Intensity generalisation: physiology and modelling of a neglected topic*

Stefano Ghirlanda

Department of Psychology, University of Bologna

Reprint of December 2, 2006

Abstract

I briefly review empirical data about the generalisation of acquired behaviour to novel stimuli, showing that variations in stimulus intensity affect behaviour differently from variations in characteristics such as, for instance, visual shape or sound frequency. I argue that such differences can be seen already in how the sense organs react to changes in intensity compared to changes in other stimulus characteristics. I then evaluate a number of models of generalisation with respect to their ability to reproduce intensity generalisation. I reach three main conclusions. First, realistic stimulus representations, based on knowledge of the sense organs, are necessary to account for intensity effects. Models employing stimulus representations too remote from the sense organs are unable to reproduce the data. Second, the intuitive notion that generalisation is based on similarities between stimuli, possibly modelled as distances in an appropriate representation space, is difficult to reconcile with data about intensity generalisation. Third, several simple models, in conjunction with realistic stimulus representations, can account for a wide array of generalisation phenomena along both intensity and non-intensity stimulus dimensions. The paper also introduces concepts which may be generally useful to evaluate and compare different models of behaviour.

First published in *Journal of Theoretical Biology* **214(2)**, pp. 389-404, 2002. Minor differences may be present compared to the printed version. Coorespondence: Stefano Ghirlanda, stefano.ghirlanda@unibo.it.

1 Introduction

One of the main tasks of animal behaviour theory is to predict how an animal will react to novel stimuli that are somewhat different from familiar ones, to which the animal's reactions are known. This is the problem of generalisation, also known as "stimulus selection" within ethology (Baerends & Krujit, 1973) and "stimulus control" within experimental psychology (Mackintosh, 1974). Adequate models of generalisation are important not only to understand behaviour mechanisms, but also to gain insight in the evolution of communication phenomena such as courtship rituals or mimicry (Enquist & Arak, 1998; Ryan, 1998; Holmgren & Enquist, 1999).

Among the many ways in which stimuli can vary, this paper considers variations in physical intensity. This choice is motivated by the relatively little attention devoted to the topic. While it has been known since Pavlov's (1927) seminal work that stimulus intensity is an important determinant of behaviour (see the data summarised below), few theories have been designed with intensity effects in mind. The main aim of this paper is to evaluate a number of existing theories with respect to their ability to account for intensity generalisation. I consider the following models: gradient-interaction theory (Spence, 1936, 1937; Hull, 1943, 1949), several theories based on the concept of similarity between stimuli (Shepard, 1987; Nosofsky, 1986; Pearce, 1987), overlap theory (Ghirlanda & Enquist, 1999), the model by Blough (1975) and feed-forward neural networks (see Haykin, 1994). These theories deliver markedly different pictures about generalisation. In particular, they make different predictions about intensity generalisation, that can thus help to distinguish among them. To compare the considered theories I introduce a common terminology and basic conceptual tools intended to facilitate model analysis and evaluation. These may be of interest beyond the scope of this paper.

2 Empirical data

2.1 Behaviour

A simple but effective way of exploring generalisation is given by the following two-step procedure, typical of experimental psychology (Mackintosh, 1974). First, animals are trained to discriminate between two stimuli. Second, their reactions to a larger set of stimuli are recorded (generalisation test). During the first step, responses to a "positive" stimulus may be, for instance, rewarded with food, while reactions to a "negative" stimulus may be unrewarded. When animals have

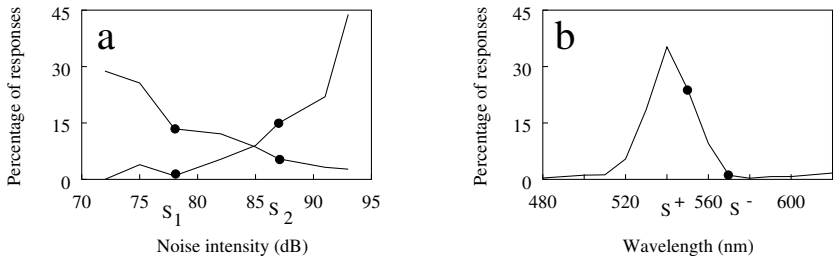

Figure 1: a) Noise-intensity generalisation in rats (data from Huff *et al.*, 1975). One group of animals was trained to react to intensity S_2 and not to react to the weaker intensity S_1. In the following generalisation test these subjects produced a monotonic response gradient increasing toward stronger intensities. A second group was trained on the opposite discrimination, yielding a reversed gradient. b) Wavelength generalisation in pigeons after discrimination between two lights of different wavelength (pecks to S^+ rewarded with food and pecks to S^- unrewarded. Data from Hanson, 1959). Note the peaked shape of the gradient, typical of non-intensity dimensions.

learned to respond to the positive stimulus and to ignore the negative one, the generalisation test is administered.

Here I discuss the case where the positive, negative and test stimuli only differ in their intensity. The main finding is that generalisation gradients following discrimination between two intensities are typically monotonic (figure 1a, Mackintosh, 1974). On the other hand, gradients obtained along most other stimulus dimensions are typically peaked (figure 1b). These latter dimensions include light wavelength, sound frequency, object size and location, and others. The notable fact that some stimuli elicit more responding than the positive stimulus, S^+, will be referred to as a response bias (other frequently used terms are "peak shift" and "supernormal stimulation"). Note that when a weak stimulus is rewarded and a strong one is not, the intensity gradient is still monotonic, but reversed (figure 1a; Zielinski & Jakubowska, 1977). Thus, it is not intensity per se the important variable, but rather the interplay of experiences along an intensity continuum.

There has been some discussion regarding the monotonicity of intensity gradients, because non-monotonic gradients are indeed found at times. Reviewing data about generalisation (Ghirlanda & Enquist, 2001), I found 23 out of 31 intensity gradients to be monotonic, to be compared with only 2 out of 38 along non-intensity dimensions ($P < 10^{-9}$, Fisher's exact probability test). This seems to warrant the claim that intensity gradients are typically "monotonic", and the

search for an explanation. The data sources for intensity gradients are: Razran (1949, three gradients extracted from 67 studies in Pavlov's laboratory), Pierrel & Sherman (1960, two gradients), Ernst *et al.* (1971, four gradients, two nonmonotonic), Thomas & Setzer (1972, four gradients, one nonmonotonic), Brennan & Riccio (1973, four gradients), Lawrence (1973, two nonmonotonic gradients), Scavio & Gormezano (1974), Huff *et al.* (1975, two gradients), Zielinski & Jakubowska (1977, six gradients), Wills & Mackintosh (1998, three nonmonotonic gradients). All these gradients (apart those in Razran, 1949, as mentioned above) are group averages. When more than one gradient per study is indicated, different experimental groups were tested after different training procedures.

The claim of monotonic intensity gradients, however, should be carefully qualified, given that no truly monotonic response gradient can exist. A first, rather trivial constraint on monotonicity is that physiological factors limit the range of intensities that an animal can react to: too high intensities will damage the sense organs, and too low intensities cannot be discriminated from absence of stimulation. A second reason is that we cannot assume that only the positive and negative stimuli determine generalisation. Animals are usually frightened by very intense stimuli and tend to ignore weak ones (due to either individual experiences or species experiences, coded in the genes). This may make it difficult to observe strong responding to very intense or very weak stimuli, when approaching such stimuli is required. A further factor that is known to influence gradient shape is the testing phase itself, during which animals can learn that the test stimuli are not reinforced. For instance, Pierrel & Sherman (1960) showed that an initially monotonic gradient (observed on the first two testing sessions) turned into a peaked one with further testing.

Despite these constraints, intensity generalisation gradients are typically monotonic over the range of intensities used in experiments, and this is what a theory of generalisation needs to explain. One may decide to drop the term "monotonic", but the evidence would stay. The question of why intensity gradients are monotonic would just be rephrased as: "Why is the peak of intensity gradients so remote from the positive stimulus, that it is typically not seen in experiments?"

While the issue of gradient shape will be the main argument of discussion in this paper, it is worthwhile to point out that stimulus intensity affects responding in other characteristic ways. Note for instance the extent of the response biases in figure 1a: responding to the extreme intensities is 2 or 3 times stronger than responding to the positive training stimulus (cf. the smaller bias in figure 1b). More generally, it can be shown that intensity dimensions produce significantly

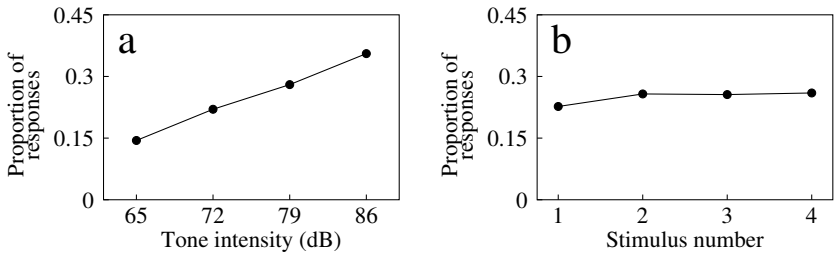

Figure 2: Comparison of generalisation after equal training on many stimuli. a) data from Scavio & Gormezano (1974) about conditioning of the rabbit's nictitating membrane response to tones of different intensities. Note that even after equal training the most intense stimulus still elicits more than twice as many responses as the weakest stimulus. b) data from Guttman (1965) about operant conditioning of key-pecking to monochromatic lights of different wavelengths. All stimuli elicited roughly the same number of responses.

stronger response biases ($P < 10^{-5}$, two-sample, two-tailed Kolmogorov-Smirnov test based on the gradients cited above Ghirlanda & Enquist, 2001).

A further area of interest is generalisation following experiences with many stimuli. Along non intensity dimensions, equal amount of positive experience with two or more stimuli typically induces almost the same rate of responding to all of the trained stimuli (Kalish & Guttman, 1957, 1959; Guttman, 1965). In contrast, it appears that equal experience with many stimuli of different intensities may result in more responding to the more intense stimuli (figure 2). This finding is not as well established as the previous ones, but it is potentially a third peculiarity of intensity dimensions which deserves further study.

2.2 Reactions of sense organs to variations in intensity

In looking for an explanation of intensity generalisation gradients, we may first note a peculiarity of intensity dimensions that is seen already in the responses of receptor cells. That is, when the physical intensity of a stimulus increases, the activation of the receptors reached by the stimulus also increases. This is at odds with what happens along non-intensity dimensions (Coren & Ward, 1989). Consider for instance photoreceptor reactions to monochromatic lights: the activation of a receptor shows a peak at some wavelength, decreasing when light wavelength either increases or decreases. Similarly, sound receptors are most sensitive to a given frequency of sound.

Thus, changes in intensity cause changes in activation patterns in the sense

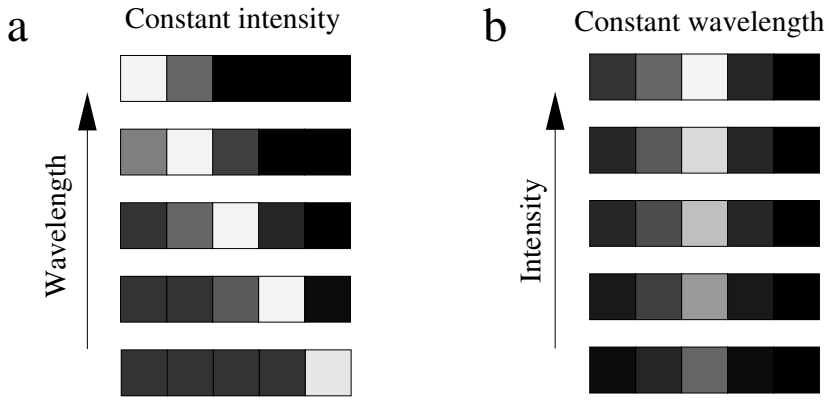

Figure 3: Schematic representation of the activation of the five types of pigeon photoreceptors to different visual stimuli. A lighter shade of grey represents a stronger activation. a) Lights of five different wavelengths, each one corresponding to the peak sensitivity of one of the receptor types (from top to bottom: 370nm, 415nm, 461nm, 514nm, 567nm). Physical intensity of stimulation is constant, and total activation of the receptors is also roughly constant. b) Lights of different physical intensities and constant wavelength (461nm). These representations do not take into account the effects of oil droplets, see Zeigler & Bischof (1993).

organs which are unlike changes along other dimensions (Ghirlanda & Enquist, 1999, see for instance figure 3). This peculiarity of intensity does not seem to be a logical necessity, as we can conceive of receptors whose sensitivity peaks at a given intensity. Indeed, the mechanisms by which receptor activation is monotonically related to physical intensity are very different for different sensory modalities (for instance, increased photon capture in photoreceptors and increased amplitude of mechanical vibrations in the ear). It is by virtue of such mechanisms that a dimension is perceived as an "intensity" dimension, rather than because the dimension is related to physical intensity. For example, human skin receptors sensitive to temperature are not more active the warmer it gets (Coren & Ward, 1989).

Of course, much more is going on in the sense organs than mentioned above, and certainly sensory physiology influences behaviour in many ways not considered here (Kandel *et al.*, 1991; Arbib, 1995). However, one of the points of this paper is that even taking into account simple properties of receptors, such as those discussed above, can be important for models of generalisation, that often have ignored the sense organs.

3 Theoretical considerations

The comparison between different models of animal behaviour is frequently hindered by the lack of a standard terminology and a common framework within which models can be analysed. In this section I introduce some basic concepts and notations to be able to describe models in a uniform fashion, at the cost of sometimes departing from traditional presentations. I will consider models of generalisation as made up of two distinct parts. The first provides a representation of stimulation, establishing a correspondence between real stimuli and the abstract objects which represent stimuli in the model. In most models stimuli are represented as points in a space with one or more dimensions, but how this representation is achieved varies considerably. To emphasize that physical stimuli and their representations are different entities, I will use uppercase letters for stimuli and the corresponding lowercase letters for their representations. The second part of a model of generalisation allows to calculate how behaviour elicited by a stimulus generalises to other stimuli. Obviously, this latter part operates on the stimulus representations, not on the stimuli themselves. Thus, modelling of both representation and generalisation contributes to the success or failure of a theory.

3.1 Representation

I will distinguish two main kinds of representation spaces. In an **object space** each dimension represents a physical characteristic of stimulation such as light wavelength, sound intensity and so on. In a **receptor space** (Ghirlanda & Enquist, 1999) each dimension represents the activation of a receptor cell (an example of such a representation has been used in figure 3). The distinction between object and receptor spaces is of interest because models adopting one or the other face different problems. Particularly relevant to this paper is the fact that in an object space dimensions are not linked to the sense organs. Thus, there is no justification for why intensity dimensions should yield different generalisation gradients. To make this distinction, a model based on an object space needs further assumptions. In contrast, receptor spaces, by encoding properties of receptors, provide an objective basis for why intensity generalisation is peculiar (according to section 2.2).

A seeming drawback of receptor spaces is that information about e.g. colors, shapes, intensities is not explicit. For instance, to calculate stimulus intensity in a receptor space we would need to sum the activation of all receptor cells (or a similar procedure, see e.g. Ghirlanda & Enquist, 1999). However, this is exactly the kind of information available to real brains. Receptor spaces force us to think about what information behaviour is based on. Object spaces are easier to use and build (it is enough to measure the physical characteristics of stimuli), but implicitly assume that the nature of sensory information has little relevance to behaviour. To give just a trivial example, a representation of monochromatic lights in terms of their wavelength does not explain why ultraviolet light cannot acquire control over human behaviour. It is instead clear that ultraviolet light does not elicit any reaction from human photoreceptors.

A third kind of representation is sometimes used, in which stimuli are broken into a number of "elements" of unspecified nature. Clearly, such an **element space** is of no practical use without assumptions (explicit or implicit) about what elements constitute any actual stimulus. It is often possible, however, to interpret the stimulus elements either as physical characteristics of stimuli or as receptor activations, making it possible to test the model in conjunction with object or receptor spaces.

3.2 Generalisation

In this paper I will consider mainly the following simple model of generalisation. Suppose that a behaviour is elicited by stimulus A with a given strength (for instance, rate of lever-pressing). We may express generalisation of such a behaviour to a stimulus B by means of a **generalisation coefficient**, $g_a(b)$ (recall that a is the representation of A). If, for instance, $g_a(b) = 1/2$ then B would elicit the behaviour with half the strength of A. A theory of generalisation can thus be viewed as a set of rules for calculating generalisation coefficients. In addition, the theory must provide a means to combine generalisation coefficients resulting from experiences with different stimuli. If only a positive and a negative stimulus, S^+ and S^-, have been experienced, reaction to any stimulus S will be a function of the two generalisation coefficients $g_{s^+}(s)$ and $g_{s^-}(s)$. The simplest possibility to combine the effects of S^+ and S^- is to form the difference

$$\delta(s) = g_{s^+}(s) - g_{s^-}(s) \tag{1}$$

with the assumption that responding to S' is predicted to be stronger than responding to S'' if $\delta(s') > \delta(s'')$. This very simple linear model, which will be referred to as the **difference model**, is interesting for a number of reasons. First, some models have precisely this form. Second, some theories provide a means to calculate generalisation coefficients, but have not addressed generalisation after discrimination training explicitly. The difference model can be a starting point to model generalisation within such theories. Third, a model may effectively behave as a difference model although this might not be apparent from its formulation. For instance, the model may provide a learning rule without any explicit statement about what will be learnt. Solution of the learning equation can show that the model's predictions can be expressed in the form equation (1), bringing insight about why the model behaves as it does.

With respect to intensity generalisation, empirical data will be correctly reproduced by a difference model if $\delta(s)$ is monotonically related to the intensity of S. More precisely, the difference must be ever-increasing when S^+ is more intense than S^-, and ever-decreasing otherwise (figure 1). Alternatively, the model may show that the peak of $\delta(s)$ is further away from s^+ along intensity dimensions, compared with non-intensity dimensions.

Of course, equation (1) is a very simple way of letting experiences with S^+ and S^- interact, adequate within the scope of this paper but not without shortcomings. For instance, g_{s^+} and g_{s^-} have the same weight in equation (1), but in general experiences with S^+ and S^- can have different importance in determining behaviour.

Moreover, equation (1) can be negative, and thus not suitable for interpretations such as probability of reaction. Both these problems can be approached by simple modifications of equation (1), such as

$$\delta'(s) = f\left(c_+ g_{s^+}(s) - c_- g_{s^-}(s)\right) \tag{2}$$

where $c_\pm$ are coefficients able to weight differently the two generalisation gradients, and f is a smooth monotonic function whose output is always positive (for instance, a logistic function). In addition to potentially solving the problems just mentioned, equation (2) describes some models more accurately than equation (1). For these reasons, I will sometimes refer to equation (2) as well as to equation (1).

4 Gradient-interaction theory

Gradient-interaction theory (Spence, 1936; Hull, 1943) is the oldest theory of generalisation that still influences how researchers think about generalisation. Stimuli are represented in an object space, that is they are located along a dimension according to the value of a physical parameter such as sound frequency or intensity. Generalisation is modelled by assuming that an "excitatory gradient" builds around the S^+ on the considered dimension (see e.g. Mackintosh, 1974). Its height at a given point along the dimension determines how the corresponding stimulus is reacted to. During the acquisition of a discrimination between S^+ and S^-, an "inhibitory gradient" is likewise assumed to build around the S^-. Predictions about generalisation are obtained by combining these two gradients. The most straightforward way of doing this (and the most used in the literature, see Hearst, 1968; Marsh, 1972; Mackintosh, 1974) is to adopt the difference model equation (1), where the generalisation coefficient $g_{s^\pm}(s)$ is the height of the $S^\pm$ gradient. I will use the term "generalisation coefficient" rather than "gradient" for consistency and to avoid confusion with empirical generalisation gradients.

A major shortcoming of gradient-interaction theory is that it provides little theoretical reasons to assume one form of g_s over another (Ghirlanda & Enquist, 1999), while it is clear that such an assumption is crucial to the theory. Actual research has employed two methods to overcome this difficulty. The first method is to identify g_s with the empirical response gradient obtained in a generalisation test following training where S alone is reinforced. One can thus train two groups of animal separately to react to only one of S^+ and S^-, and use the results from generalisation tests to predict behaviour of a third group trained to discriminate between S^+ and S^-. To my knowledge, this semi-empirical method has not

been tested along intensity dimensions. Even if we could get correct predictions, however, these would be based on empirical gradients which would remain unexplained. Moreover, along non-intensity dimensions this method has had only moderate success (Kalish & Guttman, 1957, 1959; Hearst, 1968; Marsh, 1972).

The second method used to derive predictions from gradient-interaction theory is to assume that all generalisation coefficients have a bell-shape and peak on the training stimulus. This assumption originates from Spence's (1936; 1937) work, and from the fact that bell-shaped empirical gradients prevail along non-intensity dimensions. The assumption appears problematic since a combination of bell-shaped (Gaussian) generalisation coefficients cannot result in a monotonic gradient. Nevertheless, if g_{s^+} and g_{s^-} are assumed to be very broad, relative to the distance between s^+ and s^-, we can obtain a peaked gradient whose peak is so far from s^+ as to be compatible with empirical data. But why generalisation should be broader along intensity dimensions, compared with other ones, cannot be answered by the theory. In conclusion, the main objection to gradient-interaction theory remains that the generalisation coefficients do not emerge from a theoretical proposal, but must be either assumed or derived from experiments.

It has been suggested (Perkins, 1953; Logan, 1954) that gradient-interaction theory can account for intensity effects if one takes into account that intensity gradients are also influenced from the zero-intensity stimulus. This can be conceived as an additional negative stimulus since it is never rewarded in experiments, and would push the response peak still further from the positive stimulus (Mackintosh, 1974). This proposal is unsatisfactory for at least two reasons. The first can be seen considering figure 1a (the same setup and results are also found in Zielinski & Jakubowska, 1977). Here the same two noise stimuli, S_1 (of lower intensity) and S_2, are alternatively used as the positive and negative stimulus. Both gradients are monotonic, but the gradient obtained after reinforcing S_1 cannot get its monotonicity from the presence of the zero-intensity stimulus. In fact, this would rather oppose strong responding to weak stimuli. Note that I am not claiming that the zero-intensity stimulus has no effect (see section 2.1), just that taking it into account fails to explain intensity generalisation within gradient-interaction theory. The second reason for this failure is that, were the Perkins-Logan argument true, we could get monotonic gradients along non-intensity dimensions by using two negative stimuli rather than one as customary. However, this effect has not been found so far (see e.g. Galizio, 1985, reporting data from two separate experiments).

Although gradient-interaction theory has been applied only to one-dimensional variations in stimulation, a multi-dimensional version of the theory is conceivable,

for instance by assuming bell-shaped generalisation along each dimension as well as rules to combine the contributions of different dimensions. Such a model could be applied to multi-dimensional receptor spaces. However, it would still fail to reproduce the observed differences between intensity and non-intensity dimensions, for reasons explained next.

5 Theories based on similarity and distance

The notion that responding to a novel stimulus is based on its similarity to familiar stimuli is an intuitively appealing one. In the terminology of this paper, to found a theory on similarity means simply to use the similarity of stimulus B to stimulus A as the generalisation coefficient $g_a(b)$. Some theories are explicitly based on such an assumption (Shepard, 1987), while others can be rephrased in terms of similarity. For example, within gradient-interaction theory it is possible to interpret the height of the S^+ excitatory gradient as the similarity of S to S^+. Another example is the model by Pearce (1987) discussed below.

A generalisation coefficient can be considered a measure of similarity if it has at least the following properties:

1. $g_a(a) > g_a(b)$, since B cannot be more similar to A than A itself ($b \neq a$ is assumed);

2. $g_a(b) = g_b(a)$, since A is as similar to B as B is similar to A.

Since similarity is a very general concept, lending itself to many implementations, I will discuss only a few main points relevant to existing theories. These will be illustrated further by the analysis of two specific models (Shepard, 1987; Pearce, 1987).

We need first to understand how biases are predicted in a similarity-based theory. When S departs from S^+, in a direction away from S^-, two things happen. First, the positive contribution from $g_{s^+}(s)$ (similarity to S^+) decreases because S is getting different from S^+. Second, the negative contribution from $g_{s^-}(s)$ also decreases, since S is departing from and S^- as well. The balance of these two terms decides whether responding will increase or decrease. Note that both $g_{s^+}(s)$ and $g_{s^-}(s)$ will be small for a stimulus S "very different" from both S^+ and S^-. It follows that that the predicted response to such a stimulus, $g_{s^+}(s) - g_{s^-}(s)$, will also be small, so that the gradient cannot be monotonic. However, we cannot say what "very different" means in practical terms, so that we cannot immediately

dismiss similarity-based theories based on this argument alone. Unlike gradient-interaction theory, other theories may have a means to predict a stronger displacement of the response peak along intensity dimensions.

Stimulus similarities are of course computed based on stimulus representation. The most common way of doing so is to first represent stimuli as points in an Euclidean space, and then compute their similarity as a decreasing function of Euclidean distance in the space. That is, the closer two stimulus representations are, the more similar the stimuli are taken to be (Shepard, 1957, 1987). Such a framework is faced with different difficulties according to what representation space is used. A model adopting an object space must justify why generalisation along intensity dimensions is different from generalisation along other dimensions (cf. section 3.1). This is the same difficulty encountered by gradient-interaction-theory. In a receptor space, the problem is that distances are not informative about intensities: knowing that s is at a given distance from s^+ does not tell whether s is more or less intense than s^+. In multidimensional spaces there is an infinite number of stimuli of different intensities which are all at the same distance from s^+. Thus, it seems difficult to distinguish between intensity and non-intensity dimensions in theories based on distance, unless further assumptions are made. This is why, for instance, the multi-dimensional version of gradient-interaction theory sketched above is unsuccessful also when applied to a receptor space.

A further possibility is that the structure of the space itself somehow encodes the intensity relationships among stimuli, so that the distances in such a space take into account intensity before being used to compute similarities. Whether this is at all feasible is beyond the scope of this paper.

5.1 The similarity-choice model

The most influential distance-based model is the similarity-choice model (Shepard, 1957, 1987) and its developments (e.g. Nosofsky, 1986, 1991), employed chiefly in human psychology (but see Wilkie, 1989; Cheng *et al.*, 1997, for applications to animal behavior). Similarity and distance are assumed to be connected by the following function:

$$g_a(b) = \exp(-kd^m(a,b)) \tag{3}$$

where k is a free parameter, $d(a,b)$ the distance between a and b, and m a positive number that determines how sharply similarity decreases as distance increases. The values $m = 1$ (exponential decrease of similarity) and $m = 2$ (Gaussian decrease) are typically used in the literature (but see Shepard, 1991).

This theory has been mainly concerned with data coming form identification and classification experiments, where subjects learn to attach a discrete response to each training stimulus (e.g., 'this is stimulus 1' or 'this stimulus belongs to category 1'). The model can accommodate such data very well (Nosofsky, 1986; Shin & Nosofsky, 1992; Shanks, 1995). In this context, the extent to which responses appropriate to a stimulus A are made to stimulus B is used to infer the amount of generalisation between the two stimuli A and B (see also below). It is important to note that in this paper we are dealing with a different problem: the generalisation of a continuous response to novel stimuli, after experience with only two training stimuli. In particular, response biases have not been investigated within the framework of the similarity-choice model. We can try to extend the model to this problem by using the difference model equation (1) in conjunction with the generalisation coefficient equation (3). Responding to s would thus be given by:

$$\delta(s) = \exp(-kd^m(s^+, s)) - \exp(-kd^m(s^-, s)) \tag{4}$$

The choice $m = 1$ corresponds to the widespread claim that generalisation is typically exponential (in an appropriate representation space, see below). In the present context, however, exponential generalisation coefficients are inadequate. In fact, when $m = 1$ the maximum value of equation (4) is always reached for $s = s^+$. This means that not only monotonic gradients, but response biases in general are impossible (this holds even with the more sophisticated model in equation (2)). It appears thus that, despite the successes mentioned above, exponential generalisation gradients are not a good starting point to model generalisation of continuous responses, where response biases are an ubiquitous finding. The choice $m = 2$ in equation (3) is potentially more promising, since Gaussian decrease of similarity allows for response biases. Such biases could also be systematically stronger along intensity dimensions, depending on how stimuli are represented. For instance, a smaller k might be predicted along intensity dimensions, compared with non-intensity ones. However, no such proposal is as yet available.

The representation space used in the similarity-choice model is also worth discussing. Stimuli are represented as points in a "psychological space" (or "conceptual space"). Such a space is built from human subjects' judgements of similarities between stimuli, so that the distances between stimulus representations reflect the similarity judgements (via equation (3), see Shepard, 1987). An alternative procedure (viable for animals as well as for humans) is to let subjects learn to identify a number of stimuli, and then infer distances between stimulus representations from

how often a stimulus is mistaken for another one. Note that in both cases stimulus representations are not predicted by the model, but are constructed to match behaviour. Presently, there is no way of connecting the psychological space with perceptual or brain processes. It is thus unclear whether a difference between intensity and other dimensions can emerge. Indeed, in the context of identification tasks, the construction of psychological spaces often results in object spaces, where intensity is a dimension not differing from all others (Nosofsky, 1987), and this appears to be problematic for an explanation of the phenomena considered here.

Moreover, several studies suggest that intensity effects are also important within the original scope of the similarity-choice model. For instance, intensity appears to bias similarity judgements (Tversky, 1977; Johannesson, 1997), so that an intense stimulus is rated more similar to a weak one than vice-versa. The typical approach to resolve this problem is to include additional parameters to allow for "prominence" effects (Nosofsky, 1991; Johannesson, 1997). This is often equivalent to letting k in equation (3) vary with intensity. Fitting these additional parameters accounts for the data but brings little insight about why intensity causes such a bias. Note also that after such modifications the theory is no longer based on similarity alone.

5.2 Pearce's 1987 model

Pearce (1987) presents a model of generalisation which directly relies on intensity of stimulation to predict behaviour. A buffer of limited capacity is assumed to hold a representation of all the stimulation reaching the animal at any moment. A generalisation coefficient $g_a(b)$ is defined based on the intensity of the stimulation common to the states a and b of the buffer:

$$g_a(b) = \frac{c^2(a,b)}{i(a)i(b)} \tag{5}$$

where $i(a)$ is the intensity of stimulation in situation A and $c(a,b)$ the total intensity of all stimulation present on both occasions A and B (see Pearce, 1987, for a fuller discussion of the model). Note that "occasion" A contains both the stimulus controlled by the experimenter and other stimulation in the environment (the so-called "contextual stimuli"). A formula to predict responding after discrimination training is not explicitly provided by Pearce, but a difference model based on the generalisation coefficient equation (5) seems in line with the reasoning in the paper.

This model accounts for a large number of phenomena (Pearce, 1987), but not intensity generalisation. Since equation (5) is a measure of similarity, we know from the previous section that the model cannot produce monotonic gradients (note that $g_a(a) = 1$, $g_a(b) < 1$ whenever $a \neq b$, and $g_a(b) = g_b(a)$). Nevertheless, since the model is specified in some detail, we can go further and try to calculate whether it predicts strong response biases. Consider a discrimination between, say, a tone T and absence of tone $\bar{T}$. If we adopt the difference model, responding to a tone T' is given by:

$$\delta(t') = \frac{c^2(t,t')}{i(t)i(t')} - \frac{c^2(\bar{t},t')}{i(\bar{t})i(t')}. \tag{6}$$

From equation (6) we can derive the maximum intensity compatible with responding to T' (supposed more intense than T) being stronger than responding to T (see Appendix 1):

$$i_{\max} < i(t)\left(1 + \frac{i(\bar{t})}{i(t) - i(\bar{t})}\right). \tag{7}$$

The meaning of equation (7) is that any stimulus more intense than $i_{\max}$ will certainly elicit a weaker reaction than T. Thus, the response peak must occur before $i_{\max}$. Since $i(t)$ should be considerably larger than $i(\bar{t})$ (the intensity of the "no tone" situation) the second term in parentheses is small. Thus, responding is predicted to drop at intensities rather close to $i(t)$, contrary to what is observed in, for instance, experiments of intensity generalisation after a discrimination between an auditory stimulus and silence (see e.g. Razran, 1949; Brennan & Riccio, 1973). However, equation (7) cannot rule out the possibility that the model may account for a part of the data about intensity generalisation. In fact, if $i(\bar{t})$ is not much smaller than $i(t)$, the constraint equation (7) is compatible with the gradient peak being quite far away from T. This result may be relevant to experiments where animals have to discriminate between two similar intensities (see e.g. figure 1a).

On the other hand, it is difficult to derive a more detailed prediction than equation (7). The reason is that Pearce did not commit his model to any particular representation, showing that a remarkable number of predictions could be drawn with only basic assumptions about representation. However, without more detailed assumptions, to calculate the stimulation common to two situations becomes difficult. For instance, if T' is a more intense tone than T, but of the same frequency, what is the intensity of the common stimulation, $c(t,t')$? At the level of sense organs T' and T activate the same receptors, but to different degrees. The

same problem arises also when considering non-intensity dimensions. Thus, an application of the notion of "common stimulation" to receptor spaces is problematic, because we lack a rule to determine the extent to which different activations of the same receptor count as "common". This makes it difficult to test Pearce's model with realistic representations of stimuli.

6 Overlap theory

Overlap theory (Ghirlanda & Enquist, 1999) adopts a receptor space to represent stimuli. The generalisation coefficient $g_a(b)$ assumed is the "overlap", or scalar product, between receptor activation patterns. The overlap is defined as

$$a \cdot b = \sum_i a_i b_i \tag{8}$$

where a_i represents the activation of receptor i when stimulus A is presented ($a_i > 0$), and the sum runs over all receptors. Thus, in overlap theory $g_a(b) = a \cdot b$. Note that this quantity cannot be interpreted as the similarity of a and b. The basic difference between equation (8) and a similarity measure is that $g_a(a) = a \cdot a$ is not the maximum value that g_a can reach (cf. requirement 1 in section 5). It is in fact possible that $a \cdot b > a \cdot a$. Indeed, this happens if b activates the same receptors activated by a, but to a greater extent. In general, if a stimulus changes so that receptor activation increases, the overlap with other stimuli increases. The difference model in this context reads:

$$\delta(s) = s^+ \cdot s - s^- \cdot s \tag{9}$$

that is, the overlap with s^+ contributes positively to responding, while the overlap with s^- contributes negatively. Formula equation (9) allows overlap theory to predict monotonic gradients along intensity dimensions. Consider for simplicity the trivial case of a sense organ with only one receptor. The representation of a stimulus S is then a single positive number s, increasing with increasing physical intensity of stimulation. Equation (9) reduces then to $\delta(s) = (s^+ - s^-)s$, and it is easy to see that $\delta(s)$ increases towards higher intensities if $s^+ > s^-$, while if $s^+ < s^-$ responding increases towards lower intensities (cf. figure 1a). With more realistic, multi-dimensional representations the overlap approach predicts both monotonic gradients along intensity dimensions and non-monotonic gradients along other dimensions (see Ghirlanda & Enquist, 1999, for details). In conjunction with an object space, equation (9) wrongly predicts monotonic gradients also along non-intensity dimensions.

7 Blough's (1975) model

The influential model of learning by Rescorla & Wagner (1972) was not initially concerned with generalisation, but has been extended by Blough (1975) (see also Rescorla, 1976). Blough's model is based on an element space, i.e. a stimulus is represented by the activations it elicits in an array of elements. The element activations are continuous variables. To each element, i, is assigned an "associative strength", v_i, so that responding to S is given by its total associative strength, defined as $V(s) = \sum_i s_i v_i$, where s_i is the activation of element i when S is presented (note the affinity with equation (8)). Associative strengths change in time according to a version of the so-called "delta rule", or least-mean-squares algorithm, which seems to have been independently discovered a number of times (see Widrow & Hoff (1960); Rescorla & Wagner (1972); Blough (1975); McClelland & Rumelhart (1985) and Haykin (1994); Arbib (1995) for reviews). Although this algorithm is widely known, it has not been fully evaluated as a model of animal learning and behaviour.

This model's ability to predict generalisation depends crucially on how real stimuli are assumed to activate the elements (cf. section 3.1 and section 5.2). To model monochromatic lights of different wavelength Blough assumed each element to be activated by a light of wavelength w according to

$$s_i(w) = e^{-q(w-w_i)^2} \tag{10}$$

where the parameter w_i determines to which wavelength element i is most sensitive, and q is a free parameter. In the case of wavelength generalisation, this assumption produces realistic gradients. However, if element activation is assumed to vary as in equation (10) along all dimensions, monotonic gradients cannot be produced. The lack of a means to determine how the elements are activated by physical stimuli was considered the main drawback of the model by Blough himself. Based on section 2.2, it seems natural to identify the elements with the receptors cells, so that knowledge of the sense organs is readily available in the model. For instance, a simple model of receptors reacting to variations in both wavelength and physical intensity of stimulation (I) could be:

$$s_i(w, I) = h(I)e^{-q(w-w_i)^2} \tag{11}$$

where the function $h(I)$ increases monotonically with I (see e.g. Torre *et al.*, 1995, for examples with several receptor types). Indeed, if element activation is assumed to increase monotonically along intensity dimension, the model predicts

monotonic intensity gradients. This can be easily checked by computer simulations, but it is also possible to solve the model analytically. After training on a discrimination between s^+ and s^- Blough's model behaves according to

$$V(s) = c_+ s^+ \cdot s - c_- s^- \cdot s \tag{12}$$

where c_+ and c_- are numerical coefficients whose role is only to ensure appropriate responses to s^+ and s^- (see Appendix 2). That is, in this case Blough's model is equivalent to a model based on overlaps and adopting equation (2) to predict responding (the function f in equation (2) can be easily added).

8 Artificial neural networks

Intensity generalisation gradients can be reproduced accurately by feed-forward artificial neural networks, learning by either an evolutionary process or the back-propagation algorithm (Haykin, 1994; Ghirlanda & Enquist, 1998). The simplest network architecture, referred to as a two-layer network, is that of a layer of input units connected directly to a single output unit. This is the same architecture of Blough's (1975) model, in which case the back-propagation algorithm also coincides with Blough's learning rule. Thus, overlap theory, two-layer networks, and Blough's model all agree in their predictions. Indeed Appendix 2 shows that equation (12) is the only way in which the two-layer architecture can react appropriately to two stimuli, irrespective of how behaviour is acquired (including evolutionary processes). By adding a layer of non-linear units between the input layer and the output unit we obtain three-layer networks, which are able to learn a wider class of discriminations (Haykin, 1994), but still agree with the two-layer networks in simple cases such as discriminations between two stimuli (Ghirlanda & Enquist, 1998, 1999).

It is important to note that the success of both two- and three-layer networks is conditional to adopting a receptor space to represent stimuli. Other representations which are sometimes used but yield incorrect results are: a dedicated input unit for each stimulus, the use of all-or-nothing units, the assumptions that more intense stimuli activate more units (rather than the same units to a greater extent), and object spaces.

Table 1: Ability of models to predict both intensity and non-intensity generalisation, based on different representation spaces

Model	Generalisation mechanism		Correct prediction using:		References
	Description	Mathematics	object space	receptor space	
Similarity-based theories:					
Gradient interaction	Combination of assumed or empirically obtained gradients	$g_{s^+}(s) - g_{s^-}(s)$	No	No	Spence (1936,1937) Hull (1943)
Euclidean distance	Distance from experienced stimuli	$e^{-kd^m(s^+,s)} - e^{-kd^m(s^-,s)}$ $(m=1,2)$	No	No	Shepard (1957,1982) Nosofsky (1986,1990)
Pearce's model	Intensity of shared stimulus elements	$\frac{c^2(s^+,s)}{i(s^+)i(s)} - \frac{c^2(s^-,s)}{i(s^-)i(s)}$	No	No	Pearce (1987)
Overlap theory	Overlaps with experienced stimuli	$s^+ \cdot s - s^- \cdot s$	No	Yes	Ghirlanda & Enquist (1999)
Blough's model	Agrees with overlaps after learning	$c_+ s^+ \cdot s - c_- s^- \cdot s$	No	Yes	Blough (1975)
Feed-forward networks:					
Two-layers	Not explicit (agrees with overlaps and Blough)	$f(v \cdot s) = f(\sum_i v_i s_i)$	No	Yes	
Three-layers	Not explicit (can agree with overlap and Blough, but is more powerful)	$f\left(\sum_i v_i f\left(\sum_j w_{ij} s_j\right)\right)$	No	Yes	Ghirlanda & Enquist (1998)

Symbol legend (see also text):

g_a: generalisation coefficient caused by experience with a

$c(a,b)$: intensity of elements common to both a and b

$a \cdot b = \sum_i a_i b_i$: overlap (scalar product) between a and b

v, w: weight vector and weight matrix

$d(a,b)$: Euclidean distance between a and b

$i(a)$: intensity of a

c_+, c_-: numerical coefficients (see Appendix 2)

$f(x) = \frac{1}{1+e^{-x}}$: sigmoid function

9 Discussion

Table 1 summarises the above arguments. A first important result is that no model based on an object space is able to predict generalisation along both intensity and non-intensity dimensions. A second result is that models based solely on similarity (gradient-interaction theory, the similarity-choice model and Pearce's 1988 model) appear unable to explain intensity effects satisfactorily. These conclusions are partly based on the simple difference model equation (1), and it may be possible to improve some of these models' predictions (for instance by adding additional parameters to represent intensities). However, the model by Blough (1975), overlap theory and feed-forward artificial neural networks are already able, without modification, to reproduce realistic intensity and non-intensity gradients along many dimensions.

A comparison among the successful models is interesting since, although they ultimately agree, they address stimulus control from different points of view. The model by Blough (1975) is a model of both individual learning and generalisation, whose architecture is that of a two-layer feed-forward neural network. Neural networks are *per se* computational models which give us the possibility of understanding how abstract computations can be carried out in biological systems. Feed-forward neural networks generalise naturally but need additional procedures to learn (e.g., evolutionary algorithms or learning rules such as Blough's (1975) one). Lastly, overlap theory focuses on what controls responding (the overlaps between a stimulus and the familiar ones), but it is not a model of learning or of how overlaps might be computed in nervous systems. A better understanding of generalisation is likely to come from an integration of all these approaches, rather than from favouring one over the others. For, instance, Blough's learning model cannot explain generalisation of genetically inherited responses, but neural network models can be shown to generalise very similarly independent of whether a response is inherited or learned by the individual (Ghirlanda & Enquist, 1998, see also Appendix 2).

The analysis presented in this paper can be extended in several directions. First, a complete analysis should also account for non-monotonic intensity gradients, although these are a minority. I have already mentioned at least two potentially important factors: 1) gradient shape can change from monotonic to peaked during testing and 2) responses to very weak or very strong stimuli can be partly beyond experimental control (section 2.1). The extent to which these factors can account for non-monotonic intensity gradients, and whether they can be incorporated into existing models, needs yet to be assessed. Second, one should consider

other effects of stimulus intensity on behaviour. I have already mentioned two such effects (stronger response biases and existence of response biases after equal training with many stimuli, see section 2.1), but intensity is known to have several other behavioural consequences, for instance on learning (see e.g. Feldman, 1975; Mackintosh, 1976). Models such as Blough's (1975) one can reproduce at least some of these findings (Ghirlanda, unpublished results) but more work is needed to fully evaluate this and other models. Third, one should consider other stimulus dimensions. For instance, generalisation along dimensions such as auditory click rate (Weiss & Schindler, 1981), light flicker (Magnus, 1958; Sloane, 1964) and others (Ghirlanda & Enquist, 1999) do not seem to yield bell-shaped or exponential generalisation gradients. When stimulation varies along these dimensions, as well as other such as object size, the total activation of the receptors varies. However, contrary to the intensity dimensions studied in this paper, different stimuli along the dimension activate different receptors. This calls for a more thorough analysis of such dimensions.

Moreover, all these are only a few examples of how stimulation can vary. The general problem of predicting behaviour following arbitrary variations in stimulation is likely to remain a puzzle if we do not focus on what information the brain receives from the sense organs and how this might be processed by real nervous systems. This seems an obvious conclusion, but one which has been often overlooked in models of behaviour. The main result of this and previous studies (Blough, 1975; Rescorla, 1976; Ghirlanda & Enquist, 1998, 1999) is that, if and only if stimulation is represented realistically, the same simple models can predict the existence of both monotonic and bell-shaped gradients along a wide variety of dimensions.

Finally, I remark the simplicity of the successful models. Their operation can be understood in terms of difference models of the form equation (1) or equation (2), and their implementation as connectionist models is simply that of a number of sensory units connected directly to an output unit (except of course for three-layer neural networks). It is clear that such a simple computational architecture is untenable as a full theory of animal behaviour (see e.g. Ghirlanda & Enquist, 1999, for some of the shortcomings). But it is also important to understand what is the minimal model able to reproduce a given phenomenon. Only with this understanding we can proceed to include a simple model's successful features into more complex and biologically realistic models.

10 Acknowledgements

I am grateful to Björn Forkman, Mikael Johannesson and John M. Pearce for discussion and valuable suggestions. I also thank four anonymous referees. Magnus Enquist has helped improve the whole manuscript and particularly table 1.

References

ARBIB, M. A. (ed.) (1995). *The handbook of brain theory and neural networks*. MIT Press.

BAERENDS, G. P. & KRUJIT, J. P. (1973). Stimulus selection. In: *Constraints on Learning* (HINDE, R. A. & STEVENSON-HINDE, J., eds.). New York: Academic Press.

BLOUGH, D. S. (1975). Steady state data and a quantitative model of operant generalization and discrimination. *Journal of Experimental Psychology: Animal Behavior Processes* **104**(1), 3–21.

BRENNAN, J. F. & RICCIO, D. C. (1973). Stimulus control of avoidance behavior in rats following differential or nondifferential pavlovian training along dimensions of the conditioned stimulus. *Journal of Comparative and Physiological Psychology* **85**(2), 313–323.

CHENG, K., SPETCH, M. L. & JOHNSON, M. (1997). Spatial peak shift and generalization in pigeons. *Journal of Experimental Psychology: Animal Behavior Processes* **23**(4), 469–481.

COREN, S. & WARD, L. M. (1989). *Sensation and perception*. Fort Worth, TX: Harcourt-Brace, 2 ed.

ENQUIST, M. & ARAK, A. (1998). Neural representation and the evolution of signal form. In: *Cognitive ethology* (DUKAS, R., ed.). Chicago: Chicago University Press, pp. 1–420.

ERNST, A. J., ENGBERG, L. & THOMAS, D. R. (1971). On the form of stimulus generalization curves for visual intensity. *Journal of the Experimental Analysis of Behavior* **16**(2), 177–180.

FELDMAN, J. M. (1975). Blocking as a function of added cue intensity. *Animal Learning & Behavior* **3**(2), 98–102.

GALIZIO, M. (1985). Human peak shift: Analysis of the effects of three-stimulus discrimination training. *Learning and Motivation* **16**, 478–494.

GHIRLANDA, S. & ENQUIST, M. (1998). Artificial neural networks as models of stimulus control. *Animal Behaviour* **56**, 1383–1389.

GHIRLANDA, S. & ENQUIST, M. (1999). The geometry of stimulus control. *Animal Behaviour* **58**, 695–706.

GHIRLANDA, S. & ENQUIST, M. (2001). Patterns of generalisation and response biases. Manuscript.

GUTTMAN, N. (1965). Effects of discrimination formation on generalization from the positive-rate baseline. In: *Stimulus Generalization* (MOSTOFSKY, D. I., ed.). Stanford, CA: Stanford University Press.

HANSON, H. (1959). Effects of discrimination training on stimulus generalization. *Journal of Experimental Psychology* **58**(5), 321–333.

HAYKIN, S. (1994). *Neural Networks: A Comprehensive Foundation*. New York: Macmillan.

HEARST, E. (1968). Discrimination training as the summation of excitation and inhibition. *Science* **162**, 1303–1306.

HOLMGREN, N. & ENQUIST, M. (1999). Dynamics of mimicry evolution. *Biological Journal of the Linnéan Society* **66**, 145–158.

HUFF, R. C., SHERMAN, J. E. & COHN, M. (1975). Some effects of response-independent reinforcement in auditory generalization gradients. *Journal of the Experimental Analysis of Behavior* **23**(1), 81–86.

HULL, C. L. (1943). *Principles of Behaviour*. New York: Appleton-Century-Crofts.

HULL, C. L. (1949). Stimulus intensity dynamisn (V) and stimulus generalization. *Psychological Review* **56**, 67–76.

JOHANNESSON, M. (1997). Modelling asymmetric similarity with prominence. *British Journal of Mathematical and Statistical Psychology* **53**(1), 121–139.

KALISH, H. & GUTTMAN, N. (1957). Stimulus generalisation after equal training on two stimuli. *Journal of Experimental Psychology* **53**(2), 139–144.

KALISH, H. & GUTTMAN, N. (1959). Stimulus generalisation after equal training on three stimuli: a test of the summation hypothesis. *Journal of Experimental Psychology* **57**(4), 268–272.

KANDEL, E., SCHWARTZ, J. & JESSELL, T. (1991). *Principles of neural science*. London: Prentice-Hall, 3 ed.

LAWRENCE, C. (1973). Generalization along the dimension of sound intensity in pigeons. *Animal Learning & Behavior* **1**(1), 60–64.

LOGAN, F. A. (1954). A note on stimulus intensity dynamism (V). *Psychological Review* **61**, 77–80.

MACKINTOSH, N. (1974). *The psychology of animal learning*. London: Academic Press.

MACKINTOSH, N. J. (1976). Overshadowing and stimulus intensity. *Animal Learning & Behavior* **4**(2), 186–192.

MAGNUS, D. (1958). Experimentelle untersuchung zur bionomie und ethologie des kaisermantels Argynnis paphia L. (Lep. Nymph.) I. Uber optische auslöser von anfliegereaktio ind ihre bedeutung fur das sichfinden der geschlechter. *Zeitschrift für Tierpsychologie* **15**(4), 397–426.

MARSH, G. (1972). Prediction of the peak shift in pigeons from gradients of excitation and inhibition. *Journal of Comparative and Physiological Psychology* **81**(2), 262–266.

MCCLELLAND, J. & RUMELHART, D. (1985). Distributed memory and the representation of general and specific information. *Journal of Experimental Psychology: General* **114**(2), 159–188.

NOSOFSKY, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General* **115**(1), 39–57.

NOSOFSKY, R. M. (1987). Attention and learning processes in the identification and categorization of integral stimuli. *Journal of Experimental Psychology* **13**, 87–108.

NOSOFSKY, R. M. (1991). Stimulus bias, asymmetric similarity, and classification. *Cognitive Psychology* **23**, 91–140.

PAVLOV, I. P. (1927). *Conditioned reflexes*. Oxford: Oxford University Press.

PEARCE, J. M. (1987). A model for stimulus generalization in Pavlovian conditioning. *Psychological Review* **94**(1), 61–73.

PERKINS, C. C. J. (1953). The relation between conditioned stimulus intensity and response strength. *Journal of Experimental Psychology* **46**, 225–231.

PIERREL, R. & SHERMAN, J. G. (1960). Generalization of auditory intensity following discrimination training. *Journal of the Experimental Analysis of Behavior* **3**, 313–322.

RAZRAN, G. (1949). Stimulus generalisation of conditioned responses. *Psychological Bulletin* **46**, 337–365.

RESCORLA, R. A. (1976). Stimulus generalization: some predictions from a model of Pavlovian conditioning. *Journal of Experimental Psychology* **2**, 88–96.

RESCORLA, R. A. & WAGNER, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In: *Classical conditioning: current research and theory* (BLACK, A. H. & PROKASY, W. F., eds.). New York: Appleton-Century-Crofts.

RYAN, M. (1998). Sexual selection, receiver bias, and the evolution of sex differences. *Science* **281**, 1999–2003.

SCAVIO, M. J. J. & GORMEZANO, I. (1974). CS intensity effects on rabbit nictitating membrane, conditioning, extinction and generalization. *Pavlovian Journal of Biological Science* **9**(1), 25–34.

SHANKS, D. S. (1995). *The psychology of associative learning*. Cambridge: Cambridge University Press.

SHEPARD, R. N. (1957). Stimulus and response generalization: A stochastic process relating generalization to distance in psychological space. *Psychometrika* **22**, 325–345.

SHEPARD, R. N. (1987). Toward a universal law of generalization for psychological science. *Science* **237**, 1317–1323.

SHEPARD, R. N. (1991). Integrality versus separability of stimulus dimensions: From an early convergence of evidence to a proposed theoretical basis. In: *Perception of structure* (LOCKHEAD, G. R. & POMERANTZ, J. R., eds.). Washington, DC: American Psychological Association.

SHIN, H. J. & NOSOFSKY, R. M. (1992). Similarity-scaling studies of dot-pattern classification and recognition. *Journal of Experimental Psychology: General* **121**, 137–159.

SIMMONS, G. F. (1974). *Differential Equations*. New Delhi: Tata McGraw-Hill.

SLOANE, H. N. J. (1964). Stimulus generalization along a light flicker rate continuum after discrimination training with several S-s. *Journal of the Experimental Analysis of Behavior* **7**(3), 217–221.

SPENCE, K. (1936). The nature of discrimination learning in animals. *Psychological Review* **43**, 427–449.

SPENCE, K. (1937). The differential response in animals to stimuli varying within a single dimension. *Phychological Review* **44**, 430–444.

THOMAS, D. & SETZER, J. (1972). Stimulus generalization gradients for auditory intensity in rats and guinea pigs. *Psychon.Sci* **28**, 22–24.

TORRE, V., ASHMORE, J. F., LAMB, T. D. & MENINI, A. (1995). Transduction and adaptation in sensory receptor cells. *Journal of Neuroscience* **15**, 7757–7768.

TVERSKY, A. (1977). Features of similarity. *Psychological Review* **84**, 327–352.

WEISS, S. J. & SCHINDLER, C. W. (1981). Generalization and peak shift in rats under conditions of positive reinforcement and avoidance. *Journal of the Experimental Analysis of Behavior* **35**, 175–185.

WIDROW, B. & HOFF, M. E. J. (1960). Adaptive switching circuits. In: *IRE WESCON Convention Record*, vol. 4. New York: IRE.

WILKIE, D. M. (1989). Evidence that pigeons represent Euclidean properties of space. *Journal of Experimental Psychology: Animal Behavior Processes* **15**, 114–123.

WILLS, S. & MACKINTOSH, N. J. (1998). Peak shift on an artificial dimension. *Quarterly Journal of Experimental Psychology* **51B**(1), 1–31.

ZEIGLER, H. P. & BISCHOF, H.-J. (eds.) (1993). *Vision, brain and behavior in birds*. Cambridge, Mass: MIT Press.

ZIELINSKI, K. & JAKUBOWSKA, E. (1977). Auditory intensity generalization after CER differentiation training. *Acta Neurobiologiae Experimentalis* **37**, 191–205.

A Weak response bias along intensity dimensions in Pearce's (1987) model

Here I examine the conditions under which, following a discrimination between two stimuli T and $\bar{T}$, with $i(t) > i(\bar{t})$, equation (6) predicts a stronger response to a stimulus T' than to T. Reaction to T' is stronger than reaction to T if

$$\delta(t') - \delta(t) = \frac{c^2(t,t')}{i(t)i(t')} - \frac{c^2(\bar{t},t')}{i(\bar{t})i(t')} - 1 + \frac{c^2(\bar{t},t)}{i(\bar{t})i(t)} > 0\,. \tag{13}$$

I will now provide an upper bound for $i(t')$ (assuming $i(t') > 0$). From the definition of $c(a,b)$ as the intensity of the common stimulation between situations A and B we infer

$$c(a,b) \leq \min(i(a), i(b))$$

and thus $c(\bar{t},t) \leq i(\bar{t})$, $c(t,t') \leq i(t)$. Using these relationships in equation (13) we obtain:

$$\delta(t') - \delta(t) \leq \frac{i(t)}{i(t')} - \frac{c^2(t,t')}{i(\bar{t})i(t')} - 1 + \frac{i(\bar{t})}{i(t)}\,.$$

Since we look for an upper bound we can discard the negative term containing $c^2(t,t')$:

$$\delta(t') - \delta(t) < \frac{i(t)}{i(t')} - 1 + \frac{i(\bar{t})}{i(t)}\,.$$

A necessary condition for a response bias is thus:

$$\frac{i(t)}{i(t')} - 1 + \frac{i(\bar{t})}{i(t)} > 0$$

which is equivalent to equation (7).

B Blough's (1975) model is a difference model which learns overlaps

The learning rule used in Blough (1975) is:

$$\dot{v} = \alpha(\lambda)(\lambda - v \cdot s)s \tag{14}$$

where v is the vector of associative strengths, $\dot{v}$ its time derivative, λ the reinforcement present at a given time, s the stimulation, and $\alpha(\lambda)$ an increasing function of λ which regulates the speed of learning. If two stimuli $s^{(1)}$ and $s^{(2)}$ alternate during learning, and if the change in behaviour produced by each presentation is small, we can substitute the r.h.s. in equation (14) with:

$$\dot{v} = \alpha_1 \left(\lambda_1 - v \cdot s^{(1)}\right) s^{(1)} + \alpha_2 \left(\lambda_2 - v \cdot s^{(2)}\right) s^{(2)} \tag{15}$$

where $\alpha_i = \alpha(\lambda_i)$ for short. Equation (15) can be solved by standard methods (see e.g. Simmons, 1974). It is also easy to prove that

$$v = c_1 s^{(1)} + c_2 s^{(2)} \tag{16}$$

is a solution for appropriate values of c_1 and c_2. These can be obtained by noting that equation (15) implies $\dot{v} = 0$ if

$$v \cdot s^{(i)} = \lambda_i \,. \tag{17}$$

Recall that $v \cdot s$ is the model's reaction to s, so equation (17) gives also the precise meaning of λ_i as the model's output to $s^{(i)}$ after learning. By substituting equation (16) in equation (17) we get:

$$\begin{cases} c_1 \left\| s^{(1)} \right\|^2 + c_2 \left(s^{(1)} \cdot s^{(2)} \right) = \lambda_1 \\ c_1 \left(s^{(1)} \cdot s^{(2)} \right) + c_2 \left\| s^{(2)} \right\|^2 = \lambda_2 \end{cases} \tag{18}$$

where $\left\|s^{(i)}\right\|^2 = s^{(i)} \cdot s^{(i)}$. Introducing the following matrix notation:

$$C = \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} \quad S = \begin{pmatrix} \left\|s^{(1)}\right\|^2 & s^{(1)} \cdot s^{(2)} \\ s^{(1)} \cdot s^{(2)} & \left\|s^{(2)}\right\|^2 \end{pmatrix} \quad \Lambda = \begin{pmatrix} \lambda_1 \\ \lambda_2 \end{pmatrix}$$

equation (18) are written simply as $SC = \Lambda$. The solution $C = S^{-1}\Lambda$ exists if

$$|S| = \left(\left\|s^{(1)}\right\|^2 \left\|s^{(2)}\right\|^2 - \left(s^{(1)} \cdot s^{(2)}\right)^2 \right) \neq 0$$

in which case:

$$\begin{cases} c_1 = \frac{1}{|S|} \left(\left\|s^{(2)}\right\|^2 \lambda_1 - \left(s^{(1)} \cdot s^{(2)}\right) \lambda_2 \right) \\ c_2 = \frac{1}{|S|} \left(\left\|s^{(1)}\right\|^2 \lambda_2 - \left(s^{(1)} \cdot s^{(2)}\right) \lambda_1 \right) \end{cases} \tag{19}$$

(here I do not consider the case $|S| = 0$ since it is very unlikely if the $s^{(i)}$'s model the responses of a large array of non-linear receptors). In summary, we have shown that when $v = c_1 s^{(1)} + c_2 s^{(2)}$, with the values equation (19) for c_1 and c_2, learning stops ($\dot{v} = 0$). Note however that the solution found is not unique. In fact, we can add to v any vector u such that $u \cdot s^{(1)} = u \cdot s^{(2)} = 0$ without altering the response to the experienced stimuli (that is, the relationships equation (17) will continue to hold). The precise value of u in any particular case depends on the value of v before learning, and it can be important for generalisation to novel stimuli, especially those with for which $s^{(1)} \cdot s$ and $s^{(2)} \cdot s$ are small.