Signal detection: applying analysis methods from psychology to animal behaviour

Christian J. Sumner[1], Seirian Sumner[2]

[1] Department of Psychology, Nottingham Trent University, 50 Shakespeare Street, Nottingham, NG1 4FQ.

[2] Centre for Biodiversity and Environmental Research, Department of Genetics Evolution and Environment,

Division of Biosciences, University College London, Gower Street, London, WC1E 6BT, United Kingdom

Current length: ~6500 words. 4 figures.

Keywords: animal behaviour, psychology, insects, signal detection theory, psychology

Acknowledgements

The authors would like to thank Margaret Couvillon and Francisca Segers for providing us with their data, and Mark Hauber, Kern Reeve and three anonymous reviewers for careful reading of the manuscript. The operant conditioning data was collected by Ana Alves-Pinto.

Abstract

Conspecific acceptance thresholds (CAT) [1], which have been widely applied to explain ecological behaviour in animals, proposed how sensory information, prior information and the costs of decisions determine actions. Signal detection theory ([2]; SDT), which forms the basis of CAT models, has been widely used in psychological studies to partition the ability to discriminate sensory information from the action made as a result of it. In this article we will review the application of SDT in interpreting the behaviour of laboratory animals trained in operant conditioning tasks, and then consider its potential in ecological studies of animal behaviour in natural environments. Focusing on the nestmate recognition systems exhibited by social insects, we show how the quantitative application of signal detection theory has the potential to transform acceptance rate data into independent indices of cue sensitivity and decision criterion (also known as the acceptance threshold). However, further tests of the assumptions underlying SDT analysis are required. Overall, we argue that SDT, as conventionally applied in psychological studies, may provide clearer insights into the mechanistic basis of decision making and information processing in behavioural ecology.

A. Introduction

Signal detection theory (SDT; [2]) is a theoretical framework that was developed to analyse behavioural responses of mammals (principally humans) performing a perceptual task (e.g. auditory, visual, tactile) in a laboratory. It has been applied to analyse a wide variety of psychological [3] and neuroscience [4] experiments, but it has also found application in areas as diverse as medical diagnostics, weather forecasting (see [5]), marketing [6], eye-witness testimony [7] and engineering [8].

Reeve's conspecific acceptance threshold model (CAT; [1]) sought to provide a theoretical basis for how decisions must be made by non-human animals in their natural environments. This model has been applied widely to help understand how animals make decisions in their natural environment, when decisions must be made with uncertainty [9-13]. In particular, it considered how an acting animal must repeatedly make a judgement about another animal of the same species. Guarding behaviour in a social insect nest is a good example of this: a 'guard' needs to make a decision to accept or reject an incoming individual; if it wrongly accepts a non-nestmate, this could carry costs to the colony in terms of egg dumping/social predation or stealing resources. Guarding, therefore, is effectively a series of largely independent binary (accept or reject) decisions made by an individual in response to some sensory information – e.g. the smell or appearance of an incomer. This process is remarkably similar to perceptual experiments in a laboratory, where subjects must repeatedly perform a series of very similar trials and are forced to make a (usually) binary decision on every trial.

CAT conceptualises a task as three components: production of sensory cues, recognition of sensory cues and an action. SDT does not consider cue production, but it does divide the task into the recognition of sensory cues and an action. Both consider sensory information, originating from a given object in the world, as a single noisy variable (i.e. a number). In both theories the decision-making process compares the incoming sensory variable against an "acceptance threshold" (for CAT) or "decision criterion" (for SDT).

SDT and CAT have been largely applied separately to different problems. SDT has been applied in the context of animal behaviour; for example, identifying mates of the correct species [14], choosing prey [15] and in mimicry [16]. Some theoretical studies have suggested SDT may have limitations in behavioural ecology due to its simplicity [17-19]; others call for the wider application of SDT to communication (e.g. [20, 21]) and real ecological data [22]; and there is a recent wider interest in how SDT might be applicable to understanding animal responses to modified landscapes [23, 24]. Despite this interest, there is a lack of quantitative applications of SDT in behavioural studies, or tests of the assumptions made by SDT.

Here, we first provide a tutorial-style overview of signal detection theory, and consider an example applied to animals in operant conditioning tasks. Second, we compare this briefly to Reeve's theory. Third, we apply a specific analysis in SDT to some behavioural ecology data: guarding behaviour in social insects.

We show that SDT can be useful in analysing such data: the benefits are that it transforms raw data on decision rates into indices that separately show the sensitivity of the response to both the cue and decision criterion. The decision criterion corresponds directly to the acceptance threshold in Reeve's model. On the other hand, the sensitivity gives a measure of the combined reliability of the differences between stimuli (or cues) and the reliability of the processing by the nervous system. Importantly, sensitivity does not depend on the decision made by the animal (i.e. the acceptance threshold), and vice versa. These indices therefore correspond well with typical experimental hypotheses on sensory cues. In contrast, empirical acceptance rates are a combination of both criterion and sensitivity and are therefore difficult to interpret in terms of CAT, SDT or any other quantitative theory. We argue for further development in the use of SDT in behavioural ecology, including the testing of its underlying assumptions.

B. Signal detection theory in perception.

Introduction to signal detection theory

A simple example of using SDT in experimental psychology is when testing the ability of a subject to detect a short tone pip ("beep") in a background of white noise ("ssss...."; [2]). Over repeated trials subjects are required to decide whether there was a tone present or not. Tones are typically presented randomly on 50% of those trials, whilst for the other 50% of trials there is only the background noise. Performance tends to vary from 100% correct when the amplitude of tone is large (loud), to 50% correct (chance, guessing) when the amplitude is small (faint; quiet). When the tone becomes faint, perceptions become ambiguous. Under these conditions, subjects vary in their propensity to say 'yes' or 'no'. Early researchers, seeking to probe sensory perception realised this, and saw these "decision biases" as separate to the subjects' true perceptual acuity and therefore as a barrier to measuring it accurately (for a brief history, see p22 of [25]).

SDT assumes that a subject's ability to discriminate between sensory stimuli is limited by the variability of the internal representations of those stimuli (Figure 1A). Any given stimulus could illicit a range of sensory values from one instance of the same stimulus to another, due to variability in the stimulus itself and/or noise in the nervous system. If the distributions of these representations for two different stimuli overlap, then some mistakes are inevitable. The more the distributions overlap, the more errors are made (c.f. Figure 1A and 1B).

The task of the subject is to decide which stimulus they think produced the sensation. In SDT, the subject is assumed to facilitate this by choosing a "decision criterion" (vertical lines in Figure 1A and B). Their response then only depends on which side of the criterion line a given individual sensory event lies. This decision criterion (labelled as *c* in Figure 1) is an important parameter in SDT.

In our example there are four possible outcomes: a correct identification of a tone is recorded as a 'hit', whilst failing to detect the tone is a 'miss'. Correctly identifying that there is no tone is a 'correct rejection', whilst an erroneous call for a tone that is not there is a 'false alarm' (i.e. wrongly perceiving the occurrence of a tone). In Figure 1, if the area under each curve sums to one, then the areas to the right of the criterion value indicate the probabilities of hits (shaded red) and false alarms (shaded blue), whereas the areas to the left of the criterion (not shaded) give the misses and correct rejections. The probability of hits, misses, false alarm and correct rejections vary depending on where the decision criterion is placed. As the criterion becomes more conservative (moving to the right in Figure 1), hit and false alarm probabilities drop.

The most common version of SDT makes strong assumptions about the underlying shape of the distributions associated with the subject's internal representations. It assumes that the distributions are normal (or Gaussian) in shape and of equal variance. In this equal variance SDT model, the only difference between the representations of stimulus classes are the mean values. Moreover, we can standardise the units of this difference to be the number of standard deviations separating their means (sometimes called a Z-score). Thus, SDT is not concerned with what that cue (or combination of cues) is, only the sensitivity of the representation to differences between stimuli. This is the second parameter in SDT, and is known as the sensitivity measure, "d-prime" (*d′*). A *d′* of 1 means that the two distributions are 1 standard deviation apart.

These two parameters provide useful information on the nature of the signal-action relationship. Measuring *d′* tells us about the information made available for a subject in the form of a signal or cue, on which to make a decision. The *d′* value is dependent on both how different the stimulus classes are (e.g. how loud the tone is, or how strong the pheromone is) and how clearly these differences are represented by the nervous system (e.g. how accurately neurons represent the stimulus, sometimes referred to as "internal-noise"). SDT cannot reveal which of these is the limiting factor. It simply quantifies the overall sensitivity of the subject's perception to a stimulus difference.

The decision criterion is useful because it tells us about the subject's action: what do they do with the information? Unlike *d′*, the decision criterion is always under the subject's control. The choice of criterion is likely to depend on the circumstances. Overall performance is maximised if the criterion is placed to bisect the two distributions. By convention, SDT considers this to be a criterion of zero, since the decision is not biased toward either action. In the context of a laboratory test of hearing, being unbiased makes sense since the subject is trying to get as many trials correct as possible. However, in the context of behavioural ecology, not all errors are equally costly. An extremely conservative (positive *c*) criterion can sometimes make sense, as we shall see shortly. SDT does not to seek to explain the selection pressures that set the criterion, but it does provide a simple means to measure it.

The power of the assumptions in this model is that one can unambiguously calculate the model parameters - sensitivity to the sensory cue (*d′*) and the decision criterion (*c*) - directly from empirical data on hit and false alarm rates (Figure 1).

For normal distributions of equal variance, the calculations are:

$$d' = Z(\text{hit rate}) - Z(\text{false alarm rate}) \qquad \text{(eqn 1)}$$

and

$$c = -[Z(\text{hit rate}) + Z(\text{false alarm rate})]/2 \qquad \text{(eqn 2)}$$

where hits and false alarms are expressed as a probability or proportion (from 0 to 1). The function *Z(.)* is the "inverse of the cumulative normal distribution" (sample R code is given in Supplementary Information). To illustrate *Z(.)* intuitively, the probability of a hit in Figure 1B is 0.618. This value is the area (shaded red and blue) under the normal distribution from *c* upwards (to infinity). The criterion line (grey) is 0.3 standard deviations below the mean of the distribution for stimulus class 2. Thus, $Z(0.618) = 0.3$. Conversely for the false-alarms, the criterion line is 1.7 standard deviations *above* the criterion line. This can be calculated from the blue shaded region, which has an area of 0.045, and $Z(0.045) = -1.7$.

Thus, the power of SDT in this equal variance form is that it provides a method for converting empirical the measurements of hit- and false alarm-rates into a model which separately quantifies sensory acuity (*d'*) and the decision-rule for action (*c*), with no free parameters. For a more detailed account, we highly recommend the very readable textbook by Macmillan and Creelman [25].

Using SDT to show how animals use information changes under different conditions

A good practical application of SDT is to disambiguate cue and decision-rule changes across different contexts. For example, consider a predator who hunts using auditory or olfactory cues and whose main source of food fluctuates seasonally. One might hypothesise that in periods of plenty they can afford to act only on salient cues, but in fallow periods they might need to act on the faintest sound or smell: a change in the decision criterion (*c*). Alternatively, seasonal changes in prey odour or behaviour might alter the sensory cues themselves (*d'*).

We illustrate this use of SDT using a set of operant conditioning lab experiments on auditory perception in mammals. Operant conditioning methods are widely used in animal behaviour studies to pair a behaviour with a specific behavioural response: correct (but not incorrect) responses to a stimulus are reinforced by a reward [26]. Sensory perception experiments in psychology use this same approach to measure how the ability to perform a task depends on some stimulus parameter, and then apply SDT to the data to ultimately determine the parameters and their relationships with the signal. We will examine how the sound level (loudness) of a tone pip affects the ability of ferrets to detect a tone in a background of noise; i.e. how sensitive their auditory perception is to the cue [27]. Such perceptual measurements are difficult and time consuming to collect for animal subjects, and can depending on how well the animals are trained and/or the cognitive demands of the task. We wished to establish the accuracy of measuring the sensitivity of the auditory system, and how (if at all) the methods used were affecting the data. Therefore, we varied the task in a way that should not alter *d'* if we were measuring stable perceptual sensitivity to the cue.

The ferrets (*Mustela putorius*) were trained in a forced choice task to indicate whether they heard a tone in the background noise, or just background noise (see Supplementary Information, and also Alves-Pinto et al. (2012) for details on experimental methods). After training, the level of the tone was varied systematically and the ferrets' performance in the task recorded using a computer. Ferrets then performed two different versions of the task which differed in the order by which test tones were presented. One version used the "method of limits" (ML): ferrets were tested at a different tone level on different days, with the tone level getting lower (more difficult to detect) each day. The second version used the "method of constant stimuli" (MCS): ferrets were presented with a similar range of sound levels in each behavioural session, with the sound level varying randomly from trial-to-trial. The perceptual sensitivity to presence of tones (*d'*) and how this varies with sound level should be the same for both methods, since the underlying representation of sound in the brain is unchanged. However, the decision criterion should depend on the method, and thus patterns of hits and false alarms are expected to be different.

The data show that the way hits and false alarms vary with sound level depends on the method used (Figure 2A). As expected, d' is the same irrespective of the method used (Figure 2B). In contrast, the decision criterion (Figure 2C) becomes increasingly conservative at higher sound levels for the method of constant stimuli (MCS), but remains virtually constant at all sound levels for the method of limits (ML).

Why would SDT predict the differences in decision criterion between the task variants? Recall that a criterion close to zero indicates an unbiased response which maximises overall performance for the two stimulus classes under consideration. When presented with a single sound level on each session (ML), the ferret can optimise the decision criterion so that performance overall is always maximal. When presented with a mix of sound levels on different trials of a single session (MCS) the ferret must choose a single criterion regardless of the sound level. Thus the criterion can only be unbiased at a single sound level (~60dB SPL). Overall, the robustness of the underlying cue sensitivity (*d'*) to varying methods of stimulus presentation supports the SDT model, and shows that behaviour is can be explained as a stable stimulus representation followed by an adaptable decision-making process.

Although designed for an experimental psychology question on the nature of perception, these contrasting methods have some relevance to ecology where sometimes a cue may be sporadic and variable in intensity as in MCS (e.g. stochastic variation in prey abundance [28]), and at other times it may be a cue that becomes increasingly weak as in ML (e.g. seasonal, often gradual reduction in the population density and hence the cue intensity of a migratory prey or predator [29]). Thus, measuring how the decision criterion varies with the level of an experimental cue may provide insights into the behaviour and ecology of the signaller.

This analysis rests upon the assumption of equal variance normal distributions. A way to empirically test these assumptions is to try to experimentally alter the decision criterion whilst keeping the stimuli the same. The equal variance SDT model predicts that the hit and false alarm rates will co-vary according to an exact relationship, which is determined by the normal distributions. Figure 2D shows the measurement of this relationship in one ferret (methods given in Supplementary Information). Tones are presented at a single sound level, and the ferret is encouraged to modify its decision criterion by experimentally altering the relative amounts of water available at the “yes” and “no” answer spouts from one test session to the next. Thus when (for example) 5x as much water is available at the “yes” spout, they are much more likely to answer there. Plotting hits against false alarms for different behavioural sessions with different reward ratios for the two responses makes it clear that these data follow a smooth parabola, and fit reasonably well to the curve predicted by an equal variance SDT model with a d' of 1.4. This curve is called the *receiver operating characteristic* (*ROC*). These data support that, to a good approximation, the ferret is behaving like an equal-variance SDT model. Subsequent experiments further supported this ([3, 30, 31]; also see Supplementary Information).

C. SDT analysis of behavioural ecology in insects.

Is this simple equal variance version of SDT of value in behavioural ecology? To our knowledge, there are no explicit tests to this effect; we address this below. In behavioural ecology, Reeve’s Conspecific Acceptance Threshold Model [1] is the most widely used “signal-detection” model. In the next this section we compare Reeve’s CAT model with SDTs.

SDT and the con-specific acceptance threshold model (CAT)

Reeve sought to provide a theoretical framework for how environmental and evolutionary pressures would influence an animal's decisions: i.e. how the acceptance threshold should be set. He derived a mathematical framework describing how this threshold would be optimised according to the frequency of encounters between different classes of conspecifics (e.g. nestmates and non-nestmates) and the relative costs of different kinds of errors (e.g. accepting a non-nestmate vs. rejecting a nestmate). Examples of these factors are illustrated in Figure 1C. The optimal acceptance threshold will bisect the two distributions symmetrically if the costs of different kinds of error are equal. However, if the cost of making one error is much greater than the other, an overall optimal decision is to accept more errors of one class than the other; e.g. if the costs of accepting a non-nestmate (e.g. from social parasitism) are greater than the cost of rejecting a nestmate.

Although SDT acknowledges these factors, a mathematical framework for why a criterion should take a particular value is missing from SDT. Often, the criterion is seen as a confound in determining the underlying sensory acuity. In SDT an optimal and 'unbiased' decision is one which maximises overall performance: SDT provides methods for directly calculating the decision criterion from the data and leaves the interpretation to the experimenter.

Another important difference is how SDT and the CAT are applied in practice. In CAT the underlying distributions could be normal (see [1], p413), or some other shape (Figure 1C follows Figure 3 from [1]), and there is no expectation that the variance of two normal distributions would be the same (e.g. Figure 1 in [1]). The assumption of equal-variance normal distributions is not ubiquitous in SDT, but the analytical power to separate sensitivity and criterion from a single hit and single false alarm probability depends on it.

CAT is a richer and more complete model of the process: it allows consideration of the ecological costs and benefits of decisions. However, it also has numerous free parameters. SDT, in its most common guise, is parameter free. Its analytical power follows from this simplicity. Often it is the complexities of nature which make it so interesting, and this is what CAT tries to capture. SDT is a more reductionist's approach to understanding behaviour. We now apply SDT to some published datasets on nestmate recognition in social insects, to determine the extent to which it can explain signal perception and response in an ecological context. According to SDT, hits and false alarms analysed separately cannot be used to unambiguously infer changes in sensitivity to the cue, or decision criterion because hits and false alarms are affected by both. Some recent studies [32, 33] recognise this problem but did not address it in their analysis. SDT provides the basis for separation of these factors. However, SDT has (as far as we have found) not been formally applied to analysis of insect behaviour. Therefore, we will now present test-cases based on previously published data. The primary purpose here is not to re-evaluate scientific findings but to illustrate the application of SDT in behavioural ecology, and to tentatively evaluate its value.

Recognition and acceptance in the guard behaviour of insects

The ideas behind Reeve's model have been widely applied in analysing insect behaviour [34, 35]. Of particular interest is guarding behaviour in social insect colonies, where nestmates defend their colony from conspecific intruders who may seek to steal resources or parasitise (egg dump; cuckoo) the nest [36, 37]. There is strong selection for colonies to evolve effective strategies to detect and repel non-nestmates, who would otherwise exploit the colony's resources. When a guard encounters another conspecific, they may respond by accepting them into the nest, or aggressively rejecting them. By recording these actions, the ability to recognise kin and/or alter their acceptance thresholds in response to a changing environment can be determined. Such studies have found that recognition acuity can depend on the context of the judgements, and that acceptance thresholds can shift according to the costs associated with the decision, consistent with the CAT model. Many of these studies report "hit" (e.g. successfully ejecting an intruder) or "false alarm" (inadvertently ejecting a nestmate) rates, and compare them across conditions (e.g. in the nest or away from it). Typically behavioural differences in acceptance rates across experimental conditions are quantified using statistical analysis.

Example: the effect of context on nestmate recognition in honey bees and stingless bees

We start with an example of guard behaviour in honeybees. Couvillon and colleagues [11] studied how context affects recognition in honey bees (*Apis mellifera*) and stingless bees (*Tetragonisca angustula*). Here we shall apply SDT to their data on honey bees. A similar analysis of stingless bees is provided in the Supplementary Information.

The discrimination of nestmates from non-nestmates was investigated in honey bees that were either in their natural context at the hive entrance, or in plastic test arenas away from the hive entrance. In the natural conditions, an experimenter introduced bees at the hive entrance near where the guard bees were located (the "hive" condition). Introduced bees could be nestmates from the same hive, or non-nestmates collected from other hives. Acceptance or rejection of the introduced bee by the guard bees was scored by the experimenter. In the plastic test arena, the process was similar, except that the context for recognition was manipulated to assess how the presence/absence of colony odour or guard number (1 versus 2 bees) influenced acceptance. The authors found that guard bees of *A. mellifera* made more errors in nestmate recognition when they were tested away from the hive: guards accepted more non-nestmates. The presence of colony odour reduced these errors but guard number had no effect (primary data replotted in Figure 3A). Viewed this way, the increase in false-alarms might be interpreted as indicating a change in the acceptance threshold in a CAT framework.

Figure 3B replots these data to illustrate better the accuracy of recognition. Expressed in this way the nestmate and non-nestmate recognition can be averaged to produce an overall recognition accuracy (solid line), which makes it clear that recognition becomes less accurate when guards are away from the hive (again see [11] for quantification and more detailed analysis), and when there are no odour cues. Viewed in this way, the drop in recognition rates might be interpreted as a reduction in the salience of the sensory cues.

Which is the correct interpretation? Do the data indicate a change in the acceptance thresholds applied by the guards? Or are the guards less able to discriminate nestmates from non-nestmates? The decision criterion (c) and sensitivity (d') of SDT, respectively, can tell us this.

The leftmost panel in Figure 3C shows a pictorial representation of the SDT analysis for the hive condition. Here, it is most intuitive to consider "hits" as correct acceptance of nestmates (note that this is the opposite way around to the CAT in Figure 1C where the internal variable is assumed to be dissimilarity to a nestmate; SDT is symmetrical so we are free to choose the most intuitive labelling). A "false alarm" is therefore the acceptance of a non-nestmate. Hits and false alarms correspond to the area under each probability distribution to the right of the decision criterion. Using Eqn 1 and 2, we can calculate *d′* and *c* directly from the acceptance rates:

$$d' = Z(H) - Z(FA) = Z(0.81) - Z(0.13) = 2.11$$

$$c = -[Z(H) + Z(FA)]/2 = -0.34$$

Essentially, we have "fitted a model" to the data: the model predicts the underlying sensory cues. Importantly, because we are assuming that both cues are normally distributed with equal variance, the model is parameter free. The *d′* (=2.11) value quantifies the number of standard deviations separating the means of these distributions; this is a measure of the size of the sensory cue for discriminating nestmates (RHS curve) from non-nestmate (LHS curve) bees. The *c* value (=-0.34) describes the criterion as number of standard deviations away from an optimal criterion which bisects the two distributions equally (i.e. c=0); this is the threshold at which bees will be accepted. A value of -0.34 suggests that decisions are not strongly biased. Thus, SDT provides a complete description of these data in terms of sensitivity to the sensory cues (*d′*) and criterion threshold (*c*), as an alternative to acceptance rates.

To quantify the effects odour and guard number, we now repeat this analysis for the non-hive condition experiments (Figure 3C, middle and right panels). Plotting d′ and c across all experiments provides a clear picture of the effects of hive vs non-hive conditions on nestmate discrimination by guards (Figure 3D): *d′* and *c* are both reduced compared with in-hive conditions, suggesting reduced sensitivity and an acceptance threshold that is more permissive. Using probit analysis [38] it is possible to confirm the statistical significance of changes in *d′* ($\chi^2(4)=14.6$, $p<0.01$) and *c* ($\chi^2(4)=37.1$, $p<0.0001$) across the five experimental conditions (Figure 3D shows Bonferroni corrected post-hoc comparisons with hive behaviour). Away from the hive, there is significant effect of odour on the decision criterion ($\chi^2(1)=15.8$, $p<0.0001$) but not the number of guards ($\chi^2(1)=2.54$, $p=0.11$). Changes in *d′* do not reach significance ($\chi^2(1)\leq 2.39$, $p\geq 0.122$ for comparisons between all *d′* pairs) away from the hive.

Couvillion and colleagues suggest three possible explanations for the increase in changes in behaviour away from the hive: that guards are missing salient cues; that being away from the nest results in reduced expression of aggressive behaviours by guards because of the costs of aggression are no longer outweighed by the benefits of defending a colony; or that guards away from their hive are simply not behaving normally. The SDT approach allows us to clearly separate the contributions of changes in sensitivity to sensory cues and criterion (or acceptance thresholds) in a single analysis procedure. It supports two of these explanations simultaneously: there is both a reduction in the salience of the cues *and* an overall reduction in aggressive behaviour. One important next line of inquiry would be to determine whether this reduction in sensitivity is due to lower quality physical cues, or a reduced neuronal capacity of guards to interpret the signal away from the hive. This might be achieved via independent manipulation of cues for recognition and the context of the recognition, paired with simultaneous electrophysiological recordings from neurons (e.g. [39]).

Are the assumptions underlying SDT valid in behavioural ecology?

We have shown how SDT can be applied to behavioural data from insects, and how it putatively helps us to unpack differences in sensitivity to recognition cues vs. decision criterion (acceptance thresholds). However, this comes with a caveat: we have not demonstrated that the model assumptions are valid. In particular: are the data consistent with the assumption of equal variance normally distributed sensory representation? Note that this does not require that the physical cues are normally and equally distributed; only that the variability of the internal representation of these cues in the nervous system is. Thus this must be tested through behavioural measurements.

In our study of ferrets, we were able to construct Receiver Operating Characteristics (ROCs) curves, shifting decision criterion by experimentally varying the rewards. This showed that the shape of the ROC curve was consistent with the model assumptions (Figure 1D). Another method for mapping out a ROC is to collect a measurement of subjects' confidence in their decision: "On a scale of 1-5, how sure are you that there was a tone?"

In the study of insect behaviour, a human observer must make a judgement about whether an insect "recognised" another individual, based on how an observation of behaviour. In the case of guarding, the response to a non-nestmate should be aggression, which is expressed through a number of behaviours such as movement of the mandibles, biting or stinging. A count of each type of aggressive behaviour therefore may represent the insect's confidence in determining whether it has encountered a non-nestmate or nestmate. In a study on the guarding behaviour of carpenter ants, Rossi *et al* [32] characterised the strength of the ants' responses in this manner, to a considerable level of detail. We will show how these data can be used to test the equal variance assumption made by SDT.

Figure 4A shows a subset of the data from Rossi and colleagues. It shows the proportion of occasions when guarding ants responded to an introduced ant with either no aggressive responses (0), or a single aggressive response (1), (2), (3), and so on. It is clear from this that ants performed aggressive responses more frequently when encountering non-nestmates than nestmates. By simple mathematical steps (see Supplemental Information) the number of aggressive responses can be used to derive hit and false alarm rates at different levels of confidence, and construct ROC curves.

An intuitive understanding of this analysis can be gained through visual representations of the data (Figure 4B). The more confident an animal is that it has encountered a non-nestmate, the more aggressive responses it will make. A given number of responses therefore corresponds to a certain confidence level, and in SDT a confidence level corresponds to a specific decision criterion. For example, the shaded area under the distributions in Figure 4B equals the hit (red) and false alarm (blue) rates for a confidence level corresponding to '4 actions'. For the ants, a sensory cue of this magnitude or larger warrants 4 or more aggressive responses. Notice that a higher confidence level (e.g. 9 actions; grey dotted lines in Figure 4B) corresponds to a smaller proportion of the distributions, and hence lower hit and false alarm rates.

Figure 4C shows a ROC curve constructed from these data. Each point on the plot shows the hit (aggression toward a non-nestmate) and false alarm (aggression toward a nestmate) rates, for a different criterion threshold. For example, the top-right most point shows the proportion of hits and false alarms when ants make any aggressive response (minimum 1) at all. This corresponds to the lowest level of confidence. These hit (0.82; nestmates) and false alarm rates (0.56; non-nestmates) correspond to the height of the stacked bar plots for 1 action in Figure 4A. This would be the hit and false alarm rates that we would record if the experimenter scored any response at all as a hit or false alarm. As we descend the points (leftward and downward) in Figure 4C, each successive point represents a higher level of confidence. Thus, if experimenters only scored a hit or false alarm when the ants made 4 or more aggressive responses, then the hit and false alarm rates would be 0.38 and 0.1 respectively.

Appropriately enough, the probability of observing higher confidence levels (more aggressive reactions) is lower: hits and false alarms are reduced. Together these data points prescribe a smooth parabola – a pattern hit and false alarm rates as a function of the criterion threshold.

How do we test whether this ROC is consistent with the assumptions of SDT? Recall that for the case where the sensory cues are normally distributed with equal variance, this curve takes a particular shape for a given value of *d′*. Such a dashed line is plotted in Figure 4C. The mean data values fall quite closely to them, and are within the bootstrapped 95% CIs for the hit and false alarm rates. Rossi and colleagues also examined the ants' behaviour when exposed to formic acid; analyses of these data are shown in the Supplementary Information. The Supplementary Information also considers other ways of scoring the ants' responses, such as considering the type of response made rather than just the number (which is picked here as the simplest, with the fewest assumptions). These data do not suggest any obvious failing of the model; indeed they suggest that the conformation to equal variance in these data may be no worse than for ferrets in an operant conditioning task. This analysis provides some confidence to suggest that data from behavioural ecology studies can conform to the assumptions of SDT.

D. Concluding comments

We have demonstrated the potential value of using SDT - a widely used tool in experimental psychology, engineering and medicine - in behavioural ecology studies, using the example of guarding behaviour in social insects as a test case. Our analyses suggest that SDT has the potential to be a useful, theoretically motivated basis and valid method for separating sensitivity to recognition cues. SDT does not seek to explain the constraints on setting a decision threshold whilst CAT additionally provides an account of how the costs of decisions should influence decisions. Yet, SDT is simpler to apply quantitatively to data.

In reanalysing previously published results, we have proceeded with the assumptions of the "equal variance Gaussian" model of signal detection. Strong undermining of these assumptions in psychological studies of perception is rare. The tests presented here suggest that the same may be true of insect behaviour. Nevertheless, it is important that these assumptions are tested further. Some data already exist to do this, and a logical next step is to perform a systematic and exhaustive review of the literature, and reanalysis of data where appropriate and possible. The equal-variance assumption may prove widely valid, and simple *d′* and *c* calculations may be valid from a single hit and single false alarm rate; or it may be necessary to collect data suitable for constructing ROCs in every case are more complex (e.g. unequal variance models) applied. Equally, existing data from experiments that were not designed for ROC analysis may not yield definitive conclusions of any kind. For example, the detailed, fine-scale/multi-rank scoring of guarding behaviour used for constructing ROCs of ant behaviour is not as common in studies of guarding behaviour in bee and wasp colonies, partly because these are often more difficult to observe than in ants. A systematic reanalysis of existing data may be helpful in designing new ways of recording behaviour for such analyses.

Even if the assumptions of SDT turn out not to hold in behavioural ecology, this article has illustrated some of the flaws in the current thinking. Most obviously, changes in the acceptance threshold cannot be inferred from false alarm rates alone. SDT also illustrates how hit and false alarm acceptance rates separately cannot be used to unambiguously infer changes in either acceptance thresholds or sensitivity to sensory cues (“recognition rates”). If disambiguating these factors is important, then it is necessary to find a valid quantitative model. Whilst equal variance-SDT is probably the simplest of all such models that can be envisaged, more complex models are available. An unequal variance SDT model is possible and changes the ROC in a predictable way [25]. More generally, SDT is closely related to theories of statistical decision making, which provides a theoretical base for modern machine learning (a.k.a. “artificial intelligence”; [40] methods. Bayesian methods, which are increasingly applied in behavioural ecology, neuroscience and machine learning, extend the power of decision models to account for prior knowledge and cost in a way which can encapsulate the environmental and evolutionary drivers which were formalised by the CAT model.

For the immediate future, we would argue that the value of equal-variance SDT needs exploring first. It is important to test the SDT model against a much more extensive range of empirical measurements than was possible here. It may hold in some circumstances, but more complex models may be required in others. Of course, complexity comes at the cost of more model parameters, which require more data to fit unambiguously (the current extreme example of this being deep neural networks). It is the simplicity of the SDT-model which allows for separation of sensory cue sensitivity and decision criterion with minimal data. Furthermore, following the truism that “all models are wrong”, the critical requirement is that the model complexity is matched sensibly to the experimental measurement error. Equal-variance SDT may be the solution to what researchers in the field are already searching for [32]: a better way to interpret their data in terms of recognition and action.

References

[1] Reeve, H. K. 1989 The Evolution of Conspecific Acceptance Thresholds. *Am Nat* **133**, 407-435. (DOI:Doi 10.1086/284926).

[2] Green, D. A. & Swets, J. A. 1966 *Signal detection theory and psychophysics*. Oxford, England, John Wiley.

[3] Sollini, J., Alves-Pinto, A. & Sumner, C. J. 2016 Relating Approach-to-Target and Detection Tasks in Animal Psychoacoustics. *Behav Neurosci* **130**, 393-405. (DOI:10.1037/bne0000143).

[4] Scholes, C., Palmer, A. R. & Sumner, C. J. 2015 Stream segregation in the anesthetized auditory cortex. *Hearing research* **328**, 48-58. (DOI:10.1016/j.heares.2015.07.004).

[5] Swets, J. A. 1996 *Signal Detection Theory and ROC analysis in psychology and diagnostics: Collected papers*. Hillsdale, N.J., Erlbaum Associates.

[6] Singh, S. N. & Churchill, G. A. J. 1986 Using the theory of signal detection to improve ad recognition testing. *Journal of Marketing Research* **23**, 327-336.

[7] Wixted, J. T. & L., M. 2015 Evaluating eyewitness identification procedures: ROC analysis and its misconceptions. *Journal of Applied Research in Memory and Cognition* **4**, 318-323.

[8] Schonhoff, T. & Giordano, A. 2007 *Detction and Estimation Theory*. New York, Peason.

[9] Starks, P. T., Fischer, D. J., Watson, R. E., Melikian, G. L. & Nath, S. D. 1998 Context-dependent nestmate-discrimination in the paper wasp, Polistes dominulus: a critical test of the optimal acceptance threshold model. *Anim Behav* **56**, 449-458. (DOI:DOI 10.1006/anbe.1998.0778).

[10] D'Ettorre, P., Brunner, E., Wenseleers, T. & Heinze, J. 2004 Knowing your enemies: seasonal dynamics of host-social parasite recognition. *Naturwissenschaften* **91**, 594-597. (DOI:10.1007/s00114-004-0573-1).

[11] Couvillon, M. J., Segers, F. H. I. D., Cooper-Bowman, R., Truslove, G., Nascimento, D. L., Nascimento, F. S. & Ratnieks, F. L. W. 2013 Context affects nestmate recognition errors in honey bees and stingless bees. *J Exp Biol* **216**, 3055-3061. (DOI:10.1242/jeb.085324).

[12] Engel, K. C., Manner, L., Ayasse, M. & Steiger, S. 2015 Acceptance threshold theory can explain occurrence of homosexual behaviour. *Biol Letters* **11**. (DOI:Artn 2014060310.1098/Rsbl.2014.0603).

[13] Mora-Kepfer, F. 2014 Context-dependent acceptance of non-nestmates in a primitively eusocial insect. *Behav Ecol Sociobiol* **68**, 363-371. (DOI:10.1007/s00265-013-1650-2).

[14] Price, J. J. 2013 Why is birdsong so repetitive? Signal detection and the evolution of avian singing modes. *Behaviour* **150**, 995-1013. (DOI:10.1163/1568539X-00003051).

[15] Coombs, S. 1999 Signal detection theory, lateral-line excitation patterns and prey capture behaviour of mottled sculpin. *Anim Behav* **58**, 421-430. (DOI:DOI 10.1006/anbe.1999.1179).

[16] McGuire, L., Van Gossum, H., Beirinckx, K. & Sherratt, T. N. 2006 An empirical test of signal detection theory as it applies to Batesian mimicry. *Behav Process* **73**, 299-307. (DOI:10.1016/j.beproc.2006.07.004).

[17] McNamara, J. M. & Trimmer, P. C. 2019 Sequential choices using signal detection theory can reverse classical predictions. *Behav Ecol* **30**, 16-19. (DOI:10.1093/beheco/ary132).

[18] Trimmer, P. C., Ehlman, S. M., McNamara, J. M. & Sih, A. 2017 The erroneous signals of detection theory. *P Roy Soc B-Biol Sci* **284**. (DOI:Artn 2017185210.1098/Rspb.2017.1852).

[19] Sih, A., Trimmer, P. C. & Ehlman, S. M. 2016 A conceptual framework for understanding behavioral responses to HIREC. *Curr Opin Behav Sci* **12**, 109-114. (DOI:10.1016/j.cobeha.2016.09.014).

[20] Wiley, R. H. 2006 Signal detection and animal communication. *Adv Stud Behav* **36**, 217-247. (DOI:10.1016/S0065-3454(06)36005-6).

[21] Wiley, R. H. 2013 A receiver-signaler equilibrium in the evolution of communication in noise. *Behaviour* **150**, 957-993. (DOI:10.1163/1568539X-00003063).

[22] Kikuchi, D. W., Malick, G., Webster, R. J., Whissell, E. & Sherratt, T. N. 2015 An empirical test of 2-dimensional signal detection theory applied to Batesian mimicry. *Behav Ecol* **26**, 1226-1235. (DOI:10.1093/beheco/arv072).
[23] Fletcher, R. J., Maxwell, C. W., Andrews, J. E. & Helmey-Hartman, W. L. 2013 Signal detection theory clarifies the concept of perceptual range and its relevance to landscape connectivity. *Landscape Ecol* **28**, 57-67. (DOI:10.1007/s10980-012-9812-6).
[24] Rosa, P. & Koper, N. 2018 Integrating multiple disciplines to understand effects of anthropogenic noise on animal communication. *Ecosphere* **9**. (DOI:ARTN e0212710.1002/ecs2.2127).
[25] Macmillan, N. A. & Creelman, C. D. 2005 *Detection theory: A user's guide*. Mahwah, NJ, Lawrence Erlbaum Associates.
[26] Stebbins, W. C. 1970 Principles of Animal Psychophysics. In *Animal Psychophysics: the design and conduct of sensory experiments* (pp. 1-19. Boston, MA, Springer.
[27] Alves-Pinto, A., Sollini, J. & Sumner, C. J. 2012 Signal detection in animal psychoacoustics: analysis and simulation of sensory and decision-related influences. *Neuroscience* **220**, 215-227. (DOI:10.1016/j.neuroscience.2012.06.001).
[28] Coblentz, K. E. 2020 Relative prey abundance and predator preference predict individual diet variation in prey-switching experiments. *Ecology* **101**. (DOI:10.1002/ecy.2911).
[29] Hulthen, K., Chapman, B. B., Nilsson, P. A., Vinterstare, J., Hansson, L. A., Skov, C., Brodersen, J., Baktoft, H. & Bronmark, C. 2015 Escaping peril: perceived predation risk affects migratory propensity. *Biol Letters* **11**. (DOI:ARTN 2015046610.1098/rsbl.2015.0466).
[30] Mill, R. W., Alves-Pinto, A. & Sumner, C. J. 2014 Decision criterion dynamics in animals performing an auditory detection task. *PloS one* **9**, e114076. (DOI:10.1371/journal.pone.0114076).
[31] Sumner, C. J., Wells, T. T., Bergevin, C., Sollini, J., Kreft, H. A., Palmer, A. R., Oxenham, A. J. & Shera, C. A. 2018 Mammalian behavior and physiology converge to confirm sharper cochlear tuning in humans. *Proceedings of the National Academy of Sciences of the United States of America* **115**, 11322-11326. (DOI:10.1073/pnas.1810766115).
[32] Rossi, N., Baracchi, D., Giurfa, M. & d'Ettorre, P. 2019 Pheromone-Induced Accuracy of Nestmate Recognition in Carpenter Ants: Simultaneous Decrease in Type I and Type II Errors. *Am Nat* **193**, 267-278. (DOI:10.1086/701123).
[33] Larson, J., Fouks, B., Bos, N., d'Ettorre, P. & Nehring, V. 2014 Variation in nestmate recognition ability among polymorphic leaf-cutting ant workers. *Journal of Insect Physiology* **70**, 59-66.
[34] Downs, S. G. & Ratnieks, F. L. W. 2000 Adaptive shifts in honey bee (Apis mellifera L.) guarding behavior support predictions of the acceptance threshold model. *Behav Ecol* **11**, 326-333. (DOI:DOI 10.1093/beheco/11.3.326).
[35] Couvillon, M. J., Robinson, E. J. H., Aktinson, B., Child, L., Dent, K. E. & Ratneiks, F. L. W. 2008 En guarde: rapid shifts in honeybee, Apis Mellifera, guarding behaviour are triggered by the onslaught of conspecific intruders. *Anim Behav* **76**, 1653-1658.
[36] Cini, A., Gioli, L. & Cervo, R. 2009 A quantitative threshold for nest-mate recognition in a paper social wasp. *Biol Letters* **5**, 459-461. (DOI:10.1098/rsbl.2009.0140).
[37] Pirk, C. W. W., Neumann, P. & Hepburn, R. 2007 Nestmate recognition for eggs in the honeybee. *Behav Ecol Sociobiol* **61**, 1685-1693.
[38] DeCarlo, L. T. 1998 Signal detection theory and generalized linear models. *Psychol Methods* **3**, 186-205. (DOI:Doi 10.1037/1082-989x.3.2.186).
[39] Hussaini, S. A. & Menzel, R. 2013 Mushroom Body Extrinsic Neurons in the Honeybee Brain Encode Cues and Contexts Differently. *J Neurosci* **33**, 7154-7164. (DOI:10.1523/Jneurosci.1331-12.2013).
[40] Duda, R. O., Hart, P. E. & Stork, D. G. 2001 *Pattern Classification*. New York, Wiley.

Figure 1. A and B: Showing the principles of signal detection theory with the difference in the means (d' – a measure of sensitivity or responsiveness to a particular cue), the decision criterion (c) – a metric reflecting the action taken, and the area under each distribution to the right of c, which gives the hits (red) and false alarms (blue). A case of very easy (Panel A: d'=4) and more difficult (Panel B: d'=2) discrimination between two stimuli. In all cases of SDT, the costs of errors in either direction are assumed to be equal C. Comparative example of the conspecific acceptance threshold model [1]. Shaded areas correspond to the hits and false alarms for an acceptance threshold (c=3.7) which maximises overall % of correct decisions, which is also optimal when costs of errors are equal. A lower threshold (c=1.7) corresponds to optimal if the costs of errors are unequal. In this case, the cost of deciding a stimulus from class 2 belonged to class 1 is 10x that of deciding that a stimulus from class 1 belongs to class 2.

Figure 2. SDT analysis of tone-in-noise detection in ferrets, showing how decision criterion depends on the method of data collection, but the measure of sensitivity (*d'*) is consistent; these are consistent with the predictions of signal detection theory. A. Hit and false alarm rates vs. the sound level of the tone for the two variants of data collection method. B. The SDT calculation of *d'* (cue sensitivity) for the two variants. C. The SDT calculation of the decision criterion (*c*). D. Receiver Operating Characteristics (ROC) curve measured for a ferret by varying the amount of water rewarded at each spout, for a tone at .60dB SPL (d' ~1). Points represent hit versus false-alarm rates for individual sessions (diamonds) and from all trials at a given reward condition (black squares). The grey line represents the SDT model prediction for a *d'* of 1.4.

Figure 3. SDT analysis of guard behaviour in honey bees (*A. mellifera)* in different contexts [11]. A. Original data expressed as acceptance rates for nestmates (NM; blue) and non-nestmates (NNM; red). Error bars are omitted for simplicity. B. The same data expressed as proportion of correct responses (NM acceptance is unchanged. NNM proportion correct = 1 – acceptance rate). Yellow points and line show the mean of NM and NNM accuracy. C. SDT models of behaviour, showing the probability distributions of sensory values, and the position of the decision criterion (== acceptance threshold). Left plot shows the behaviour at the hive entrance. Shaded areas indicate the areas associated with acceptance, corresponding to the proportion of hits (nestmates) and false-alarms (non-nestmates). Middle and right plots show the behaviour in the non-hive context with and without odour respectively. Both plots superimpose distributions (solid lines) and criterion the guard conditions (dashed lines), which exactly overlie each other in the no odour conditions (right most panel). D. SDT statistics for each condition. Left plot shows the sensitivity to sensory cues (d'); Right plot shows the decision criteria (*c*). Post-hoc significance tests (Bonferonni corrected) compared with the hive are shown above the points and a significant effects of odour on criterion (probit analysis described in main text) are indicated below the points.

Figure 4. Testing the assumptions of SDT using data on guard behaviour in ants [32]. A. Original data expressed as the proportion of different aggressive responses grouped according to the number of responses made (so 0 = no response, 1 = 1 aggressive response etc.). B. A diagrammatic SDT

representation of ant responses to nestmates (red) and non-nestmates (blue), where the vertical dashed lines show different decision criterion, each corresponding to a different confidence level with a different number of actions (a selection are chosen for clarity). Shaded areas indicate the hits (red) and false alarms (blue) at the criterion level corresponding to 4 or more actions. Probability scale is in arbitrary units (AU). C. ROC analysis of the data. Points show the hit and false alarm rates for a given criterion, with error bars showing bootstrapped 95% CI. Solid grey line shows the ROC curve predicted by an equal variance SDT model with a d’=0.85. Dotted line shows d’=0.