Ambiguity produces attention shifts in category learning

Miguel A. Vadillo,1,2 Cristina Orgaz,3 David Luque,4,5 and James Byron Nelson6

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 Boiles 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

© 2016 Vadillo et al. This article is distributed exclusively by Cold Spring Harbor Laboratory Press for the first 12 months after the full-issue publication date (see http://learnmem.cshlp.org/site/misc/terms.xhtml). After 12 months, it is available under a Creative Commons License (Attribution-NonCommercial 4.0 International), as described at http://creativecommons.org/licenses/by-nc/4.0/.

Miguel A. Vadillo, Cristina Orgaz, David Luque, and James Byron Nelson.

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

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). Although 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; Uemgoer 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 were 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, non-salient 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>
<thead>
<tr>
<th>Table 1. Design summary of Experiment 1</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Only categorization</td>
</tr>
</tbody>
</table>

**Group reversed**

|                                | AX—2                          | XB—1                          |
|--------------------------------|-------------------------------|
|                               | XA—1                          | XC—1                          |
|                               | XB—2                          | XD—2                          |

**Group same**

|                                | AX—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.
Categorization accuracy is summarized in Table 2. A Group × Cue Experiment 1 Results nevertheless trained concurrently with an interference treatment for cues that were not directly involved in interference, but were stated above, the category assignments were only reversed for attention shifts away from the cue to the context. Furthermore, as when a cue is involved in an interference treatment, part of the attention shifts away from the cue to the context. Furthermore, as when a cue is involved in an interference treatment (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, $F_{(1,98)} = 2.61, P = 0.11, \eta^2_p = 0.032$, 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 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, $F_{(1,98)} = 2.61, P = 0.11, \eta^2_p = 0.032$, 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.

### 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, $F_{(1,98)} = 5.14, P = 0.026, \eta^2_p = 0.045$. 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, $F_{(1,98)} = 54.22, P < 0.001, \eta^2_p = 0.364$, and Cue, $F_{(1,98)} = 40.51, P < 0.001, \eta^2_p = 0.321$. 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, $F_{(1,98)} = 2.61, P = 0.11, \eta^2_p = 0.032$, 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.

### Table 2. Categorization accuracy in Experiments 1 and 2

<table>
<thead>
<tr>
<th>Experiment</th>
<th>Stage</th>
<th>Group reversed</th>
<th>Group same</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>A/B</td>
<td>C/D</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0.915(0.009)</td>
<td>0.910(0.008)</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>0.847(0.010)</td>
<td>0.917(0.011)</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>0.913(0.012)</td>
<td>0.907(0.013)</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>0.842(0.012)</td>
<td>0.928(0.010)</td>
</tr>
</tbody>
</table>

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.
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 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, $F_{(1,203)} = 121.04, P < 0.001, d_{h} = 0.77$, Group, $F_{(1,203)} = 44.76, P < 0.001, d_{h} = 0.94$, 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, $F_{(1,203)} = 6.91, P = 0.009, d_{h} = 0.37$, 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, $F_{(1,203)} = 11.66, P = 0.001, d_{h} = 0.48$, 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.

**Table 3.** Cue attentional advantage in Stage 1B

<table>
<thead>
<tr>
<th>Experiment</th>
<th>Group Reversed</th>
<th>Group Same</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A/B C/D</td>
<td>A/B C/D</td>
</tr>
<tr>
<td>1</td>
<td>26.71 (6.43)</td>
<td>30.54 (7.54)</td>
</tr>
<tr>
<td>2</td>
<td>43.07 (5.76)</td>
<td>31.98 (6.58)</td>
</tr>
</tbody>
</table>

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.

![Figure 2](https://www.learnmem.org/137/Learning%20%26%20Memory/137/fig2.png)

**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.

---

Attention and ambiguous cues

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...
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, $F_{(1,301)} = 6.85$, $P = 0.009$, $d = 0.32$, and Group, $F_{(1,301)} = 4.56$, $P = 0.034$, $d = 0.24$. 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, $t_{(303)} = 3.71$, $P < 0.001$, $d = 0.43$. For cues C/D, this difference was non-significant, $t_{(303)} = 0.87$, $P = 0.383$, $d = 0.10$. 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 $d$ 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.

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 A or B cues appeared to the left of X and the first two trials where the A or B cues appeared to the right of X. Epochs were calculated separately for trials where the dot appeared on either the A or B cues, and on B trials and the C cues, respectively.

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-inch 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 containing 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 containing 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. 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 correction 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 instructions in the dot-probe task, 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 (< 1.50 msce), 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.

Materials and Methods
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 4 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 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
Lawrence Erlbaum Associates, Mahwah, New Jersey, USA.

Received November 19, 2015; accepted in revised form January 21, 2016.
Ambiguity produces attention shifts in category learning

Miguel A. Vadillo, Cristina Orgaz, David Luque, et al.

Access the most recent version at doi:10.1101/lm.041145.115

Supplemental Material
http://learnmem.cshlp.org/content/suppl/2016/03/10/23.4.134.DC1.html

Creative Commons License
This article is distributed exclusively by Cold Spring Harbor Laboratory Press for the first 12 months after the full-issue publication date (see http://learnmem.cshlp.org/site/misc/terms.xhtml). After 12 months, it is available under a Creative Commons License (Attribution-NonCommercial 4.0 International), as described at http://creativecommons.org/licenses/by-nc/4.0/.

Email Alerting Service
Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or click here.

To subscribe to Learning & Memory go to:
http://learnmem.cshlp.org/subscriptions

© 2016 Vadillo et al.; Published by Cold Spring Harbor Laboratory Press