Stable Individual Differences in Occasion Setting

Steven Glautier and Ovidiu Brudan

Southampton University

6438 words 15/01/18

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 Experiment 1 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. Our results were repeated in two replications and supported by Bayesian analyses. 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. 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. 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. Procedurally, presenting a conditioned stimulus (CS) alone, without the unconditioned stimulus (US) it was previously paired with, is known as extinction. 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. Actually 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 changes 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 consideration of these practical implications for the Discussion to focus here on two

[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.

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 $V > 0$ presentation of the CS activates the US representation – informally presentation of the CS leads to an expectation of the US through spreading activation. However, V may also be less than zero, in which case inhibitory, rather than excitatory learning has taken place. When there is an inhibitory CS-US association presentation of the CS effectively suppresses expectation of the US. Inhibitory learning takes place when an expected US fails to occur and is of particular concern in the current analysis because US expectancy is violated when extinction begins.

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

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 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 A after an acquisition phase involving a series of $A\!:\!A+$ trials. Asymptotically $V_{A:} + V_A \rightarrow 1$ and ${}^{V_{A:}}\!/_{V_A} = {}^{\alpha_{A:}}\!/_{\alpha_A}$. Following acquisition there is an extinction phase involving $B\!:\!A-$ trials. At the start of extinction $V_{B:} = 0$ and $\Sigma V = V_{B:} + V_A > 0$. 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_A declines. Learning (and responding) during extinction stops when $V_{B:} + V_A = 0$ and at this point $B\!:$ is inhibitory ($V_{B:} < 0$) while A remains excitatory ($V_A > 0$). The presence of inhibitory $B\!:$ is said to protect A from extinction.

Despite the fact that the presence of an inhibitory CS during extinction can protect a CS from extinction (e.g. Rescorla, 2003) there are few experiments which have found evidence for context inhibition developing during extinction, calling into question the proposal that protection from extinction could occur. 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 find extinction context inhibition effects 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). The 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_A was paired with the US during an acquisition phase to produce an $A \rightarrow US$ association (left) and then, in a new context, context B, CS_A was extinguished by presentation without the US. According to the Rescorla-Wagner model, as explained above, this produces an inhibitory association between CS_A and context B (bottom right). In contrast, in the occasion-setting model, it is assumed that context B forms an inhibitory link with the $A \rightarrow US$ 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_A is presented outside of context B the $A \rightarrow US$ 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 $A \rightarrow US$ 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 $CS_A \rightarrow US$ relation then it should not affect another CS-US relation (e.g. $CS_B \rightarrow US$). 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 each of the identically designed experiments reported on below had two parts. Part 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. In Part 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. It was predicted that the 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 $A+$ and $AB-$ 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 B the second-order associative structures formed by the non-inhibitors will have an excitatory input to the US representation from B (right-hand side, top). There is also an excitatory input to the US representation from A, but this association is gated closed by the presence of B. In contrast the inhibitors will have excitatory inputs to the US representation from both A and B (right-hand side, bottom). Consequently we predicted stronger responses in the AB compound test for the inhibitors than for the non-inhibitors.

Experiment 1

Participants took part in a two-part computer-based experiment. Both parts involved predictive learning tasks. Participants viewed a series of on-screen trials and they 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. In Part 1 participants learned to predict coloured flashes on the computer screen on the basis of cues provided in the form of visually distinctive objects 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, 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. Participants had to press a key to indicate their expectation, with respect to the coloured flashes, on each trial before the objects passed the sensor. There were three 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. Full details of the task are given in Glautier et al. (2013). Part 1 contained acquisition, extinction, and response recovery phases. Responses made in a summation test carried out at the end of the extinction phase were used to classify participants as either non-inhibitors or inhibitors.

In Part 2 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...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. Full details of the task are given in Glautier et al. (2013, Experiment 1). Part 2 contained feature negative training, reversal learning for the feature, and the *AB* compound test in which we examined differences between the participants classified as either non-inhibitors or inhibitors on the basis of their responses in the summation test from Part 1. Inhibitors were defined as those participants who completely suppressed responding to cue G when it was presented in context B after extinction.

Method

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

Participants. Twenty-eight participants were recruited by word of mouth by student volunteer research assistant JR from a local community in Wiltshire, UK. Their mean age was 21 years (range 16-30) and they included 11 males. They were paid £4 on completion.

Apparatus. The experiment was run on a laptop computer with a screen measuring 30.5 cm x 19.2 cm (W x H). The screen was run at 60hz in 32 bit colour mode with pixel resolutions of 1440 x 900. The display was controlled by a computer programs written by the first author in Microsoft C# language and used Microsoft XNA Game Studio Version 3.1 for rendering of the experimental scenario.

Design and procedure. Table 1 shows the design for Part 1. The columns in the table show the trial types that were presented in each stage of the task. Different stages took place in different contexts wherein the screen background changed. Acquisition, extinction, and the response recovery test took place the three different

contexts i.e. it was an ABC design. Characters A, B, C, and G indicate the identity of the cue object that was present on each trial. Characters X, Y, and Z indicate the trial outcomes. The screen backgrounds that served the roles of contexts A:, B:, and C: were selected randomly without replacement from four possibilities for each participant. The cue objects serving the roles of A, B, C, 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. Trial order was randomised within each stage separately for each participant subject to the constraint that there could be no more that four repeats of a trial type in a single back-to-back sequence.

Table 2 shows the design for Part 2. The columns in the table show the trial types that were presented in each stage of the task. Stages were not differentiated by context changes. Characters A, B, C, D, 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 A, B, C, D, 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 onscreen their right-left location was randomised. Trial order was randomised within each stage separately for each participant subject to the constraint that there could be no more that four repeats of a trial type in a single back-to-back sequence. We were interested primarily in the feature negative training (A+, AB- 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 (C-, CD+ trials) to ensure that participants attended to the identity of the cues.

Analysis. Data processing in the Results and Data analyses sections presented below were undertaken using R, JAGS, and associated packages (Plummer, 2017; R

Core Development Team, 2012).

Results

Acquistion, extinction, summation test, and recovery. Figure 3a shows the proportion of x-responses made in response to cue A during the acquisition, extinction, and response recovery test phases. The proportion of x-responses increased during the acquisition phase, decreased during extinction, and then recovered on the first trial of the recovery test. The proportion of x-responses to cue A was higher on the first trial of the recovery test than it was on the last block of extinction. Figure 3b shows the proportion of x-responses made to cue G during the acquisition and summation test phases. Responses increased during the acquisition phase and were reduced during the summation test.

Feature negative discrimination and feature reversal. Figure 4a shows that the feature negative and the feature positive discriminations were acquired successfully. The outcome was rated more likely on trials with cue A and on trials with the cue compound CD than on trials with the cue compound AB and on trials with cue C. The reversal training discrimination was also acquired successfully; cue B attracted higher outcome ratings than did cue M (Figure 4b). Of particular interest was survival of the feature negative discrimination following reversal training. The difference in response to cue A and the cue compound AB after reversal training is shown in Figure 5a where it can be seen that cue B maintained it's capacity to suppress responses to cue A. Responses to cue A were higher on the A+ reminder trials than on the AB- feature negative survival test trials.

AB compound test. Figure 6a shows responses to the AB compound after feature reversal as a function of inhibition group. The Inhibition group produced higher ratings for the outcome than the No-Inhibition group. There were 18 participants classified as inhibitors out of 28.

Experiment 2

Differences from Experiment 1 are noted below.

Method

Participants. Fifty-two participants were recruited from the University of Southampton Highfield campus by word of mouth and posted advertisement. Their mean age was 20.6 years (range 18-39) and they included 13 males. Course credit was given for participation.

Apparatus. The experiment was run on personal computers with screens measuring 41 cm x 26 cm (W x H). The screens were run at 75hz in 32 bit colour mode with pixel resolutions of 1440 x 900.

Results

Acquistion, extinction, summation test, and recovery. Figure 3c shows the proportion of x-responses made in response to cue A during the acquisition, extinction, and response recovery test phases. The proportion of x-responses increased during the acquisition phase, decreased during extinction, and then recovered on the first trial of the recovery test. The proportion of x-responses to cue A was higher on the first trial of the recovery test than it was on the last block of extinction. Figure 3d shows the proportion of x-responses made to cue G during the acquisition and summation test phases. Responses increased during the acquisition phase and were reduced during the summation test.

Feature negative discrimination and feature reversal. Figure 4c shows that the feature negative and the feature positive discriminations were acquired successfully. The outcome was rated more likely on trials with cue A and on trials with the cue compound CD than on trials with the cue compound AB and on trials with cue C. The reversal training discrimination was also acquired successfully; cue B attracted higher outcome ratings than did cue M (Figure 4d). Of particular interest was survival of the feature negative discrimination following reversal training. The difference in response to cue A and the cue compound AB after reversal training is shown in Figure 5b where it can be seen that cue B maintained it's capacity to suppress responses to cue A. Responses to cue A were higher on the A+ reminder trials than on the AB- feature

negative survival test trials.

AB compound test. Figure 6b shows responses to the AB compound after feature reversal as a function of inhibition group. The Inhibition group produced higher ratings for the outcome than the No-Inhibition group. There were 31 participants classified as inhibitors out of 52.

Data analyses

Two main Bayesian analyses were carried out on the critical data from the AB compound test. The analyses were based on examples given in Kruschke (2015) and Lee and Wagenmakers (2014). Data was combined over both experiments for these analyses. Bayesian analyses were chosen due to their inherent suitability for sequential updating of parameter estimates over repeated runs of an experiment (Dienes, 2011; Kruschke, 2013; Wagenmakers, Lodewyckx, Kuriyal, & Grasman, 2010) and to obtain full distributional information on the parameters of interest. 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. 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/.

Analysis 1

Figure 7 shows a graphical representation of the JAGS model that was used for Analysis 1 in which the posterior distributions for the ratings obtained in the AB compound test were estimated for the non-inhibitor and the inhibitor groups. 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 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 answer 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.

Analysis 2

Figure 9 shows a graphical representation of the JAGS model that was used for Analysis 2 in which the effect size (δ) was estimated for the difference in the AB compound test ratings between the non-inhibitor and inhibitor groups. As previously the ratings of the non-inhibitors (x_i) were assumed to come from a Gaussian distribution with mean μ and standard deviation σ. In contrast, in this analysis, the ratings of the inhibitors (y_i) were assumed to come from a Gaussian distribution with mean $\mu + \alpha$ and standard deviation σ where $\alpha = \delta \times \sigma$. The data from both groups was standardised on the mean of the inhibitors using σ_{xy} and the priors on μ and δ were Cauchy distributions with location and scale parameters 0 and 1, respectively; the prior on σ was a half-Cauchy with location and scale 0 and 1, respectively (Lee & Wagenmakers, 2014).

Figure 10 shows the prior and posterior distributions for the effect size δ. The mean of the posterior distribution (δ) was 0.55 ($\sigma = 1.01$) and the 95% credible interval boundaries for δ were $[0.11, 0.99]$. The Savage-Dickey Bayes factor shown in the Figure is the ratio of the prior and posterior distributions at $\delta = 0$; the probability that the effect size $\delta = 0$ is 3.61 less likely given the data, substantial evidence in favour of the alternative hypothesis (Jeffreys, 1961).

General discussion

In Part 1 of each of the current experiments we provided reliable demonstrations of ABC renewal effects – restoration of an extinguished response consequent to a change of context (Figures 3a and 3c). We also clearly showed that context inhibition

developed during extinction, due to the suppression of responding to cue G which occurred when G was presented in the extinction context (Figures 3b and 3d). A reasonable objection to this claim is that a control condition was not included, the suppressed response to cue 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 (cue G, Figures 2 and 3 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 experiments and over both tests on cue G, as done in Glautier et al., the mean rating for cue G was 0.23 (SE=0.04, n=80).

However, again repeating the findings of Glautier et al., 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 this classification of participants, as inhibitors and non-inhibitors, indicated that different associative structures were being used to resolve the ambiguity in the predictive value of cue A during extinction. Importantly, in Part 2 of each of the current experiments, we observed that the classification established in Part 1 enabled prediction of performance in a test in Part 2. Specifically, a feature negative discimination survived reversal training of the feature to a greater extent among the non-inhibitors than amongst the inhibitors, consistent with the logic of the proposed associative structures as described in the Introduction (Figure 6). Our conclusion that non-inhibitors and inhibitors differ on feature negative survival after reversal training of the feature is supported by replication in two experiments and in Bayesian analyses of the data combined over the two experiments. These analyses showed that the posterior distributions representing

the responses on the feature negative survival test differed substantially for the two groups (Figure 8). We estimated an effect size index for the difference between the two groups as $\delta = 0.55$ (95% credible interval $[0.11, 0.99]$, Figure 10) which in Cohen's terms is a medium effect size (Cohen, 1988).

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 AB compound test after reinforcement of feature B than participants using second-order strategies. 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). However, this observation has also been discussed in relation to stimulus configuration and rule-based learning (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 a) configural learning or b) rule learning. Perhaps distinctions along one or both of these axes fit the facts just as well differences in the 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 ($A+/AB-$ trials) an excitatory association between cue A and the US is formed alongside an inhibitory association between cue B and the US. In contrast, in Pearce's model an excitatory association between cue A and the US is formed alongside an inhibitory association between a configural cue AB and the US. The effect of reinforced feature trials in the Rescorla-Wagner model has already been covered (page 8) and it clearly implies a strong response should then be seen when the AB compound is presented for test. In Pearce's model the reversal training will result in an excitatory representation forming between B and the US, thus there are now three associations ($A^+ \rightarrow US$, $AB^- \rightarrow US$, and $B^+ \rightarrow US$ where the plus and minus signs indicate excitatory and inhibitory respectively) which need to be taken into account to understand the predicted response in the AB compound test. The response to the AB compound is determined by the state of the $AB^- \rightarrow US$ association and by generalisation between the cues A, B, and the AB configuration. Strong generalisation from B would cause loss of the feature negative discrimination but if there was little or no generalisation then the $B+$ 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'. Presumeably this classification could be used to explain the Part 2 results (c.f. Figure 6) but would a configural with limited generalisation group be predicted to show weak inhibition in the Part 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 $B{:}A^- \rightarrow US$, and responding to the novel test $B{:}G$ depends on generalisation between $B{:}G$ and these associations. $B{:}G$ shares one common element with $A{:}G$ and one common element with $B{:}A$ and a response is predicted if the absolute value of excitatory association

exceeds that of the absolute value of inhibitory association and 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 Part 1 results is incompatible with a configural account of the group differences in Part 2.

A similar line of argument can be used to determine whether or not the observed group differences could be due to differences in rule-based learning. It has been proposed that human learning is better described in terms of rule-based rather than associative mechanisms (e.g. Mitchell, De Houwer, & Lovibond, 2009) and Shanks and Darby (1998) found evidence that rule-based learning was associated with better performance in a negative patterning discrimination. A negative patterning discrimination involves $A+$, $B+$, and $AB-$ trials and, in fact, these are exactly the trial types to which our participants were exposed in Part 2 (c.f. Table 2). Of course in our design the $B+$ trials were presented after the $A+/AB-$ trials but supposing that our non-inhibitors were better labelled 'rule-followers', parsing the trials as a single negative patterning discrimination, and the inhibitors labelled 'associative' then perhaps this could explain the Part 2 results – rule-followers should show a weak response on the AB test. But would this lead to a prediction of weak inhibilintion in a rule-following group in Part 1? Once again the answer is clearly no. An appropriate context dependent extinction rule would be 'no X outcomes in context B', hence a response should be at least as unlikely in a rule-following group (AKA non-inhibitor) as in an associative group (AKA inhibitor). This conflicts with our observations therefore a rule-based account of the group differences in Part 1 results is incompatible with a rule-based account of the group differences in Part 2.

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). 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. Extinction based on excitatory cue compounds should produced the greatest benefits for those identified as inhibitors.

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

Table 1

Design for Part 1. Each stage contained the trial types in the numbers indicated (68 trials in total) with order randomised within stage.

		Feature negative survival	
Feature negative	Feature reversal	Reminder	Test
A+ x10	B+ x8	A+ x2	AB- x2
AB- x10	M- x8		
C- x10			
CD+ x10			

Table 2

Design for Part 2. Each stage contained the trial types in the numbers indicated (60 trials in total) with order randomised within stage.

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.

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

Bouton, M. E. & Bolles, R. C. (1979). Contextual control of the extinction of conditioned fear. *Learning and Motivation*, *10*, 445–466.

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.

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.

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.

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.

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

Choy, Y., Fyer, A. J., & Lipsitz, J. D. (2007). Treatment of specific phobia in adults. *Clinical Psychology Review*, *27*, 266–286.

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

Conklin, C. A. & Tiffany, S. T. (2002). Applying extinction research and theory to cue-exposure addiction treatments. *Addiction*, *97*, 155–167.

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.

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.

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.

Dienes, Z. (2011). Bayesian Versus Orthodox Statistics: Which Side Are You On? *Perspectives on Psychological Science*, *6*, 274–290.

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.

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. eprint: /oup/backfile/content_public/journal/cercor/18/1/10.1093/cercor/bhm044/2/bhm044.pdf

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.

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

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

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.

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*.

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.

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*.

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.

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.

Holland, P. C. (1992). Occasion setting in pavlovian conditioning. *Psychology of Learning and Motivation-Advances in Research and Theory*, *28*, 69–125.

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.

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.

Kruschke, J. K. (2013). Bayesian Estimation Supersedes the t-test. *Journal of Experimental Psychology-General*, *142*, 573–603.

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.

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.

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.

Mitchell, C. J., De Houwer, J., & Lovibond, P. F. (2009). The propositional nature of human associative learning. *Behavioral and Brain Sciences*, *32*, 183–198.

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

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.

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.

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.

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.

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.

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

Rescorla, R. A. (2003). Protection from extinction. *Learning & Behavior*, *31*, 124–132.

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.

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.

Shanks, D. R. & Darby, R. J. (1998). Feature- and rule-based generalization in human associative learning. *Journal of Experimental Psychology-Animal Behavior Processes*, *24*, 405–415.

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

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.

Swartzentruber, D. (1995). Modulatory mechanisms in pavlovian conditioning. *Animal Learning & Behavior*, *23*, 123–143.

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.

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.

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.

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.

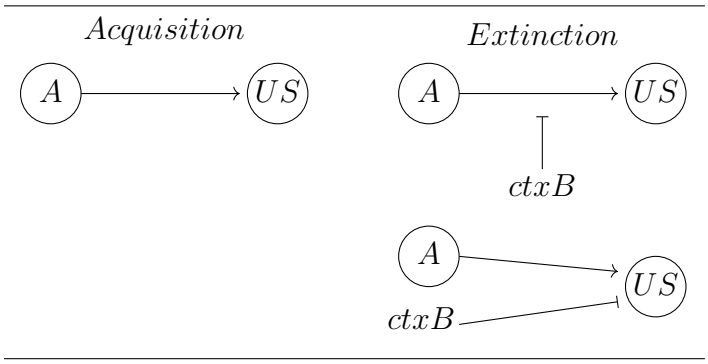


Figure 1. Illustration of first order and second order associative structures formed during acquisition and extinction (ctxB=context B). The Rescorla-Wagner model model suggests first order associations are formed during extinction (bottom-right) whereas an occasion setting model suggests second order associations are formed (top-right). See introductory text starting on pages 5-8 for further details.

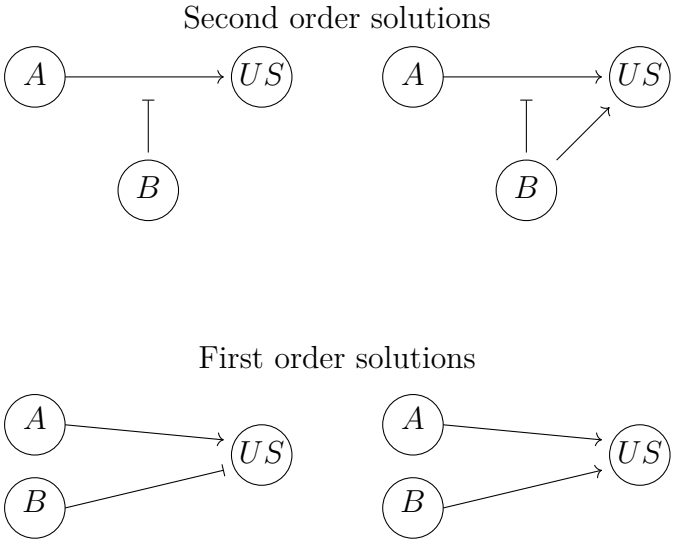
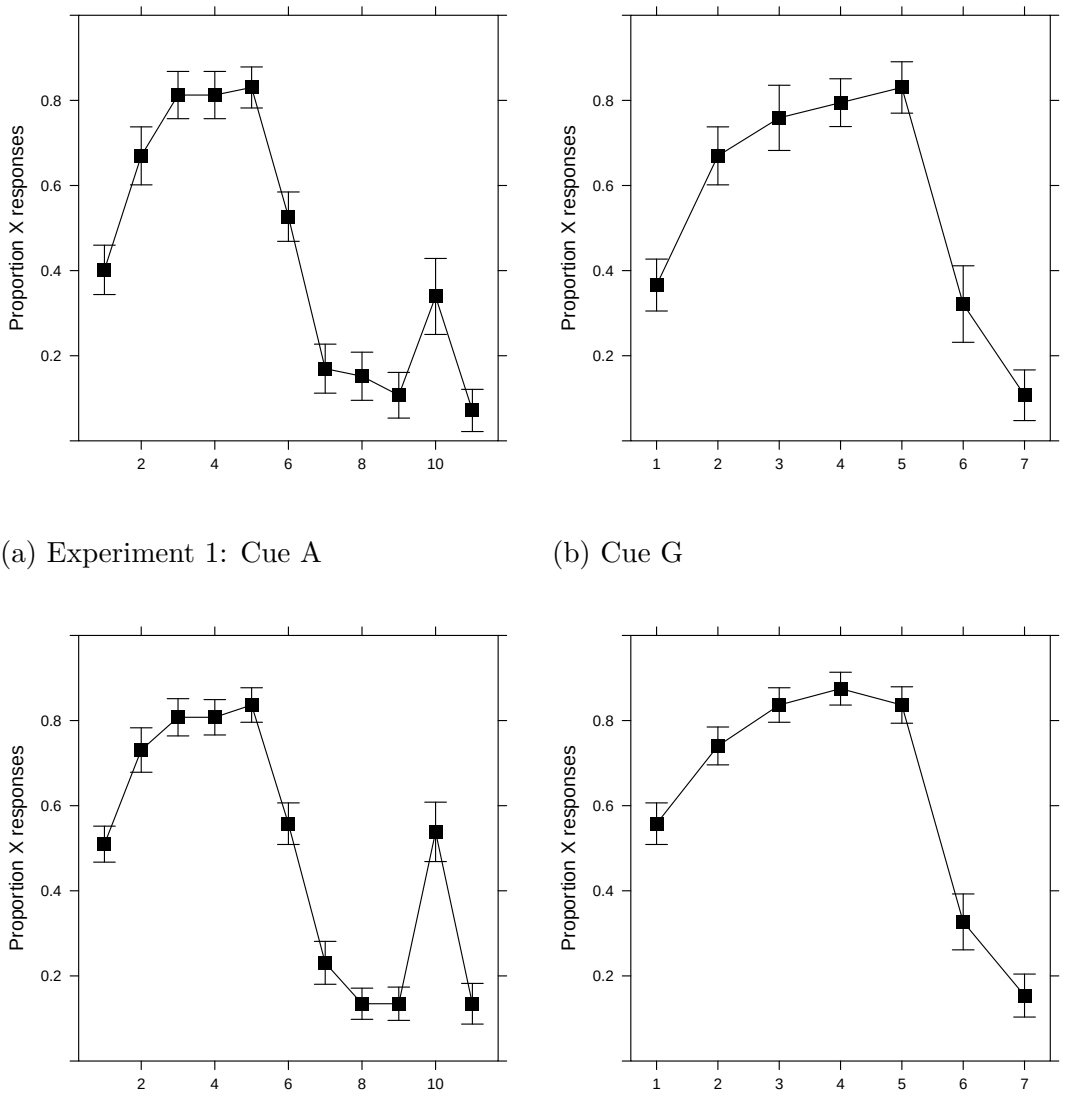


Figure 2. Status of first and second-order associative structures following training in a feature negative discrimination (left-hand side) and after reinforcement of the feature, B+ trials (right-hand side). See introductory text starting on page 8 for further details



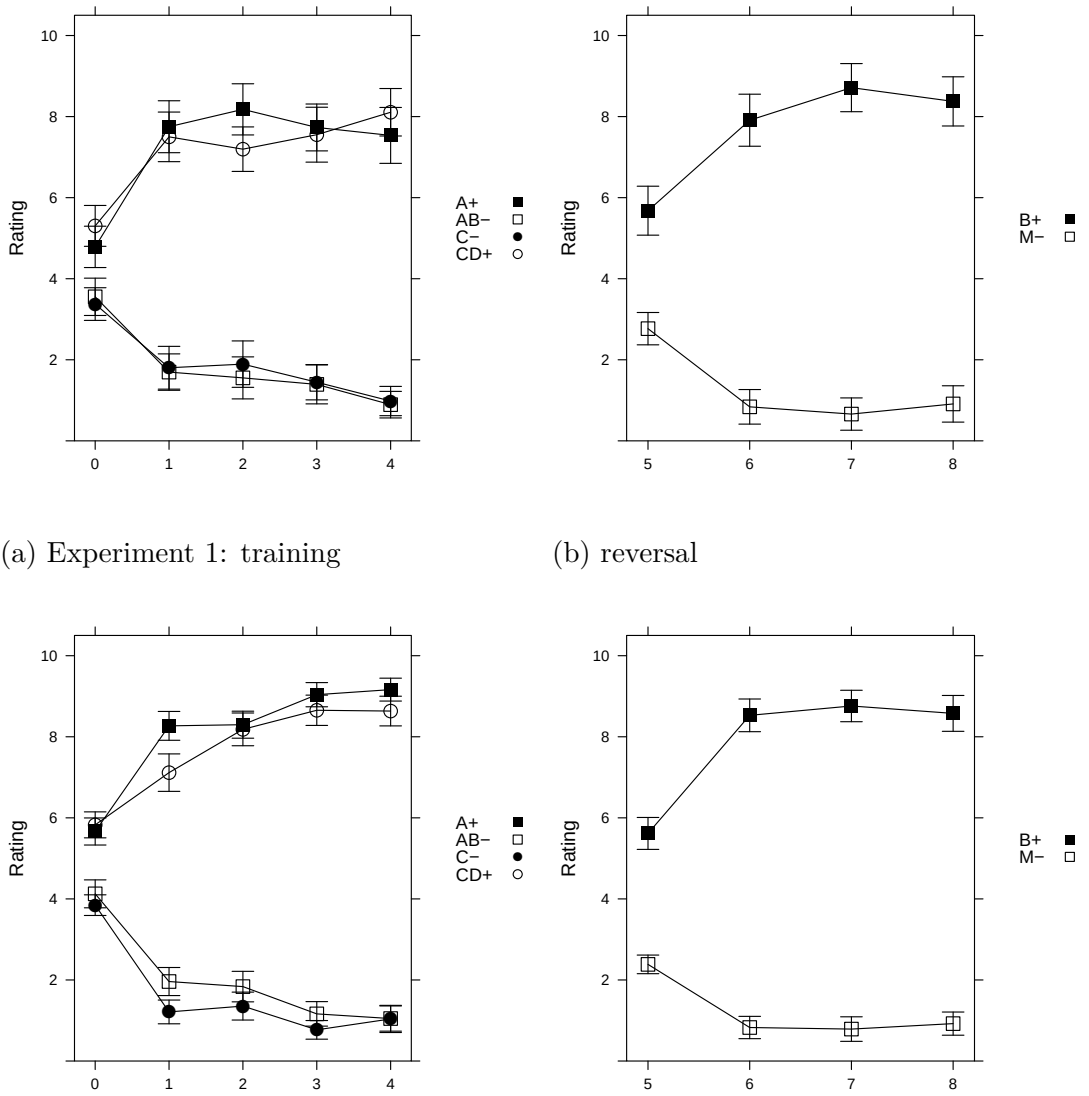
(a) Experiment 1: Cue A

(b) Cue G

(c) Experiment 2: Cue A

(d) Cue G

Figure 3. Experiments 1 and 2 x-responses to cues A and G during acquisition, extinction, summation, and recovery test phases. Means ± 1 standard error.



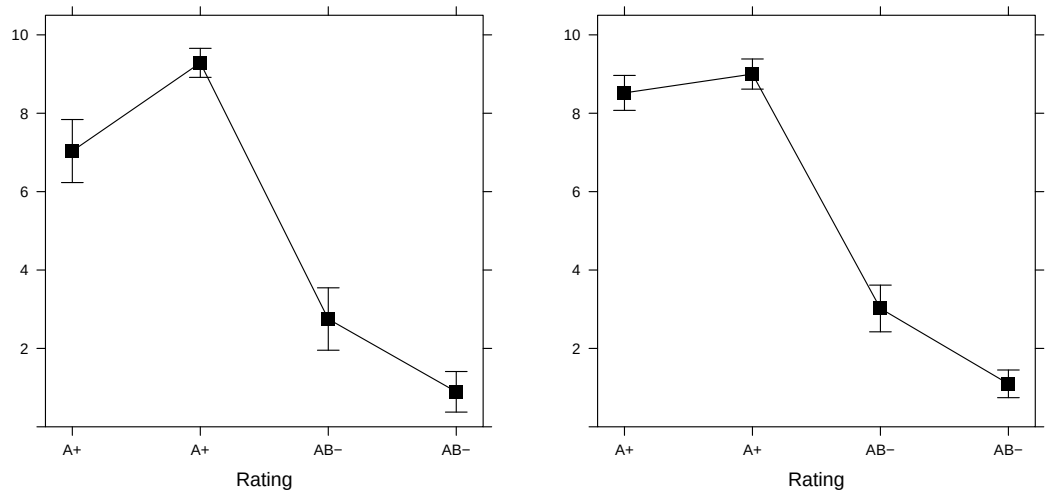
(a) Experiment 1: training

(b) reversal

(c) Experiment 2: training

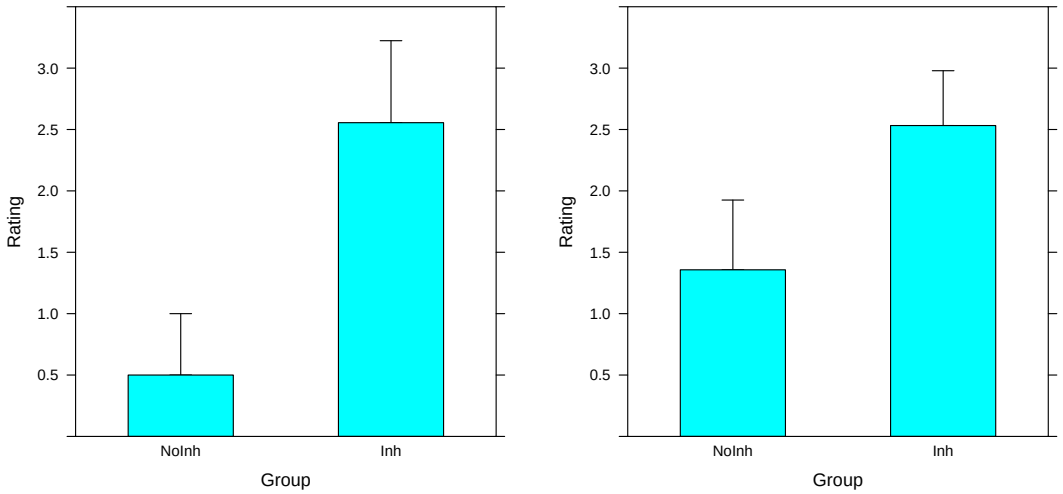
(d) reversal

Figure 4. Experiments 1 and 2 outcome-likelihood ratings during feature negative training and feature reversal phases. Means ± 1 standard error.



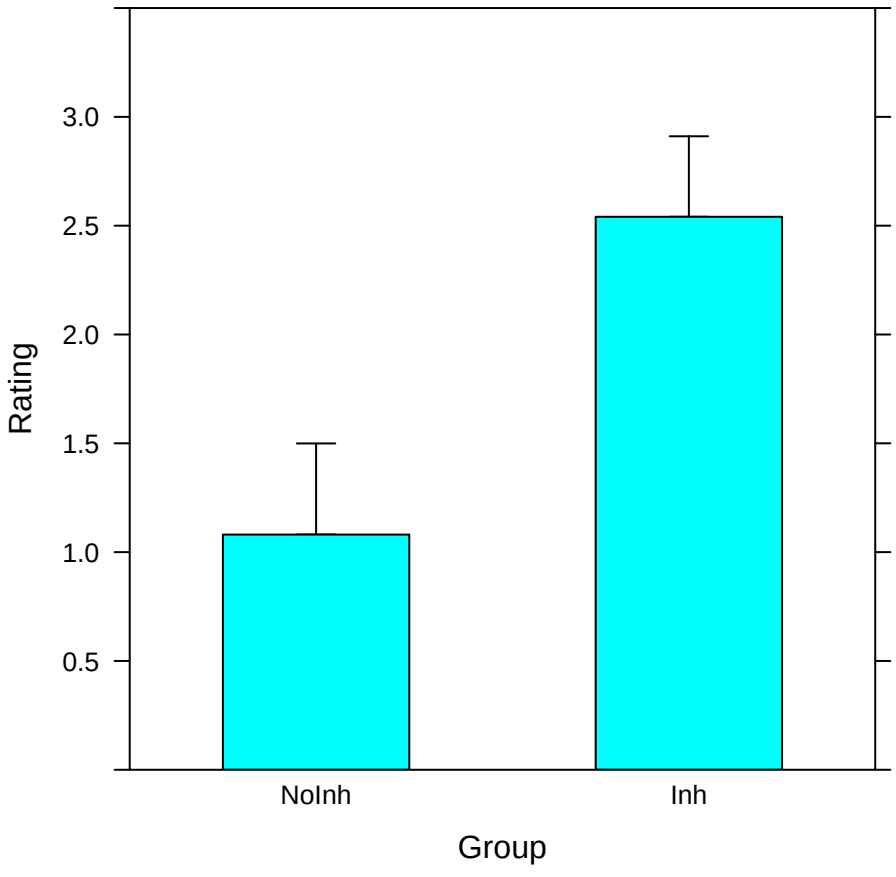
(a) Experiment 1 (b) Experiment 2

Figure 5. Experiments 1 and 2 test for survival of feature negative discrimination following feature reversal. Means ± 1 standard error.



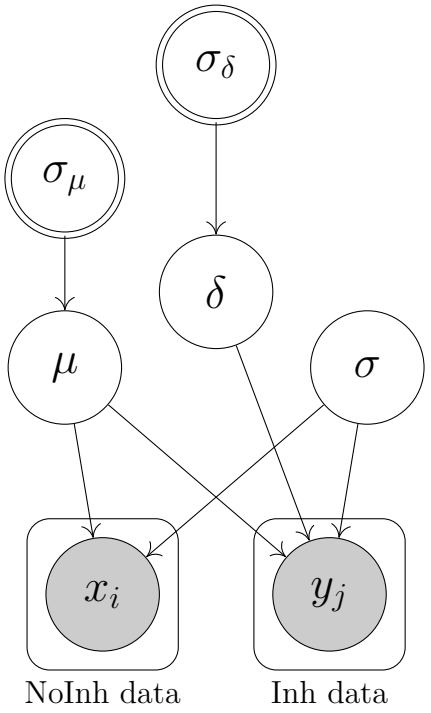
(a) Experiment 1

(b) Experiment 2



(c) Combined Experiments 1 and 2

Figure 6. Top. Experiments 1 and 2 responses to AB compound in feature negative survival test for participants classified as non-inhibitors and inhibitors. Bottom. Data combined over Experiments 1 and 2. Means ± 1 standard error.



$$\sigma_\delta \leftarrow 100 \times \sqrt{2} \times \sigma_{xy}$$

$$\sigma_\mu \leftarrow 100 \times \sigma_{xy}$$

$$\delta \sim \text{Gaussian}(0,\ {}^1\!/_{\sigma_\delta^2})$$

$$\mu \sim \text{Gaussian}(\mu_x, {}^1\!/_{\sigma_\mu^2})$$

$$\sigma \sim \text{Uniform}(\sigma_{xy} \times {}^1\!/_{100},\ \sigma_{xy} \times 100)$$

$$x_i \sim \text{Gaussian}(\mu,\ {}^1\!/_{\sigma^2})$$

$$y_j \sim \text{Gaussian}(\mu + \delta,\ {}^1\!/_{\sigma^2})$$

Figure 7. Graphical model for Analysis 1. 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 ${}^1\!/_{\sigma^2}$

.

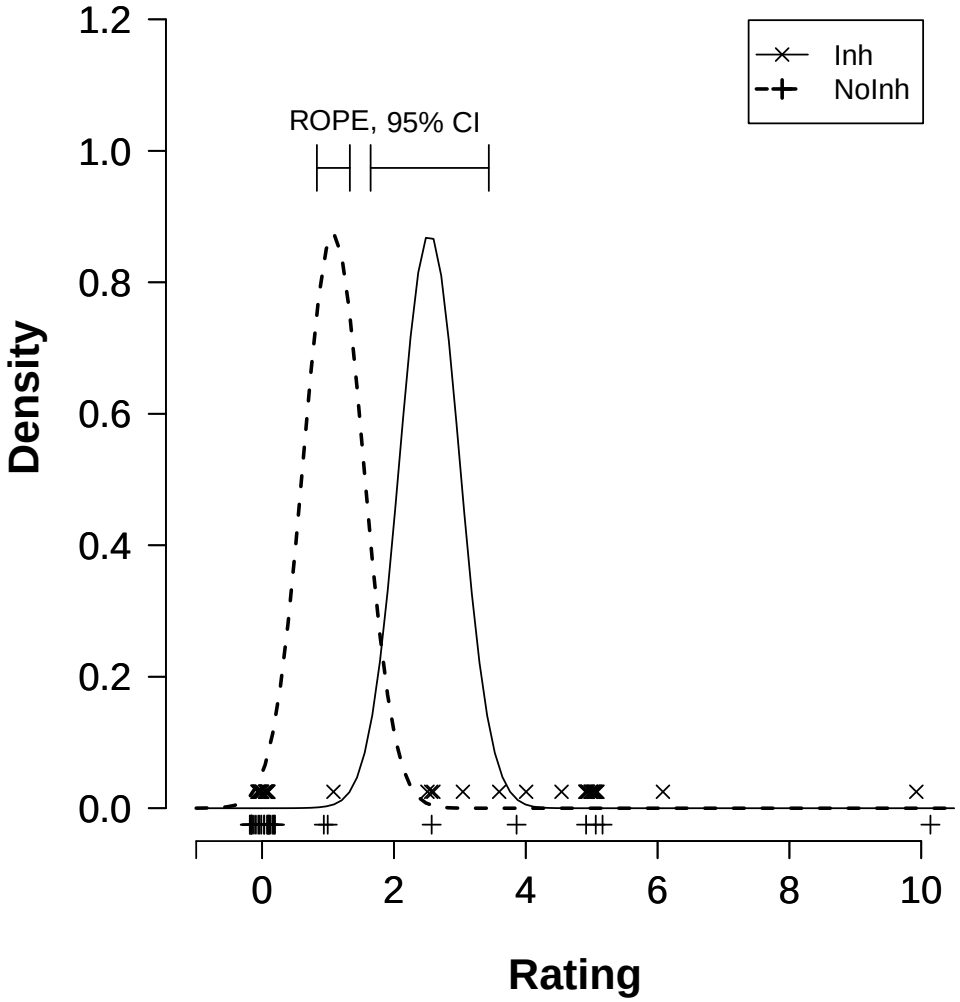


Figure 8. Jittered raw data and posterior distributions for the ratings given by the non-inhibitors and inhibitors in the AB compound test. The region of practical equivalence and the 95% Bayesian credible interval are given around the means of the distributions for the non-inhibitors and the inhibitors, respectively.

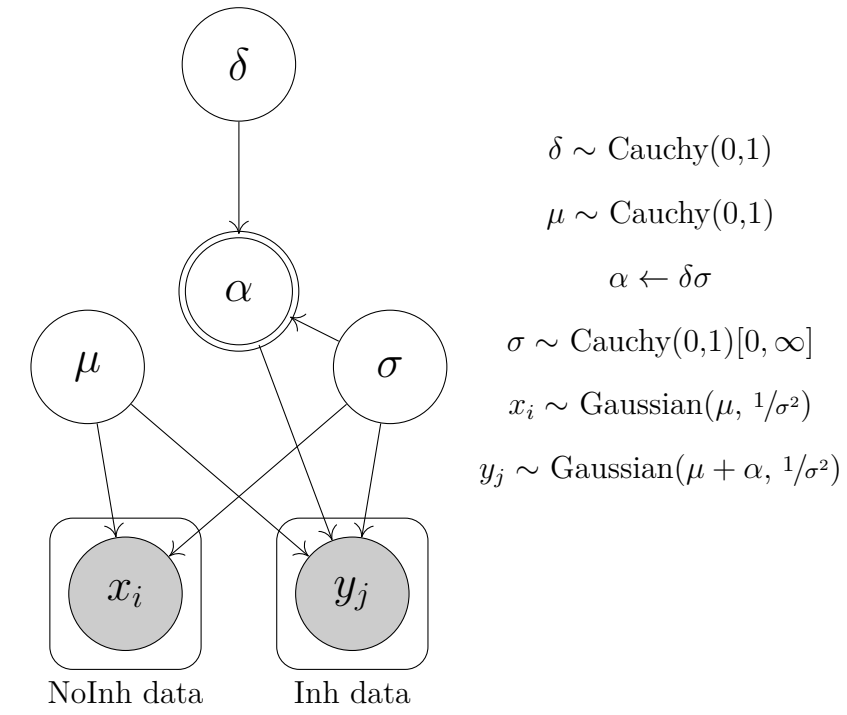


Figure 9. Graphical model for Analysis 2. The parameterisation of Gaussian distributions in JAGS is in terms of mean and precision where precision is $1/\sigma^2$

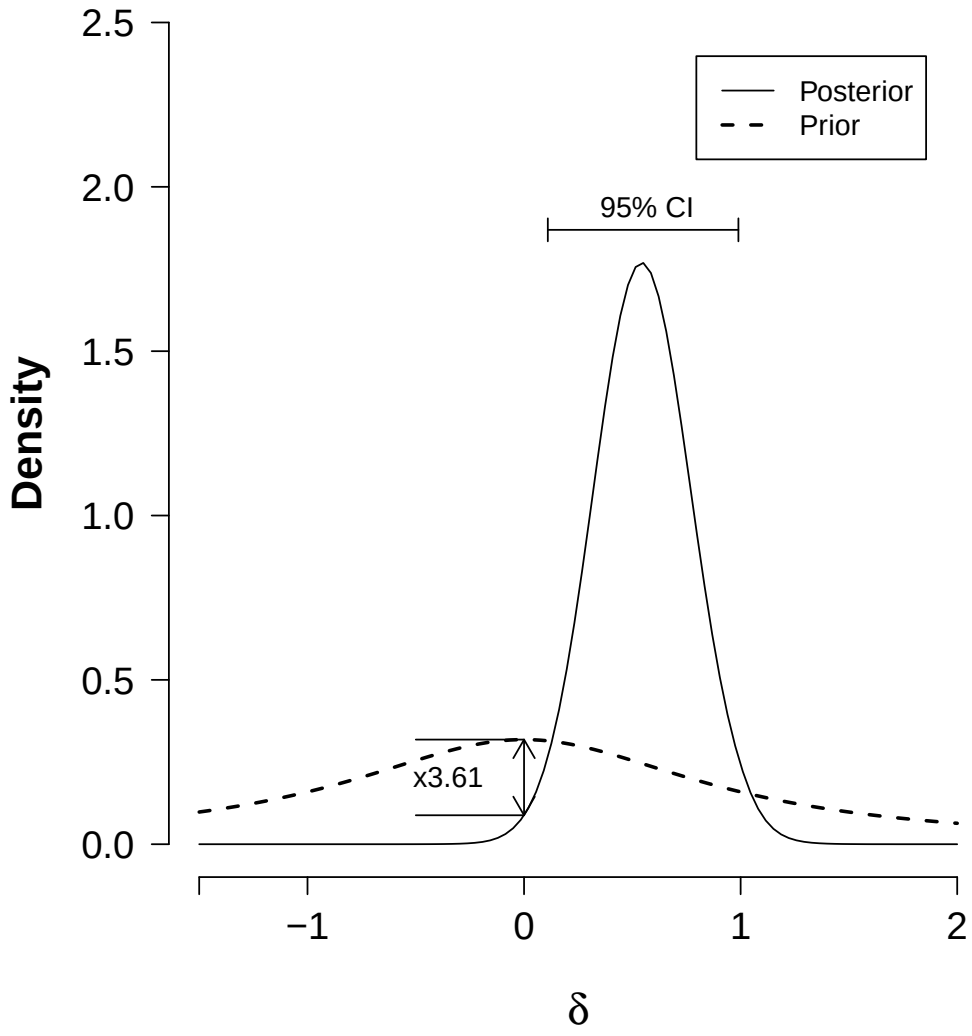


Figure 10. Prior and posterior distributions for the effect size (δ) on the difference between the group means for non-inhibitors and inhibitors in the AB compound test. The relative values of the distributions is given at $\delta = 0$ and the 95% Bayesian credible interval is given around the mean of the posterior distribution.