiveness of reinforcement and non-reinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning II: Current research and theory* (pp. 64–69). New York: Appleton Century Crofts.

Robbins, T. W., Gillan, C. M., Smith, D. G., de Wit, S., & Ersche, K. D. (2012). Neurocognitive endophenotypes of impulsivity and compulsivity: Towards dimensional psychiatry. *Trends in Cognitive Sciences*, *16*, 81–91. doi:https://doi.org/10.1016/j.tics.2011.11.009

Shanks, D. R., Charles, D., Darby, R. J., & Azmi, A. (1998). Configural processes in human associative learning. *Journal of Experimental Psychology-Learning Memory and Cognition*, *24*, 1353–1378. doi:10.1037/0278-7393.24.6.1353

Shiban, Y., Pauli, P., & Mühlberger, A. (2013). Effect of multiple context exposure on renewal in spider phobia. *Behaviour research and therapy*, *51*, 68–74. doi:10.1016/j.brat.2012.10.007

Shiban, Y., Schelhorn, I., Pauli, P., & Muehlberger, A. (2015). Effect of combined multiple contexts and multiple stimuli exposure in spider phobia: A randomized clinical trial in virtual reality. *71*, 45–53. doi:10.1016/j.brat.2015.05.014

Swartzentruber, D. (1995). Modulatory mechanisms in Pavlovian conditioning. *Animal Learning & Behavior*, *23*, 123–143. doi:10.3758/bf03199928

Thomas, B. L. & Ayres, J. J. B. (2004). Use of the ABA fear renewal paradigm to assess the effects of extinction with co-present fear inhibitors or excitors: Implications for theories of extinction and for treating human fears and phobias. *Learning and Motivation*, *35*, 22–52. doi:10.1016/s0023-9690(03)00040-7

Trask, S., Thrailkill, E. A., & Bouton, M. E. (2017). Occasion setting, inhibition, and the contextual control of extinction in Pavlovian and instrumental (operant) learning. *Behavioural Processes*, *137*, 64–72. doi:https://doi.org/10.1016/j.beproc.2016.10.003

Wagenmakers, E.-J., Lodewyckx, T., Kuriyal, H., & Grasman, R. (2010). Bayesian hypothesis testing for psychologists: A tutorial on the Savage-Dickey method. *Cognitive Psychology*, *60*, 158–189. doi:10.1016/j.cogpsych.2009.12.001

Wagner, A. R., Brandon, S. E., Mowrer, R. R., & Klein, S. B. (2001). A componential theory of pavlovian conditioning. In *Handbook of contemporary learning theories* (pp. 23–64). Mahwah, N.J.: Lawrence Erlbaum Associates.

Weisstein, E. W. (2017). Normal difference distribution. Retrieved October 13, 2017, from http://mathworld.wolfram.com/NormalDifferenceDistribution.html

Williams, D. A. (1995). Forms of inhibition in animal and human learning. *Journal of Experimental Psychology-Animal Behavior Processes*, *21*, 129–142. doi:10.1037/0097-7403.21.2.129

Young, M. E., Wasserman, E. A., & Johnson, J. L. (2000). Positive and negative patterning in human causal learning. *Quarterly Journal of Experimental Psychology*, *53B*, 121–138.

	Acquisition	Extinction	Summation test	Recovery test
Context	A:	B:	B:	C:
	$D \rightarrow X$ x10	$D \rightarrow Z$ x8		$D \rightarrow Z$ x2
	$E \rightarrow Y$ x10	$E \rightarrow Y$ x8		
	$F \rightarrow Z$ x10	$F \rightarrow Z$ x8		
	$G \rightarrow X$ x10		$G \rightarrow Z$ x2	

Table 1

Design for Experiment 1. Three different experimental contexts were used (A:, B:, and C:) along with four different cues (D, E, F, and G) which were arranged to signal outcomes (X, Y, and Z) as indicated. Outcomes X and Y were different coloured flashes, outcome Z was no-outcome. The experiment was run in four consecutive phases. Each phase contained the trial types in the numbers indicated (68 trials in total) with order randomised within block. The acquisition phase was divided into 5 blocks (1...5) with two trials of each type in each block. The extinction phase was divided into 4 blocks (6...9) again with two trials of each type in each block. The summation test consisted of two single trial blocks (10,11) as did the recovery test (12,13). See text for further details.

		Feature negative survival	
Feature negative	Feature reversal	Reminder	Test
$I+$ x10	$J+$ x8	$I+$ x2	$IJ-$ x2
$IJ-$ x10	$M-$ x8		
$K-$ x10			
$KL+$ x10			

Table 2

Design for Experiment 2. Five different cards were used, represented in the table as I, J, K, L, and M. The cards signalled one of two outcomes, either win (+) or lose (-). The experiment was run in four consecutive phases with the trial types indicated in the table (60 trials in total) with order randomised within block. The feature negative phase was divided into 5 blocks (1...5) with two trials of each type in each block. The feature reversal phase was divided into 4 blocks (6...9) again with two trials of each type in each block. The feature negative survival phases consisted of four single trial blocks (10...13). The $I+$ reminder trials were presented in blocks 10,11 and the $IJ-$ test trials were presented in blocks 12,13. See text for further details.

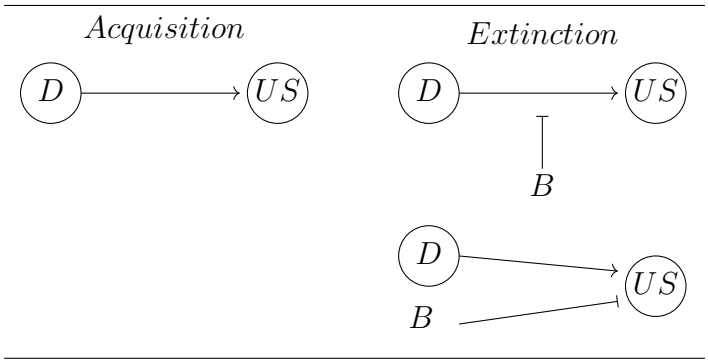


Figure 1. Illustration of first order and second order associative structures formed during acquisition and extinction (D=cue D, B=context B). Arrow headed lines represented excitatory links, stopped lines represent inhibitory links. The Rescorla-Wagner model suggests first order associations are formed during extinction (bottom-right) whereas an occasion setting model suggests second order associations are formed (top-right). See introductory text starting on pages 5-8 for further details.

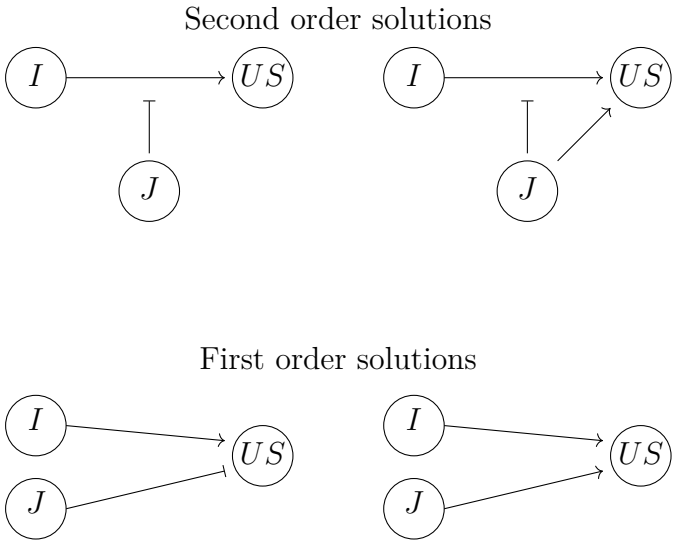
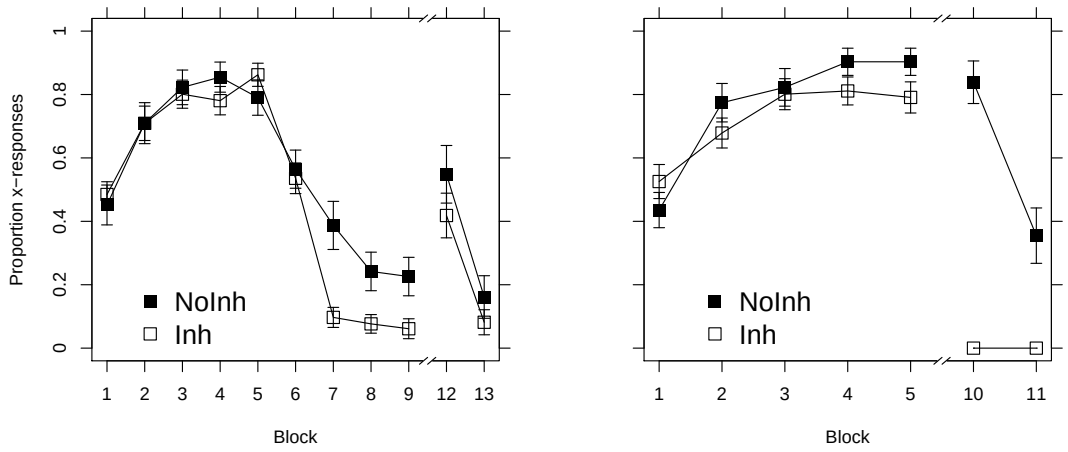
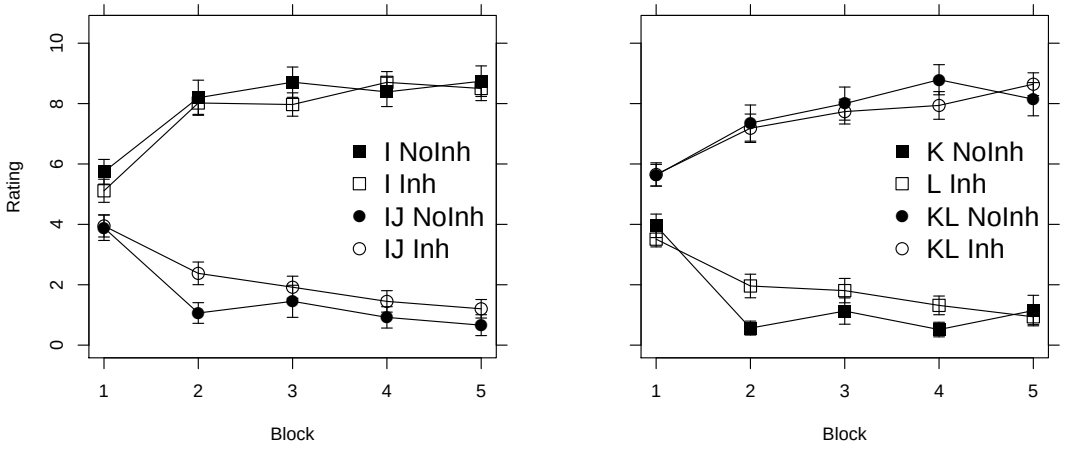


Figure 2. Status of first and second-order associative structures following training in a feature negative discrimination (I+/IJ- trials, left-hand side) and after reinforcement of the feature (J+ feature reversal trials, right-hand side). See introductory text starting on page 8 for further details



(a) Cue D (b) Cue G

Figure 3. Proportion of trials within blocks on which participants produced an x-response during Experiment 1 for participants classified as inhibitors (Inh) or as non-inhibitors (NoInh) during the summation test on cue *G*. Left-hand side shows responses to cue D during acquisition (blocks 1-5), extinction (blocks 6-9), and recovery test (blocks 12, 13) phases. Right-hand side shows responses to cue G during acquisition (blocks 1-5) and summation test (blocks 10, 11) phases. Means ± 1 standard error.



(a) Feature negative (b) Feature positive

Figure 4. Mean outcome ratings within block during Experiment 2. The left-hand side shows progression of the feature negative discrimination and the right-hand side shows the feature positive discrimination that were learned during blocks 1-5 for participants classed as inhibitors (Inh) or as non-inhibitors (NoInh) in the summation test on cue G. Means ± 1 standard error.

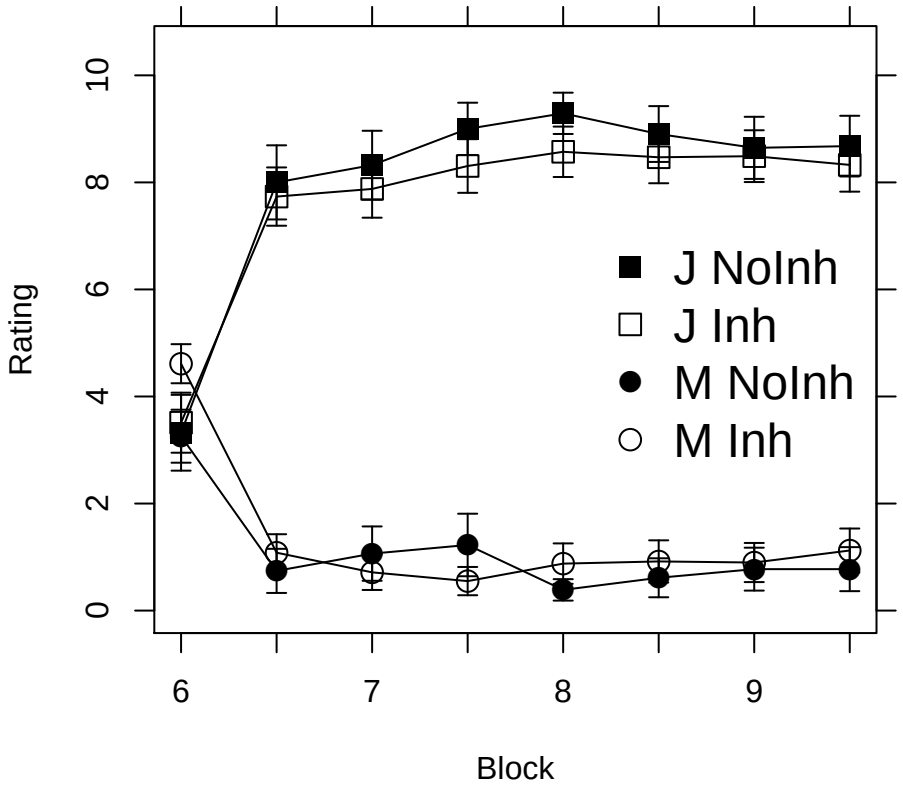


Figure 5. Mean outcome ratings within block during Experiment 2 during the reversal phase (blocks 6-9) for participants classed as inhibitors (Inh) or as non-inhibitors (NoInh) in the summation test on cue G. Each block contained two trials which are plotted separately. Means ± 1 standard error.

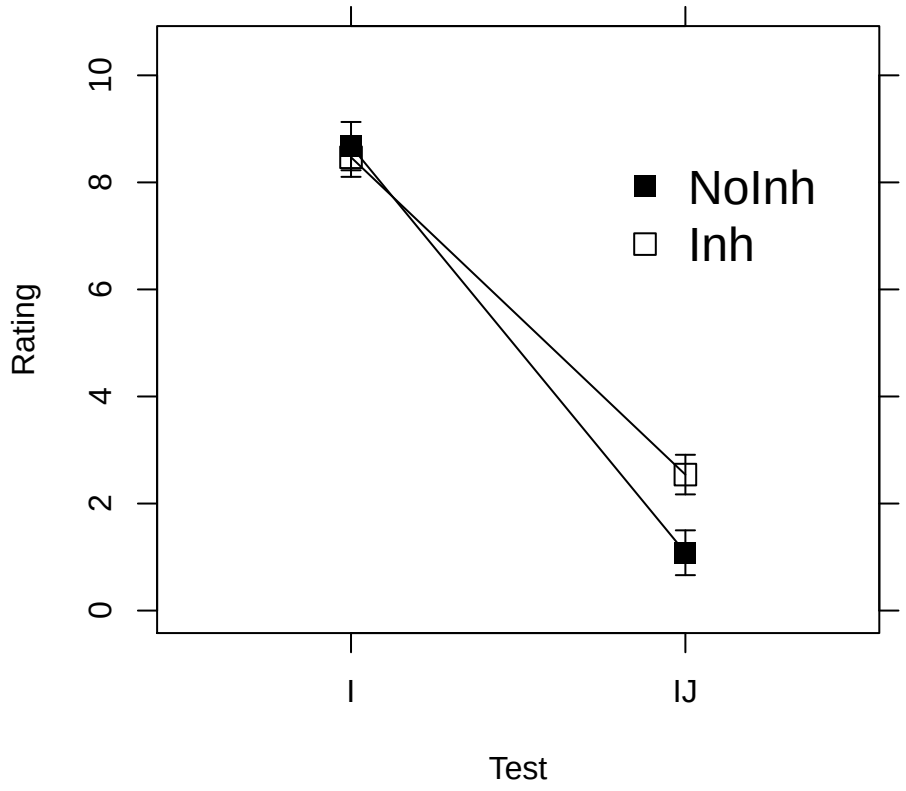


Figure 6. Mean outcome ratings within blocks during the Experiment 2 feature negative survival test for participants classified as inhibitors (Inh) or as non-inhibitors (NoInh) during the Experiment 1 summation test on cue G. Ratings averaged over blocks 10 and 11 for the reminder test on cue I, and over blocks 12 and 13 for the critical feature negative test on the IJ cue compound. Means $\pm$ 1 standard error.

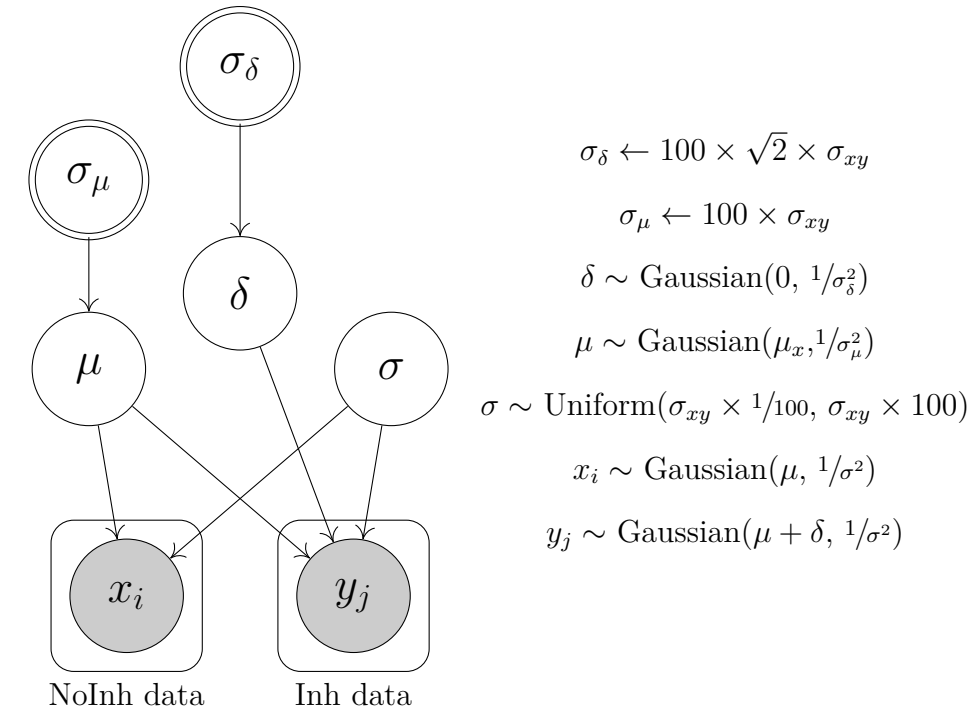


Figure 7. Graphical model for analysis to obtain posterior distributions (Figure 8) for the ratings of inhibitors and non-inhibitors in the feature negative survival test on IJ. σ_{xy} is pooled standard deviation of observed data from both groups, μ_x is the mean of the data from the non-inhibitors group. The parameterisation of Gaussian distributions in JAGS is in terms of mean and precision where precision is $1/\sigma^2$.

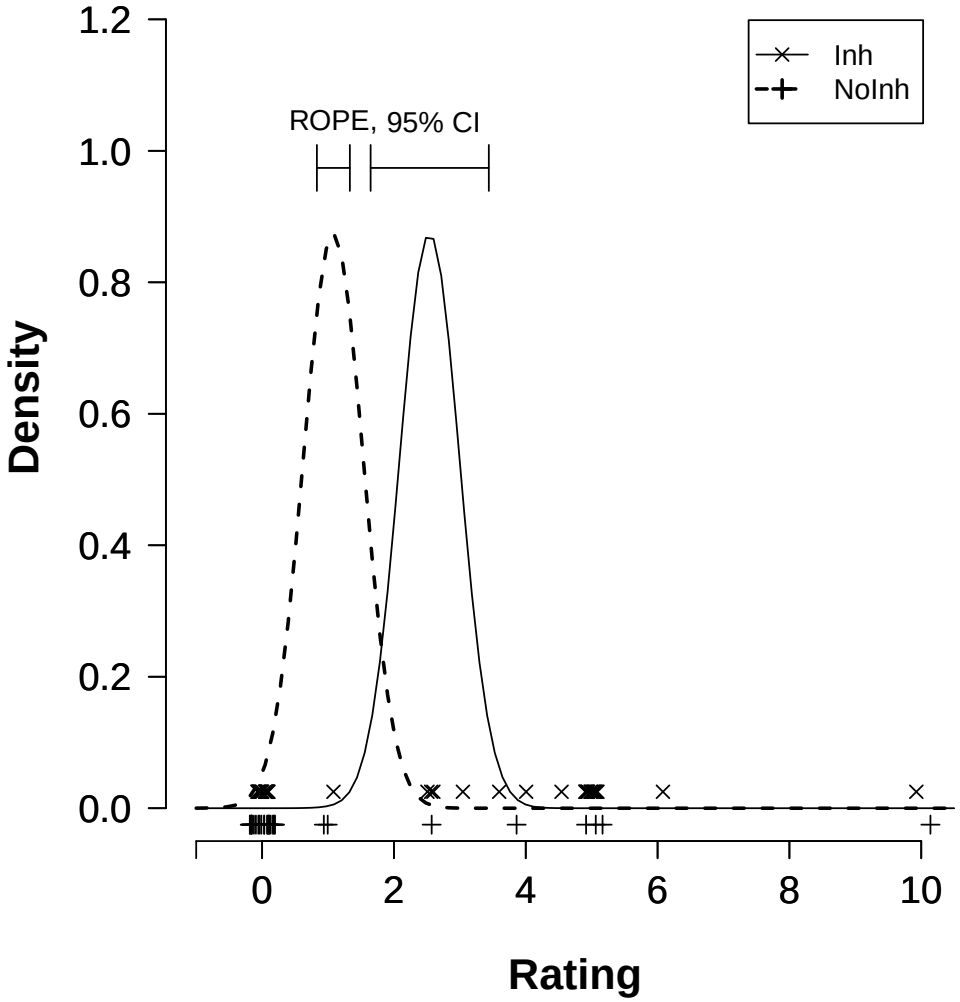


Figure 8. Jittered raw data and posterior distributions for the ratings given by the non-inhibitors and inhibitors in the feature negative survival test on IJ. The region of practical equivalence and the 95% Bayesian credible interval are given around the means of the distributions for the non-inhibitors and the inhibitors, respectively. The posterior distributions generated according the model specification illustrated in Figure 7.