Research

Ambiguity produces attention shifts in category learning

Miguel A. Vadillo,[1,2] Cristina Orgaz,[3] David Luque,[4,5] and James Byron Nelson[6]

[1]Department of Primary Care and Public Health Sciences, King's College London SE1 1UL, United Kingdom; [2]Department of Experimental Psychology, University College London WC1H 0AH, United Kingdom; [3]Departamento de Psicología Básica, Universidad Nacional de Educación a Distancia, Madrid 28040, Spain; [4]School of Psychology, University of New South Wales, Sydney NSW 2052, Australia; [5]Departamento de Psicología Básica, Instituto de Investigación Biomédica de Málaga (IBIMA), Universidad de Málaga, Málaga 29071, Spain; [6]Facultad de Psicología, Universidad del País Vasco, San Sebastián 20018, Spain

It has been suggested that people and nonhuman animals protect their knowledge from interference by shifting attention toward the context when presented with information that contradicts their previous beliefs. Despite that suggestion, no studies have directly measured changes in attention while participants are exposed to an interference treatment. In the present experiments, we adapted a dot-probe task to track participants' attention to cues and contexts while they were completing a simple category learning task. The results support the hypothesis that interference produces a change in the allocation of attention to cues and contexts.

[Supplemental material is available for this article.]

One of the interesting research topics in current cognitive psychology is the study of how people and animals adapt to changes in their environment without forgetting memories of related events that might conflict with their current knowledge (e.g., Anderson 2003; Speekenbrink and Shanks 2010; McClelland 2013; Vadillo et al. 2013; Smith and Bulkin 2014). Among other mechanisms, an intriguing possibility is that changes in attention to cues and contexts can facilitate new learning and protect previous knowledge from interference. A popular effect known as highlighting provides an excellent example to understand how this process works (see Kruschke 2009; Sewell and Lewandowsky 2012). In a typical highlighting experiment, participants are first exposed to a series of exemplars with two different features, A and B, that have to be classified as members of category 1. During the second stage of the experiments, participants continue seeing AB-1 exemplars now intermixed with a second set of exemplars with features A and C that must be classified as members of category 2. The result of these experiments is that cue C becomes strongly predictive of category 2. For instance, if participants are shown an exemplar with cues B and C, they are much more likely to classify it as a member of category 2 than as a member of category 1.

A popular explanation for this highlighting result is that during the first stage, both A and B become very strongly associated with category 1. Then, during the second stage, the first time the participant sees an AC exemplar, the presence of A induces him or her to believe that the correct category is 1. This prediction results in an error. To try to minimize this error, attention shifts away from cue A to cue C. As a result, the association between C and category 2 develops easily and previous knowledge about A-1 is protected from interference. The reduction in attention to A when C is present prevents further learning about A (Kruschke 2009; Wills et al. 2014).

This process is relevant to extinction, one of associative learning's oldest phenomena (for review, see Dunsmoor et al. 2015). Since the seminal work conducted by Ivan Pavlov, it is well-known that extinction does not erase previous conditioning. If animals experience a number of pairings of an initially neutral conditioned stimulus (CS) with a biologically significant unconditioned stimulus (US) they will eventually respond to the CS. This response then declines in extinction when the animals are exposed to presentations of the CS without the US. However, the absence of responses to the CS does not indicate that the CS–US association has been deleted. For instance, strong responding can reappear if the CS is presented in a new context. This "renewal" of conditioned responding has been extensively replicated over species and experimental paradigms (e.g., Bouton and Bolles 1979; Bouton and King 1983) and has been the object of substantial research in human learning (e.g., Rosas et al. 2001; Vadillo et al. 2004; Nelson et al. 2011b; Cobos et al. 2013).

The standard account of renewal and related effects assumes that once an association is learned, any subsequent learning of a new association involving the same CS becomes context-dependent (Bouton 1993, 1997), particularly when the associations conflict with each other (Nelson and Callejas-Aguilera 2007). In the case of extinction, once the organism has learned the connection between the CS and the US, the subsequent presentations of the CS without the US do not delete the original memory of the CS–US association. Instead, those presentations are assumed to result in the creation of an inhibitory CS–US association (e.g., Wagner 1981) that counteracts previous learning. Most important, this second association is supposed to be context-dependent, which means that both associations will counteract each other only if the CS is presented in the context where extinction took place. In any other context, the inhibitory CS–US association will not be functional and, therefore, the

Corresponding author: miguel.vadillo@kcl.ac.uk
Article is online at http://www.learnmem.org/cgi/doi/10.1101/lm.041145.115.

original responding to the CS will reappear. Note that the process allows the organisms to store two conflicting associations by making one of them context-dependent.

Although the study of extinction in classical conditioning and human predictive learning has developed independently of research on category learning, the mechanisms that are assumed to explain why extinction becomes context-dependent are consistent with the attentional processes that seem to be involved in category learning. It has been suggested that when participants are exposed to information that conflicts with their previous knowledge, as in extinction, attention to the context is aroused (Bouton 1997; Kruschke 2001, 2011; Rosas et al. 2006; Nelson et al. 2013). This aroused attention to the context would explain why subsequent learning depends so strongly on the context. These theories—both conditioning and category learning theories—assume that contexts are naturally attended in order to select the current meaning of ambiguous stimuli (e.g., Rosas et al. 2006). Though these theories have been applied to discrete cues, the reasoning applies equally well to contextual stimuli that are present on every trial and incidental to the task until the point that conflicting associations are acquired. Attention shifts from the ambiguous cue to the context because this might reduce the amount of error produced by the new contingency that the cue holds with the to-be-predicted outcome (either a category or an US).

However, there is an important difference between models of extinction and attentional models of category learning. Within the former approach, once attention to the context is aroused, it has been assumed to remain henceforth (Rosas et al. 2006). That is to say, any subsequent learning that takes place in that context will be context-dependent regardless of whether or not such learning produces conflicting associations. This means that if a cue is extinguished in one context, then any other association that is learned in that context will also be context-dependent, even if it is not directly involved in an extinction treatment (e.g., Rosas and Callejas-Aguilera 2006, 2007; but see Nelson et al. 2011a). The effect is robust enough in some predictive-learning tasks that once extinction has occurred in a task, contextual control arises even with contexts and stimuli not involved with extinction, and even with contexts and stimuli encountered in a different task encountered later (i.e., Rosas and Callejas-Aguilera 2006). In contrast, some attentional models of category learning assume that the level of attention paid to the context can vary depending on the specific configuration of cue and context (e.g., Kruschke 2001; see also George and Kruschke 2012; Uengoer et al. 2013). For instance, if cue A is undergoing an interference treatment in context X, and cue B is not, participants might shift attention toward X to reduce error in the presence of AX, but not when presented BX because paying attention to B does not produce a prediction error.

Although the theories of extinction discussed here assume that changes in attention are responsible for the development of context-dependent associations, none of the experiments conducted so far has taken a direct measure of attention other than using differential rates of learning as an index (Nelson et al. 2013). The goal of the present experiments was to provide a more direct measure of attention, and examine whether learning conflicting associations produces any general increase in attention to task stimuli as proposed by Rosas et al. (2006), or whether that attention is more specific to the stimuli that produce the conflict, as predicted by attentional models of categorization (Kruschke 2001, 2011).

The experiments use simultaneous measures of learning and attention in a simple category-learning task. In each trial, participants were asked to classify combinations of cues into two different categories. Concurrently, they were also tested with a dot-probe task (MacLeod et al. 1986) devised to measure the relative amount of attention paid to each of the cues. In a standard dot-probe task, participants are presented with two cues for a brief period of time (usually around 200 msec) and immediately afterward a dot appears on one of them. Participants are instructed to locate the dot as soon as possible by pressing one key if the dot appears on one stimulus (e.g., the one of the left) and a different key if the dot appears on the other stimulus (e.g., the one on the right). The typical result of the dot-probe task, and similar attentional cueing tasks (e.g., Posner et al. 1978), is that reaction times are shorter when the target dot is presented on salient or task-relevant stimuli (Koster et al. 2004). An advantage of these attentional tasks is that they can be easily combined concurrently with tasks to track how participants deploy attention to cues while they are doing something else (Raes et al. 2010; Vogt et al. 2010; Le Pelley et al. 2013).

The design of Experiment 1 is shown in Table 1. Participants were asked to categorize several combinations of cues, each containing an informative and perceptually salient cue (A–D) and a nonpredictive and nonsalient cue (X) playing the role of the context. Participants were instructed to categorize each pair of stimuli (AX, BX, CX, or DX) as members of categories 1 and 2 using the up/down arrow keys. For participants in Group Same, the cue-category assignments remained constant throughout the whole experiment. In Group Reversed, the category assignments reversed for half of the cues (A and B) on test. Although the design of the experiment departs in some important respects from the standard design used to study extinction (e.g., a punctuate, nonsalient cue plays the role of a context; and participants are trained with two conflicting outcomes, instead of one outcome and its absence) it preserves the important components: Participants have to learn contradictory information in different moments of the experiment and, in addition to the predictive cues (A–D) that signaled the correct outcome, they could nevertheless pay attention to a nonpredictive and nonsalient cue (X).

During Stage 1A participants only completed the categorization task. Starting from Stage 1B, the categorization task was combined with a dot-probe task to track their attention to cues across the experiment. Figure 1 depicts a summary of the structure of each trial combining the dot-probe and the categorization tasks. We expected an attentional advantage to emerge for the predictive cue. Reaction times in the dot-probe task should be lower when the dot appears on the attended predictive cue than on the unattended contextual stimulus. The goal of the experiment was to determine whether that attentional preference would be attenuated when the cue-category assignments changed. This attenuation would be consistent with the hypothesis that

Table 1. Design summary of Experiment 1

	Stage 1A 4 blocks × 8 trials	Stage 1B 16 blocks × 16 trials	Stage 2 6 blocks × 16 trials
	Only categorization	Categorization + dot probe	
Group reversed			**XA—2** **XB—1** XC—1 XD—2
Group same		XA—1 XB—2 XC—1 XD—2	**XA—1** **XB—2** XC—1 XD—2

Letters A–D denote different cues with distinctive colors and shapes. X denotes a dark rectangle playing the role of a contextual cue. Numbers 1 and 2 refer to the correct categories associated with each pair of cues.

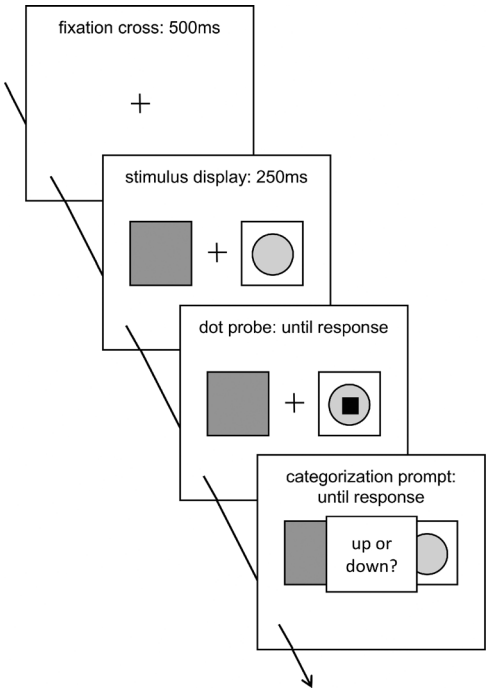

Figure 1. Schematic of the sequence of events in a standard trial from Stages 1B and 2. During Stage 1A the sequence of events was identical, except for the omission of the dot probe.

when a cue is involved in an interference treatment, part of the attention shifts away from the cue to the context. Furthermore, as stated above, the category assignments were only reversed for half of the cues. This feature of the design allowed us to test whether a similar increase in attention to the context took place for cues that were not directly involved in interference, but were nevertheless trained concurrently with an interference treatment (cues C/D in Table 1).

Results

Experiment 1

Categorization accuracy is summarized in Table 2. A Group × Cue mixed analysis of variance (ANOVA) on categorization accuracy during Stage 1 revealed that participants in Group Reversed were slightly more accurate than participants in Group Same, $F_{(1,98)} = 5.14$, $P = 0.026$, $d_s = 0.45$. However, there was no effect of Cue and no Group × Cue interaction, suggesting that within each group, participants were equally accurate in their categorization of A/B and C/D trials. A similar analysis of categorization accuracy during Stage 2 yielded significant main effects of Group, $F_{(1,98)} = 54.22$, $P < 0.001$, $d_s = 1.49$, and Cue, $F_{(1,98)} = 40.51$, $P < 0.001$, $d_z = 0.64$, and a significant interaction between both factors, $F_{(1,98)} = 26.34$, $P < 0.001$, $\eta^2_p = 0.212$. As could be expected, categorization accuracy remained high for participants in Group Same, but declined for participants in Group Reversed. Furthermore, this decrease was more marked for the categorization of cues A/B than for cues C/D.

As we were interested in the relative attention devoted to cues in relation to contexts, we constructed "cue attention advantage" scores by subtracting the response time epochs (see Materials and Methods for the definition of epochs) to predictive cues from those of the contextual cue. No reaction-time data were collected in Stage 1A, as the dot-probe task was not used here. In Stage 1B, participants came to attend to the cues more than the contexts over trials, reflected as a faster response to the cue and a positive context-minus-cue difference (see Table 3). A Group × Cue ANOVA on cue attention advantage revealed no significant effects, all Fs < 1. An epoch-by-epoch analysis of these data is available in the Supplemental Material.

Figure 2A shows cue attention advantages during Stage 2. The interested reader can find an analysis of the raw reaction time data in the Supplemental Material. Furthermore, all the raw data from this and the following experiment is available at https://osf.io/bhv2j/. As can be seen in Figure 2A, reaction times suggest that reversing the A/B outcomes caused a loss of the cues' attentional advantage (i.e., more attention to the context) on A/B trials, with little effect on C/D trials. However, the data were noisy and the attentional advantage for the A/B cues in the group for which the outcomes were not reversed was unexpectedly low in the first epoch (we should point out, however, that the A/B and C/D trials were functionally interchangeable in this group). A Group × Cue × Epoch ANOVA on these data revealed a main effect of Epoch, $F_{(2,196)} = 3.32$, $P = 0.038$, $\eta^2_p = 0.033$, and a marginally significant Cue × Epoch interaction, $F_{(2,196)} = 2.86$, $P = 0.060$, $\eta^2_p = 0.028$. Most important, the Group × Cue interaction also approached traditional levels of statistical significance, $F_{(1,98)} = 3.45$, $P = 0.066$, $\eta^2_p = 0.034$, suggesting that the decline in cue advantage in Group Reversed was more marked for cues A/B than for cues C/D. Separate Group × Epoch ANOVAs on cue advantages for A/B and C/D revealed that the main effect of Group approached significance for cues A/B, $F_{(1,98)} = 2.61$, $P = 0.11$, $d_s = 0.32$, but not for cues C/D, $F_{(1,98)} < 1$. In the case of A/B, the main effect of Epoch was also statistically significant, $F_{(1.88,184.54)} = 5.99$, $P = 0.003$, $\eta^2_p = 0.058$. The rest of effects were far from statistical significance, largest $F_{(2,196)} = 1.27$, $P = 0.283$. These results suggest that when informative cues become ambiguous, the attentional advantage that these cues have over the uninformative contextual cues is lost. The effects overall, however, were small and the crucial Group × Cue interaction fell short of what is typically considered reliable.

Experiment 2

Experiment 2 was a conceptual replication of Experiment 1, with some minor changes. In Experiment 2 we used a larger sample to increase our power. Additionally, we included two contextual stimuli. A factor that might have limited the strength of our manipulation in Experiment 1 is that the context was a constant, nonsalient, and noninformative cue. This might have caused participants to ignore cue X to such an extent that changes in relative attention during Stage 2 were difficult to detect. During Stage 1B

Table 2. Categorization accuracy in Experiments 1 and 2

		Group reversed		Group same	
Experiment	Stage	A/B	C/D	A/B	C/D
1	1	0.915(0.009)	0.910(0.008)	0.874(0.016)	0.875(0.016)
	2	0.847(0.010)	0.917(0.011)	0.967(0.010)	0.974(0.006)
2	1	0.913(0.012)	0.907(0.013)	0.923(0.008)	0.919(0.010)
	2	0.842(0.012)	0.928(0.010)	0.960(0.005)	0.961(0.006)

Mean proportion of correct responses in the categorization task for trials involving cues A/B or cues C/D, collapsed across blocks. The numbers between parentheses denote the standard errors of the mean.

Table 3. Cue attentional advantage in Stage 1B

	Group reversed		Group same	
Experiment	A/B	C/D	A/B	C/D
1	26.71 (6.43)	30.54 (7.54)	30.63 (7.03)	27.14 (7.94)
2	43.07 (5.76)	31.98 (6.58)	30.16 (6.39)	33.65 (7.05)

Attentional advantage was computed by subtracting participants' reaction time when the dot probe was presented on cues A–D from their reaction time when the dot was presented on contextual cue X. The numbers between parentheses denote the standard errors of the mean.

of Experiment 2 participants were exposed to two different contextual cues, X and Y, both of which were relatively nonsalient and irrelevant for the categorization task. The number of training trials was the same and hence the number of trials in which the stimuli occurred with context X was halved relative to Experiment 1. Finally, we extended the length of Stage 2 with two additional blocks of trials to improve the sensitivity of our dependent variables. The full design of the experiment is summarized in Supplemental Table S1 of the Supplemental Material.

We restricted the analysis of Stage 1 to the A–D cues when they occurred with X, as these were the combinations of interest on the test. During Stage 1A there were no trials with context X. Therefore our analyses were restricted to Stage 1B. As can be seen in Table 2, during Stage 1B participants' accuracy in the categorization task was similar for both cues and groups. A Group × Cue ANOVA on categorization accuracy revealed no main effects or interaction, largest $F_{(1,203)} = 1.94$, $P = 0.165$. In contrast, a similar Group × Cue ANOVA on categorization accuracies from Stage 2 revealed a significant effect of Cue, $F_{(1,203)} = 121.04$, $P < 0.001$, $d_z = 0.77$, Group, $F_{(1,203)} = 44.76$, $P < 0.001$, $d_s = 0.94$, and a Cues × Group interaction, $F_{(1,203)} = 113.79$, $P < 0.001$, $\eta^2_p = 0.359$. This interaction shows that categorization performance dropped in Group Reversed with respect to Group Same, although more so for cues A/B than for cues C/D.

The analysis of reaction times to the dot probe during Stage 1B reveals that participants developed an attentional preference for the predictive cues A–D over the contextual cue (see Table 3). Responses tended to be systematically faster when the dot appeared on cues A–D than when it appeared on cue X. A Group × Cue ANOVA on attentional advantage for cues yielded no significant main effects or interactions, largest $F_{(1,203)} = 2.07$, $P = 0.152$.

The most interesting results for our present purposes are those related to dot-probe performance during Stage 2. Figure 2B shows mean attentional advantages for cues across different groups, epochs, and cue combinations. As can be seen, these data suggest that, overall, attentional scores were larger for participants in Group Same than for participants in Group Reversed. Furthermore, this difference tends to be somewhat larger for cues A/B than for cues C/D. These impressions were confirmed by a Group × Cue × Epoch ANOVA, which yielded a main effect of Group, $F_{(1,203)} = 6.91$, $P = 0.009$, $d_s = 0.37$, and a marginally significant Group × Cue interaction, $F_{(1,203)} = 3.82$, $P = 0.052$, $\eta^2_p = 0.018$. All other main effects and interactions were nonsignificant, largest $F_{(3,609)} = 2.05$, $P = 0.105$. Follow-up ANOVAs conducted separately for cues A/B and cues C/D revealed that there was a main effect of Group for cues A/B, $F_{(1,203)} = 11.66$, $P = 0.001$, $d_s = 0.48$, but not for cues C/D, $F_{(1,203)} = 1.574$, $P = 0.211$. Overall, these results converge to the conclusion that the contingency reversal of cues A/B lead to a decrement in the attentional preference for these cues over the contextual cue X. The effect tended to be smaller (indeed, nonsignificant) for cues C/D, which were trained concurrently with cues A/B but were not directly involved in a contingency reversal. Unfortunately, as in Experiment 1, the crucial Group × Cue interaction missed full statistical significance ($P = 0.052$) despite the larger sample size and the inclusion of two extra testing blocks.

Combined analysis of Experiments 1 and 2

The results of our experiments converged to the conclusion that the decline of attentional advantage was more marked for the cues directly involved in the contingency reversal (A/B) than for other cues trained concurrently (C/D). However, the critical Group × Cue interaction was only marginally significant in both cases (*P* values equal to 0.066 and 0.052, in Experiments 1 and 2, respectively). These consistently improbable interactions

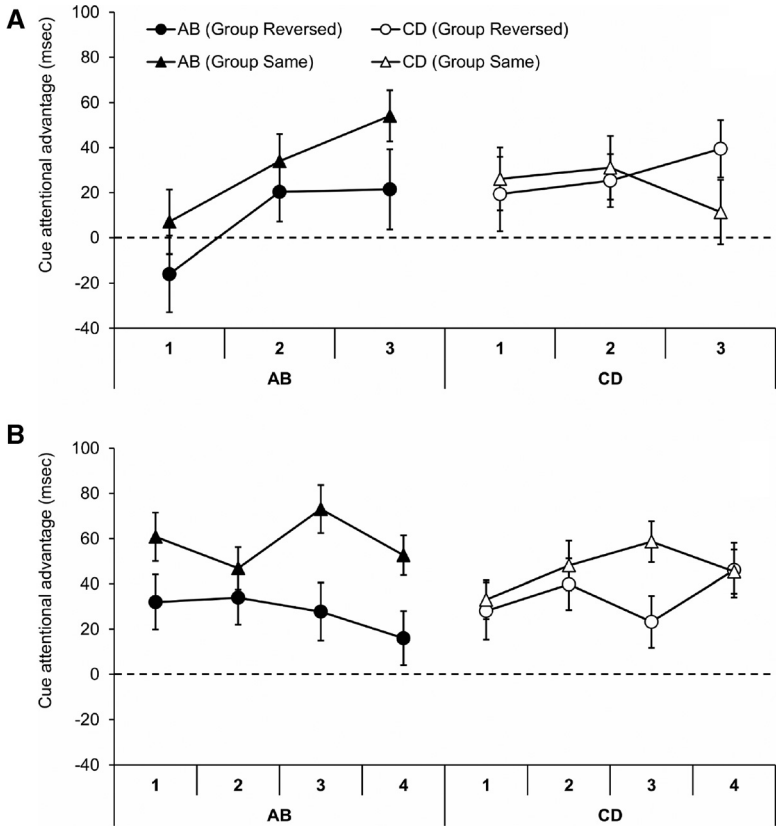

Figure 2. Mean attentional advantage for cues A/B and C/D during Stage 2 test in Experiments 1 and 2 (*A* and *B*, respectively). Attentional advantage was computed by subtracting participants' reaction time when the dot probe was presented on cues *A–D* from their reaction time when the dot was presented on contextual cue X. Error bars denote the standard error of the means. Each epoch comprises data from two blocks of trials.

strongly suggest that a true interaction might exist, although they also cast doubts on the reliability of our dependent variable or the statistical power of our samples. To overcome these problems, we conducted a high-powered combined analysis of the data from both experiments. Given that Experiment 1 included only six testing blocks (collapsed in three epochs), while Experiment 2 comprised eight testing blocks (collapsed in four epochs), we averaged all reaction-time data across blocks.

Figure 3 shows the results of the combined analyses. As can be seen, overall cue advantages tended to be lower for participants in Group Reversed than for participants in Group Same. However, this decrease is steeper for cues A/B than for cues C/D. An Experiment × Group × Cue mixed ANOVA yielded significant main effects of Experiment, $F_{(1,301)} = 6.85$, $P = 0.009$, $d_s = 0.32$, and Group, $F_{(1,301)} = 4.56$, $P = 0.034$, $d_s = 0.24$. The main effect of Cue was far from statistical significance, $F_{(1,301)} < 1$. Neither the Experiment × Group interaction, $F_{(1,301)} < 1$, or the Experiment × Cue interaction, $F_{(1,301)} < 1$, reached statistical significance. Most important, the critical Group × Cue interaction that did not reach full significance in the individual analyses was now statistically significant, $F_{(1,301)} = 7.26$, $P = 0.007$, $\eta^2_p = 0.024$, and did not interact with experiment, $F_{(1,301)} < 1$. Across experiments, the attentional advantage for cues A/B was significantly different between Groups Reversed and Same, $t_{(303)} = 3.71$, $P < 0.001$, $d_s = 0.43$. For cues C/D, this difference was nonsignificant, $t_{(303)} = 0.87$, $P = 0.383$, $d_s = 0.10$. The combined analysis also provides some hint as to why the Group × Cue interaction failed to reach full statistical significance in each individual experiment. If this interaction is seen as a between-groups (Reversed vs. Same) difference of a within-participants (A/B versus C/D) difference, then its effect size measured in Cohen's d_s units is 0.31. Even with the relatively large samples used in Experiments 1 and 2, our power to detect this effect in a two-tailed test was only 0.34 and 0.60, respectively. In contrast, the observed power of the aggregate analysis is 0.93.

Discussion

The results of the present series of experiments suggest that exposure to information that contradicts previous beliefs produces a shift in attention. Specifically, when predictive cues changed their meaning and were assigned to new categories, participants' attentional preference for the predictive cues relative to concomitant, less conspicuous, and nonpredictive contextual cues was disrupted. This result is consistent with the predictions of some attentional models of category learning. For instance, according to the EXIT model devised by Kruschke (2001) attention shifts away rapidly from any cue that produces a prediction error. This feature of the model allows it to explain a significant number of associative learning phenomena such as highlighting and blocking (Kruschke 2001, 2009). In our experiments, on the first trials of Stage 2, participants' previous knowledge of the categories associated with cues A/B should produce a prediction error. According to the EXIT model, this prediction error can be partially ameliorated by diverting attention away from those cues. As a result the relative amount of attention available to process the nonpredictive background cues increases.

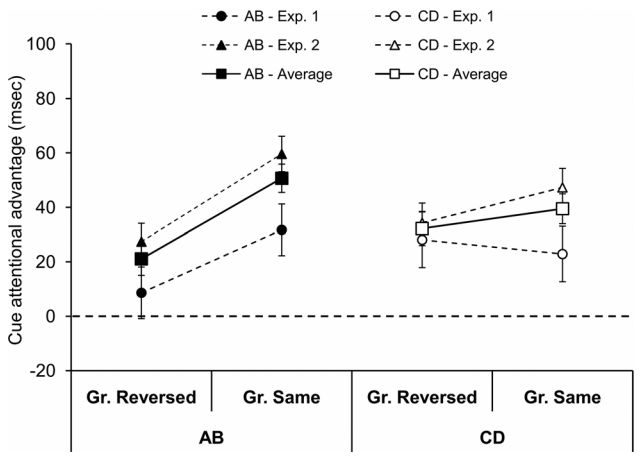

Figure 3. Mean attentional advantage for cues A/B and C/D during Stage 2 test in Experiments 1 and 2 collapsed across epochs. The series with the larger markers denotes the average cue advantage across experiments. Error bars denote the standard error of the means.

As discussed in the Introduction, the fact that prediction errors can produce these changes in attention to cues and contexts might provide an insight into the mechanisms of one of the most popular areas of research in animal and human conditioning. Since the days of Pavlov, it is known that the extinction of conditioned responses cannot be explained in terms of unlearning. Even if these responses are reduced to negligible levels, there are some manipulations that can promote a partial or full recovery of conditioned responding (Pavlov 1927; Rescorla and Heth 1975; Bouton and Bolles 1979; Bouton and King 1983). These findings suggest that extinction does not consist of the removal of previously learned associations, but of the learning of new associations that counteract previous learning. The fact that conditioned responses can reappear in some circumstances shows that whatever people and animals learn during extinction is not always generalized to new situations. Traditionally, the recovery of conditioned responding has been attributed to changes in the context from the extinction phase to the testing phase (Bouton 1993, 1997). From this point of view, the associations learned during extinction are bound together with a representation of the context in which they were learned. These associations are only expressed when the cue or conditioned stimulus is presented specifically in that context. As a result, if the cue is presented in a context that differs somehow from that in which extinction was learned, conditioned responses will reappear.

This theoretical view has become the standard explanation for many learning effects related to extinction and to interference in general, both within the animal conditioning tradition and in human learning research. It also provides an excellent background to understand why psychological treatments for anxiety and fear disorders (usually based on ideas taken from the extinction and interference literatures) sometimes fail to prevent the resurgence of clinical symptoms (Milad and Quirk 2012; Vervliet et al. 2013). However, a missing detail in this general framework is explaining why the associations learned during extinction and interference become context-dependent in first instance.

In a brief concluding comment, Bouton (1997) speculated that when extinction takes place and an organism discovers that the meaning of a cue is ambiguous, it might begin to pay more attention to the context in case changes in the context prove useful to predict when the CS will be followed by the US and when not. The idea that changes in attention can explain why interfering information becomes context-dependent has stimulated an interesting series of experiments during the last years (Rosas and Callejas-Aguilera 2006, 2007; Nelson et al. 2013; Bernal-Gamboa et al. 2014). But, unfortunately, none of these studies has directly measured changes in attention while participants are exposed to conflicting information. Instead, they have used attention to explain the effect of a context switch and the effect of the context switch to infer attention. Here we have demonstrated the context switch effect while measuring attention independently. Our

results confirm that learning interfering information does produce a change in attention, so that part of the attention usually allocated to predictive cues diverts to nonpredictive and nonsalient cues analogous to the contexts of associative learning experiments (in particular the type used by Rosas and Callejas-Aguilera 2006). Ideally, future research should follow up our results using alternative procedures to measure attentional processes (e.g., Livesey et al. 2009; Wills et al. 2014; Glautier and Shih 2015; Luque et al. 2015).

Attentional models of category learning not only provide an interesting framework to fully understand how attention is deployed to cues and contexts in learning effects related to interference and extinction. They also make some specific predictions that stand in contrast with those made by some theories of extinction and interference (i.e., ATCP by Rosas et al. 2006). According to EXIT (Kruschke 2001) the amount of attention paid to a cue can depend on the specific exemplar or configuration where that cue is presented. A specific cue, A, can be attended in the presence of a given cue, B, but not in the presence of another cue, C. From this point of view, the relative increase in attention to the context that we observed in the present experiments should only be observed for configurations of cues and contexts that were directly involved in a prediction error. That is to say, if the reversal of categories in trials AX and BX produced a shift in attention toward X, this need not affect the deployment of attention to X when other configurations, like CX or DX, are presented, because these cues were never involved in a category reversal. This prediction is partially confirmed by our results. We observed a shift in attention to cues C and D following the reversal of cues A and B. The effect on C/D, however, tended to be smaller than the effect that the manipulation had on cues A and B and, in fact, did not reach full statistical significance in the high-powered combined analysis. Although the prediction that aroused attention to the context is configuration-independent has received some empirical support (Rosas and Callejas-Aguilera 2006, 2007; Bernal-Gamboa et al. 2014), some of that supporting evidence has been difficult to replicate (Nelson et al. 2011a). Experiments exploring effects unrelated to extinction or interference have traditionally found that the attention paid to previously predictive stimuli transfers quite well to new situations and new compounds (Le Pelley and McLaren 2003; Lochman and Wills 2003; Griffiths and Le Pelley 2009; Livesey et al. 2009), but some experiments suggest that the amount of attention paid to cues can be configuration dependent (George and Kruschke 2012). We hope that our results and the methods used in the present series of experiments will contribute to the resolution of this exciting debate.

Materials and Methods

Experiment 1

One hundred psychology students from UNED volunteered to take part in the experiment. Half of them were randomly assigned to Group Reversed and the other half to Group Same. Previous research with the experimental task used in this study (Le Pelley et al. 2013, Experiment 3) suggested that this sample size was large enough to detect any modulation of attentional effects in a mixed experimental design. Participants conducted the experiment in small groups in a computer room with individual cubicles. The experiment was conducted on PCs with 15-in TFT monitors set at a resolution of 800 × 600. The experimental program was written in MATLAB using Cogent 2000 and Cogent Graphics (www.vislab.ucl.ac.uk/cogent.php) to present stimuli and record participants' responses.

Four differently colored geometrical shapes (a set of blue [red–green–blue: 0, 0, 255] diagonal lines, a purple ellipse [255, 50, 255], a yellow triangle [255, 255, 0], and a light blue cross [0, 255, 255]) were randomly assigned to cues A–D. For all participants, context cue X was a dark green square (150, 150, 0). These cues were presented against a black background. As shown in Figure 1, these stimuli were embedded in one of two white boxes with size 290 × 290 pixels.

The procedure of the experiment is very similar to the one used in Experiment 3 of Le Pelley et al. (2013). Stage 1A was a pretraining phase to allow participants to gain familiarity with the categorization task without the additional complexity of the dot-probe task. The sequence of events within each trial was similar to the one represented in Figure 1, except that the dot probe was not presented. After a white fixation cross (20 × 20 pixels), participants were presented two rectangles at both sides of the screen, each contained either a predictive or contextual cue. The assignment of distinctive cues A–D to the left/right of the rectangle containing X was randomized across trials. Participants were instructed to categorize each pair of stimuli (AX, BX, CX, or DX) using two different responses, 1 and 2. If participants thought that the correct category for a pair of stimuli was 1, they were instructed to press the up arrow key. If they thought that the correct category was 2, they were instructed to press the down arrow key. Although participants were not told in advance the correct category for each pair of stimuli, they were given corrective feedback after every incorrect response and were instructed to use this feedback to learn the correct cue-category associations. During Stage 1A participants were exposed to four blocks of trials, each block contained eight trials (four cue types [A–D] × 2 left/right cue locations) presented in random order.

Immediately after Stage 1A participants were presented with a set of instructions alerting them to the addition of the dot-probe task. During Stage 1B participants performed the same categorization task, but now they also had to report the location of a dot probe that could appear either on the cue or the contextual cue stimulus. The sequence of events within each trial is represented in Figure 1. After the fixation cross, participants were presented with the two stimuli at both sides of the screen. After 250 msec the dot probe appeared on top of the left or the right stimulus. The dot probe was a 30 × 30 pixel red square with red–green–blue values of (255, 0, 0). Participants were instructed to report the location of the dot as fast as possible using the left/right arrow keys. Furthermore, they were instructed to ignore the identity of the stimuli until they had responded to the location of the dot. If participants' response was incorrect, the trial did not proceed until they entered the correct response. After responding to the location of the dot, the dot probe disappeared from the screen and a categorization prompt appeared in the center of the screen. Participants were instructed to use the up/down arrow keys to categorize the stimuli as they had done during Stage 1A. During Stage 1B participants completed 16 blocks of trials, each containing 16 trials (4 trial types × 2 cue locations × 2 dot-probe locations). There was no interruption between Stages 1B and 2. Except for the differences in the design summarized in Table 1, the procedure was identical during Stage 2 testing. The outcome assignment for cues A and B was reversed in Group Reversed, and remained the same in Group Same. Participants completed 6 blocks of 16 trials each before finishing the experiment.

The dot-probe data from Stage 1B and test were initially screened to eliminate responses that could have been initiated before the stimulus was presented (<150 msec), and responses clearly indicative of inattention to the task (>1.5 sec). Reaction times (RT) from trials where the position of the dot was incorrectly reported were also eliminated. Then, the mean and standard deviation of the remaining scores were calculated for each subject. Scores that were more than three standard deviations away from the mean were removed. The data were then collapsed into Epochs for the A/B trials and the C/D trials. An Epoch consisted of eight trials of each type. For example, the first Epoch for dot-probe responses when the dot appeared on either the A or B cue consisted of the average of the first two trials where the A or B cues appeared to the left of X and the first two trials where the cues appeared to the right of X. Epochs were calculated separately for trials where the dot appeared on the predictive cue (A–D) and trials when the dot appeared on the contextual cue (X). Excluded data were simply not included in the averages.

Experiment 2

Two hundred and five psychology students from UNED volunteered to take part in Experiment 2. Ninety-nine of them were randomly assigned to Group Reversed and 106 to Group Same. Participants conducted the experiment in the same setting used in Experiment 1. As shown in Supplemental Table S1, the design is very similar to the one used in Experiment 1. The main difference is that during Stage 1B participants were exposed to two different contextual cues, X and Y. A dark green rectangle and a dark red (128, 0, 0) rectangle were used as context cues. With the addition of the new contextual cue, the 256 Stage 1B trials were divided into 8 blocks. Each block now consisted of 32 trials (4 trial types × 2 contexts × 2 cue locations × 2 dot locations). During Stage 2 testing only contextual cue X was presented in all trials. Unlike in Experiment 1, Stage 2 comprised eight blocks of trials (instead of six). Apart from these differences, all the details in the procedure and design were identical to Experiment 1.

Acknowledgments

The authors would like to thank Andy Wills for his valuable comments on earlier versions of this article. Participation of J.B.N. was supported, in part, by grant PSI2014-52263-C2-2-P from the Spanish Ministry of Science and Innovation and grant IT-694-13 from the Basque Government.

References

Anderson MC. 2003. Rethinking interference theory: executive control and the mechanisms of forgetting. *J Mem Lang* **49:** 415–445.

Bernal-Gamboa R, Rosas JM, Callejas-Aguilera JE. 2014. Experiencing extinction within a task makes nonextinguished information learned within a different task context-dependent. *Psychon Bull Rev* **21:** 803–808.

Bouton ME. 1993. Context, time, and memory retrieval in the interference paradigms of Pavlovian learning. *Psychol Bull* **114:** 80–99.

Bouton ME. 1997. Signals for whether versus when an event will occur. In *Learning, motivation, and cognition: the functional behaviorism of Robert C. Bolles* (ed. Bouton ME, Fanselow MS), pp. 385–409. American Psychological Association, Washington, DC.

Bouton ME, Bolles RC. 1979. Contextual control of the extinction of conditioned fear. *Learn Motiv* **10:** 445–466.

Bouton ME, King DA. 1983. Contextual control of the extinction of conditioned fear: tests for the associative value of the context. *J Exp Psychol Anim Behav Process* **9:** 248–265.

Cobos PL, González-Martín E, Varona-Moya S, López FJ. 2013. Renewal effects in interference between outcomes as measured by a cued response reaction time task: further evidence for associative retrieval models. *J Exp Psychol Anim B Behav Process* **39:** 299–310.

Dunsmoor JE, Niv Y, Daw N, Phelps EA. 2015. Rethinking extinction. *Neuron* **88:** 47–63.

George DN, Kruschke JK. 2012. Contextual modulation of attention in human category learning. *Learn Behav* **40:** 530–541.

Glautier S, Shih S. 2015. Relative prediction error and protection from attentional blink in human associative learning. *Q J Exp Psychol (Hove)* **68:** 442–458.

Griffiths O, Le Pelley ME. 2009. Attentional changes in blocking are not a consequence of lateral inhibition. *Learn Behav* **37:** 27–41.

Koster EHW, Crombez G, Verschuere B, De Houwer J. 2004. Selective attention to threat in the dot probe paradigm: differentiating vigilance and difficulty to disengage. *Behav Res Ther* **42:** 1183–1192.

Kruschke JK. 2001. Toward a unified model of attention in associative learning. *J Math Psychol* **45:** 812–863.

Kruschke JK. 2009. Highlighting: a canonical experiment. In *The psychology of learning and motivation* (ed. Ross B), Vol. 51, pp. 153–185. Academic Press, San Diego, CA.

Kruschke JK. 2011. Models of attentional learning. In *Formal approaches in categorization* (ed. Pothos EM, Wills AJ), pp. 120–152. Cambridge University Press, Cambridge, UK.

Le Pelley ME, McLaren IPL. 2003. Learned associability and associative change in human causal learning. *Q J Exp Psychol B* **56:** 68–79.

Le Pelley ME, Vadillo M, Luque D. 2013. Learned predictiveness influences rapid attentional capture: evidence from the dot probe task. *J Exp Psychol Learn Mem Cogn* **39:** 1888–1900.

Livesey EJ, Harris IM, Harris JA. 2009. Attentional changes during implicit learning: signal validity protects a target stimulus from the attentional blink. *J Exp Psychol Learn Mem Cogn* **35:** 408–422.

Lochmann T, Wills AJ. 2003. Predictive history in an allergy prediction task. In *Proceedings of EuroCogSci 03: the European Cognitive Science Conference* (ed. Schmalhofer F, Young RM, Katz G), pp. 217–222. Lawrence Erlbaum Associates, Mahwah, New Jersey, USA.

Luque D, Morís J, Rushby JA, Le Pelley ME. 2015. Goal-directed EEG activity evoked by discriminative stimuli in reinforcement learning. *Psychophysiology* **52:** 238–248.

MacLeod C, Mathews A, Tata P. 1986. Attentional bias in emotional disorders. *J Abnorm Psychol* **95:** 15–20.

McClelland JL. 2013. Incorporating rapid neocortical learning of new schema-consistent information into complementary learning systems theory. *J Exp Psychol Gen* **142:** 1190–1210.

Milad MR, Quirk GJ. 2012. Fear extinction as a model for translational neuroscience: ten years of progress. *Annu Rev Psychol* **63:** 129–151.

Nelson JB, Callejas-Aguillera JE. 2007. The role of interference produced by conflicting associations in contextual control. *J Exp Psychol Anim Behav Process* **33:** 314–326.

Nelson JB, Lombas S, León SP. 2011a. Concurrent extinction does not render appetitive conditioning context specific. *Learn Behav* **39:** 87–94.

Nelson JB, Sanjuan MC, Vadillo-Ruiz S, Pérez J, León SP. 2011b. Experimental renewal in human participants. *J Exp Psychol Anim Behav Process* **37:** 58–70.

Nelson JB, Lamoureux JA, León SP. 2013. Extinction arouses attention to the context in a behavioral suppression method with humans. *J Exp Psychol Anim Behav Process* **39:** 99–105.

Pavlov IP. 1927. *Conditioned reflexes*. Oxford University Press, London.

Posner MI, Nissen MJ, Ogden WC. 1978. Attended and unattended processing modes: the role of set for spatial location. In *Modes of perceiving and processing information* (ed. Pick HL, Saltzman IJ), pp. 137–157. Erlbaum. Hillsdale, NJ.

Raes AK, Koster EHW, Van Damme S, Fias W, De Raedt R. 2010. Aversive conditioning under conditions of restricted awareness: effects on spatial cueing. *Q J Exp Psychol (Hove)* **63:** 2336–2358.

Rescorla RA, Heth CD. 1975. Reinstatement of fear to an extinguished conditioned stimulus. *J Exp Psychol Anim Behav Process* **1:** 88–96.

Rosas JM, Callejas-Aguilera JE. 2006. Context switch effects on acquisition and extinction in human predictive learning. *J Exp Psychol Learn Mem Cogn* **32:** 461–474.

Rosas JM, Callejas-Aguilera JE. 2007. Acquisition of a conditioned taste aversion becomes context dependent when it is learned after extinction. *Q J Exp Psychol (Hove)* **60:** 9–15.

Rosas JM, Vila NJ, Lugo M, López L. 2001. Combined effect of context change and retention interval on interference in causality judgments. *J Exp Psychol Anim Behav Process* **27:** 153–164.

Rosas JM, Callejas-Aguilera JE, Ramos-Álvarez MM, Fernández-Abad MJ. 2006. Revision of retrieval theory of forgetting: what does make information context-specific? *Int J Psychol Psychol Ther* **6:** 147–166.

Sewell DK, Lewandowsky S. 2012. Attention and working memory capacity: insights from blocking, highlighting, and knowledge restructuring. *J Exp Psychol Gen* **141:** 444–469.

Smith DM, Bulkin DA. 2014. The form and function of hippocampal context representations. *Neurosci Biobehav Rev* **40:** 52–61.

Speekenbrink M, Shanks DR. 2010. Learning in a changing environment. *J Exp Psychol Gen* **139:** 266–298.

Uengoer M, Lachnit H, Lotz A, Koenig S, Pearce JM. 2013. Contextual control of attentional allocation in human discrimination learning. *J Exp Psychol Anim Behav Process* **39:** 56–66.

Vadillo MA, Vegas S, Matute H. 2004. Frequency of judgment as a context-like determinant of predictive judgments. *Mem Cognit* **32:** 1065–1075.

Vadillo MA, Orgaz C, Luque D, Cobos PL, López FJ, Matute H. 2013. The role of outcome inhibition in interference between outcomes: a contingency-learning analogue of retrieval-induced forgetting. *Brit J Psychol* **104:** 167–180.

Vervliet B, Craske MG, Hermans D. 2013. Fear extinction and relapse: state of the art. *Annu Rev Clin Psychol* **9:** 215–248.

Vogt J, De Houwer J, Moors A, Van Damme S, Crombez G. 2010. The automatic orienting of attention to goal-relevant stimuli. *Acta Psychol (Amst)* **134:** 61–69.

Wagner AR. 1981. SOP: a model of automatic memory processing in animal behavior. In *Information processing in animals: memory mechanisms* (ed. Spear NE, Miller RR), pp. 5–47. Erlbaum, Hillsdale, NJ.

Wills AJ, Lavric A, Hemmings Y, Surrey E. 2014. Attention, predictive learning, and the inverse base-rate effect: evidence from event-related potentials. *Neuroimage* **87:** 61–71.

Received November 19, 2015; accepted in revised form January 21, 2016.