PERGAMON

Behaviour Research and Therapy 38 (2000) 967–983

BEHAVIOUR RESEARCH AND THERAPY

www.elsevier.com/locate/brat

# Protection from extinction in human fear conditioning

Peter F. Lovibond*, Natasha R. Davis, Ailie S. O'Flaherty

*School of Psychology, University of New South Wales, Sydney, NSW 2052, Australia*

Received in revised form 27 May 1999

**Abstract**

Two experiments examined the ability of an added stimulus to interfere with extinction of a target excitatory fear stimulus (a predictor of shock) in human autonomic conditioning. Both experiments demonstrated disruption of extinction when the added stimulus was inhibitory (a predictor of no shock, or safety signal). Subjects showed a return of fear when the target stimulus was tested alone, on both self-reported shock expectancy and skin conductance measures. The second experiment also demonstrated disruption of extinction when the added stimulus was excitatory. This result suggests that protection from extinction may occur even when the added stimulus is not inhibitory. Additional factors that may contribute to protection from extinction include context-specificity, occasion-setting and external inhibition. The results highlight the role that concurrent stimuli play in extinction, and emphasise the need to keep concurrent stimuli as similar as possible to the desired transfer context in practical applications of extinction such as exposure therapy for anxiety. 

## 1. Introduction

Protection from extinction refers to interference with extinction of a target stimulus by an inhibitory stimulus. Ordinarily, when a Pavlovian conditioned stimulus (CS) has been established as a predictor of an unconditioned stimulus (US), presentation of the CS without the US leads to extinction: the loss of the ability of the CS to produce a conditioned response (CR). However, when the excitatory CS is accompanied during extinction trials by an inhibitory CS — a predictor of the absence of the US — it may retain the ability to generate a

* Corresponding author. Fax: +61-2-9385-3641.
*E-mail address:* p.lovibond@unsw.edu.au (P.F. Lovibond).

0005-7967/00/$ - see front matter 
PII: S0005-7967(99)00121-7

CR when subsequently tested alone (e.g. Soltysik, Wolfe, Nicholas, Wilson & Garcia-Sanchez, 1983).

Protection from extinction has important implications for any clinical procedure involving extinction or unlearning. For example, exposure therapy is often assumed to involve a process analogous to extinction. Exposure is employed in the treatment of many clinical disorders, and is a primary feature of the treatment of anxiety disorders. Even if an anxiety reaction has been established via some other route than direct conditioning (e.g. Rachman, 1977), there is reason to believe that the same mechanisms that underlie extinction of a conditioned reaction may be operative (Lovibond, 1993; Reiss, 1991). Any inhibitory stimulus or 'safety signal' present during exposure could therefore undermine the effectiveness of the treatment. One widely recognised source of safety is the therapist: exposure to anxiety-provoking cues in the presence of the therapist may lead to rapid symptom reduction but a return of anxiety when the patient subsequently confronts the stimuli alone. Clinical researchers are discovering more subtle sources of safety — both in the external environment and in the behaviour of the patient — that appear to interfere with exposure treatment (e.g. Salkovskis, Clark & Gelder, 1996; Wells et al., 1995).

The phenomenon of protection from extinction has also been influential in shaping theories about the conditions for unlearning and new learning. For example, most contemporary theoretical models of associative learning anticipate it. The influential model of Rescorla and Wagner (1972) assumes that an inhibitory stimulus has a negative associative strength which cancels out the positive associative strength of the target excitatory stimulus, leading to little or no discrepancy between what is predicted by all cues (close to zero) and what actually happens on extinction trials (zero). Thus there is little change in associative strength of the stimuli, and in particular the target CS should retain a positive associative strength, revealed when it is tested alone. Such an outcome is not anticipated by traditional stimulus-response models of learning, which predict that presentation of the target CS without the CR should lead to successful extinction. Protection from extinction is, however, highly compatible with attributional and expectancy-based models of learning: if a subject attributes the absence of an expected outcome to the inhibitory stimulus, then there is no reason to revise the causal status of the target stimulus (see also Salkovskis et al., 1996; Wells et al., 1995).

Given the theoretical and practical significance of the phenomenon, it is surprising that there appear to be no demonstrations of protection from extinction in human subjects. Furthermore, the animal literature has not yet firmly established the conditions under which the effect will occur. For example, a wide variety of appetitive and aversive conditioning preparations and temporal arrangements have been employed, with no clear pattern of outcomes. The inhibitory stimulus has been presented prior to the target (Kamin & Szakmary, 1977), during the target (Baum & Jacobs, 1989; Kamin, 1969, pp. 283–285) and after the target (Johnston, Clayton & Seligman, 1972, cited in LoLordo & Rescorla, 1966; Seligman & Johnston, 1973; Soltysik et al., 1983). While most experiments have shown protection from extinction, several experiments have failed to find the effect, and at least one has shown the opposite effect (Hawk & Riccio, 1977). Almost all studies have compared a protection condition with conventional CS alone extinction. To our knowledge, only one study has examined whether it is necessary for the added stimulus to be inhibitory. Calton, Mitchell and Schachtman (1996) reported that a previously extinguished CS may act to protect a target CS from extinction, as compared to a

target alone condition and to target plus another excitatory stimulus, but not when compared to target plus a novel stimulus. However, not all learning theories assume that a previously extinguished CS has active inhibitory properties, so it is not clear whether the Calton et al. (1996) results would also be obtained with a conventional conditioned inhibitor.

Given the ambiguity of the available data, the present experiments were designed with two aims: (1) to demonstrate protection from extinction in human subjects and (2) to evaluate whether it is necessary for the added stimulus to be inhibitory. The preparation employed was autonomic conditioning with an aversive US (electric shock), in order to maximise the applicability of the findings to the practical domain of extinction of anxiety.

## 2. Experiment 1

The first experiment was designed to demonstrate protection from extinction in human autonomic conditioning. A within-subject design was employed to maximise statistical power (see Table 1). In the *acquisition* phase, subjects were exposed to a conditioned inhibition discrimination (A+/AE−). These trials were intended to establish E as a signal for safety, since its presence was associated with omission of the shock that otherwise occurred after A. The two stimuli to be extinguished, C and D, were paired with shock in this phase. During the *extinction* phase, both C and D were presented without shock, but the target stimulus C was accompanied by the safety signal E, whereas the comparison stimulus D was presented alone. Protection from extinction would be observed in the *test* phase if subjects responded more strongly to C than to D.

The procedure allowed trial-by-trial recording of an objective measure of autonomic arousal (skin conductance) as well as a self-report measure: online ratings of shock expectancy. The self-report expectancy measure was assumed to reflect the learning of causal knowledge, as posited by attributional/expectancy models of learning (e.g. Reiss, 1991; see also Lovibond, 1998). The autonomic conditioning literature shows a close correspondence between expectancy ratings and autonomic measures of learning, encouraging the view that they arise from a

Table 1
Design of experiment 1[a]

| Phase | | |
|---|---|---|
| Acquisition | Extinction | Test |
| A+ (2) | A+ (2) | A+ (1) |
| B− (2) | B− (2) | B− (1) |
| AE− (4) | | |
| C+ (2) | CE− (3) | C− (1) |
| D+ (2) | D− (3) | D− (1) |

[a] Letters A–E refer to conditioned stimuli; + and − refer to the presence and absence, respectively, of electric shock after the CS; numbers in parentheses give the number of trials of each type; within each phase, trial types were intermixed.

common learning process (see Davey, 1987; Dawson & Schell, 1985; Lovibond, 1992, 1993). This expectancy–autonomic correspondence, which is not always seen between self-report and autonomic measures in clinical studies, supports the view that moment-by-moment expectancy of an aversive outcome is the proximal cognitive basis for anxiety. Thus, it was anticipated that protection from extinction would be manifest on both the expectancy and autonomic measures in the present experiment.

### 2.1. Method

#### 2.1.1. Subjects

Subjects were 26 undergraduate students, 19 females and seven males, who volunteered for the experiment in partial fulfilment of a course requirement.

#### 2.1.2. Apparatus

The apparatus was the same as that described in Hinchy, Lovibond and Ter-Horst (1995). In brief, subjects were tested singly in a darkened room. One auditory and four visual stimuli served as CSs. The auditory stimulus was a 70-dB, 500-Hz sine wave tone played through a speaker located on the floor and to the right of the subject. The visual stimuli were coloured blocks of blue, red, white and green presented on a 30-cm colour computer monitor approximately 100 cm in front of the subject. The blocks were approximately 6 $cm^2$ and appeared in the centre of the screen. The monitor was also used to display all instructions.

Skin conductance was measured through electrodes attached to the second and third fingers of the subject's nonpreferred hand, and was recorded on a paper chart via a constant voltage pre-preamplifier and Grass preamplifier and amplifier. Shock electrodes were attached to the proximal and medial segments of the first (index) finger of the same hand. Subjects used their preferred hand to record their subjective moment-to-moment expectancy of shock on a continuous, 180°, rotary dial attached to the arm of seat. The semicircular dial was labelled *expectancy of shock in the next few moments*, with the left extreme labelled *certain no shock*, the centre position *uncertain* and the right extreme *certain shock*. An IBM-compatible computer was used to present all stimuli and to record the shock expectancy ratings (at 0.5-s intervals).

#### 2.1.3. Procedure

The general procedure followed that described in Hinchy et al. (1995). Subjects were fitted with electrodes, and were led through a work-up procedure to select a "definitely uncomfortable, but not painful" shock level. They were then taken into the experimental room and given spoken and written instructions about the purpose of the experiment and the use of the expectancy pointer. They were told that they should move the expectancy pointer as often as they wished, so that the pointer always indicated their current degree of expectancy of shock at the end of the CS. Subjects were specifically directed to work out which auditory and visual stimuli led to shock and which did not. They were informed that shocks would only occur during the last 0.5 s of the time a stimulus was presented.

The conditioning procedure followed that developed by Lovibond (1992). Throughout the experiment, CS durations varied randomly between 10 and 30 s, while shock durations were always 0.5 s. Intertrial intervals varied randomly between 20 and 50 s (mean = 30 s). Five CSs

were employed, designated A to E. For all subjects, red served as CS A, blue served as CS B and the tone served as CS E. For half of the subjects, white served as CS C and green as CS D, and for the other half these stimuli were reversed.

A within-subjects design was employed, with three phases (see Table 1). Within each phase, trial order was random except that there were no more than two consecutive presentations of the same trial type. In the *acquisition* phase, subjects were exposed to two trials on which A was followed by shock and two trials on which B was followed by no shock. Subjects also received four trials on which A was combined with the auditory cue E in a simultaneous compound, followed by no shock. These trials were designed to establish A as a predictor of shock and E as a predictor of shock omission (safety signal or conditioned inhibitor). Finally, subjects were exposed to two trials of C and D followed by shock. These stimuli served as the excitatory cues for extinction.

In the *extinction* phase, the target cues C and D were each presented three times, without shock. During these extinction trials, C was accompanied by the putative safety signal E. In the *test* phase, C and D were each presented alone, so that they could be directly compared with each other. No shock was presented after C and D. A continued to be paired with shock and B with no shock throughout the *extinction* and *test* phases in order to provide continuity over phases, maintain exposure to the shock and provide reference points for comparison with the test stimuli. Throughout the experiment, the number of trials administered for each combination of stimuli was based on previous research in the laboratory, and the need to minimise the total number of trials so that subjects would still show electrodermal responsivity at the time of the *test* phase.

Upon completion of the experiment, subjects were given a questionnaire designed to assess knowledge of the relationships between the CSs and shock (Dawson & Reardon, 1973). However, since two of the stimuli had changed their associative status during the experiment, it was considered more appropriate to determine whether subjects had learned the stimulus relations on the basis of the expectancy ratings they had given during the *acquisition* phase, before any cues were extinguished. Attention was restricted to the critical A+/AE− discrimination. Accordingly, subjects were classified as aware of the stimulus relations if their mean expectancy rating on the second and third A+ trials was at least 20 points higher than their mean expectancy rating on the third and fourth AE− trials. The value of 20 points was chosen as a relatively liberal criterion for awareness, so as to exclude the minimum number of subjects from the analysis (see Lovibond, 1992).

### *2.1.4. Scoring and analysis*

The electrodermal measure was change in tonic skin conductance level (SCL), calculated for each trial as the difference between the mean SCL during the CS (excluding the first 10 s) and the mean SCL for the 10-s baseline period immediately prior to CS onset (Lovibond, 1992). SCLs were hand scored from paper chart recordings. In order to eliminate the large individual differences between subjects in the magnitude of their SCL change scores, these data were mean-corrected. This procedure, described in detail in Lovibond (1992), expresses the SCL change score on each trial as a proportion of the mean SCL change score for that subject over all trials. Thus a score of 1 indicates a typical response for that subject, 0 indicates no response and a score above 1 indicates a higher than typical response. The advantage of mean

correction over range correction (Lykken, Rose, Luther & Maley, 1966) is that the mean is a more stable value to use for calibration than the difference between the largest and smallest responses.

Expectancy ratings were transformed to a 0–100 scale where 0 is certain no shock, 50 uncertain and 100 certain shock. A single expectancy score was calculated for each CS presentation based on the average level across the presentation excluding the first 10 s. For 10-s CS presentations, both the SCL measure and the expectancy measure were based on a single value taken at the end of the 10 s (Lovibond, 1992). Data for each phase were analysed by a set of planned, orthogonal contrasts using a multivariate, repeated measures model (O'Brien & Kaiser, 1985). Bonferroni correction was used to control for type 1 errors across the two dependent measures ($\alpha = 0.05/2$).

### 2.2. Results

Four subjects were classified as unaware of the A+/AE− contingencies based on their *acquisition* phase expectancy ratings. Because previous research has consistently shown that unaware subjects fail to discriminate between cues on either the expectancy or skin conductance measures (e.g. Dawson & Schell, 1985; Lovibond, 1992), statistical analyses were restricted to the remaining 22 subjects.

#### 2.2.1. Expectancy ratings

The top panel of Fig. 1 shows that across the *acquisition* phase, subjects learned to discriminate between trials on which A was presented alone and trials on which it was combined with E. Expectancy ratings to the final A+ trial in the acquisition phase were reliably higher than to the average of the final two AE− trials ($F(1, 21) = 101.9$, $p < 0.05/2$). Subjects also learned to expect shock after the two target cues C and D. On the second trial, expectancy ratings to C and D were reliably higher than to the safe cue B ($F(1, 19) = 37.5$, $p < 0.05/2$). There was no difference in shock expectancies to C and D on the second trial ($F < 1$).

In the *extinction* phase, combining C with E greatly reduced expectancy ratings on the first trial by comparison with D ($F(1, 21) = 55.5$, $p < 0.05/2$). Averaged over CE and D, expectancies dropped from trial 1 to trial 3 ($F(1, 21) = 23.0$, $p < 0.05/2$), confirming that extinction had occurred. By the third trial, the difference between CE and D was no longer significant ($F(1, 21) = 2.0$, $p < 0.05/2$). Expectancy ratings to the nonshocked cue B increased somewhat in comparison to the *acquisition* phase, but B was still clearly discriminated from the shocked cue A ($F(1, 21) = 57.0$, $p < 0.05/2$).

In the *test* phase, there was a clear difference between A and B ($F(1, 21) = 96.6$, $p < 0.05/2$). There was also a clear protection from extinction effect for stimulus C. Expectancy ratings to D were consistent with the course of extinction in the previous phase, but expectancies to C increased almost to the level of the consistently shocked cue A when E was omitted. The difference between C and D was highly significant ($F(1, 21) = 35.6$, $p < 0.05/2$).

#### 2.2.2. Skin conductance

The bottom panel of Fig. 1 shows that the pattern of electrodermal responding was broadly similar in most respects to the shock expectancy ratings, but with greater variability. By the

end of the *acquisition* phase, subjects responded more strongly to A than to the compound AE, but this difference was only marginally significant ($F(1, 21) = 4.5$, $0.05 > p > 0.05/2$). Subjects responded more strongly to the two shocked target cues C and D than to the safe cue B on the second trial ($F(1, 21) = 11.3$, $p < 0.05/2$). There was no difference between C and D ($F < 1$).

In the *extinction* phase, in contrast to the pattern observed on the expectancy measure, there was no apparent effect of adding E to C on the electrodermal measure. There was no difference between CE and D on either the first or the third extinction trial ($F$s < 1). Responding to both CE and D declined from the first to the third trial, consistent with extinction, but this trend did not reach significance ($F(1, 21) = 2.1$, $p > 0.05$). Throughout the *extinction* phase, the

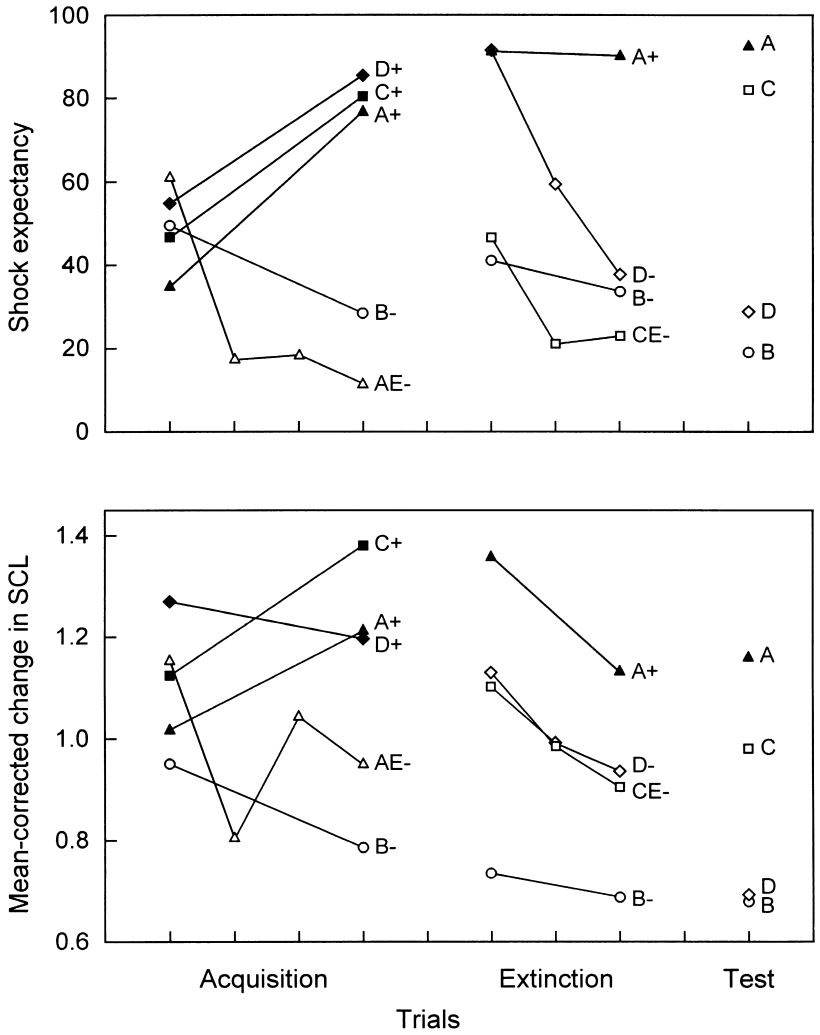


Fig. 1. Mean shock expectancy ratings (top panel) and change in skin conductance level (bottom panel) as a function of trials in experiment 1.

discrimination between the consistently shocked cue A and the consistently non-shocked cue B was maintained ($F(1, 21) = 25.8$, $p < 0.05/2$).

In the *test* phase, electrodermal responding to all cues was somewhat lower than in previous phases, consistent with the overall habituation effect typically observed over trials on this measure. The discrimination between A and B was reduced but was still significant ($F(1, 21) = 6.8$, $p < 0.05/2$). Importantly, the protection from extinction pattern observed on the expectancy measure was also evident on the electrodermal measure. Responding to D continued to fall, while responding to C increased when stimulus E was omitted. The difference between C and D was marginally significant ($F(1, 21) = 5.7$, $0.05 > p > 0.05/2$).

### *2.3. Discussion*

The primary outcome of experiment 1 was that subjects showed higher shock expectancy and higher skin conductance during stimulus C, which had been combined with the inhibitory stimulus E during extinction, than during D, which had been presented alone. The experiment thus demonstrated protection from extinction on both measures. This result indicates that an important empirical finding in basic animal research is also demonstrated by human subjects, and provides further support for the applicability of associative conditioning models to human learning (Shanks, 1994). In particular, the results suggest that protection from extinction may occur in clinical settings such as exposure treatment for anxiety disorders.

The general concordance between the autonomic and self-report measures supports cognitive models of associative learning and anxiety that emphasise the role of expectancy processes in the generation of anxiety. However, one interesting dissociation between the measures was observed during the *extinction* phase. Whereas the addition of the putative inhibitory stimulus E greatly reduced shock expectancy to C by comparison with D, no such effect was apparent on the skin conductance measure. One possible account of the skin conductance data relies on the observation that this measure is sensitive to novelty and attentional factors as well as to anxious arousal (e.g. Öhman, 1983). Thus, the novelty of the CE combination may have elevated skin conductance and cancelled out any inhibitory effect of E. It is also important to note that there has been no successful demonstration of transfer of conditioned inhibition in human conditioning. That is, a conditioned inhibitor does not suppress either shock expectancy or skin conductance to a separately trained excitor to a greater extent than does a novel comparison stimulus (Grings, Carey & Schell, 1974; McNally & Reiss, 1982, 1984; Wilkinson, Lovibond, Siddle & Bond, 1989). Therefore at least part of the suppression of expectancy ratings to C by the addition of E may also have been induced by an associatively neutral stimulus. This possibility raises the question: to what extent does the protection from extinction observed in this experiment depend on the inhibitory training given to the added cue? Experiment 2 addressed this question.

## 3. Experiment 2

Experiment 1 demonstrated protection from extinction of a target shock cue by an added inhibitory cue, but it did not establish that the inhibitory training given to the added cue was

necessary. Furthermore, the return of responding to C could have been due to generalisation decrement from the *extinction* phase to the *test* phase. That is, extinction of C may have actually occurred on the CE trials, but this extinction failed to generalise to C tested alone. Experiment 2 pitted the generalisation decrement account against the Rescorla–Wagner account, by combining the control cue D with an excitatory cue — that is, another stimulus previously paired with shock (A). In other respects the design was the same as experiment 1 (see Table 2).

According to the generalisation decrement account, there should be an equally large loss of extinction when D is tested alone as when C is tested alone. According to the Rescorla and Wagner (1972) account, however, the addition of the excitatory cue A should have the opposite effect on D as the addition of the inhibitory cue E has on C. This model predicts that the summation of two excitors should lead to 'super extinction' of both stimuli, due to the extremely high discrepancy between what is predicted by available cues (effectively two USs, or a double strength US) and what occurs — namely nothing. Some evidence for this proposition was provided by Wagner, Saavedra and Lehmann (cited in Wagner & Rescorla, 1972), who found greater extinction of a target excitatory cue in the presence of a strong compared to a weak excitor. Thus, if protection from extinction depends on the associative strength of the added stimulus, it was expected that a similar or stronger difference between C and D would be observed in the present experiment than in experiment 1. The experimental procedure followed that of experiment 1, except that the conditioned stimuli were colour pictures (see Lovibond, Siddle & Bond, 1993). Compounds were formed by presenting two pictures simultaneously in the top and bottom halves of one colour slide.

### 3.1. Method

#### 3.1.1. Subjects

Subjects were 20 undergraduate students, 12 females and eight males, who volunteered for the experiment in partial fulfilment of a course requirement.

#### 3.1.2. Apparatus

The apparatus was the same as that used in experiment 1, with the exception that the stimuli

Table 2
Design of experiment 2[a]

| Phase | | |
|---|---|---|
| Acquisition | Extinction | Test |
| A+ (2) | A+ (2) | A+ (1) |
| B− (2) | B− (2) | B− (1) |
| AE− (4) | | |
| C+ (2) | CE− (3) | C− (1) |
| D+ (2) | DA− (3) | D− (1) |

[a] Notation as for Table 1.

used as CSs were colour pictures depicting a flower, a mushroom, a mountain, a tree and a lake. The pictures were projected by a Kodak Carousel projector onto a screen approximately 2 m in front of the subject. The image size was approximately 97 × 66 cm. Pictorial stimuli were chosen to facilitate comparison with other research in the laboratory that employed pictures of fear-relevant stimuli (snakes, spiders; e.g. Lovibond et al., 1993). Our pilot work suggested that subjects could learn to use a visual cue as easily as an auditory cue to modulate responding to a target visual cue predicting shock.

*3.1.3. Procedure*

The general procedure was the same as for experiment 1. Stimuli A and B were the flower or mushroom, counterbalanced over subjects. Similarly, stimuli C and D were the tree or mountain, counterbalanced over subjects. Stimulus E was the lake for all subjects. Single stimuli consisted of one colour picture in either the top or bottom half of the screen, randomised over trials. Compound stimuli consisted of two colour pictures, one in the top half and one in the bottom half of the screen, with position randomised over trials (see Lovibond et al., 1993). The design was identical to that of experiment 1, except that the control stimulus D was accompanied during extinction by stimulus A, which had been paired with shock during acquisition (see Table 2).

## *3.2. Results*

Two subjects were classified as unaware of the A+/AE− contingencies based on their *acquisition* phase expectancy ratings. Results are presented for the remaining 18 subjects.

*3.2.1. Expectancy ratings*

The top panel of Fig. 2 shows that the pattern of expectancy ratings during the acquisition and *extinction* phases was very similar to that obtained in experiment 1. By the end of acquisition, subjects gave higher shock expectancy ratings to A than to AE ($F(1, 17) = 143.4$, $p < 0.05/2$) and to C and D combined than to B ($F(1, 17) = 33.5$, $p < 0.05/2$). There was no difference in expectancy ratings to C and D ($F < 1$).

During extinction, subjects maintained their discrimination between the consistently shocked stimulus A and the consistently non-shocked stimulus B ($F(1, 17) = 107.8$, $p < 0.05/2$). Expectancies to the DA compound were approximately the same as to both D and A alone during acquisition, and dropped rapidly across the three non-shocked extinction trials. By contrast, addition of the putative inhibitory stimulus E again reduced expectancies to the target stimulus C, but expectancy ratings did not decline substantially further over trials. This pattern led to a significant difference between DA and CE on the first extinction trial ($F(1, 17) = 20.9$, $p < 0.05/2$), and a significant difference in the opposite direction by the third extinction trial ($F(1, 17) = 8.1$, $p < 0.05/2$). There was a significant reduction in expectancy from the first to the third trial, averaged over DA and CE ($F(1, 17) = 39.3$, $p < 0.05/2$), which was clearly due primarily to the DA trials.

In the *test* phase, the discrimination between A and B remained relatively strong ($F(1, 17) = 23.6$, $p < 0.05/2$). Shock expectancy ratings for stimulus C when presented alone were again very close to those of the shocked stimulus A. However, in this experiment, expectancy

ratings for the control stimulus D were also very high, leading to a non-significant difference between C and D ($F(1, 17) = 2.3$, $p > 0.05/2$).

*3.2.2. Skin conductance*

The bottom panel of Fig. 2 shows that the pattern of electrodermal responding was again broadly similar to that found for the shock expectancy ratings. In this experiment there was slightly better discrimination between the shocked and non-shocked cues in the *acquisition* phase. Thus, by the end of acquisition, subjects showed greater electrodermal responses to A than to AE ($F(1, 17) = 7.7$, $p < 0.05/2$), and to the average of C and D than to B ($F(1, 17) = 10.5$, $p < 0.05/2$). There was no significant difference between C and D ($F < 1$). In the

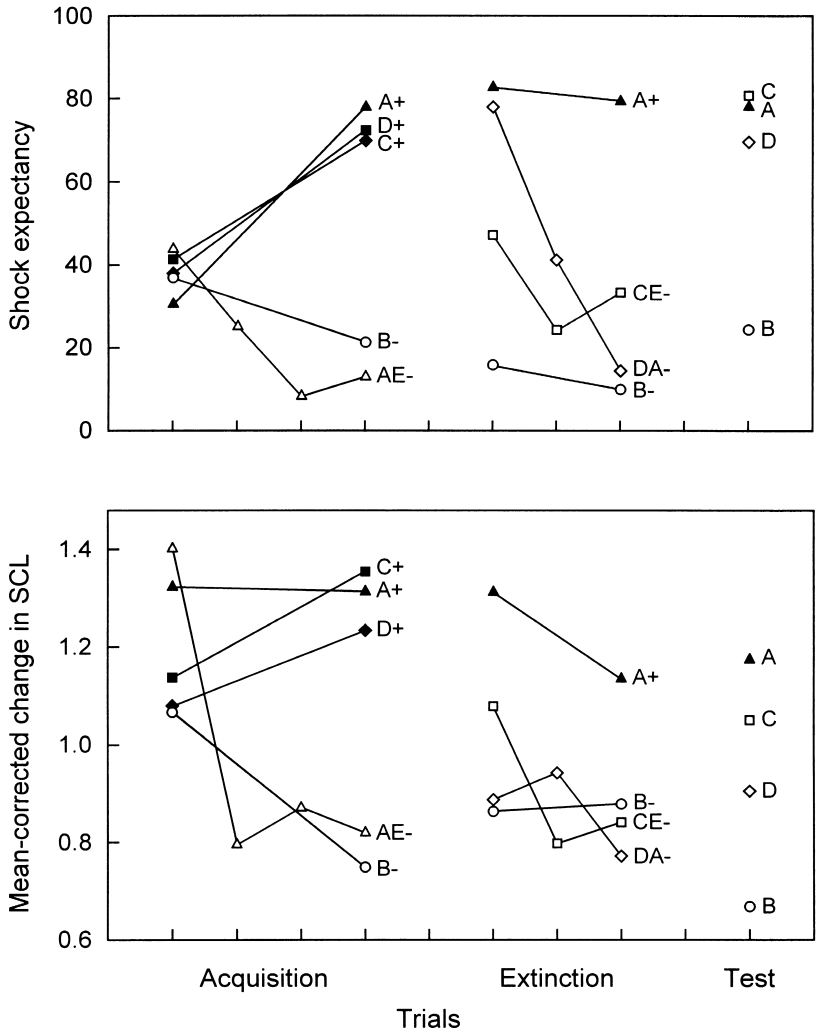


Fig. 2. Mean shock expectancy ratings (top panel) and change in skin conductance level (bottom panel) as a function of trials in experiment 2.

*extinction* phase, electrodermal responses to the two compounds were both low and similar to the safe cue B. As in experiment 1, but again unlike the expectancy measure, subjects failed to discriminate between the compound involving a putative safety signal (CE) and the compound without such a stimulus (DA). In this experiment, DA actually contained two previously shocked stimuli, but elicited smaller electrodermal responses than either of these stimuli had alone during acquisition. This pattern was confirmed by the statistical analysis. There was no difference between CE and DA on the first trial ($F(1, 17) = 2.3$, $p > 0.05$) or on the third trial ($F < 1$). Responses on the third trial were somewhat lower than on the first trial, averaged over CE and DA, but this difference did not reach significance ($F(1, 17) = 2.6$, $p > 0.05$).

The *test* phase electrodermal data again followed the expectancy data. Responses to both C and D were higher when tested alone than they had been in compound during extinction. The difference between C and D was suggestive of a greater protection from extinction effect for C, but unlike experiment 1, this difference did not approach significance ($F(1, 17) = 1.8$, $p > 0.05$). Although mean electrodermal responding to A remained substantially higher than to B in the *test* phase, this difference did not quite reach significance ($F(1, 17) = 3.3$, $p > 0.05$).

### 3.3. Discussion

In this experiment, subjects again demonstrated protection from extinction of the target cue C that had been combined with an inhibitory cue during extinction. However, they showed the same effect for the comparison cue D that had been paired with an excitatory cue during extinction. That is, subjects showed a return to expecting shock after both C and D, despite the very different associative histories of the cues with which they had been paired during extinction. The results do not appear to have been affected by the use of colour pictures as CSs in this experiment: a highly similar pattern was observed on both the expectancy and skin conductance measures for all common trials across the two experiments (that is, all trials except those involving D during the *extinction* and *test* phases). In particular, a similar protection from extinction effect was observed to stimulus C across the two experiments. Experiment 2 allowed a direct comparison between a cue extinguished with an inhibitor and a cue extinguished with an excitor, and the conclusion seems clear-cut: the added cue does not need to be inhibitory to induce protection from extinction. This result favours generalisation decrement or similar accounts, rather than accounts that depend on the safety properties of the added cue.

As in experiment 1, a similar pattern was observed in the *test* phase on both the expectancy and skin conductance measures. This concordance suggests that whatever process or processes underlie protection from extinction feed into both of these response systems. However, there was again a dissociation between the measures in the *extinction* phase. As in experiment 1, subjects gave lower shock expectancy ratings to the target cue E, which was paired with an inhibitor, than to the comparison cue D, which in this experiment was paired with an excitor. However a comparable effect was not observed on the skin conductance measure. Unlike experiment 1, the skin conductance pattern cannot be attributed to the novelty of the CE compound, since the DA compound was equally novel. Empirically, the data are consistent with previous research showing that a conditioned inhibitor fails to summate with a new excitor. However these data go further in that they also demonstrate a failure of an excitor to

summate with another excitor to produce a larger response than to either stimulus alone. Paradoxically, subjects gave a smaller skin conductance response on the first presentation of the DA compound than to either of these stimuli at the end of acquisition. Subjects also failed to show additivity on the expectancy measure, but their expectancy ratings to DA were not lower than to D or A separately. This measure may also have been constrained by a ceiling effect. The implications of these results will be considered further below.

## 4. General discussion

Both experiments demonstrated a clear protection from extinction effect, on both the expectancy and skin conductance measures, when a stimulus predictive of shock was accompanied by another stimulus during extinction trials. The target stimulus continued to elicit high shock expectancy and high tonic skin conductance when presented alone after extinction. Thus it appears that extinction of human fear conditioning is relatively easily disrupted by a stimulus change during the non-shocked trials.

It has generally been assumed that only a conditioned inhibitory stimulus, or safety signal, will interfere with extinction (e.g. Soltysik et al., 1983; Wagner & Rescorla, 1972). However the results of experiment 2 suggest that this may not be the case, and that even a conditioned excitatory stimulus may disrupt extinction. Two immediate questions arise: what processes underlie disruption by the excitatory stimulus and could these processes also account for disruption by the inhibitory stimulus? Several potential explanatory processes are suggested by the animal conditioning literature.

First, as described earlier, the failure of extinction to transfer from the compound trials to the target cues tested alone could be thought of in terms of generalisation decrement. There are now sophisticated models which describe generalisation between compounds and elements (e.g. Pearce, 1987). There is also a well-established literature showing that after extinction of fear conditioning, a change from the context of extinction can lead to a return of fear to the target stimulus (Bouton & Ricker, 1994; Gunther, Denniston & Miller, 1998). In this literature, it has been shown that the extinction context does not need to be inhibitory for such 'context-specificity' of extinction to occur (Bouton & King, 1983). Bouton (1993) has suggested that extinction does not eliminate the original excitatory association, but rather establishes an additional inhibitory association that suppresses performance. He also suggests that such additional learning is more easily disrupted by contextual changes (including the passage of time) than is the original learning. Procedurally, the added cue in a protection from extinction design (whether inhibitory, excitatory or neutral) constitutes a unique context in which extinction occurs. This context is absent when the target stimulus is subsequently tested alone. Thus, context-specificity of extinction learning is a plausible account of the disruption of extinction observed in the present experiments, both by the inhibitory stimulus and by the excitatory stimulus. In clinical terms, context-specificity may constrain the transfer of new learning whenever there is a stimulus change (whether it involves interoceptive, exteroceptive or temporal cues) between the therapeutic situation and the real-life situation the therapy is intended to transfer to.

A second, related process that may be involved in disruption of extinction is occasion-

setting. Negative occasion-setting refers to the ability of a stimulus to suppress responding to a target excitatory stimulus. It is established in the same way as conditioned inhibition: when the target stimulus is presented alone it is followed by, say, shock, but when it is accompanied by the negative occasion setter, it is not followed by shock. Negative occasion-setting is typically observed when the occasion setter precedes the target stimulus in time (Holland, 1986). Unlike conditioned inhibition, occasion setting appears to be relatively specific to the target stimulus and hence does not transfer well to other excitors. Although in the present experiments the added cues did not precede the excitatory target cues in time, there is reason to suppose that occasion-setting may have been learned. Specifically, the explicitly trained safety cue E failed to transfer its inhibitory properties to the target excitor C. Interestingly, the ability of a stimulus to act as an occasion-setter does not seem to depend on its own direct association with the US. For example, a stimulus may have an excitatory association with the US but still act to negatively modulate the association between another excitatory stimulus and the US (e.g. Rescorla, 1985). Therefore, in experiment 2, both the excitatory cue A and the inhibitory cue E may have acquired the ability to modulate the association between their target excitors and shock. When presented without these modulatory cues present, the target cues again activated an expectancy of shock and hence anxiety.

A final process which may have contributed to the present results is external inhibition. Pavlov (1927) used this term to describe the ability of another stimulus, usually a novel stimulus, to disrupt the normal performance elicited by either an excitatory or inhibitory target stimulus. It is applicable to the *extinction* phase of the present experiments, where either of the added cues, particularly the excitatory cue A, could be seen as an external inhibitor. Thus the failure of the AD compound in experiment 2 to demonstrate summation on the skin conductance measure — to provoke a reaction as large or larger than to either cue alone — could be seen as due to mutual external inhibition. External inhibition is usually thought to operate on performance rather than learning, and this analysis is consistent with the finding that a similar effect was not observed on the cognitive expectancy measure. However it is still possible that the added cues A and E also disrupted learning of extinction, and that this effect contributed to the observed failure of extinction to the target stimuli observed in the *test* phase. Clinically, external inhibition is most likely to occur when there is a strong or novel stimulus present during exposure therapy — for example, when the patient's attentional resources are occupied by a distracting cue or even by excessive anxiety.

In conclusion, there are several potential processes that may contribute to disruption of extinction. The primary candidate processes that may account for disruption by an inhibitory stimulus include cancelling the discrepancy between what is expected and what occurs on extinction trials, context-specificity of extinction, and negative occasion-setting. Disruption by an excitatory stimulus might be due to external inhibition, context-specificity of extinction and negative occasion-setting. Clearly further research is necessary to determine the relative contribution of these processes. We are presently investigating both occasion-setting and context-specificity in human fear conditioning. Preliminary results suggest that human subjects learn occasion-setting relationships easily, and demonstrate context-specificity of both excitatory and inhibitory learning. It is also important to note that clear evidence for transfer of conditioned inhibition to a novel excitor is not yet available (e.g. Wilkinson et al., 1989).

More effective inhibitory training procedures (e.g. Williams, 1996) may lead to a successful demonstration of the superiority of an inhibitory stimulus in protection from extinction.

The primary practical implication of the present research is that the extinction of fear is relatively easily disrupted in human subjects, and that such disruption may be achieved not only by safety signals but even by danger signals. It must be acknowledged that the results were obtained in a laboratory setting with an objectively mild aversive event. However, the correspondence between many laboratory learning effects and clinical phenomena (e.g. Mineka, 1985) supports the potential clinical applications of the laboratory model. On this basis, the present findings reinforce the importance of considering contextual stimuli when using extinction-like procedures in a clinical application such as exposure therapy for anxiety (Bouton, 1988; Gunther et al., 1998). External, internal and temporal stimuli present during exposure may all potentially interfere with the process of extinction or with the transfer of extinction to another context. The general concordance between the self-report expectancy measure and the skin conductance measure in the present research, together with other evidence concerning verbal processes in conditioning, suggests that the desired extinction outcome may be maximised by manipulating both the physical stimuli the subject is exposed to and the interpretation of these experiences encouraged by verbal or cognitive therapy (Lovibond, 1993; Zinbarg, 1990).

## Acknowledgements

This research was supported by grant A79530195 from the Australian Research Council. We would like to thank Mark Bouton for his helpful comments on the manuscript.

## References

Baum, M., & Jacobs, W. J. (1989). Loud noise potentiates conditioned fear in extinction using a CER (lick suppression) paradigm in rats. *Bulletin of the Psychonomic Society*, *27*, 449–451.

Bouton, M. E. (1988). Context and ambiguity in the extinction of emotional learning: implications for exposure therapy. *Behaviour Research and Therapy*, *26*, 137–149.

Bouton, M. E. (1993). Context, time and memory retrieval in the interference paradigms of Pavlovian learning. *Psychological Bulletin*, *114*, 80–99.

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., & Ricker, S. T. (1994). Renewal of extinguished responding in a second context. *Animal Learning & Behavior*, *22*, 317–324.

Calton, J. L., Mitchell, K. G., & Schachtman, T. R. (1996). Conditioned inhibition produced by extinction of a conditioned stimulus. *Learning and Motivation*, *27*, 335–361.

Davey, G. C. L. (1987). *Cognitive processes and Pavlovian conditioning in humans*. Chichester: Wiley.

Dawson, M. E., & Reardon, D. P. (1973). Construct validity of recall and recognition postconditioning measures of awareness. *Journal of Experimental Psychology*, *98*, 308–315.

Dawson, M. E., & Schell, A. M. (1985). Information processing and human autonomic classical conditioning. In P. K. Ackles, J. R. Jennings, & M. Coles, *Advances in psychophysiology*, *Vol. I*. Greenwich, CT: JAI Press.

Grings, W. W., Carey, C. A., & Schell, A. M. (1974). Comparison of two methods for producing response inhibition in electrodermal conditioning. *Journal of Experimental Psychology*, *103*, 658–662.

Gunther, L. M., Denniston, J. C., & Miller, R. R. (1998). Conducting exposure treatment in multiple contexts can prevent relapse. *Behaviour Research and Therapy*, *36*, 75–91.

Hawk, G., & Riccio, D. C. (1977). The effect of a conditioned fear inhibitor (CS-) during response prevention upon extinction of an avoidance response. *Behaviour Research and Therapy*, *15*, 97–101.

Hinchy, J., Lovibond, P. F., & Ter-Horst, K. M. (1995). Blocking in human electrodermal conditioning. *Quarterly Journal of Experimental Psychology*, *48*, 2–12.

Holland, P. C. (1986). Temporal determinants of occasion setting in feature-positive discriminations. *Animal Learning & Behavior*, *17*, 269–279.

Kamin, L. J., & Szakmary, G. A. (1977). Configural-like effects observed in the course of second-order conditioning in rats. *Learning and Motivation*, *8*, 126–135.

Kamin, L. J. (1969). Predictability, surprise, attention and conditioning. In B. A. Campbell, & R. M. Church, *Punishment and aversive behavior* (pp. 279–296). NY: Appleton-Century-Crofts.

LoLordo, V. M., & Rescorla, R. A. (1966). Protection of the fear-eliciting capacity of a stimulus from extinction. *Acta Biologica Experimentalis*, *26*, 251–258.

Lovibond, P. F., Siddle, D. A. T., & Bond, N. W. (1993). Resistance to extinction of fear-relevant stimuli: preparedness or selective sensitization? *Journal of Experimental Psychology: General*, *122*, 449–461.

Lovibond, P. F. (1992). Tonic and phasic electrodermal measures of human aversive conditioning with long duration stimuli. *Psychophysiology*, *29*, 621–632.

Lovibond, P. F. (1993). Conditioning and cognitive-behaviour therapy. *Behaviour Change*, *10*, 119–130.

Lovibond, P. F. (1998). The role of propositional knowledge in human autonomic conditioning: retrospective revaluation induced by experience and by instruction. Submitted for publication.

Lykken, D. T., Rose, R., Luther, B., & Maley, M. (1966). Correcting psychophysiological measures for individual differences in range. *Psychological Bulletin*, *66*, 481–484.

McNally, R. J., & Reiss, S. (1982). The preparedness theory of phobias and human safety signal conditioning. *Behaviour Research and Therapy*, *20*, 153–159.

McNally, R. J., & Reiss, S. (1984). The preparedness theory of phobias: the effects of initial fear level on safety-signal conditioning to fear-relevant stimuli. *Psychophysiology*, *21*, 647–652.

Mineka, S. (1985). Animal models of anxiety-based disorders: their usefulness and limitations. In A. H. Tuma, & J. D. Maser, *Anxiety and the anxiety disorders* (pp. 199–244). Hillsdale, NJ: Lawrence Erlbaum Associates.

O'Brien, R. G., & Kaiser, M. K. (1985). MANOVA method for analyzing repeated measures designs: an extensive primer. *Psychological Bulletin*, *97*, 316–333.

Öhman, A. (1983). The orienting response during Pavlovian conditioning: Contingency detection, stimulus significance and expectancies for events to follow. In D. A. T. Siddle, *Orienting and habituation: perspectives in human research* (pp. 315–369). Chichester: Wiley.

Pavlov, I. P. (1927). *Conditioned reflexes*. New York: Oxford University Press.

Pearce, J. M. (1987). A model for stimulus generalization. *Psychological Review*, *94*, 61–73.

Rachman, S. (1977). The conditioning theory of fear-acquisition: a critical examination. *Behaviour Research and Therapy*, *15*, 375–387.

Reiss, S. (1991). Expectancy model of fear, anxiety and panic. *Clinical Psychology Review*, *11*, 141–153.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black, & W. F. Prokasy, *Classical conditioning. II. Current research and theory* (pp. 64–99). New York: Appleton-Century-Crofts.

Rescorla, R. A. (1985). Conditioned inhibition and facilitation. In R. R. Miller, & N. E. Spear, *Information processing in animals: conditioned inhibition* (pp. 199–326). Hillsdale, NJ: Erlbaum.

Salkovskis, P., Clark, D. M., & Gelder, M. (1996). Cognition-behaviour links in the persistence of panic. *Behaviour Research and Therapy*, *34*, 453–458.

Seligman, M., & Johnston, J. (1973). A cognitive theory of avoidance learning. In F. J. McGuigan, & B. Lumsden, *Contemporary approaches to conditioning and learning*. Washington: Winston.

Shanks, D. (1994). Human associative learning. In N. J. Mackintosh, *Animal learning and cognition*. San Diego: Academic Press.

Soltysik, S. S., Wolfe, G. E., Nicholas, T., Wilson, W. J., & Garcia-Sanchez, J. L. (1983). Blocking of inhibitory

conditioning within a serial conditioned stimulus-conditioned inhibitor compound: maintenance of acquired behavior without an unconditioned stimulus. *Learning and Motivation*, *14*, 1–29.

Wagner, A. R., & Rescorla, A. R. (1972). Inhibition in Pavlovian conditioning: application of a theory. In R. A. Boakes, & M. S. Halliday, *Inhibition and learning* (pp. 301–336). London: Academic Press.

Wells, A., Clark, D. M., Salkovskis, P., Ludgate, J., Hackmann, A., & Gelder, M. (1995). Social phobia: the role of in-situation safety behaviours in maintaining anxiety and negative beliefs. *Behavior Therapy*, *26*, 153–161.

Wilkinson, G. M., Lovibond, P. F., Siddle, D. A. T., & Bond, N. (1989). Effects of fear-relevance on electrodermal safety signal learning. *Biological Psychology*, *28*, 89–104.

Williams, D. A. (1996). A comparative analysis of negative contingency learning in humans and nonhumans. In D. Shanks, K. J. Holyoak, & D. L. Medin, *The psychology of learning and motivation: causal learning*, *Vol. 34* (pp. 47–88). San Diego, CA: Academic Press.

Zinbarg, R. E. (1990). Animal research and behavior therapy. I. Behavior therapy is not what you think it is. *The Behavior Therapist*, *13*, 171–175.