Stimulus Equivalence and the Origins of Reasoning, Language, and Working Memory

Mike Oaksford

The ability to pass tests of stimulus equivalence seems to mark an important distinction between animals and humans that is tied to language. Most other animals are unable to reliably pass equivalence tests. Even linguistically trained chimps seem unable to pass them (Dugdale & Lowe, 1990). Moreover, pre-linguistic children show similar poor performance. Sidman (1990) argues that an innate logic (Fodor, 1975) would confuse explanandum with explanans, i.e., logical ability needs to be explained by more primitive behaviour. We argue that abductive reasoning ability is more primitive in that it is required to learn basic stimulus-stimulus relations. Moreover, we propose that extending this ability beyond the temporal limitations of our neural equipment requires the ability to sustain a representation in working memory and this may be facilitated by associating it with an endogenous rehearsable response, i.e., a speech sound. In sum, extending the temporal range over which associations can be learned using associated sound may tie together the origins of reasoning, language and working memory.

Keywords: Stimulus equivalence, reasoning, eduction, abduction, language, working memory

In this paper, I explore the implications of the findings on stimulus equivalence for the origins of reasoning abilities. The argument has its roots in a commitment to Lloyd Morgan’s canon as expressed at the end of a recent paper by Nick Chater and I (Oaksford & Chater, 2003, p.312): “Following Lloyd Morgan’s (1894) canon, if some function can be explained at the lower level, that is probably the level at which it should be explained, since there seems little point attributing people with more high-level cognitive equipment than is needed to explain their behaviour.” This point was made in the context of a closing discussion of dual process approaches to human reasoning (e.g., Evans & Over, 2004; Stanovich & West, 2000). On this view, System 1 processes are automatic, unconscious, and based on implicitly acquired world knowledge. System 2 processes are controlled, analytic, and based on explicitly acquired formal rules. The evolutionarily older System 1 is also usually regarded as largely associative and involving connectionist style processing. The evolutionarily more recent System 2 is usually regarded as involving explicit formal rules operating over symbolic representations. Nick Chater’s and my probabilistic approach (see Oaksford & Chater, 2001, 2007 for summaries) is frequently cited as abjuring System 2 processes. However, we have explicitly argued against this characterisation (Chater & Oaksford, 1999; Oaksford & Hahn, 2007) and as the above quotation demonstrates, concentrating on the System 1 level, following Lloyd-Morgan’s cannon, is a methodological commitment not a theoretical commitment to the non-existence of System 2.

Once a theoretical commitment is made to
these two levels, the question arises of how they are related? In particular, if System 1 is phylogenetically older, presumably the newer System 2, evolved to meet some need not met by System 1. Moreover, given that System 2 is involved in explicit verbal reasoning of a type seen mainly in humans (see Hurley & Nudds, 2006), this need is most likely tied to language. We argue in this paper that the findings on stimulus equivalence provide insight into the difference between Systems 1 and 2 and critical clues about how System 2 might have co-evolved with language and working memory.

The paper is organised as follows. I first outline some further distinctions between System 1 and System 2 based on the nature of the representations and processes involved and how they are used to make different types of inference. We then observe that one particular type of inference, predictive inference or eduction, seems to be the target of basic animal learning processes. We then introduce the evidence from the matching to sample paradigm showing that full stimulus equivalence behaviour seems to be tied to linguistic ability. We argue that basic animal learning processes that allow eductive inferences also rely on a limited form of abductive inference which requires keeping possible predictors of an event in mind. We argue that this process in temporally bounded by the decay characteristics of the brain’s neural hardware. We suggest that organisms that have the ability to generate endogenous stimuli, e.g., sounds, that can be associated with representations of external events and that can be rehearsed have an adaptive advantage in being able to maintain representations of possible predictors in mind for longer. Being able to maintain representations in mind also allows people to interrogate their long term memory for world knowledge and store the results as representations of different possibilities over which further, more abstract reasoning processes may develop.

**System 1 and System 2 Inferences**

Studies of humans reasoning usually begin with an account of the logic of the conditional statement, if...then. However, in natural language this term can be used to describe a variety of real world relations that can be involved in a variety of inferences (Barwise & Perry, 1983). It is knowledge of these real world relations that it is assumed animals and humans learn in standard learning paradigms. So dogs can learn that if a bell sounds, then food is about to arrive or that if it presses the lever, then food will arrive. The latter instrumental case is no different to the similar instrumental relations that people exploit in their everyday lives, e.g., if I turn the key, then the car starts. It is a common assumption that such information is what makes up our world knowledge stored in long term memory and which is the core of System 1 processes. Such knowledge can be represented in connectionist systems (McClelland & Rumelhart, 1986) or rule based systems with associated strengths (Holland, Holyoak, Nisbett, & Thagard, 1986).

How do we acquire and what do we do with such knowledge? Peirce (1931) identified a variety of inference types that apply to such relations. So one can learn about keys starting cars by inductive inferences from instances of ignition key turning events; one can predict what will happen when you turn the key in eductive inference; and one can explain why the car started in an abductive inference. Of course, not all inductively learned relations that allow good eductive inferences allow good abductive inferences, e.g., if the cock crows, then the sun will rise, is useful in prediction but not in explanation. Learning that the cock’s crow does not cause the sun to rise requires intervention to see what happens when the cock does not crow (Pearl, 1988). Such intervention is beyond most non-human organisms (but see, Hurley & Nudds, 2006). This would also appear to be true of deductive inferences, the last of Peirce’s quattuorvirate of inference types. Deductive inferences are different
in that they only rely on the form of the conditional used to describe the relation and not on the relation itself. The System 1/System 2 division seems to involve allowing that System 1 processes are adequate for all but *deductive* inference, which requires System 2 processes operating along more abstract and general lines.

This section has proposed a possible division between System 1 and System 2 on the inferences that they might support. Typically, however, System 1 processes are regarded as associative and hence bidirectional. In the next section, we describe the work that showed that System 1 processing is actually sensitive to prediction rather than just association.

**System 1: Learning and Education**

In a standard animal conditioning experiment, a pre-existing response, e.g., salivation, is gradually bought under the control of a conditioned stimulus (a “CS”, e.g., a bell sounding) by being paired with an unconditioned stimulus (a “US”, e.g., food presentation). This is classical conditioning. Alternatively, a particular response may be elicited by providing a reward each time an action is performed which approximates the desired response. So a rat may gradually be trained to press a lever to receive food, by providing food for each successive behaviour as it more closely approximates lever pressing. This is instrumental conditioning. In many experiments these procedures are combined. For example, a rat may first be trained to press a lever for food (instrumental conditioning). The lever pressing response is then subsequently treated as the unconditioned response in a classical conditioning experiment. Until the late 1960’s it was thought that for animals to learn the appropriate conditioned responses, all that mattered was how often the CS and the US were paired together. In the language of contemporary cognitive approaches to learning theory (Dickinson, 1979; Mackintosh, 1974; Sutton & Barto, 1981), it is the frequency of pairings which determines the strength of the connection between the representation of the CS and the representation of the US.

Such simple associationist interpretations of conditioning were however questioned by an experiment conducted by Rescorla (1968). This experiment established that simply pairing the US and the CS was not sufficient for conditioning to occur. In addition, the CS must be a good predictor of the US. This was established by demonstrating no learning when the CS does not predict the US although the CS and US have been paired many times. This is what Rescorla’s (1968) ingenious paradigm achieved. He used a conditioned suppression procedure where the dependent variable is the suppression of a previously learned response, in this case lever pressing. Rats trained to press a lever for food were presented with a series of 2 minute tones. An interval in which no tone was presented separated each tone presentation. Four groups of rats then received shocks during the tone with a probability of 0.4, i.e., P(shock|tone) = 0.4. The critical manipulation was to vary the predictability of the tone (CS) for the shock (US) by varying the probability of receiving a shock in the intervals between tones, i.e., P(shock|no tone). Four conditions were used P(shock|no tone) = 0.4 or 0.2 or 0.1 or 0.0. When P(shock|no tone) = 0.4 the rat is as likely to receive a shock when the tone is absent as when it is present. The tone is therefore completely non-predictive of the shock. The lower P(shock|no tone) becomes the more predictive of the shock the tone becomes. Given that each group receives the same number of tone presentations the frequency of tone-shock pairings can therefore be kept constant while CS-US predictability is varied. To measure the effects of this training the rats were returned to the lever pressing task and the suppression of lever pressing evoked by the tone measured. The dependent variable used was the *suppression ratio*, i.e., the number of lever presses during a 2 minute tone divided by the sum of the number of lever presses during the tone period plus the
number of lever presses in the preceding 2 minutes. Thus a suppression ratio of 0.0 means a total suppression of responding and a ratio of 0.5, no suppression. Since the shock (US) suppresses responding, the more the rat has learned about the CS-US relationship the lower the suppression ratio should be. As can be seen from Figure 1, Rescorla (1968) found that the greatest suppression was found for the lowest $P(\text{shock} | \text{no tone})$ group, providing unequivocal evidence that predictability, not simple contiguity is required for learning CS-US relationships.

![Figure 1](image)

**Figure 1** Median suppression ratios as a function of $P(\text{shock} | \text{no tone})$ with $P(\text{shock} | \text{tone})$ held constant at 0.4. In the control condition, the same tone presentations were given but in the absence of shocks. (After Rescorla, 1968).

These results and others (e.g., Kamin, 1968) prompted the development of more sophisticated associationist models, incorporating more elaborate ways of adjusting the association strength between the representations of the CS and the US (e.g., Rescorla & Wagner, 1972; Mackintosh, 1975; Pearce & Hall, 1980). Many of these accounts implement the $\Delta P$ rule (Cheng & Novick, 1992), i.e., the CS-US relation in Rescorla’s experiment are learned when $P(\text{shock} | \text{tone}) - P(\text{shock} | \text{no tone}) > 0$. This is the constraint on the degree to which CS-US relationships is learned. However, there may be more to prediction than this statistical sense. For example, the door to our coffee room, which is supposed to be locked, is unlocked around 80% of the time when I want a coffee, so the conditional *if I push the door, I get into the coffee room* is a good rule for me to follow because $P(\text{I get into the coffee room} | \text{I push the door}) = 0.8$, i.e. it is quite high. Moreover, I will be happy to predict that I get into the coffee room given I push the door with probability 0.8. Now it may be the case that 80% of the time when I don’t push the door, I still get in because the door has been propped open, i.e., $\Delta P = 0$. However, I am still going to push the door when it is closed predicting with a degree of belief of 0.8, or thereabouts, that I will get in.

This distinction mirrors the difference between instrumental vs. classical conditioning. In performing an action to achieve a goal, an animal is mainly concerned with the probability of what happens when it performs the action. An animal acting in the real world simply does not have the time to hang around observing what happens when it doesn’t perform the action. Only animals in Skinner boxes have this luxury provided by their captivity. On the other hand, classical conditioning depends on observing the pairings of events to see if one predicts the other. Consequently, what happens when the putative cause does not occur is important. Animate causation, the notion that things only happen in the world as a result of our or other organisms actions upon it was the Scholastic model handed down by Aristotle which was overthrown by modern science. But it suggests that the naive or folk concept of a good predictive relation in the world, that might be described by a conditional, *if $p$ then $q$*, is given solely by $P(q|p)$. The more protracted process of picking up predictive causal relations in classical conditioning involves observation of cases where $p$ does not occur and as for cocks crowing and
the sun rising, this may require interventions of which most animals are not capable.

This idea that two quantities are important in prediction, which may be different in the instrumental and classical case, is related to a similar proposal concerning argument strength (Oaksford & Hahn, 2007). If we describe a predictive relation using a conditional, \( p \Rightarrow q \), these quantities are \( P(q|p) \) and something like \( \Delta P \) (see Hattori & Oaksford [2007] for a list of further contenders).

The proposal that System 1 processes are associationist needs some modification by these kinds of results, i.e., the information that is stored in conditioning may require information about \( P(q|p) \) and \( P(q|\neg p) \) (\( \neg \) = not). As Over, Hadjichristidis, Evans, Handley, and Sloman (2007), have observed, these quantities both affect people’s conditional degrees of belief but to very different extents. The characterisation of Oaksford and Chater’s (2007) probabilistic approach as requiring just System 1 (Evans & Over, 2004) suggests that verbal reasoning in humans can be carried out using just the mechanisms that allow a non-linguistic organism to learn about and predict its environment to guide its actions. I now argue that the results derived from work on stimulus equivalence suggests that this is probably not the case and that there is a profound distinction between the processes involved in verbal reasoning and those involved in animal learning.

**Stimulus Equivalence**

We now look at some work in the learning literature which indicates that what separates animals and humans, what make them mere learners but us “reasoners”, is the ability to make limited abductive inferences and that this ability may, perhaps not surprisingly, be related to language use. We begin by examining the task where most of the interesting evidence has emerged - the “equivalence task”. This is the experimental paradigm which seems to mark an important extra-linguistic behavioural discontinuity between humans and animals. In the following section we suggest that this discontinuity points to a direct link between the emergence of abductive reasoning and language development.

Within the framework of the experimental analysis of behaviour (Skinner, 1938, 1969, see also, Lowe, Richelle, Blackman & Bradshaw, 1985; Blackman & Lejeune, 1990) work on stimulus equivalence with humans (Sidman, 1971; Sidman & Tailby, 1982) has been prominent. We will return to the definition of stimulus equivalence after describing the experimental paradigm used in testing for its presence or absence.

The paradigm used is a variant on the matching-to-sample procedures common in many areas of behavioural research with both animals and humans. The basic task is illustrated in Figure 2. A five window array is employed which

**Figure 2** The basic structure of the matching-to-sample task. A sample first appears in the centre box of the five window array (a). After touching the centre box, the comparison stimuli appear in the outer boxes (b) and the subject must learn to select the comparison which the experimenter has decided goes with the sample.
may be presented on a computer touch screen. Stimuli are presented in the windows and subjects must learn various arbitrary associations. For example, they may have to learn that Y goes with RED. The symbol “Y” first appears in the central sample window. Once a subject touches the Y, the comparison stimuli appear in random positions in the outside comparison windows. The subject must then touch the correct comparison to obtain a reward.

What interests behaviour analysts about this task is that once these relations have been trained, certain emergent and untrained behaviours seem to come for free. For example, if without further reward, the stimuli are reversed, so that RED now appears in the sample window and “Y” and another arbitrary symbol “X” now appears in the comparison windows, subjects typically select “Y”. So if you train $Y \Rightarrow \text{RED}$, you seem to get $\text{RED} \Rightarrow Y$ for free! A further emergent behaviour arises as follows. If $Y \Rightarrow \text{RED}$ and $\text{RED} \Rightarrow \$ are trained, then when $Y$ appears in the sample window and $\$ and £ in the comparison windows subjects typically select $. These emergent behaviours are described as symmetry and transitivity respectively because of the obvious connection to the logical properties of relations. A relation that is symmetric, transitive and reflexive ($Y \Rightarrow Y$) is an equivalence relation, hence the “equivalence” task.

Sidman (1990) has argued that, while the definitive tests have not been carried out, on balance the evidence is in favour of equivalence not being a uniquely human phenomenon. However, Dugdale and Lowe (1990) argue that the standard equivalence procedures were not carried out in any of the studies cited by Sidman (1990). These studies have failed to demonstrate symmetry with pigeons (e.g., Lipkens, Kop & Matthijs, 1988), cebus monkeys (D’Amato, Salmon, Loukas, & Tomie, 1985), and rhesus monkeys or baboons (Sidman, Rauzin, Lazar, Cunningham, Tailby & Carrigan, 1982). Dugdale & Lowe (1990) therefore attempted a definitive test by using the language-trained chimpanzees (Pan troglodytes), Sherman, Austin and Lana at the Language Research Centre in Atlanta (Rumbaugh, 1977; Savage-Rumbaugh, 1986). Performance on the symmetry test trials were at chance levels for all three chimps. Dugdale and Lowe (1990) concluded:

“If one of our closest relatives in the animal kingdom is unable to pass these tests, then what chance is there for our more distant relatives? What is of even greater significance is that the chimpanzees we tested are unique among all animals in that they have a history of training that is quite unprecedented in its extent and complexity. Sherman and Lana are arguably the most “test-wise” of all animals. Before this experiment they must surely have ranked among the favourites to be the first non-humans likely to pass a standard symmetry test. Despite all this, however, they did not.” (Dugdale & Lowe, 1990: 123).

This evidence is quite striking because if symmetry had been observed, the most plausible explanation would have been the test-wiseness of the subjects. The failure to observe symmetry even with these subjects is therefore all the more compelling as evidence that symmetry is a uniquely human phenomenon.

Not only is symmetry apparently uniquely human it also appears to be tied to the development of language. There are two sources of evidence for this conclusion. The first derives from studies with mentally handicapped children (Devany, Hayes & Nelson, 1986). Devany et al. (1986) investigated the equivalence task performance of 2–4 year old children with learning difficulties who showed spontaneous speech and signing, and 2–4 year old children with learning difficulties with no verbal skills. Their performance was compared with normal 2 year old controls. The normal controls and the learning difficulty/language group passed the equiv-
The second source of evidence derives from developmental work. Within the experimental analysis of behaviour tradition it has been shown that some classical animal learning phenomena are in evidence in pre-linguistic infants (Lowe, Beasty & Bentall, 1983; Bentall, Lowe & Beasty, 1985). For example, on fixed interval reinforcement schedules, where, e.g., a rat may be reinforced every minute, animals reveal “scalloping” in their cumulative record (an additive record of responses, e.g., bar presses, plotted against time). That is, after reinforcement their rate of responding drops off gradually, reaches a minimum and then picks up again as the time for the next re-enforcer comes around. Humans, on the other hand, do not show scalloping. Instead they appear to adopt one of two hypotheses. Subjects may believe that they have to generate a certain number of responses during the period to receive the re-enforcer, this leads to a consistent and high response rate, i.e., no scalloping. Alternatively, subjects may believe that they will be re-enforced after a fixed time if they respond, this leads to a consistent and low response rate, i.e., again no scalloping. Pre-linguistic infants, however, do reveal the classical FI scallop. Moreover, they do not show symmetry on equivalence tests (Beasty & Lowe, 1985; Lowe & Beasty, 1987).

However, there have been some more recent reports of potential counter-evidence to the reliance of symmetry on language. For example, Schusterman and Kastak (1993, 1998, see also Schusterman, Reichmuth, & Kastak, 2000) have shown that sea lions seem capable of establishing equivalence relations and related cognitive abilities. Carr, Wilkinson, Blackman, and McIlvane (2000) have also shown that people with little linguistic ability could establish equivalence. Moreover, in their review paper, O’Donnell and Saunders (2003) also concluded that ability for language is perhaps not a necessary condition for establishing equivalence relations. However, despite this evidence, profound quantitative differences would appear to remain between humans and nonhuman animals in the flexibility and arbitrariness (such as across modalities) of equivalence relations they can establish. Consequently, the hypothesis that language marks a major shift in this ability, if not a hard and fast qualitative division, can still be sustained.

Having reviewed some of the evidence on the equivalence task, I now suggest that there is a plausible logicist interpretation of these results.

### Logic and Stimulus Equivalence

A variety of possible explanations of the equivalence task are discussed by Sidman (1990). He discusses the possibility of an inferential account first, i.e., the possibility that these behaviours emerge because the relationships trained are regarded as instances of rules possessing inferential relations with which subjects could be expected to be familiar. However, he dismisses this view because many relations, for example “is greater than”, are not symmetric and many are not transitive, for example “lives next door to”. In consequence there is no reason to assume that an arbitrary relation possesses both of these properties.

More importantly with respect to the emergence of reasoning, Sidman (1990) also argues that the logic of the conditional, \( p \rightarrow q \), cannot explain this behaviour. This is certainly the case if the contingencies trained are regarded as rules that are interpreted in terms of the logic of material implication (\( \rightarrow \)). Only two valid inference rules attach to the material conditional, *modus ponens* (MP) and *modus tollens* (MT):

\[
\text{(MP)} \quad \frac{p \supset q, p}{\therefore q} \\
\text{(MT)} \quad \frac{p \supset q, \neg q}{\therefore \neg p}
\]

MP can capture the standard predictive relation learned during the matching to sample procedure but MP and MT cannot explain symmetry. However, if these relations are interpreted as *biconditionals* (“\( \equiv \)”), i.e., \( p \Rightarrow q \text{ AND if } q \Rightarrow p \)
q then p, as many researchers argue may occur in other experimental contexts (e.g., Johnson-Laird & Byrne, 1991), then all these behaviours would be expected. The biconditional is material equivalence after all. So if Y $\Rightarrow$ RED is interpreted as a true biconditional — Y $\equiv$ RED — then both Y $\supset$ RED and RED $\supset$ Y are true by definition. Consequently MP on each rule gives the standard predictive relation and symmetry respectively. Moreover, if Y $\equiv$ RED and RED $\equiv$ $\&$ are true then so are Y $\equiv$ $\&$ and $\&$ $\equiv$ Y as can be readily demonstrated by truth tables (which we leave as an exercise for the reader). In sum, a familiar explanatory manoeuvre in the psychology of reasoning literature would seem adequate to account for the basic equivalence task findings.

This explanation is not countenanced by Sidman, who suggests that this would be to conflate an explanandum with an explanans. The emergence of logic is something to be explained, not something that can be invoked to explain some more primitive behaviour:

“Our experience ... [of] ... relations tells us whether they possess the properties of equivalence relations. This is not a case in which logical derivations account for behaviour; rather behavioural derivations justify the logic.” (Sidman, 1990: 105)

This is essentially a similar commitment to Lloyd Morgan’s canon that I mentioned in the opening paragraph.

However, many logicists (Fodor, 1975; Macnamara, 1986), who hold that logic is at the core of cognitive processes, suspect that logic is not acquired as a matter of experience but is rather innate. If so, then logic should be regarded as an un-analysable primitive — just as, for example, the reinforcing effects of a US are regarded as an explanatory primitive (Sidman, 1990). If this is the case then invoking logic in the explanation of these behaviours seems to be no more questionable than invoking the concept of reinforcement. That such a logicist position is not adopted by Sidman may seem paradoxical as he ultimately claims that equivalence is a fundamental stimulus function, i.e., an un-analysable primitive of learning theory. However, following my own commitment to Lloyd Morgan’s Canon, I also argue that the emergence of logic is a phenomenon to be explained at the lower level rather than a fundamental innate property that explains equivalence behaviour.

There seems to be at least three reasons for denying the innateness of equivalence either conceived of as a fundamental stimulus property or as a manifestation of an innately specified logical competence. First, if equivalence were a fundamental stimulus function then it should not be unique to humans. Since the balance of evidence favours an interpretation of equivalence as a species-specific human phenomenon, equivalence is unlikely to be a fundamental stimulus function. This conclusion survives the more recent evidence for limited equivalence behaviour in sea-lions and linguistically impaired humans. If equivalence were a fundamental stimulus function, there should be only small quantitative differences in this ability between species. Second, if equivalence is found in humans because of our ability to apply an innately specified logic, then pre-linguistic children should reveal symmetry. Unlike Universal Grammar (Chomsky, 1980), different linguistic communities do not require the parameters of their logic to be set by experience — the logic of the language of thought is wholly independent of experience (Fodor, 1975). In consequence, it should be in evidence in behavioural tests prior to the emergence of language. Since the emergence of equivalence is tied to the development of language it is not likely to be a manifestation of an innately specified logic. Finally, the explanatory manoeuvre of appealing to a biconditional interpretation is itself suspect. It is anomalous that a bi-directional, equivalence interpretation rather than the material conditional is adopted especially given that the material conditional is more consistent with subjects’ history
of reinforcement.

While the equivalence task cannot be explained in terms of prior experience with real world relations or in terms of innate logical relations, it may be explained in terms of peoples' experience of drawing eductive and abductive inferences. This is the possibility we now explore.

Prediction, abduction and language

From an inferential perspective, behaviour on the equivalence task can be very simply described in terms of eductive and abductive inference. Subjects are learning predictive relations such as \( Y \Rightarrow \text{RED} \) during the training phase and are being asked to make an abductive inference when tested for symmetry. When tested for transitivity they simply chain together causal rules in either a predictive or explanatory direction. But why is language apparently related to the ability to make appropriate abductive inferences? We now suggest that language and abduction are intimately related because language, or its preadaptations (Bickerton, 2000), allows the limitations of our physiology to be transcended in learning and forming representations of new predictive relations.

In learning a new predictive relation a limited form of abduction plays a crucial role. This is because most often a US must be associated with a CS that precedes it in time. The occurrence of the US is therefore a cue to abductively infer that the CS is a possible cause of the US. In the artificial environment of the laboratory, the relevant CS is contrived to be the most salient stimulus. However, in the normal environment salience and relevance judgements (Dickinson, 1979) have to be made in order to ensure that the most plausible cause of the US is selected from the preceding events that are potential causal candidates. i.e., an abductive inference is required. Of course, this is a potentially anthropocentric reconstruction of what a non-human animal may be doing. Such an animal is not performing a conscious inference but needs to carry out a procedure which has the characteristics of an abductive inference.

![Figure 3](image)

The temporal relationship between CS and US for conditioning to occur and its dependence on the decay function. The onset of the a CS and US causes the activation \( A(\text{cs}), A(\text{us}) \), of their corresponding representations to rapidly reach asymptote after which \( A(\text{cs}) \) and \( A(\text{us}) \) begin to decay back to zero. Conditioning can only occur if both the representations of the CS and the US are active, So, while CS2 will be conditioned to the US, CS1 will not be conditioned to the US.

The stimulus selected also relies on temporal contiguity. Take for example, a dog learning that a bell sounding (CS) predicts the arrival of food (US). If the CS precedes the US by more than a certain fixed time interval, then no learning occurs (Rachlin, 1976). This contiguity effect has been explained in terms of the decay functions of the neural representations of the CS and US (Sutton & Barto, 1981; see Figure 3). The bell sounding (CS) activates its neuronal representation causing its activation \( A(\text{cs}) \) to reach a maximum. Once the bell ceases, the activity of its neuronal representation gradually decays. Only if its representation is still active will the activation of the neural representation of the US \( A(\text{us}) \) cause the strengthening of the connection between the representation of the CS and
the representation of the US. In figure 3 it can be seen that this depends on the decay function which is determined by the physiological characteristics of the neural hardware. When the stimulus is withdrawn the activity of its representation gradually decays. Only if the US is presented before the activity of the CS’s representation has decayed to zero can conditioning occur.

The physiologically determined decay function would appear to place a temporal limit on an organism’s ability to draw abductive inferences, i.e., to maintain representations of possible candidates for the cause of the US. If the US occurs more than a certain time after the CS (4 seconds, in the case of the rabbits’ nictating eyelid response, for example), then the representation of the CS won’t be available to support the abductive inference that the appearance of the US may be explained by the CS’s prior occurrence. There are several species-specific adaptations that, however, allow this limited abductive capability to be transcended. For example, the appetitive behaviour of rats would appear to allow them to abduct to the correct cause of sickness many hours after the ingestion of the food that caused it (Garcia, Ervin & Koelling, 1966). Nevertheless, animals do not seem to have a general solution to the temporal contiguity problem.

With such a narrow temporal window on the past, animals are limited in their ability to learn predictive relations. Many predictive relations that are learned involve instantaneous cause-effect relations, like the impact of a thrown stone causing a pane of glass to shatter (Chater & Oaksford, 2006). However, many other relations that humans have clearly learned are not like this, e.g., if you put manure on a plant, it will grow better; if you combine flower, eggs, and sugar in equal quantities and heat them you’ll get a cake. A great deal of “folk” knowledge of agriculture, cooking and other practical skills must have emerged from the ability to keep track of possibly many chance combinations and examining their consequences some time later. To learn these more long distance dependencies will clearly have required our ancestors to hold onto the past for more extended periods of time than the decay functions on their neural hardware will allow.

However, a caveat is required here. For periodic behaviours, animals can exploit endogenous oscillations in the nervous system that occur at many different time scales, e.g., circadian rhythms, in order to time an appropriate response (see, e.g., Gallistel [1990] for an extensive review). These endogenous oscillations can be entrained to initiate a response at the appropriate time.

A more general purpose mechanism may be to elicit an explicit endogenous response signalling the presence of salient events. Such responses may be repeated or rehearsed thereby maintaining the representations of these events over time. A suitable endogenous response is the emission of a sound. If this sound is associated with the representation of a particular stimulus then emitting the noise may activate the representation in the absence of the stimulus itself. In this way the representations of particular events (CSs) may be kept active for indefinite periods, thereby allowing them to become appropriately associated with representations of their effects (USs) independent of the temporal response characteristics of the underlying neural hardware. The ability to associate a sound with a stimulus may mean that the temporal window in which an organism exists may be extended into the past and into the future in a way that transcends the physiological response properties of individual neurons.

The ability to associate an endogenously produced response, i.e., a sound or a name, may therefore function to extend an organism’s abductive abilities. Phylogenetically, the ability of an organism to extend its learning capacities by using such “secondary signalling systems” (e.g., Vygotsky, 1962) may provide the pre-adaptation that underpins the development of language and
interpersonal communication. Ontogenetically, the emergence of language may provide the resources required to enable more efficient learning of the contingencies in the world. As this ability develops more explicit abductive reasoning would be expected to emerge such that the neonate begins to employ the contingencies it has learned to make explicit abductions. This would lead to the expectation that the ability to pass symmetry tests, i.e., the ability to make abductive inferences, should be correlated with linguistic ability. And this is just what has been observed in the work on the equivalence task we have reviewed. Moreover, consistent with this position, Dugdale & Lowe (1990) have also demonstrated that when children are found who are unable to pass symmetry tests explicit instructions to name the stimuli facilitates this ability.

The view that naming is central to the emergence of equivalence classes has been prominent in the behaviour analytic literature and was originally suggested by Horne and Lowe (1996). This theory is framed in the language of behaviour analysis, i.e., it attempts to extend the limited conceptual framework of Skinner’s *Verbal Behaviour* to account for the relation between words and what they denote. In contrast, the present account sees naming as a mechanism for maintaining mental representations of objects or events in the world over extended periods of time. Naming may not have evolved to serve this function but from a cognitive perspective it is a consequence of using sound to communicate that names should have this effect. While preferring to eschew “just so” speculation about what may or may not have occurred on the Savannah 200,000 years ago, one can imagine naming as arising from warning calls. Initially at least, an arbitrary noise cannot possess a warning function unless it is paired with the presence of a fear eliciting US or it can elicit a representation of an object which already elicits a fear reaction. From a cognitive perspective, learning sound-object pairings so that the sound can be used in social communication entails that the sound elicits, and so can be used to maintain, a representation of the object the sound denotes. This maintenance function must in turn be the principle way in which the status of the sound as a warning is subsequently learned in a social group. A neonate must learn to flee without necessarily experiencing pairings of the sound with the actual US (a predator). However, they can only learn to make the appropriate sound if at some later time they can pair the sound, which they initially produce by spontaneous imitation, with the subsequent appearance of the US. Of course, other calls may intervene and be imitated and rehearsed and so serve as possible candidates to be paired with the US, but only the warning call specific to that predator should reliably predict its appearance.

More temporally separated events can then be learned by similar mechanisms. For example, it may be that the presence of the predator forces various other possible prey animals to take cover. These latter organisms might also be suitable prey items for our troop of primitive language users. Consequently, learning the long distance relationship between the presence of the predator and the subsequent availability of food would be very useful. None of this would be possible unless there were some mechanism of keeping representations active beyond the temporal limits of the decay function of our neural hardware.

**System 1 and System 2 Revisited**

Extending the ability to learn by using endogenous responses to extend an organism’s temporal window on the world further into the past still does not yet provide a reasoning system from which logical abilities could emerge. The system thereby created allows an organism to keep representations of external environmental events alive for longer. We now suggest that the next step in the emergence of System 2 is turning this ability inwards to retain the results of interrogat-
ing System 1, which contains the products of the learning process in long term memory for world knowledge.

Figure 4 shows a possible implementation of dual process theory (Evans & Over, 2004). The operations over long term memory (LTM) for world knowledge are the System 1 processes. In the System 1 box we show a simple system of two alternative dependencies, \( p \Rightarrow q \) and \( r \Rightarrow q \), which are context sensitive, that is, if the defeaters, \( d_1 \) or \( d_2 \), are present \( q \) does not occur. Here we represent this knowledge as a constraint satisfaction neural network. The connections with arrowheads are facilitatory connections and those with circular ends are inhibitory connections. Thus both \( p \) and \( r \) will activate the \( q \) unit, whereas \( d_1 \) and \( d_2 \) will inhibit it. We can think of the dependencies under the following translation: \( p \): turn key, \( q \): car starts, \( r \): jump start, \( d_1 \): petrol tank empty, \( d_2 \): battery flat. This is the knowledge acquired by learning about starting cars.

Once information about the external world can be retained over extended periods, then this trick can be turned inwards to store the results of interrogating LTM. So in System 2, the first line shows the result of interrogating LTM with the probe \( p \), i.e., clamping on the \( p \)-node, the result is the activation of \( q \)-node. By the association of names for these representations the result of this interrogation can be stored in working memory. For example, the result of the first interrogation tells you that from what you know if you turn the key (\( p \)) the car starts (\( q \)). Another possibility you might want to explore is under what conditions you can turn they key while the car does not start. This is shown in the second line in System 2 which represents what happened when the \( p \)-node was clamped on and the \( q \)-node clamped off, i.e., the \( d_1 \)-node came on.

System 2 allows organisms that possess it to explore the possibilities implicit in their knowledge of the dependencies in the world which they acquired by learning. Oaksford and Chater (in press) show how System 1 can be provided with a probabilistic interpretation and how the representations in System 2 may be annotated with probabilities to show how probable each possibility may be. Of course, in the psychology of reasoning the notion that we reason by considering the possibilities that a linguistic expression of a dependency may allow is a common place, i.e., it is the core of the mental models theory of reasoning.

Once such a system is in place it is perhaps only a short step to proposing that people can on occasion simply use the surface linguistic expression of a dependency to recreate just the possibilities that are usually stored in System 2 without actually interrogating System 1. At this point, purely logical reasoning can perhaps emerge.

While at an early stage in looking for evidence of such an account of the origins of System 2, such a hypothesis makes various testable predictions. For example, comparatively, it would
be predicted that the between species quantitative difference in equivalence performance may be correlated with the use of complex sounds for communication and this may in turn be related to an animals’ problem solving/reasoning abilities. Moreover, phonological abilities should correlate with reasoning and problem solving abilities in humans, which is seems to be the case (Gilhooly, Logie, & Wynn, 1999). However, the critical studies remain to be conducted.

Conclusion
In this paper, I have argued that the results on Sidman’s stimulus equivalence paradigm provides the key link in showing how System 2 may have emerged from System 1. System 1 which results from the basic process of learning allows predictive inference or deduction that allows organisms a limited ability to predict their environment. The matching to sample paradigm showed that full stimulus equivalence behaviour seems to be limited by linguistic ability. I argued that this is because animal learning processes rely on a limited form of abductive inference which requires keeping candidate causes of an effect in mind. However, I argued that this process in temporally bounded by the decay characteristics of the brain’s neural hardware. It was suggested that organisms that have the ability to generate endogenous stimuli, e.g., sounds, that can be associated with representations of external events and that can be rehearsed have an adaptive advantage in being able to maintain representations of possible predictors in mind for longer. Being able to maintain representations in mind also allows people to interrogate their long term memory for world knowledge and store the results as representations of different possibilities over which further, more abstract reasoning processes may develop.

References


of the Experimental Analysis of Behavior, 80, 131–157.


(Received 12 April 2008)
(Accepted 2 June 2008)
Mike Oaksford
MIKE OAKSFORD is Professor of Psychology and Head of School at Birkbeck College, University of London. He has authored over one hundred scientific publications in psychology, philosophy, and cognitive science, and has written or edited six books. He is Consulting Editor for Psychological Review, Memory and Cognition, and Thinking and Reasoning and has been Associate Editor for QJEP and Consulting Editor for JEP:LMC. His research explores the nature of reasoning, argumentation, and the influence of emotion on cognition.