View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Open Research Exeter

Livesey, E.J[1,2]. and McLaren I. P.L[1,3]. (2009). Discrimination and Generalization Along a Simple Dimension: Peak Shift and Rule-Governed Responding. *Journal of Experimental Psychology: Animal Behavior Processes*, 35, 554-565.

1. University of Cambridge, Cambridge, UK

2. University of Sydney, Sydney, Australia

3. University of Exeter, Exeter, UK

This post-print may differ from the final published version.

Correspondence:
Evan Livesey
School of Psychology
University of Sydney
NSW 2006
Australia
Fax: +61 2 9036 5223
Email: evanl@psych.usyd.edu.au

Abstract

In two experiments, participants learned to discriminate between a pair of simply related, but very similar colors, in a two-choice categorization task. They were then tested over a wider range of isoluminant hues. Over these test values, both experiments yielded a post-discrimination gradient that was initially peak-shifted but became monotonic through the course of testing. In Experiment 2, the presence of this early peak shift and subsequent change in gradient form were related to participants' inability to verbally characterize the difference between the training stimuli. This suggests a transition from generalization based on simple physical similarity to generalization based on a verbalizeable rule, as a consequence of additional relevant information becoming available during test. An explanation appealing to both associative and strategically controlled verbal processes provides an accurate account of the results.

Peak shift is a widely and consistently reported phenomenon in animal learning, in particular as a consequence of discrimination learning along a stimulus dimension. Most laboratory animals that are rewarded for responding to one stimulus value (S+) and given less or no reinforcement for responding to another very similar value (S-) tend to show a peak preference for stimulus values similar to S+ but slightly removed along the relevant stimulus dimension, away from S-. In Hanson's (1957) original demonstration of the effect, pigeons trained on a spectral discrimination with 550nm S+ and 560nm S- showed a strong preference for a test stimulus value of 540nm over 550nm. Importantly, however, even when peak responding is displaced from the trained stimulus, as in the case of peak shift, conditioned responding tends to decline rapidly to baseline at more extreme values along the dimension. Hanson's pigeons, for instance, showed a sharp decline in responding for 530nm and shorter wavelengths.

Mackintosh (1997; see also Wills and Mackintosh, 1998) has focussed attention on this observation in discussing relational learning in animals, as it suggests that discriminative behavior is often controlled by elementary physical stimulus properties rather than abstract relationships between stimuli, even when the stimuli are conceptually related in a very simple manner. The extreme stimulus values which elicit less responding are less similar to S+ in terms of their absolute physical properties, even though, in many instances they are easier to classify in terms of their abstract or symbolic relationship to the training stimuli. For instance, two visual stimuli consisting of very similar shades of grey used as S+ and S- might be quite physically dissimilar to a very dark (almost black) or very bright (almost white) stimulus, but the abstract relationship between those

training stimuli (one being brighter than the other) could readily be applied to either extreme and provide a conceptual basis for classifying the stimuli. However, as Mackintosh (1997, 2000) has noted, given this and similar scenarios many laboratory animals appear to respond only on the basis of physical similarity, and do not show much or any preference for stimulus values that are physically quite dissimilar to S+. To this extent, discrimination learning in animals appears to be reasonably well accounted for by an appeal to associative learning mechanisms. Spence's (1937) analysis of generalization from overlapping gradients of conditioned excitation and inhibition provides an in-principle account of why one should expect peak shift. These basic generalization processes in combination with an associative learning algorithm based on prediction error can provide very accurate quantitative fits of post-discrimination gradients in a variety of situations (Blough, 1975; Ghirlanda & Enquist, 1998; Enquist & Ghirlanda, 2005).

In contrast, most human categorization experiments using simple stimulus dimensions have produced monotonic gradients where some differential responding is evident for the training stimuli but accuracy continues to increase the further one moves along the dimension. Response accuracy is highest at the most extreme values of the test range rather than peaking at stimulus values close to the training stimuli (S). Such gradients have been obtained on numerous stimulus dimensions including stimulus location (La Berge, 1961), auditory frequency (Cross & Lane, 1962), lifted weight (Capehart & Pease, 1968), brightness (Hebert, 1970; Thomas, Lusky & Morrison, 1992; Wills & Mackintosh, 1998), and line orientation (Thomas et al, 1992). While the authors of several of these studies claim to provide results consistent with the peak shift effect, the lack of a decline

in accuracy towards more extreme stimulus values provides a clear and consistent contrast with the animal literature, and has been observed over diverse stimulus dimensions.

There are of course some significant procedural differences between animal discrimination learning experiments and categorization in humans, not least of all the response requirements, which typically involve single response free-operant or go/no-go tasks for animals and two-alternative forced choice for humans. However Blough (1973) found peak shifted gradients using a two-choice procedure with pigeons, suggesting that having a response choice alone does not necessarily yield monotonic gradients. As will be discussed below, categorization procedures in humans can also produce truly peak-shifted gradients with more complex stimulus dimensions.

Instead, these results point strongly to the influence of cognitive strategies based on the relationship between the training stimuli and mediated by some relevant response rule. There is little doubt that human participants are able to learn abstract relationships and so give responses which are not grounded in the absolute physical properties of the objects and events to which they refer. This assumption forms an integral part of some theories of discrimination learning based on comparison with a decision criterion, such as Thomas' (1993) particular variant of Adaptation Level theory. In most of the experiments cited above, the training stimuli were sufficiently different to be easily distinguishable from one another from the outset. Indeed, where acquisition results are reported it seems the discriminations were acquired very rapidly. This makes it all the

more plausible that human participants would identify and use a rule based on the abstract stimulus relationship that was relevant for that discrimination. It is reasonable to assume that strategic decisions of this sort are most likely to be initiated by the participant when the specific task requirements strongly imply that an identifiable categorical rule might be useful, as is the case in two-choice categorization of relatively simple stimuli varying on a dimension whose characteristics are very easily stated.

This highlights a major distinction between the theoretical processes invoked to explain how animals and humans solve basic successive discriminations. On the one hand, associative mechanisms give a reasonably accurate account of conditioned discriminative behavior in animals. On the other hand, classification by human participants appears to be much more flexible and can engage reasoning and rule abstraction. Discriminative judgements could thus be based on inductive processes which allow the derivation of rules from episodic or working memory. Abstraction of relative stimulus properties might be considered to be a process, possibly verbally mediated, which proceeds via comparisons to stored representations of past events. However this doesn't completely rule out the involvement of more primitive learning processes in human discrimination learning. Following some dual-process theories of human learning (e.g. McLaren, Green and Mackintosh, 1994) it might be assumed that discriminative learning occurs at a more fundamental associative level, in parallel with the development of higher order response strategies. Associations between the stimulus and correct response may not be manifest when a participant is applying cognitive strategies - associatively driven action may well be the kind of behavior that the executive control available in human cognition can

actively override. However, there remains the possibility that in situations where executive processes are restricted, evidence for more simple associations may emerge. More specifically for the experiments that follow, one might instead expect discrimination learning to be driven largely by associative processes in situations where an effective response rule is difficult to verbalize or apply confidently, such as when the actual physical relationship between the discriminative stimuli is difficult to accurately verbalize. In such cases, post-discrimination generalization should more closely reflect the peak-shifted gradients seen in animal experiments.

This analysis is given qualified support by the observation that human categorization along more complex stimulus dimensions tends to agree better with an associative analysis and does not typically result in monotonic gradients. For instance, Wills and Mackintosh (1998) found very similar peak-shifted post-discrimination gradients produced by pigeons and humans along a dimension constructed of abstract shapes of varying spatial arrangement and frequency distribution. McLaren and Mackintosh (2002) also report a peak-shifted gradient for two-choice categorization along a morphed face dimension. Similarly, when the dimensional relationship between stimuli is masked by task requirements, the results appear to be more in favor of a peak shift effect. In an unpublished study, Aitken (1996; also reported briefly in Mackintosh, 1997) showed that a categorization task which normally produced monotonic gradients in response accuracy could be made to yield peak-shifted gradients by changing the task to one involving speeded, cued reactions, thereby making discrimination learning incidental to performance of the task.

However, the question remains whether there is any evidence for associative processes influencing the classification of stimuli that are more simply related along a single physical dimension. Within the realm of human post-discrimination generalization along very simple stimulus dimensions there are very few studies that bear directly on the question of separable learning processes. There is, however, some evidence in favor of this notion. Jones and McLaren (1999) trained human participants to categorize two intermediate levels of brightness of a green stimulus, and then tested across six stimuli covering a much wider range of intensities, including 'Near' stimuli, which were slightly brighter and dimmer than the trained stimuli, and 'Far' stimuli, which were much brighter and dimmer. Under normal conditions, participants produced monotonic gradients (accuracy highest for 'Far' stimuli) as would be expected from a rule-based analysis, and indeed reported using the correct rule in a post-experiment questionnaire. However, when the number of training trials was halved, or when the contingency between the training stimuli and the correct response was reduced, participants were generally unable to identify the appropriate rule, performed substantially worse and produced peak shifted gradients (accuracy highest for 'Near' stimuli). Jones and McLaren's attempts to restrict the opportunity to form an appropriate response strategy would thus appear to have yielded generalization gradients that are more in line with an associative analysis. The experiments reported below used a similar strategy, though as will be discussed, analysis of the results evolved towards testing a slightly different hypothesis.

Following a similar rationale to Jones and McLaren (1999), the following experiments examined the generalization gradient produced after a difficult discrimination, but with the important difference that this time the discrimination was between two very similar hues rather than two stimuli varying in brightness. In this respect, the experiments more closely follow the methodology used in pigeon experiments that require the birds to discriminate between different wavelengths of light. Thus, these hue stimuli were essentially a series of approximately isoluminant colors designed to capture the subjective properties of varying light wavelengths. Experiment 1 set out to establish the nature of the post-discrimination gradients produced when the initial discrimination was particularly difficult, sufficiently so to prevent many participants from verbally identifying the real difference between the stimuli. In doing so, it was hoped that the discrimination would be learned via the more gradual acquisition of associative strength to each discriminative stimulus, followed by associatively driven generalization during the test phase.

Experiment 1

The color values used for discriminative and transfer stimuli in Experiment 1 are shown in Table 1. Participants were initially trained to discriminate between two very similar shades of green, one slightly more yellow, the other slightly more blue. The colors were of medium saturation to make it difficult to distinguish between them when presented successively. The stimuli were presented as part of a simple two-choice categorization task, but were also interspersed with an unrelated filler task involving the categorization of abstract patterns. Thus the opportunity for relative comparison of the colors was

minimised. The intention was to reduce the opportunity for rule-formation during the training phase by removing the information necessary to confidently identify and apply such a rule.

The argument that verbally-mediated strategies will produce monotonic gradients extends only insofar as the verbal labels that govern such a strategy are actually relevant. For instance, the test range used in Experiment 1 terminates at hues corresponding to a medium saturation greenish yellow and greenish blue. The defining difference between the two green training hues was their relative yellowness or blueness. Hence so long as these relative characteristics are noticed and actively employed by the participant, they can be applied confidently to the entire test range. On the other hand if the terminal test values corresponded to hues more likely to be labelled orange and purple than yellow and blue then one would expect rule-mediated categorization to also yield a decline in accuracy at the extremes of the test range.

[Table 1 about here]

The test phase itself consisted of a series of randomized blocks, each containing presentations of every test stimulus. Each of these blocks contains considerably more information about the relevant stimulus dimension than the training stimuli. For this reason, further block-by-block analyses were conducted, comparing sequential sections of the test phase. These analyses were a precaution to check that exposure to the wider

range of test values had not altered responding. As will become evident, they turned out to be vital in interpreting the results of both experiments.

Method

Participants and Apparatus

Thirty-two undergraduate and graduate students from the University of Cambridge served as participants in the experiment. Each was paid £5 for their participation, and tested individually in a dimly lit room. All participants reported having normal color vision. The experiment was run on a Macintosh Power PC with a 15 inch color CRT monitor, and was programmed using REALbasic software.

Stimuli

Each hue stimulus consisted of a uniform colored square measuring 5.5 x 5.5 cm appearing individually in the centre of an otherwise black screen. A set of 15 such stimuli were created by modifying the Hue parameter in a set of color coordinates based on Hue, Saturation, and Brightness (see Table 1 for the color properties of each stimulus value). The Hue values were equally spaced and ranged from .1701 (a greenish yellow) to .5208 (a greenish blue) in approximately equal steps (adjusted slightly to fit the nearest exact Red, Green, and Blue coordinates for displaying on screen). The Hue parameter is designed to keep the Saturation (or vividness) and Value (or brightness) of the colors approximately equal, and these separate values were set at 0.5 and 0.75 respectively. The

colors were therefore considered to be sufficiently isoluminant when presented on screen. Of the 15 stimulus values, the 7th and 9th were used as training stimuli.

A concurrent task was used to prevent direct comparison of stimuli over successive trials. Hue stimuli were interleaved with complex patterns comprised of small abstract shapes. These filler stimuli were similar to those used as discriminative stimuli in other studies (e.g. Wills and Mackintosh, 1998; Livesey & McLaren, 2007). These filler stimuli differed in the identity and distributional frequency of their component shapes such that the relative number of each shape predicted the correct response in a probabilistic fashion (for a more detailed account of their construction, see Livesey & McLaren, 2007). Each stimulus consisted of 36 colored abstract shapes on a black background. The shapes were arranged in 6 x 6 array, to fit within the same area on screen as the hue stimuli. Each shape was made up of two primary colors and the relationship between any given shape and the correct response was randomised so that no single color predicted which response to make on the filler trials.

Procedure

On arrival, participants were given written and verbal instructions informing them about the nature of the task. A brief description of the hue and filler stimuli was given, along with instructions about the required responses, the feedback given after each response, and the time limit on each trial. Participants were told that the correct response depended entirely on the visual appearance of each stimulus, but were not given any information about the relationship between the hue stimuli other than the fact that they would be

uniform colored squares. They were informed that a test phase with no response feedback would be given at the end of the experiment. They were also informed that there was no relation between the hue and pattern stimuli.

Discrimination training: Training consisted of alternating presentations of hue stimuli and filler stimuli. Stimulus values 7 and 9 were assigned left and right key-press responses respectively for half the participants and right and left respectively for the other half. Trial order was randomized within blocks of 12 trials, containing three presentations each of stimulus values 7 and 9 and three each of the corresponding filler trials, with the condition that trials alternate between hue stimuli and filler stimuli. There were 8 blocks of 12 trials in total. These were presented as two blocks of 48 trials with participants instructed to take a short rest between the two blocks.

On each trial, participants were given 4 sec to respond to each stimulus by making either a left or right key response (pressing 'x' or '.' respectively) on a standard computer keyboard. Feedback after each response consisted of a 'correct' or 'WRONG' message appearing in the centre of the screen, followed by an inter-trial interval of 1.5 sec. Failure to respond in 4 sec resulted in the trial timing out and a 'no response' message appearing as feedback.

Transfer test: A total of 15 stimulus values were presented (color values shown in Table 1), with an equal number of interleaved filler trials. Trial order was randomized within blocks of 30 trials, containing one trial of each of the 15 transfer stimuli and 15 filler

stimuli. There were 6 blocks, or 180 trials in total, presented as two blocks of 90 trials, again with a short rest between the two blocks. No feedback was given during the test phase.

Analysis: For both Experiments, all analyses were conducted using repeated measures analysis of variance (ANOVA), except where otherwise indicated. Further to these analyses, the presence of a peak shift was verified using one-tailed paired-samples t-tests comparing peak accuracy with accuracy for training stimuli (S) and comparing peak accuracy with accuracy for the terminal stimuli at the extremes of the test range. These t-tests were one-tailed because the test is based on finding the peak shifted away from S (or the terminal test point), and then comparing it to performance at S (or the terminal test point), hence the difference cannot be expected to be in the other direction. The alpha level for statistical significance was $p < .05$ for all analyses.

Results

Accuracy across the training phase, divided into four successive blocks, rose from 45.8% in the first block, to 63.3%, 67.6%, and 66.9% in each of the subsequent blocks. Analyses over these points yielded a significant effect of block ($F_{3,93} = 13.646$, $p < .001$, $\eta^2 = .306$). Mean accuracy across the whole training phase was significantly above chance ($t_{31} = 3.152$, $p = .004$). While there is clearly some acquisition of the discrimination, mean accuracy was still well below the maximum by the end of training. In contrast, participants learned the filler task discrimination fairly quickly, with accuracy rising from 62.1% in the first block to 85.1% in the final block.

Following the same convention as similar previous studies (Jones & McLaren, 1999; Wills & Mackintosh, 1998), post-discrimination gradients were analysed by collapsing the full test range around S and measuring test accuracy at ordinal distances from S (as shown in Table 1). Thus positions 6 and 10 are one position removed away from S and towards the extremes of the test range, 5 and 11 are two positions removed from S, and so forth so that the terminal test values 1 and 15 are 6 positions removed from S. All statistical analyses were conducted on the collapsed gradients.

Response accuracy during the test phase, collapsed in this fashion, is shown in Figure 1. The rise in accuracy from the training stimuli to stimulus position 3 is very pronounced. There is also a slight decline in accuracy from position 3 to the extremes of the test range at position 6. Analyses across the test points revealed a significant main effect of stimulus position ($F_{6,186} = 2.83$, $p = .012$, $\eta^2 = .084$). Looking at the peak shift itself, t-tests revealed that there is a significant increase from S to 3 ($t_{31} = 3.411$, $p = .001$) but the decline to point 6 is not significant ($t_{31} = .831$, $p = .206$). Thus, averaged over the full test phase, there was a numerical peak shift but the decline at the extreme values of the test range is far from being significant and could plausibly be attributed to chance. As suggested earlier, this is certainly not an uncommon result.

[Figure 1 about here]

Of much greater interest, however, is the change in the form of the gradient from the initial blocks of the test phase to the later blocks. Two blocks of test stimuli were deemed the least number of trials over which meaningful accuracy averages could be made (thus for the collapsed gradient, a single test position constituted an average of four individual trials). Hence, initial analyses were conducted by breaking the test phase into three equal cycles (all block-by-block analyses refer to these 3 cycles). A repeated-measures ANOVA with test cycle (1-3) and stimulus position (0-6) as within-subjects factors yielded a significant main effect of stimulus position ($F_{6,186} = 2.966$, $p = .009$, $\eta^2 = .087$) and a marginally significant interaction between stimulus position and cycle ($F_{12,372} = 1.631$, $p = .081$, $\eta^2 = .050$). The main effect of cycle was not significant ($F_{2,62} = .374$, $p = .689$, $\eta^2 = .012$). Following these results separate analyses were conducted on each test cycle individually, as shown in Figure 2. Again, all statistical analyses were conducted on the collapsed gradient shown in Figure 2(i), although the full gradients across all test values are also provided in Figure 2(ii) for completeness.

Looking at Figure 2(i), the post-discrimination gradient over the first cycle of the test phase is more distinctly peak-shifted. For this first third of the test phase, the difference in accuracy between the training stimuli and the peak at position 3 was significant one-tailed ($t_{31} = 1.852$, $p = .037$), while the difference between position 3 and the terminal position 6 was also significant ($t_{31} = 2.156$, $p = .019$). Thus, over the first third of the test phase, there is a clear decline in accuracy towards the extremes of the test range and a pattern that is strongly suggestive of a peak shift effect. For this first test block, both sides of the uncollapsed gradient peak in accuracy at positions 3 steps removed from the

nearest training stimulus and both show an orderly rise in accuracy and subsequent fall towards the extremes of the test range, as can be seen in Figure 2(ii).

[Figure 2 about here]

The post-discrimination gradient over the remaining blocks of the test phase increasingly resembles a negatively accelerated monotonic curve. In the second test block, accuracy peaks at a position 2 removed from S. However, unlike the first test block, where accuracy gradually rose and fell on both sides of the gradient, here the peak is due to a dramatic rise in accuracy for a single test point on one side of the test range (stimulus value 5) as shown in Figure 2(ii). Aside from this single test point, the generalization gradient for the second test block is roughly monotonic except for a decline in accuracy at stimulus value 1. By the third test block, accuracy appears to be monotonically increasing away from S and there is certainly no evidence of a decline in accuracy towards the extremes of the test range. Accuracy at position 6 is higher or equivalent to accuracy at any other position along the gradient (accuracy at position 5 is numerically slightly higher but the difference does not approach significance; $t_{31} = .892$). In this final test block, accuracy was monotonic and increasing away from S on both sides of the test range.

Accuracy for S itself appears to drop from the first block of test trials to the second block and then recover in the final block. However, a restricted ANOVA looking solely at accuracy for S across the 3 blocks did not find a significant effect of test cycle ($F_{2,62} =$

1.334, p = .271, η^2 = .041) and only a marginally significant quadratic trend ($F_{1,31}$ = 3.088, p = .089, η^2 = .091). Aside from this, the most notable change across the three cycles of the test phase is the change in accuracy from positions 3 onwards. In order to confirm that these differences were reliable, cycle by cycle analyses over the four most extreme test positions were also conducted. Over these points, the most informative analysis is the interaction between linear trends using an ANOVA with stimulus position (3-6) and test cycle (1-3) as within-subjects factors. If a change in generalization really did result in a progressively more monotonic gradient, then the change in accuracy from 3 to 6 should progressively change from a decline to an increment. This stimulus position by cycle interaction as a linear trend was significant ($F_{1,31}$ = 4.07, p = .05, η^2 = .116), suggesting the relationship between accuracy at these test points changed across the course of testing, as the generalization gradient became more monotonic.

Discussion

The results from Experiment 1 might be considered the exception to the norm for two-choice categorization experiments. The post-discrimination generalization gradient appeared to be peak-shifted to begin with, but then became monotonic through the course of the test phase. One explanation for this effect is that some associative control is acquired in the training phase, but as the relationship between the stimuli becomes verbally identifiable through exposure to the full test range (during which the participant notices that the stimuli vary in hue), participants change their response strategy. The distinct rise in accuracy at the extremes of the test range is consistent with this idea. So too is the apparent drop in accuracy for the trained stimuli, to chance level accuracy, and

then its recovery by then end of the test phase, as discrimination shifts from a basis in the absolute stimulus properties of the training stimuli to a relative judgement of hue using a decision boundary centred approximately on the properties of the training stimuli. Such an argument assumes that the associative basis for responding is relatively weak, but sufficient to enable participants to use an otherwise untestable strategy (since the test is given with no feedback). This account is plausible, provided that verbally mediated processes are extremely flexible and can be voluntarily engaged without any genuine form of reinforcement or feedback. In humans, there is little evidence to suggest that this is not the case.

Alternatively, one might assume that the discrimination has been governed by rules from the outset but that the response rule is not immediately applied by the participant to stimuli at the extremes of the test range because they are particularly novel. By this account, one might assume that a decision bound is established in training and participants learn to use this within a given context. Stimuli further away from the decision bound are typically easier to classify, except that its use is confined to a given set of stimuli. Performance at the extremes of the test range is initially poor because of a failure to engage the correct rule, given that these stimuli are quite dissimilar to the training stimuli. However, given that the test stimuli were deliberately chosen to fall within a classifiable range under the most relevant rule, it is not clear why the rule should be particularly difficult to engage for the more extreme stimuli, despite their novelty. A case could certainly be made for novelty slowing the decision process down (perhaps because participants become more cautious under novelty) but it is not so easy to account

for the poor accuracy to these stimuli, given the fairly relaxed time constraints in this task.

Yet another explanation might be that generalization continues to be based solely on associative processes but that learning continues through the test phase. At least two possible accounts could be given under this interpretation. The first is that associative strengths continue to change during the test phase, which is in effect a period of extinction. Whilst this fits well with the drop in accuracy for S over test blocks, it seems counter-intuitive that the associative strengths of the more extreme test stimuli should actually increase in the absence of reinforcement. Perhaps a more plausible argument is that changes in either dimensional attention or stimulus associability might occur during the test phase, and that these may have some effect on performance. Mackintosh (1975) left open the possibility that stimulus associability might affect performance as well as learning, an idea for which Le Pelley, Suret, & Beesley (in press) have recently provided supportive evidence, while Kruschke (1992) assumed that attentional changes would indeed affect performance as well as learning. However, one might justifiably expect the opposite result. During the test phase, variations in the relevant dimension are continually non-reinforced, suggesting that units capturing variations along this dimension should decrease in associability.

Unlike these alternative accounts, the dual-process explanation implies that the change in generalization gradient should coincide with or immediately follow verbal identification of the relevant stimulus dimension. Thus participants will only produce the transition

during the test phase if they fail to correctly identify the difference between the training stimuli but nevertheless gradually learn to distinguish between them associatively. The relevant stimulus dimension is obvious in the test phase and would presumably encourage participants who had not devised an alternate strategy of which they were particularly confident to change their pattern of responding. In relation to these assumptions, it would be useful to determine whether the participants who most strongly demonstrated the effect were those that had failed to identify the correct stimulus relationship during the training phase. Unfortunately this cannot be achieved with any degree of precision for the results of Experiment 1 - while participants were given a very general post-experiment questionnaire, there were no specific questions addressing the relationship between the hues in the training phase.

Excluding Participants

It should be noted that no participants were excluded from Experiment 1 even though several did not display any evidence of acquiring the discrimination during training. We argued that the procedure used in this experiment necessitated a low learning rate to prevent immediate rule learning. Thus the discrimination in question has deliberately been made very difficult and acquisition of associative control would be expected to be relatively slow. Therefore, using an exclusion criterion may result in the exclusion of the participants who are most likely to show the effect of interest. Nevertheless, we can see that an argument can be made for using an exclusion criterion – no prediction about generalization is relevant to participants who have learned nothing during training. Even

if such a participant was to guess a response strategy during the test phase, the probability of doing so correctly should be chance.

[Figure 3 about here]

For similar experiments in the past, we have typically excluded participants who fail to meet a criterion of at least 55% correct over the last half of training, as having failed to learn the discrimination (e.g. Livesey & McLaren, 2007). Applying this criterion retrospectively to the data from Experiment 1 rules out 13 of the 32 participants. Although the trends of most interest do not reach statistical significance for the remaining sample of 19 participants, the pattern of results in this subset is still consistent with the overall findings, as shown in Figure 3, and the performance level is much higher. Accordingly, we resolved to use this exclusion criterion for Experiment 2.

Experiment 2

The notion that verbal identification of the stimulus dimension was important for changes in generalization was examined in Experiment 2 by giving a more structured questionnaire that asked specifically about the training stimuli and the difference between them. Participants were given a similar, very difficult discrimination and a post-discrimination generalization test across the relevant dimension. They then completed a post-experiment questionnaire, which asked specifically "Did you notice a difference between the colored square stimuli that belonged to different categories in the training phase? If so, describe the difference between them." By this stage the relevant test

dimension had been made very obvious. Any vague identification of the real stimulus relationship, even one of which the participant was not particularly confident, would have been confirmed in the test phase. Any report of not noticing a difference between the training stimuli, or describing an incorrect or irrelevant relationship between them, could therefore be taken as a genuine failure to identify the difference rather than one of which the participant had not been confident enough to report. If anything, this method of classifying participants may overestimate awareness of the difference between the training stimuli because the relevant stimulus dimension becomes very obvious during the test phase, before the post-experiment questionnaire was administered. Nevertheless, it was necessary to administer the questionnaire last so to avoid influencing the categorization data.

For Experiment 2, the exclusion criterion discussed previously was used from the outset. Participants who failed to perform above 55% accuracy over the second half of training were excluded from analysis. Analysis on the remaining participants was conducted in a similar fashion to Experiment 1. Data from the test phase was divided into equal thirds (referred to as blocks 1-3) to assay changes in post-discrimination generalization. Participants were classified according to whether they had accurately reported the difference between the training stimuli. Identification was classified as correct if any reference was made to there being more or less of the colors green, blue or yellow. Conversely, identification was considered incorrect if only an irrelevant difference was reported or if no difference was reported.

Different training stimuli and a different range of test colors were used in Experiment 2. The colors used in Experiment 1 crossed over two major color category boundaries as the test stimuli ranged from yellow to blue. The training stimuli were quite close to what some observers perceive to be relatively "pure" green. Judgements of the point at which the stimulus dimension ceases to be, say, more green and becomes more yellow are rather subjective and change noticeably from one observer to the next. Nonetheless, for most observers these changes occur at intermediate test values either side of the trained stimuli. This may have unexpected effects on the way these colors are judged and categorized. Thus to show the result is not restricted to this situation, a narrower (but also more saturated and therefore more vivid and discriminable) set of test values were used that crossed at most only one major color boundary.

Experiment 2 was run in two replications, the first using group testing, the second using individual testing. As will be discussed below, the results of the two replications concurred very closely.

Method

Experiment 2A

Participants and Apparatus

Thirty students from the University of Cambridge participated in the experiment as part of a practical class. The experiment was run on Macintosh iMac computers with 15 inch color LCD displays. Presentation of stimuli and measurement of responses was

programmed using REALbasic software. Testing was conducted in a large classroom, with participants spaced approximately 1m apart and approximately 0.5 m from the displays.

Stimuli

The hue-based stimuli were constructed as in Experiment 1, except the RGB values for each of the training and test stimuli were given a constant Green and Red input (191, and 51 respectively) and a Blue input that changed in successive steps across the dimension from 51 to 191 (which also correspond to successive steps in Hue, as shown in Table 1). Eight stimuli were constructed across this range, though the two central stimulus values 4 and 5, used as the training stimuli, were made more similar (blue components of 116 and 126 respectively). Color components for the training and test stimuli are reported in Table 1. These values appear on screen as colors ranging from a fairly pure green to a greenish blue, or cyan. The stimuli were kept constant in terms of predefined saturation and value. More subjective differences in brightness or vividness are difficult to control for but within this range, they were judged to be roughly constant. This time, all test stimuli fell within one major color transition, from green to blue, rather than straddling transitions from yellow to green to blue.

Procedure

Participants were given similar written instructions to those in Experiment 1, outlining the main procedural points and task requirements, but not revealing the nature of the

difference between the colors. Additionally, participants were explicitly informed that there was a real, physical and consistent difference between the training stimuli, even though it might not be immediately obvious to them.

Discrimination training again consisted of alternating presentations of hue-based stimuli and filler stimuli. Stimulus values 4 and 5 were assigned left and right responses respectively for half the participants and right and left respectively for the other half. Trial order was randomized within blocks of 12 trials, containing three presentations each of stimulus values 4 and 5 and three each of the corresponding filler trials, with the condition that trials alternate between hue stimuli and filler stimuli. This time, there were 12 blocks of 12 training trials, as pilot testing suggested 8 blocks was insufficient for learning the slightly more difficult discrimination. These were presented as two blocks of 72 trials with participants instructed to take a short rest between the two blocks.

Transfer test: The stimulus values shown in Table 1 were presented, with an equal number of interleaved filler trials. Trial order was randomized within blocks containing one of each of the hue-based stimuli and corresponding filler stimuli. Again, there were 6 presentations of each test stimulus in total, though for the purpose of analyses these were divided into three equal and successive blocks.

Questionnaire: A post-experiment questionnaire was given at the completion of the test phase. Questions were asked specifically about both the hue-based and filler stimuli

(these questions were clearly delineated to prevent participants confusing one with the other). The questions of most relevance were:

"Did you notice a difference between the colored square stimuli that belonged to different categories in the training phase? If so, describe the difference between them."

"What strategies did you use to make your judgements about the colored square stimuli in the training phase and the test phase?"

Not surprisingly, nearly every participant reported using a relevant color-based strategy in the test phase (as a partial answer to the second of these questions). The first question was used to divide participants according to 'reported training difference' for some of the analyses to follow.

Experiment 2B

Participant, Apparatus and Stimuli

Seventy-two students at the University of Cambridge participated in the experiment, each tested individually in a dimly lit room. The experiment was run using a Power PC Macintosh and 17 inch display positioned approximately 0.5 m from participants. The stimuli were constructed in the same way as those in Experiment 2A.

Procedure

The procedure for Experiment 2B was identical to 2A except that participants were tested individually, and were given verbal instructions that reiterated the main points of the written instructions.

Results and Discussion

Eight participants from Experiment 2A and 20 participants from Experiment 2B did not meet the criterion of 55% correct in the second half of training, and were excluded. All analyses reported below are for the remaining 74 participants (n = 22 and 52 for replications A and B respectively). Of these 74 participants, 42 (11 and 31 from replications A and B respectively) accurately reported the difference between the training stimuli, and 32 (11 and 21 respectively) reported an incorrect or no difference between them. The proportions of participants reporting the difference accurately did not significantly differ between replication (χ^2, 1df = 0.582, p > .4).

[Figure 4 about here]

Figure 4 shows acquisition over the course of training for the two replications, and split according to reported training difference (correct report vs incorrect report or no report). Overall, accuracy rose from 52.6% during the first block of 12 trials to 76.0% in the final block of trials. In comparison, accuracy for the filler task rose from 56.6% in the first block to 83.8% in the final block. Again it can be seen that the hue-based discrimination was learnable but was generally quite difficult. Mean accuracy in training, analysed using a repeated-measures ANOVA with replication and reported training difference as between-subjects factors, yielded significant effects of both reported training difference ($F_{1,70}$ = 8.934, p = .004, η^2 = .112) and replication ($F_{1,70}$ = 4.530, p = .037, η^2 = .063).

There was no interaction ($F_{1,70}$ = .136). On the whole, participants who accurately reported the difference between the training stimuli performed better during training, and participants from replication A performed better than those from replication B.

[Figure 5 about here]

Figure 5 shows the collapsed and full post-discrimination gradients for each replication of Experiment 2. A repeated-measures ANOVA of test accuracy over the collapsed test positions, with stimulus position (1-4) as a within-subjects factor and reported training difference and replication as between-subjects factors, yielded highly significant main effects of stimulus position and reported training difference, and an interaction between stimulus position and reported training difference (lowest $F_{3,210}$ = 5.963, p = .001, η^2 = .078). Neither the main effect for replication nor any interaction with it was significant (highest $F_{3,210}$ = .354).

Due to the lack of any obvious differences between the post-discrimination results of the two replications, analyses were then collapsed across replication. Results from the test phase were divided into three successive blocks. Analyses were conducted on all participants with reported training difference as a between-subjects factor, and additional analyses on the correct difference and incorrect/no difference groups separately. The data for all participants divided into three blocks can be seen in Figure 6. Results broken down by reported training difference (either correct or incorrect/none) are shown in Figure 7.

[Figure 6 about here]

Examining Figure 6(i), there is once again a numerical peak shift effect in the first test block, which progressively became more monotonic throughout the test phase. Examining the first test block, repeated measures contrasts were initially used (rather than one-tailed t-tests) to examine whether differences between S and the peak, and between the peak and terminal test position, interacted with reported training difference. These analyses confirmed the difference between positions 2 and 3 was significant ($F_{1,72}$ = 5.674, p = .02, η^2 = .073) as was the interaction with reported training difference ($F_{1,72}$ = 9.854, p = .002, η^2 = .120). The difference between S and position 2 was also highly significant ($F_{1,72}$ = 45.814, p < .001, η^2 = .389), though the interaction with reported training difference was not significant ($F_{1,72}$ = .995, p = .332, η^2 = .014). This indicates a peak shift effect in the first block of the test phase, as was found in Experiment 1.

As is evident in Figure 6, the peak shift effect disappears progressively across blocks of the test and, looking at Figure 7, nearly all of this change can be attributed to the incorrect/no difference participants. A restricted ANOVA was used to examine changes at the extremes of the test range, with stimulus position (2 vs 3) and test block (1-3) as within-subjects factors and reported training difference as a between-subjects factor. Within this analysis, an interaction between linear trends in test block and stimulus position identifies systematic changes in the difference between test positions 2 and 3 across the test phase. This linear trend interaction was highly significant ($F_{1,72}$ = 18.704,

$p < .001$, $\eta^2 = .206$). Importantly, the three-way interaction between this linear trend interaction and reported training difference was also significant ($F_{1,72} = 9.478$, $p = .003$, $\eta^2 = .116$), indicating that the gradient for the incorrect/no difference participants changed more dramatically across the test blocks.

[Figure 7 about here]

Looking solely at the incorrect/no difference participants, the collapsed gradient over the full test phase was actually numerically peak shifted. But while the difference between S and position 2 was significant ($t_{31} = 3.519$, $p < .001$), the decline from position 2 to position 3 was not ($t_{31} = 1.039$, $p = .154$). More importantly however, in the first block of the test phase there was a significant difference between positions 2 and 3 ($t_{31} = 2.669$, $p = .006$) and between positions 0 and 2 ($t_{31} = 3.717$, $p < .001$), indicating a peak shift effect. As can be seen in Figure 7(i), the significant decline in accuracy at the extremes disappears over blocks 2 and 3. There was a significant linear trend in the difference in accuracy between stimulus positions 2 and 3 across the blocks of the test phase ($F_{1,31} = 14.716$, $p = .001$, $\eta^2 = .322$). The trends in response accuracy for participants who did not correctly report the difference between the training stimuli suggest a significant shift from declining at the extremes to being monotonic.

Aside from performing at much higher accuracy, the correct difference participants also show a subtly different set of gradients, as predicted. There is no evidence of peak shift in the first or any block of the test phase, as the collapsed gradients were monotonic

across all blocks (gradients across the full dimension were at least roughly monotonic across all blocks, although there was a slight and non-significant decrease at one end in the first two blocks). Similar comparisons of the two most extreme positions (2 and 3 steps from S) across the blocks did not yield a significant linear trend ($F_{1,41} = 1.741$, $p = .194$, $\eta^2 = .041$).

In Experiment 2, there is a clear relationship between whether a participant accurately reported noticing the relationship between the training stimuli, and the pattern of results they produce on test. Those that did not report the training difference accurately, showed a clear peak shift effect in the initial stages of training, which became progressively more monotonic throughout the test phase. Those that accurately reported the training difference yielded monotonic gradients throughout the test phase. Thus the pattern of results effectively agreed very closely to that predicted on the basis of a combination of associative learning and simple verbal rules. In other words, the results are completely consistent with a dual process account, under which discrimination learning proceeds via associative mechanisms but, at the same time, participants may intentionally employ strategies to govern response selection.

General Discussion

These experiments demonstrate that a genuine peak shift effect is attainable in simple category learning with human participants using stimuli that vary more in hue than other qualities such as brightness. As such, it may be the first demonstration of its kind, as the

majority of similar studies using two-choice categorization have produced monotonic post-discrimination gradients. In fact, other techniques for assaying post-discrimination generalization, most notably absolute identification procedures (i.e. using same/different judgements) have, on the whole, also failed to provide unequivocal demonstrations of peak shift on simple stimulus dimensions, at least in the absence of strongly biased test ranges (see, for instance, Thomas, 1993). The results of Experiments 1 and 2 are therefore particularly interesting in that they yielded true peak shift effects in the initial stages of testing that cannot readily be attributed to changes in adaptation level (Thomas, 1993) or other explanations associated with biased test ranges. The initial discrimination was sufficiently difficult to prevent many participants from accurately identifying the relevant stimulus dimension. On test, participants initially responded most accurately to stimuli similar to, but removed along the dimension from the training stimuli, with a significant rise and subsequent significant fall in accuracy moving from the training stimuli to the extremes of the test range. This result is consistent with an associative analysis, where generalization is based on physical similarity to the two trained stimuli. There is little reason to assume that rule-based responding would yield anything other than monotonic generalization gradients across these test ranges.

While most human studies using two-choice categorization along a simple stimulus dimension have not yielded results consistent with the animal literature on discrimination learning, there remains a very plausible argument that common associative learning mechanisms occur in both, which are merely masked in human categorization by the use of simple verbal strategies. The results of Experiments 1 and 2 are certainly consistent

with this argument as they reveal a transient peak shift effect in the initial stages of testing. Therefore, under certain circumstances, there are clearer similarities between post-discrimination generalization in humans and other animals even along relatively simple stimulus dimensions. Taken together with the gradients produced using more complex stimulus dimensions, these data suggest an important role for associative processes in human discrimination learning, and are consistent with associative models of dimensional discrimination and generalization. Experiment 2 yielded a clear concordance between verbal report of the relational properties of the training stimuli and monotonic generalization. This suggests that differences in the generalization gradients observed in human categorization and animal discrimination learning are not just procedural in origin.

The approach to dimensional discrimination taken by several associative models (Blough, 1975; Ghirlanda & Enquist, 1998; McLaren & Mackintosh, 2000, 2002) is particularly amenable to explaining peak shift and similar post-discrimination phenomena. If one were to examine each gradient reported here independently, then these models might in principle account for all of the results, including those which were monotonic. But to do so, one would have to make the post hoc assumption that there were much broader underlying stimulus representations for the trials where monotonic generalization was observed than for the trials where peak-shift was observed. However, the within-subjects nature of these gradient changes makes this explanation implausible, as we have no basis for assuming that stimulus representations for a single participant should change in a manner that would predict the changes in generalization observed here. Instead, we

argue that the monotonic post-discrimination gradients observed by the end of the test phase and the results of similar studies in the past, are more a result of rule-governed and verbally mediated responding than of associative processes. In support of this claim, participants in Experiment 2 who could accurately characterise the difference between the training stimuli in terms that would be useful for developing an appropriate response strategy did not show the peak shift effect, but rather produced monotonic generalization gradients from the outset of the test phase. The change in gradient form, although significant over all participants, was entirely due to those who had failed to accurately identify the relationship between the training stimuli. There was thus a strong relationship between showing an initial peak shift effect and whether or not the information necessary to form an appropriate cognitive strategy had been identified.

This concordance also adds some weight to the notion that stimulus *relations* are more "conceptual" than the physical properties of a single stimulus, and are reified only by abstract thought. Although this is assumed to some degree in previous literature (e.g. Jones & McLaren, 1999; Wills & Mackintosh, 1998), one could argue that there is nothing particularly conceptual about "more lightness" or "more blue" when such relations control test categorization and that they could be learned in exactly the same way as other stimulus properties. But in this case, such relations only appear to govern categorization when participants become fully aware of the relationships between the stimuli, whereas the specific shades used as training stimuli and their categorical outcomes could be learned without any cogent relationship or difference being identified.

If one accepts this argument, then it is clear that verbal rule formation is a very flexible process. A simple relationship between the stimuli becomes more obvious in the test phase, but its relevance to the earlier discrimination is never substantively verified – the transitions seen here occur without feedback, and seem to be a product of the participant's own untested hypotheses, possibly supported by learning that has occurred via associative mechanisms. When a participant begins using a rule based on color in the absence of feedback, they might be equally likely to get the rule correct (e.g. "respond left for bluer stimuli") or incorrect ("respond right for bluer stimuli"). However, accuracy at the end of the test phase was clearly above chance even for those who reported not noticing the color difference during training. This suggests that whatever information is learned about the physical characteristics during training is available when the rule is formed. This might be simply due to the participant observing their own decisions and responses (e.g. "I have been tending to choose left when the square is a little bluer, so maybe blue means respond left").

While the peak shift effects observed here and in previous experiments highlight the similarities between discrimination learning processes in humans and other animals under very specific conditions, the changes in behavior observed in Experiments 1 and 2 also suggest that human performance on discrimination learning tasks can operate via disparate mechanisms. Associative learning is often assumed to be governed by automatic processes, which can be reliably modelled with formal algorithms and, at least in principle, imply that learned behavior should follow a predictable course. Strategic cognitive processes, including verbal reasoning and rule abstraction, may not be

characterised so easily. In the case of post-discrimination generalization, such processes do not appear to require feedback or specific instruction in order to be engaged as the basis for stimulus classification. Associative processes presumably, therefore, play a sub-ordinate role in discriminative control in humans, though potentially one of immense importance.

Conclusion

The literature on peak shift and other post-discrimination generalization effects in humans is, for the most part, either consistent with numerous interpretations or not easily reconciled with an associative analysis at all. However, rather than concluding that associative processes do not occur in human discrimination learning, it may well be the case that cognitive strategies, verbal mediation, and relative stimulus comparison readily obscure the behavioral consequences of more fundamental learning mechanisms. Under conditions where the relationship between the discriminative stimuli is obvious and easily characterized, the development of simple response strategies seems to govern generalization to related stimuli. However, when such a strategy is not so easy to devise, evidence emerges of learning and generalization based on the physical similarity of the stimuli in accord with more primitive associative processes. These results are consistent with the claim that very simple rule-governed generalization can mask other discriminative processes. They add significant weight to this argument, which has been directly supported by only a handful of previous studies (e.g. Aitken, 1996; Jones and McLaren, 1999) within the current paradigm.

References

Aitken, M. R. F. (1996). *Peak shift in pigeon and human categorisation.* Unpublished doctoral thesis, University of Cambridge, Cambridge, UK.

Blough, D. S. (1973). Two-way generalization peak shift after two-key training in the pigeon. *Animal Learning & Behavior, 1*, 171-174.

Blough, D. S. (1975). Steady state data and a quantitative model of operant generalization and discrimination. *Journal of Experimental Psychology: Animal Behavior Processes, 1*, 3-21.

Capehart, J., & Pease, V. (1968). An application of adaptation-level theory to transposition responses in a conditional discrimination. *Psychonomic Science, 10*, 147-148.

Cross, D. V., & Lane, H. L. (1962). On the discriminative control of concurrent responses: The relations among response frequency, latency, and topography in auditory generalization. *Journal of Experimental Analysis of Behaviour, 5*, 487-496.

Enquist, M., & Ghirlanda, S. (2005). *Neural Networks and Animal Behavior*. Oxford: Princeton University Press.

Ghirlanda, S., & Enquist, M. (1998). Artificial neural networks as models of stimulus control. *Animal Behaviour, 56*, 1383-1389.

Hanson, H. M. (1957). Discrimination Training Effect on Stimulus Generalization Gradient for Spectrum Stimuli. *Science, 125*, 888-889.

Hebert, J. A. (1970). Context effects in the generalization of a successive discrimination in human subjects. *Canadian Journal of Psychology, 24*, 271-275.

Jones, F., & McLaren, I. P. L. (1999). Rules and associations. In *Proceedings of the Twenty-First Annual Conference of the Cognitive Science Society*. Mahwah, NJ: Erlbaum.

Kruschke, J. R. (1992). ALCOVE: An Exemplar-Based Connectionist Model of Category Learning. *Psychological Review, 99*, 22-44.

LaBerge, D. (1961). Generalization gradients in a discrimination situation. *Journal of Experimental Psychology, 62*, 88-94.

Le Pelley, M. E., Suret, M., & Beesley, T. (in press). Learned predictiveness effects in humans: A function of learning, performance or both? *Journal of Experimental Psychology: Animal Behavior Processes*.

Livesey, E. J., & McLaren, I. P. L. (2007). Elemental Associability Changes in Human Discrimination Learning. *Journal of Experimental Psychology: Animal Behavior Processes, 33*, 148-159.

Mackintosh, N. J. (1975). A theory of attention: Variations in the associability of stimuli with reinforcement. *Psychological Review, 82*, 276-298.

Mackintosh, N. J. (1997). Has the wheel turned full circle? Fifty years of learning theory, 1946-1996. *Quarterly Journal of Experimental Psychology Section a-Human Experimental Psychology, 50*, 879-898.

Mackintosh, N. J. (2000). Abstraction and Discrimination. In C. Heyes & L. Huber (Eds.), *The Evolution of Cognition* (pp. 123-142). Cambridge, MA: MIT Press.

McLaren, I. P. L., Green, R. E. A., & Mackintosh, N. J. (1994). Animal learning and the explicit/implicit distinction: Or why what we think of as explicit for us can be implicit for them. In N. Ellis (Ed.), *Implicit and Explicit Learning of Languages* (pp. 313-332). New York: Academic Press.

McLaren, I. P. L., & Mackintosh, N. J. (2000). An elemental model of associative learning: I. Latent inhibition and perceptual learning. *Animal Learning & Behavior, 28*, 211-246.

McLaren, I. P. L., & Mackintosh, N. J. (2002). Associative learning and elemental representation: II. Generalization and discrimination. *Animal Learning & Behavior, 30*, 177-200.

Spence, K. W. (1937). The differential response in animals to stimuli varying within a single dimension. *Psychological Review*, *44*, 430-444.

Thomas, D. R. (1993). A model for adaptation-level effects on stimulus generalization. *Psychological Review, 100*, 658-673.

Thomas, D. R., Lusky, M., & Morrison, S. (1992). A Comparison of Generalization Functions and Frame of Reference Effects in Different Training Paradigms. *Perception & Psychophysics, 51*, 529-540.

Wills, S., & Mackintosh, N. J. (1998). Peak shift on an artificial dimension. *Quarterly Journal of Experimental Psychology Section B- Comparative and Physiological Psychology, 51*, 1-32.

Table 1. Color properties of the stimulus values used in Experiments 1 and 2. Red, Green, and Blue (RGB) values represent color proportions out of 255 on a standard 24bpp computer display. Hue is represented a continuous variable from 0 to 1.

Experiment 1

	Stimulus Value														
	1	2	3	4	5	6	**7**	8	**9**	10	11	12	13	14	15
	6	5	4	3	2	1	**S_A**		**S_B**	1	2	3	4	5	6
Hue	.170	.196	.221	.245	.271	.295	**.319**	.346	**.370**	.396	.420	.444	.471	.495	.521
R	189	174	160	146	131	117	**103**	95	**95**	95	95	95	95	95	95
G	191	191	191	191	191	191	**191**	191	**191**	191	191	191	191	191	179
B	95	95	95	95	95	95	**95**	102	**116**	131	145	159	174	188	191

Experiment 2

	Stimulus Value							
	1	2	3	**4**	**5**	6	7	8
	3	2	1	**S_A**	**S_B**	1	2	3
Hue	.333	.357	.381	**.411**	**.423**	.452	.476	.500
R	51	51	51	**51**	**51**	51	51	51
G	191	191	191	**191**	**191**	191	191	191
B	51	71	91	**116**	**126**	151	171	191

Figure 1

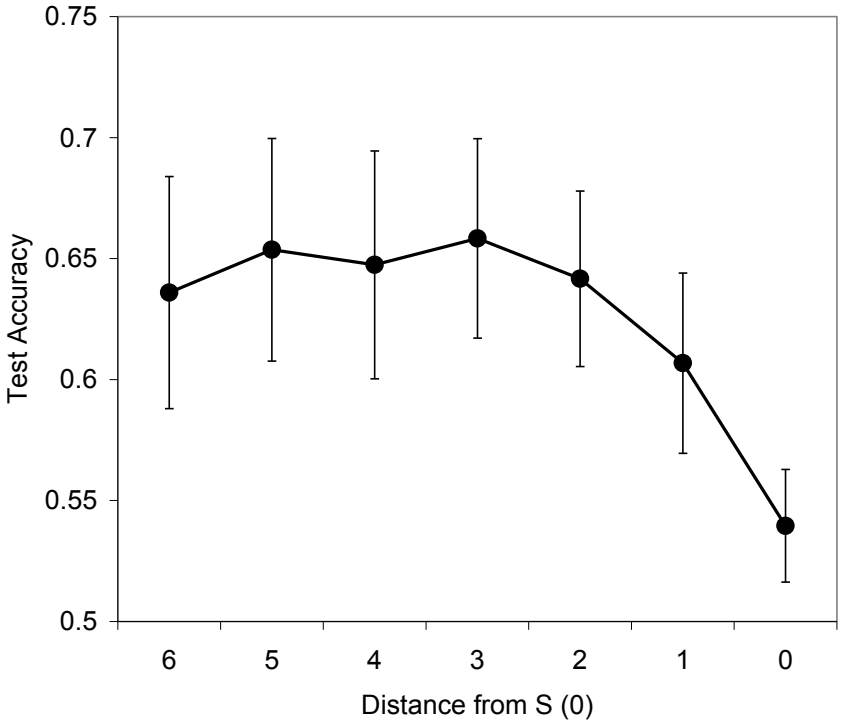

Figure 1. Mean test accuracy ± SEM for Experiment 1 plotted as a function of distance from the nearest training stimulus (S).

Figure 2

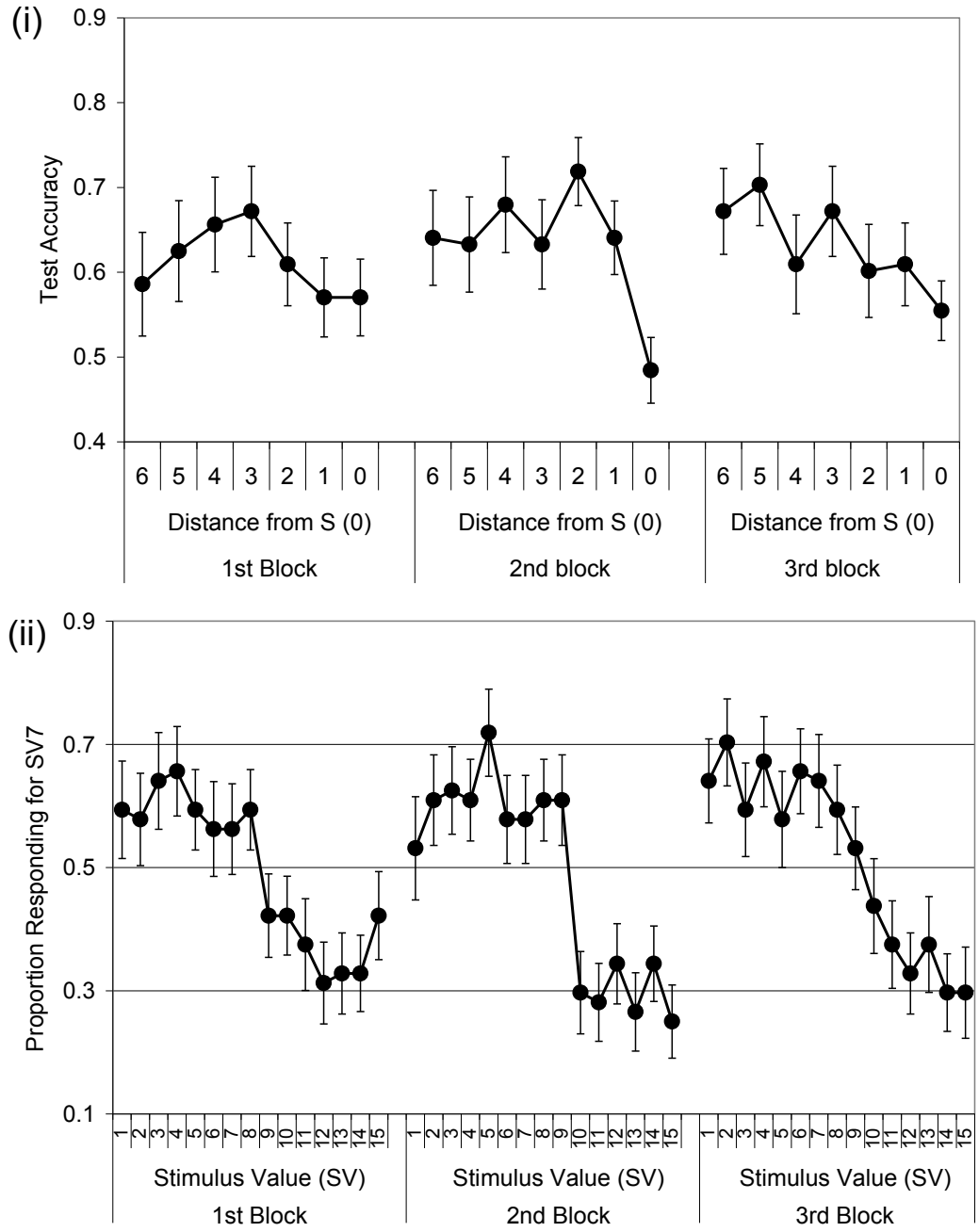

Figure 2. Post-discrimination gradients from Experiment 1 averaged separately over successive blocks of the test phase. Data are expressed as (i) mean test accuracy ± SEM plotted as a function of distance from the nearest training stimulus (S), and (ii) the proportions of responses that were appropriate for stimulus value 7 ± SEM plotted over the uncollapsed test range.

Figure 3

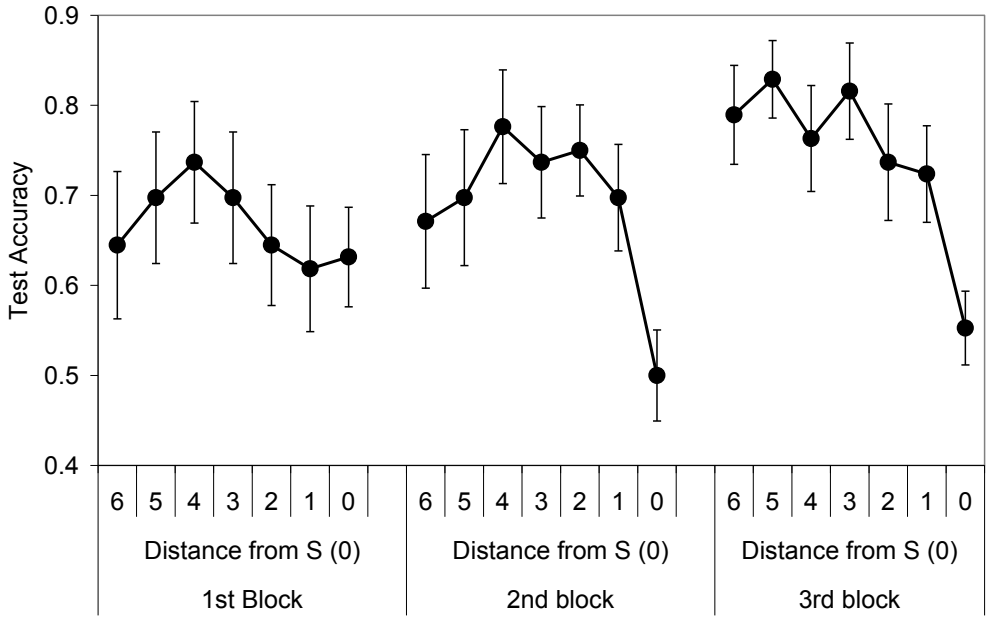

Figure 3. Test accuracy ± SEM plotted as a function of distance from S, over successive thirds of the test phase, for the 19 participants who performed above 55% accuracy over the second half of training (i.e. the exclusion criterion used in Experiment 2).

Figure 4

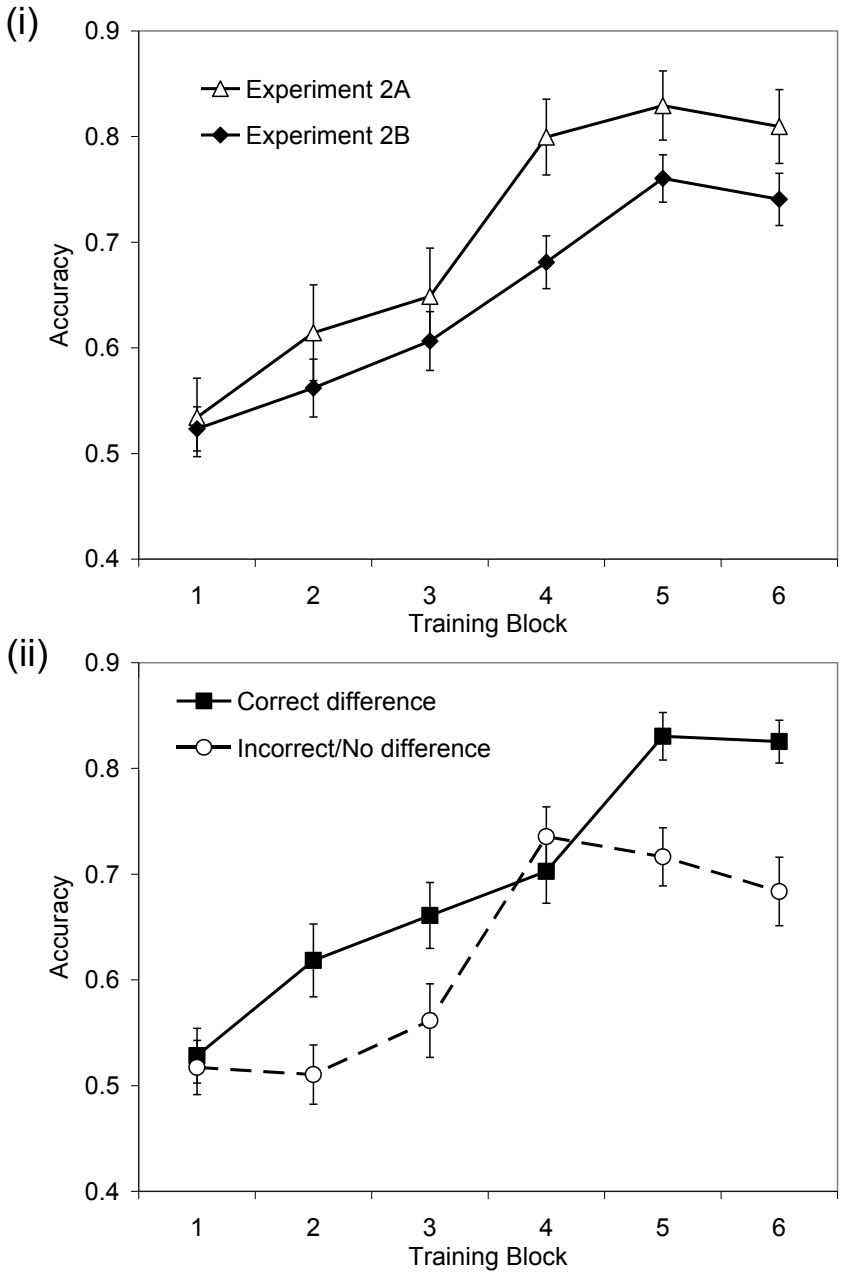

Figure 4. Percentage correct responses ± SEM during discrimination training. Data points indicate accuracy for successive blocks of 12 hue-based discrimination trials, split according to (i) replication (Experiments 2A and 2B), and (ii) reported training difference (correct report vs incorrect or no report).

Figure 5

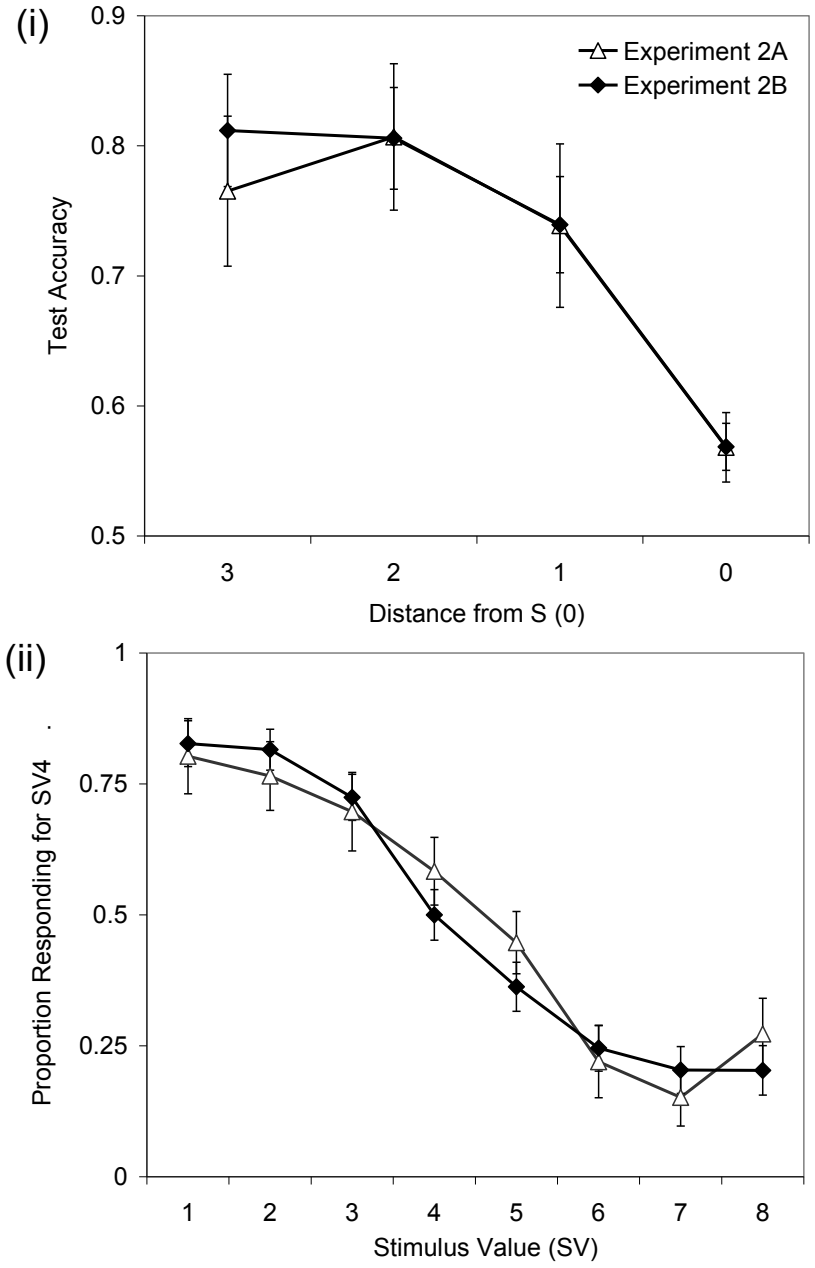

Figure 5. Post-discrimination gradients from the two replications of Experiment 2. Data are expressed as (i) mean test accuracy ± SEM plotted as a function of distance from the nearest training stimulus (S), and (ii) the proportions of responses that were appropriate for stimulus value 4 ± SEM plotted over the uncollapsed test range.

Figure 6

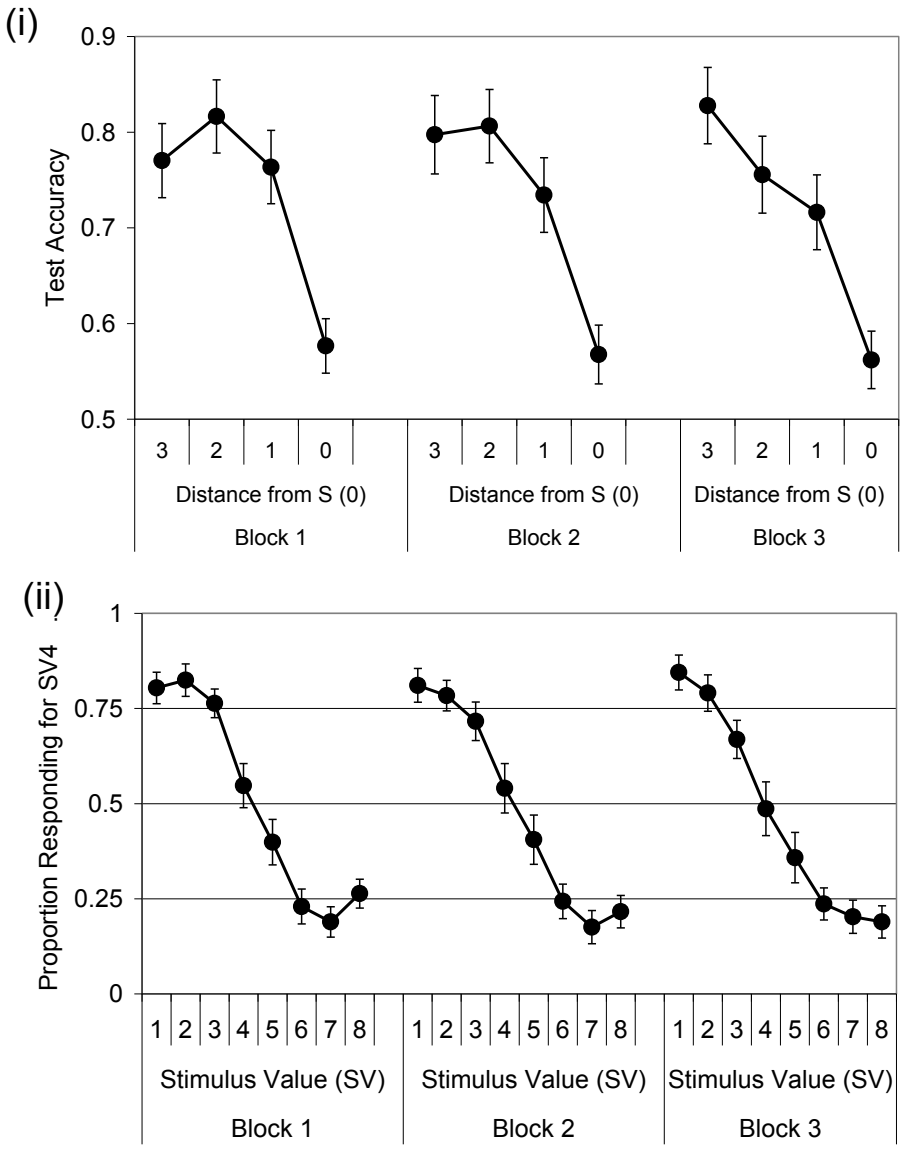

Figure 6. Post-discrimination gradients from Experiment 2 averaged separately over successive thirds (blocks 1-3) of the test phase. Data are expressed as (i) mean test accuracy ± SEM plotted as a function of distance from the nearest training stimulus (S), and (ii) the proportions of responses that were appropriate for stimulus value 4 ± SEM plotted over the uncollapsed test range.

Figure 7

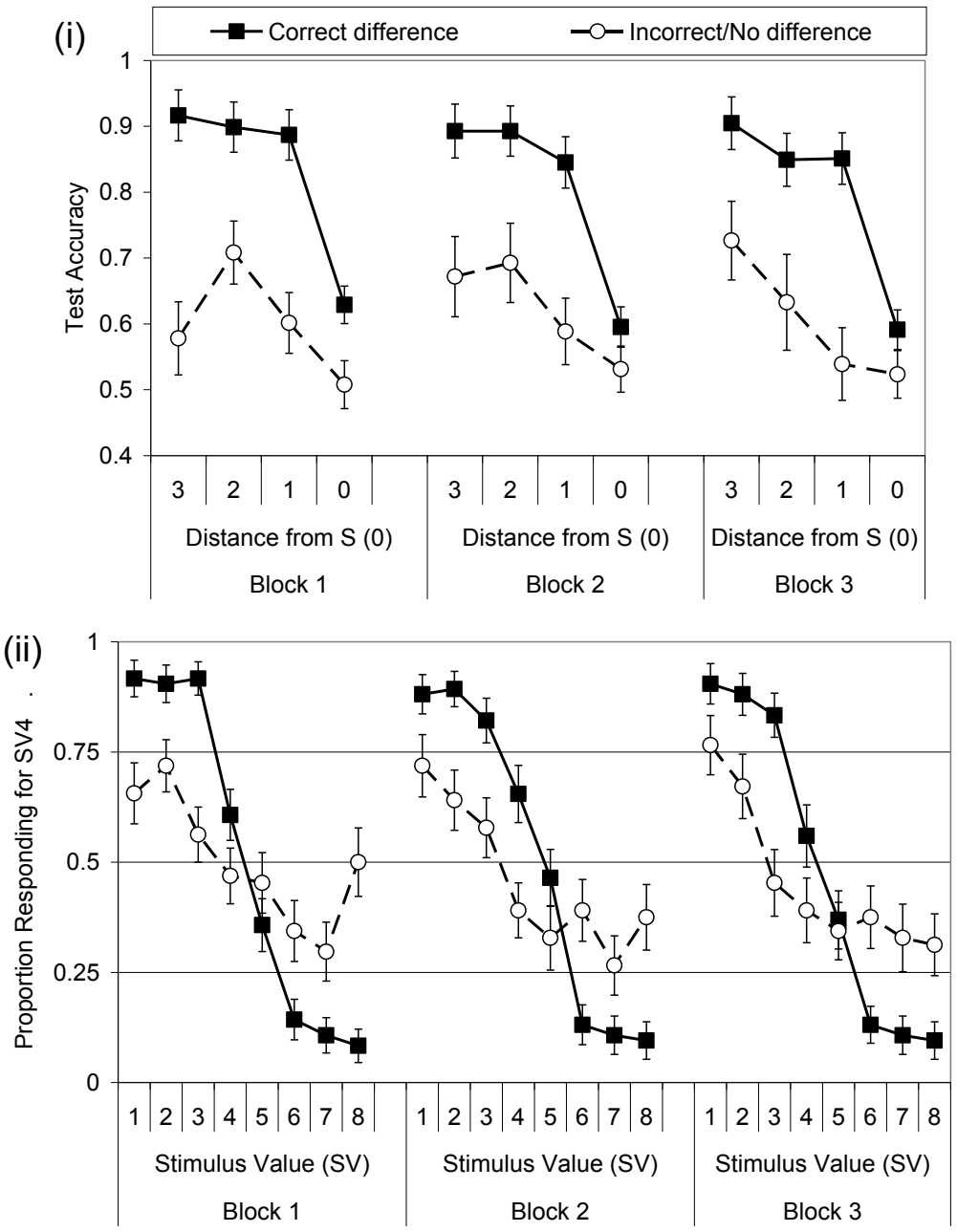

Figure 7. Post-discrimination gradients from Experiment 2, with participants divided according to reported training difference, averaged separately over successive thirds (blocks 1-3) of the test phase. Data are expressed as (i) mean test accuracy ± SEM plotted as a function of distance from the nearest training stimulus (S), and (ii) the proportions of responses that were appropriate for stimulus value 4 ± SEM plotted over the uncollapsed test range.