The Discrimination and Representation of Relative and Absolute Number in Pigeons and Humans

A thesis

submitted in fulfilment

of the requirements for the degree of

Doctor of Philosophy

in Psychology

by
Lavinia Chai Mei Tan

2009

Keywords: bisection, counting, numerical competence, numerosity discrimination, numerical representation, pigeon.

Acknowledgements

It would not be possible to thank all the people have helped me throughout my PhD sufficiently; however, there are several individuals who have been most influential in helping me reach this point and shaping my experience as a PhD student, and whose contributions I must acknowledge.

First, I would like to thank my supervisors, Randy Grace and Anthony Mclean, both of whom are responsible for igniting the behavioural spark in me, and have kept it burning with their support and inspiration over the years. My PhD experience would have been infinitely duller were it not for Randy's boundless wisdom and his contagious, endless enthusiasm for research, data and the general pursuit of knowledge; I have gained so much from his guidance and mentorship in matters experimental, analytical and academic. Also, I would not have ventured into behavior analysis were it not for Anthony's initial encouragement and interest in my undergraduate studies, and I am inexpressibly grateful for Anthony's dependable background presence as well as his always sagely, well-timed advice and aid.

I would also like to acknowledge other faculty and staff, in particular Neville Blampied, Rob Hughes, and Paul Russell, for their interest, assistance and support, and the animal technicians, Trish Meatchem, Neroli Harris, Silvana Da Costa, Fiona Patterson, who play an invaluable role in caring for the operant laboratory and its inhabitants.

I would not be without the friendship of my peers and colleagues: particularly Liz Kyonka, who has counseled, celebrated, and commiserated with me; Abby Morgan for her assistance and support inside and outside of university; Mark Berg, who showed me how Experimenter works; and of course Celia Lie and Lincoln Hely, whose comradery, although largely remote, is nonetheless appreciated. It is a real privilege to be part of the NZABA family, and I am grateful for the support and guidance of Doug Elliffe, Brent Alsop and Dave Harper.

Lastly, but not leastly, I would like to thank all my friends and family for their support and tolerance during my PhD, I would not have made it were it not for their love and patience.

Abstract

THE DISCRIMINATION AND REPRESENTATION OF RELATIVE AND ABSOLUTE NUMBER IN PIGEONS AND HUMANS

By Lavinia Chai Mei Tan

University of Canterbury

The ability to discriminate relative and absolute number has been researched widely in both human and nonhuman species. However, the full extent of numerical ability in nonhuman animals, and the nature of the underlying numerical representation, on which discriminations are based, is still unclear. The aim of the current research was to examine the performance of pigeons and humans in tasks that require the discrimination of relative number (a bisection procedure), and absolute number (in a reproduction procedure). One of the main research questions was whether numerical control over responding could be obtained, above and beyond control by temporal cues in nonhuman animals, and if so, whether it was possible to quantify the relative influences of number and time on responding. Experiment 1 examines nonhuman performance in a numerical bisection task; subjects were presented with either 2 and 6, 4 and 12, or 8 and 24 keylight flashes across three different conditions, and were required to classify these flash sequences as either a "large" or "small" number, by pecking the blue or white key, respectively. Subjects were then tested with novel values within and 2 values higher and lower than the training values. Experiments 2-4 investigate responding in a novel numerical reproduction procedure, in which pigeons were trained to match the number of responses made during a production phase to the number of keylight flashes (2, 4, or 6) in a recently completed sample phase. Experiments 2 and 2A examined discrimination performance when the temporal variables, flash rate and sample phase duration, were perfectly correlated (Experiment 2) or only weakly correlated (Experiment 2a) with flash number. Acquisition of performance in the numerical reproduction procedure was investigated in Experiment 3.

For Experiments 1-3, hierarchical regression analyses showed significant control by number over responding, after controlling for temporal cues. Additionally, positive transfer to novel values both within and outside the training range was obtained when the temporal organization of test sequences

was similar to baseline training. Experiment 4 investigated the effects of increasing or decreasing the retention interval (RI) on performance in the reproduction procedure, and found this produced a response bias towards larger numbers, contrary to predictions based on previous RI research, and suggested responding was not affected by memorial decay processes. The structure of the representation of number developed by subjects in the bisection and reproduction procedures was investigated using analyses of responding and response variability in Chapters 2 and 6, respectively. Bisection points obtained in Experiment 1 were located at the arithmetic, not geometric mean of all three scales, and coefficients of variation (CVs) obtained in both the bisection and reproduction experiments tended to decrease as flash number increased. Additionally, analyses of the acquisition data found differences in average response number was better fit by a linear than logarithmic scale. These results show that responding did not conform to scalar variability and is largely inconsistent with previous nonhuman research. Together these results suggest responding appeared to be based on a linear scale of number with constant generalisation between values, similar to that associated with human verbal counting, rather than a logarithmic scale with constant generalisation or a linear scale with scalar generalisation between values. Experiment 5 compared pigeons' and humans' verbal and nonverbal discrimination performance with numbers 1-20 in analogous bisection, reproduction and report tasks. Human verbal and nonverbal performance in the three tasks was similar and resembled nonhuman performance, although verbal discriminations were more accurate and less variable. The main findings from Experiments 1 and 2A were replicated with humans; bisection points were located at the arithmetic mean, average response number increased linearly as sample number increased, though there was a tendency to underestimate sample number, and decreasing CVs were also obtained for values less than 8. An additional, interesting finding was that CVs showed scalar variability for values greater than 8, suggesting a less exact representation and discrimination process was being used for these values. Collectively, these five experiments provide new evidence for a nonverbal ability to discriminate relative and absolute number with increasing relative accuracy resembling human verbal counting in both human and nonhumans.

Table of Contents

List of Figures

List of Tables

1 Chapter 1: Introduction

Can nonhuman animals understand number? What is the extent of their numerical abilities? Researchers have been investigating numerical competence in a wide variety of animal species for about 100 years. There are a variety of reasons why nonhuman numerical abilities are worthy of investigation. Animals may have evolved an ability to discriminate number due to the ecological advantages it affords; the ability to monitor the number of predators and competitors, and the quantity of food in foraging patches and the number of young would likely enhance an individual's survival and reproduction rate. Additionally, higher-order numerical abilities in nonhuman animals may require the development and comprehension of a complex, abstract, concept, and it is unclear whether a capacity for language is necessary for this. When considering the evolutionary continuum of cognitive abilities, it is highly likely that human and nonhuman numerical capacities may overlap somewhat. This possibility is worthy of investigation, and some research has studied similarities and differences in numerical understanding in adult and young humans and nonhuman animals, in particular asking whether nonhuman abilities can resemble those possessed by adult humans or children, and whether discriminations are based on similar numerical processes in both humans and nonhumans.

The investigation of numerical competence has intrigued researchers with a wide range of backgrounds, including comparative psychology, development psychology, ethology and learning theory, cognitive psychology, and more recently neuropsychology. Consequently, paradigms and methodologies used in this area are somewhat eclectic, borrowing and applying approaches from the different areas. For instance, the habituation-discrimination procedure, commonly used in developmental psychology (Fantz, 1964; also see Cohen & Cashon, 2003 for review) has been used to investigate spontaneous numerical understanding and ability (e.g. in infants, Starkey, Spelke & Gelman, 1990; Wynn, 1992; in nonhuman animals, Hauser, Macneilage & Ware, 1996; Flombaum, Junge & Hauser, 2005, Hauser & Carey, 2003), while research in the animal laboratory has adapted

timing and category learning match-to-sample procedures to investigate responding in numerical discrimination tasks. Although the heterogeneity of research in this area makes direct comparison and comprehensive evaluation of the range of experiments somewhat difficult, the diversity can be seen to reflect the complex nature of their unifying focus, the understanding of number.

The current research aims to provide a systematic investigation of the ability of nonhuman animals and humans to perform a variety of numerical discriminations. The performance of pigeons in relative and absolute numerical discrimination tasks, and the acquisition of performance in the latter, are examined in experiments presented in Chapters 2-4. Chapters 5 and 6 examine the possible cognitive processes underlying these numerical judgments, through manipulations of retention interval and analyses of response variability. Finally, performance of humans and nonhumans in analogous tasks is compared to investigate similarities and continuity in numerical abilities in Chapter 7. Chapter 8 summarises and integrates the main findings, introduces a numerical discrimination model that is able to account for much of the results, and provides suggestions for future research.

This chapter introduces general theories and principles relating to numerical research. Section 1.1 describes and characterises the range of numerical abilities recognised and investigated in humans and nonhumans animals. Section 1.2 discusses the requirements and principles of true counting behaviour that have been proposed by Davis and Perusse (1988). Section 1.3 examines theories of numerical representation and describes three current models of numerical discrimination. The investigation of numerical competence in nonhuman animals and the challenge of isolating numerical control over behaviour relative to other potentially confounding cues in this type of research is discussed in Section 1.4. Finally, Section 1.5 provides a summary of the current research.

1.1 Numerical abilities

In order to begin examining what numerical abilities nonhuman animals may possess, an outline of the abilities recognised in humans,and their associated prerequisites is necessary.

Numerical abilities range from simple to very complex, and although these can be difficult to characterize precisely, they can be broken down to four main processes situated along a continuum of varying difficulty (Davis & Perusse, 1988).

The simplest type of ability is the ability to discriminate relative numerosity. Relative numerosity discriminations are dichotomous "more or fewer" judgments that do not necessarily require, but do not exclude an understanding of absolute number. This ability may serve as a base skill for more complicated numerical abilities. A large amount of research has been dedicated to relative numerosity judgments in nonhuman animals, as it is likely they would possess this ability. Numerical ability is of significant biological importance; survival would be greatly enhanced by the ability to determine and select a foraging patch containing more food and fewer predators (Honig & Stewart, 1989). Although a true concept of number may not be required to discriminate successfully the relative numerosity of any given two sets, testing over many trials with several different values may be informative about numerical concepts and representation.

Following the ability to discriminate relative numerosity are more complex skills that involve the discrimination of absolute number. Subitising is the next numerical process, and is slightly more complicated than relative numerosity discriminations. This ability involves the near-instantaneous tagging or labelling of stimuli and is generally limited to up to six items (Davis & Perusse, 1988). Subitising appears to be more of a perceptual than cognitive process, and may rely more on pattern recognition than actual numerical discrimination in humans (Von Glaserfield, 1982). Generally, subitising is a process that can only be applied to simultaneously-presented visual stimuli. However, it is entirely possible that it can be applied to stimuli presented sequentially in alternative modalities as well. For instance, most people are capable of discriminating and reproducing rhythmic regularities without any labeling or counting, e.g. the fa-la-la-la-las in the Christmas carol, "Deck the Halls" (Davis & Perusse, 1988).

It has been proposed that the understanding of the association between a numerical label

and the subitised set does not occur until later in cognitive development (Wynn, 1992a), however there is some disagreement; others have argued that counting ability must be established before subitising, with numerically meaningful tags, can occur (Gelman & Gallistel, 1978). If subitising provides a necessary foundation for the development of more complex numerical abilities, children learning to count should show some proficiency for instant recognition of small numbers. However Gelman and Gallistel (1978) found no evidence supporting this hypothesis, and concluded that subitising, in humans at least, is a post-counting ability.

Estimation is a similar process to subitising, involving the rapid enumeration of larger quantities of generally more than 6 items. It has been argued that the same perceptual processes are used in estimation and subitising(Kaufman et al., 1949), however this seems unlikely given the possible variations in patterns with larger quantities. A differing view is that estimation is a much more complex ability that does not develop until after subitising and counting skills have been acquired; one must be able to count up to that number in order to estimate it effectively, and consequently this requires an understanding of cardinality and ordinality (Klahr & Wallace, 1973).

Counting is the precise discrimination of the absolute number of any given set of items. There is a substantial leap in complexity between the previously mentioned numerical skills and counting, consequently specific criteria must be met before these skills can be attributed to counting.

1.2 Principles of counting

Gelman and Gallistel (1978) proposed five principles that are necessary for counting to occur. The first four principles involve the procedure of counting, or how to count. Firstly, each item to be counted must be associated with a distinct numerical tag. This principle is known as the *one-to-one correspondence principle*. The application of this principle involves two

processes, *tagging* and *partitioning*. Tagging is the application of a unique tag to each of the items as they are counted. Verbal labels, however, are not necessary; any series of symbols or behaviour may serve as tags. It has even been suggested that one process by which items are tagged may even be the activation of nodes in short-term memory (Davis & Memmott, 1982). Additionally, Köhler (1950, in Davis and Perusse, 1988) proposed that animals might use a system of inner marks to 'think unnamed numbers'. Partitioning occurs once an item has been tagged. Partitioning is the separation of items that have been counted from those that are yet to be counted and involves the transfer of the counted item from one category to the other during the counting process. An item has been counted when a tag is applied to that item and consequently is unavailable for subsequent use with that set. In order for a correct count to be achieved, partitioning and tagging must start at the same time and end together.

The *stable-order principle* requires that numerical tags be applied in a fixed order. Gelman and Gallistel (1978) noted that this principle is the most difficult for children to grasp. Because verbal tags are arbitrary in nature, a child's counting ability relies upon their capacity to remember a particular sequence of these verbal tags, a task that increases in difficulty as the set size increases.

The third principle is c*ardinality*; the final tag in a series represents the numerosity of the whole set. Besides being able to apply tags to a set of items in a fixed sequence, the individual must recognise the numerically descriptive property of the cardinal tag. It should be noted that an understanding of cardinality is necessary, but not sufficient for counting. It is possible to learn a set of tags without knowing about the ordered relationship between them.

The *order-irrelevance principle* states that items can be tagged in any order, as long as each item is tagged only once. Understanding this principle involves an implicit understanding that the cardinal number of a set remains the same for any order of enumeration, that each tag is temporarily applied to an item, and that the tags are independent of the items themselves.

The *abstraction principle* states that the preceding four principles can be applied to any

set of entities, whether physical or non-physical. This can be interpreted as entailing a true concept of number; an understanding that number is an abstract dimension that is not limited to any specific modality or item. Consequently, an understanding of the abstraction principle should allow the transfer of numerical discriminations between 1) simultaneous and sequential stimulus presentations; 2) sensory modalities; 3) perceived and performed numbers (Davis & Perusse, 1988).

Several problems arise in the attribution of counting abilities to nonhuman species. Firstly, there is contention over the requirement of a verbal representation of number, which is often assumed when using a human definition of counting. However, many enumerative processes would be excluded if this restrictive definition were used. Davis and Perusse (1988) proposed the term "protocounting" for situations where alternative numerical processes such as relative numerosity, subitising, estimation can be excluded as explanations of behaviour, but there is no evidence for true counting (involving an absolute sense of number). The term protocounting would also include some more advanced skills that fall short of counting, such as absolute number discriminations.

1.3 Numerical representation and models of counting

A number of procedures used in research on numerical competence involve the presentation of different stimuli, which vary in numerosity, and an associated response, which may differ depending on the numerosity of the stimulus presented. The stimulus presentation and response phases also may be separated by a brief delay. Accurate discrimination of number and appropriate responding to number in these tasks would require a mechanism that allows the perception and input of numerical information, the retention of that information in memory, and the mapping of number into the appropriate response output. In other words, it is likely a mental representation of number is developed and used to determine responding in numerical tasks.

Different models have been proposed to explain how number might be discriminated and represented nonverbally. Gelman and Gallistel (1978; Gallistel & Gelman, 1992, 2000) initially proposed that the process of nonverbal numerical representation resembled and adhered to the same principles as verbal counting- integer symbol models. Numerical values are represented as neuronal symbols, called numerons. Numerons are discrete and arbitrary, and organised in a list form. Enumeration involves assigning a symbol in the list to each presented item, with one-to one correspondence,and always proceeding in the same order. The critical aspect of their model is that both verbal and nonverbal counting representations are based on the same principles and operate over both small and large ranges of numbers. Gelman and Gallistel acknowledge that an analogue representational system would perform similarly to a symbolic system, as their features are similar; the mode-control model, described below, resembles an integer symbol model in many respects.

1.3.1 The mode-control model

One type of model that assumes an analogue magnitude-based representation of number is the mode-control, or pacemaker-accumulator model (Meck & Church, 1983), illustrated in Figure 1.1, below. This was developed through the application of an information-processing model for temporal discrimination to numerical discriminations in nonhuman animals (Meck, Church & Gibbon, 1985), but has also been applied to human counting behaviour (Gallistel & Gelman, 1992). The mode-control model consists of a pacemaker, a mode switch, an accumulator, working and reference or long-term memory and a comparator. The pacemaker generates pulses at a constant rate, which is gated into the accumulator part of the model by the mode switch. The accumulator value can be transferred into working memory and then stored in long-term memory when a response is reinforced. When making a numerousness judgment, a comparator can be used to compare the current accumulator value in working memory to an

exemplar or prototype value in long-term memory. The switch can operate in 3 modes as a response to a stimulus: 1) a run mode, where the switch stays open for the whole trial regardless of the duration of the stimulus; 2) a stop mode, where the switch stays open for an extended interval corresponding to the duration of the stimulus, such that the total pulses in the accumulator is a measure of duration; and 3) an event mode, where each stimulus presentation results in the switch opening and closing after a relatively fixed delay. Thus, this system can be used to measure duration by using the run or stop modes, or number by using the event modes, and assumes a common representation of time and number.

The mode-control model also conforms to four of Gelman and Gallistel's five (1978) criteria for counting. Each event/stimulus presentation results in a constant increment to the accumulator, conforming to the one-to-one principle. The accumulator must go through every smaller increment to reach a particular number, such that each "tag", or increment in the accumulator, is always applied in the same order, consistent with the stable order principle. The cardinality principle is conserved, as the accumulator value at the end of the enumeration process is always equal to the number of events. Finally, theoretically, the accumulator can be used to enumerate any sort of object or event, consistent with the abstraction principle.

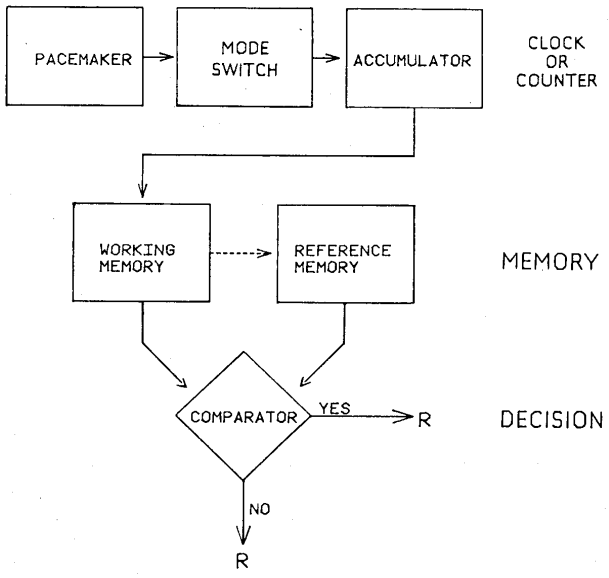

Figure 1.1. The mode-control model of timing and counting, from Meck and Church (1983)

The neural network model

Dehaene and Changeux (1993) proposed a neural network model to describe nonverbal numerical representation, shown in Figure 1.2. The core feature of their model was a numerosity detection system, which consists of three modules, 1) an input retina, which receives information about the size and location of the presented stimuli; 2) a topographical map in which each object is represented by a set pool of neurons, normalising size and configuration, and 3) a map of numerosity detectors which sums the outputs from the topographical map. To explain the discrimination of the numerosity of auditory stimuli, Dehaene and Changeux include an echoic auditory memory, which also provides input for the numerosity detectors. Thus, their model accounts for the discrimination of number in both visual and auditory modalities.

Up to five simultaneously presented objects can be represented on the input retina, which is then coded over a topographically organized sheet of input neuronal clusters which then project onto a two-dimensional topographical map, consisting of 9x50 sets of neuron clusters which code for location and normalises for size. The neuronal clusters in the map are activated by input in such a way that perceived objects of different size are represented by a similar number of active neuronal clusters. Each of these clusters then projects to the numerosity detectors.

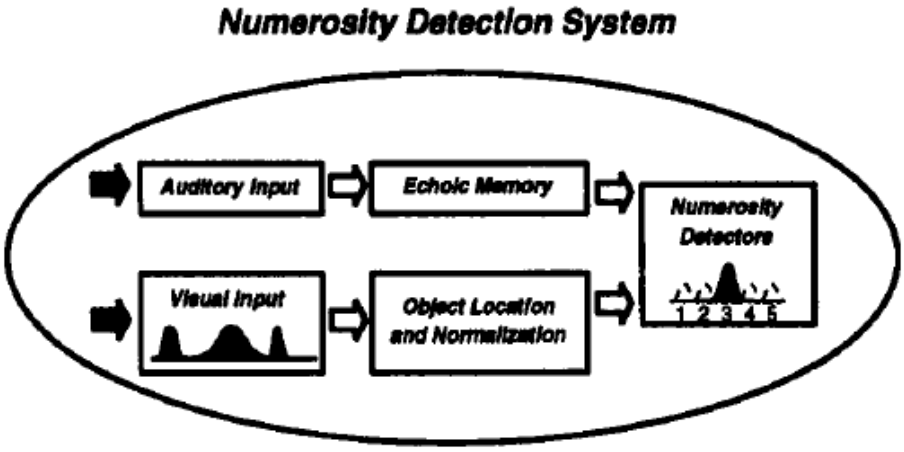

Figure 1.2. Functional description of the numerosity detection system of the neural network model, from Dehaene and Changeux (1993).

The numerosity detectors consist of two types of units, summation units, which receive the normalized visual or auditory input and relay that information to numerosity units via excitatory or inhibitory connections. Each summation unit sums the activity from the whole location map, which is proportional to the numerosity, and is activated if the total excitation exceeds the threshold; thus when the input numerosity exceeds a certain limit. Summation units project topographically to numerosity units, and connections are organized such that each numerosity unit is only activated when their associated summation unit is active, and summation units for higher numerosities are not. Thus, each numerosity unit only responds to a limited range of activation for the total normalized values, meaning they are only activated by input equalling that particular numerosity, not more or less.

The numerosity detection system then sends information to a motor output system, which produces the relevant response associated with the numerosity, determined by its association with an external reward input. That is, visual and auditory numerical stimuli elicit responding from the organism; those responses that are reinforced are strengthened, and those that are not are eliminated.

The modules described thus far provide the necessary components for the detection of numerosity. Dehaene and Changeux (1993) propose an additional memory module that allows the comparison of two numerosities (see Figure 1.3). One numerosity can be held in memory and compared with another numerosity being processed; a point-to-point matching module calculates the similarities (and differences) between the two numerosities, and connections with the motor output system allows the system to produce an output depending on the particular relations of the two inputs, e.g. respond if numerosity 1 is larger than numerosity 2.

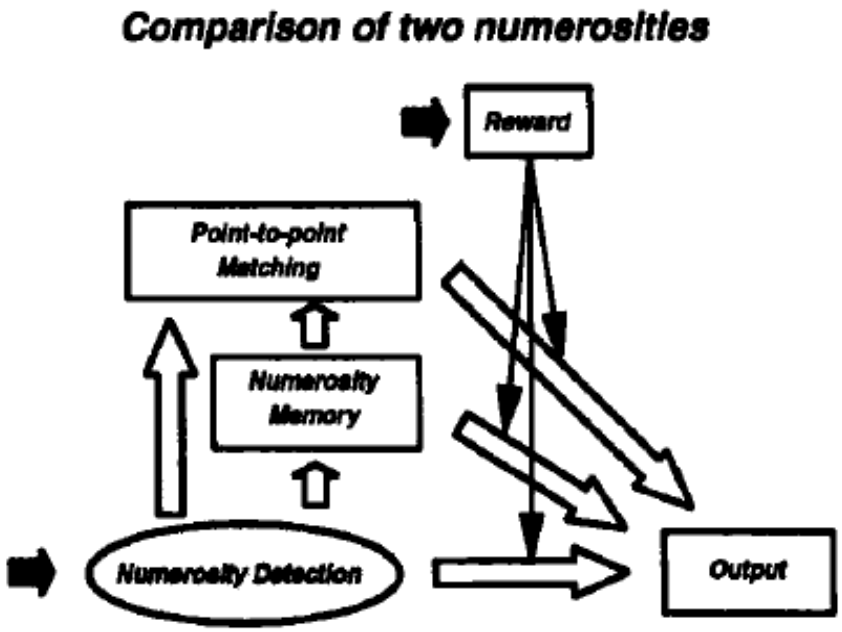

Figure 1.3. Functional description of numerosity comparison in the neural network model, from Dehaene and Changeux (1993).

Then, either using the numerosities stored in long-term memory or currently being presented, the auto-evaluation system tries to reconstruct the action (see Figure 1.4). An action-matching module then produces an internal positive or negative reward signal depending on whether the action was matched successfully, or not. Dehaene and Changeux claim that this process allows the system to discover that an increase in numerosity results in addition, whereas a decrease in numerosity results in subtraction. However, this reasoning seems circular given that addition and subtraction are, by definition, a respective increase or decrease numerosity; the system does not really "discover" anything and appears to be superfluous. It is unclear why an understanding of addition and subtraction cannot merely emerge from the re-comparison of numerosities before and after an object has been removed or added.

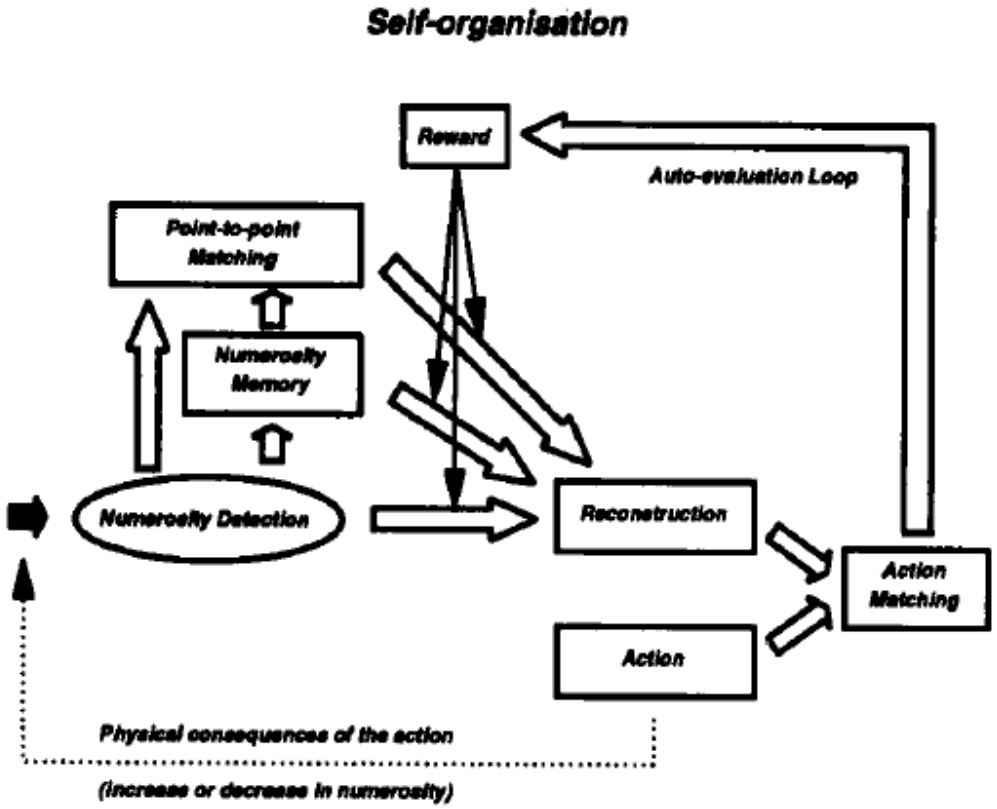

Figure 1.4. Functional description of self-organisation in the neural network model, from Dehaene and Changeux (1993)

The mode control and neural network models share two core similarities, both predicting analogue representations of number that conform to Weber's law. There are also some critical differences between the two models. The summation units in the neural network model are only active when the number of events exceeds a particular threshold, while the event mode in the mode-control model is always active, regardless of the number of events. Additionally, in the mode control model numerical representations are developed through a serial process, whereas in the neural network model, representations are perceived and acquired in parallel. The parallel processing of the neural network model would impose restrictions on the number of objects that can be perceived, due to limits in memory capacity, similar to those with the object-file model. The neural network model is more limited in other respects; unlike the mode-control model, it is limited to discriminations of auditory and visual stimuli, and cannot explain temporal discriminations.

The object-file model

It is generally agreed that analogue magnitude is the core representation for large numbers (< 8), but there is some disagreement about whether this is also true for smaller numbers, or whether a more exact numerical representation operates over this range. One such model of representation is the object-file model (Feigenson, Carey & Spelke, 2002; Feigenson & Carey, 2005; Feigenson, 2008).

Object files were originally proposed as an attentional mechanism for object tracking by Treisman (Kahneman, Treisman & Gibbs, 1992). Items are encoded individually into separate files which operate as the representation of items in the array (Simon, 1997; Uller, Carey, Huntley-Fenner & Klatt, 1999) and are not numerical representations *per se*, but are merely mental indicators which can be used for the discrimination of small sets. It is unclear how much information about objects' characteristics are included in object files; some versions of the model claim object files are imagistic, while others claim that they are arbitrary representations of items (Uller et al., 1999; Feigenson et al. 2002). Discrimination is accomplished by assessing one-to-one correspondence between the files in two models (e.g. in same-different comparisons), or between attentional indices and objects in the array.

Symbolic and object-file models differ in a variety of ways (Uller et al., 1999). They differ in how events are represented, and also the algorithm in which they detect differences between the representations being discriminated. The applicability of the object-file model is much more limited than symbolic models. Because object-files form part of the visual attentional system, this model is not applicable to the numerical discrimination of non-visual stimuli, whereas symbolic models of numerical representatation can theoretically represent number presented in any modality. In symbolic models, the number of both sets of items are stored in short term memory symbolically and compared by determining whether symbols match. Conversely, the object-file model constructs separate individual object files for each set

of items (e.g. two representations with two object files in each), and checks for one-to-one correspondence between the object files in the two representations. Consequently, the memory demands for the object-file model are much greater than for an analogue or symbolic model.

The object-file model also predicts an upper-limit on numerical representations, since only a maximum of four items can be simultaneously tracked in a visual array (Trick & Pylyshyn, 1994) and so cannot account for the discrimination of numbers larger than four. Findings that supported this prediction were obtained in research examining the spontaneous representation of number. Rhesus monkeys were able to discriminate possible from impossible outcomes with numerical values up to up to 4, but failed with a 4 vs. 5 comparisons (Hauser & Carey, 2003), and also select the larger of two sequentially presented sets of food items with comparisons of values up to 5 vs. 3, but failed with values that were larger than 4 and 5 (Hauser, Carey & Hauser, 2000). Additionally, some research with numerical discriminations in children also has found distinct set-size limits on discriminative ability (Feigenson, 2008, Feigenson & Carey, 2005; Feigenson, Carey, & Hauser, 2002; Feigenson, et al., 2002; Uller et al., 1999; Xu, 2003). However, there is also conflicting evidence against object-file models. Other researchers have found no changes in performance and accuracy across numerical ranges spanning the object-file limits with objects presented both simultaneously or in sequence (Beran, 2004; Beran, 2007; Beran, Taglialatela, Flemming, James & Washburn, 2006; Cantlon & Brannon, 2006; Cordes, Gallistel, Gelman, & Whalen, 2001; Hanus & Call, 2007).

Another possibility is that two numerical systems exist; an object-file system for the representation of small numbers and an approximate analogue-magnitude system for numbers outside of the object-file limit (e.g. Feigenson, Dehaene & Spelke, 2004; Xu, 2003). However this is not the most parsimonious hypothesis, given that a single mechanism, the analogue-magnitude model is able to account for both large and small numbers successfully. The analogue-magnitude model can also predict the difference in performance across the ranges predicted by the object-file model, if certain assumptions are made about the limits on the

discriminability and variability of certain ratios and numbers (e.g. a ratio of 1:3, but not 4:5 is easily discriminable due to the increasing similarity in number or overlapping variability.

1.3.4 The representation of number

There are three forms of subjective numerical scales that are thought to be developed and used in numerical discriminations; one is a precise representation of number which is limited to human verbal counting processes and two possible approximate numerical representations that is found in human and nonhuman nonverbal discriminations of number. The precise representation of number consists of a linearly spaced number scale with constant generalisation or variability between values and allows exact, 1:1 mapping between objective and subjective number (see Panel A in Figure 1.5). It is the typical scale structure that would be used by an adult human who possessed a true concept of number and was proficient in numerical discriminations. The approximate representations are shown in Panels B and C in Figure 1.5, respectively; a logarithmic numerical scale with constant generalisation between values and a linear numerical scale with increasing generalisation between values.

Both types of approximate representations possess a common feature; as numerical magnitude increases, the representations of individual numbers become increasingly more variable, due to the compressed scaling in the logarithmic scale, and increasing generalisation between numbers in the linear scale. Consequently, both scales predict a distinguishing characteristic of responding in nonverbal numerical procedures; scalar variability, when response variability increases proportionally to mean numerical magnitude, such that relative response variability remains constant as number increases. This contrasts with binomial variability, obtained with humans in verbal counting procedures, where relative response variability decreases as number increases (e.g. Brannon, 2006; Cordes, Whalen, Gallistel & Gelman, 2001).

A more in-depth review of numerical representation and response variability is provided

in Chapter 6. To date, scalar variability is an almost universal finding in nonhuman responding in numerical tasks (although for one exception, see Machado & Rodrigues, 2007), and thus it is not clear whether it is possible for nonhuman animals to develop a more precise representation of number that may more closely resemble that used by humans when counting verbally.

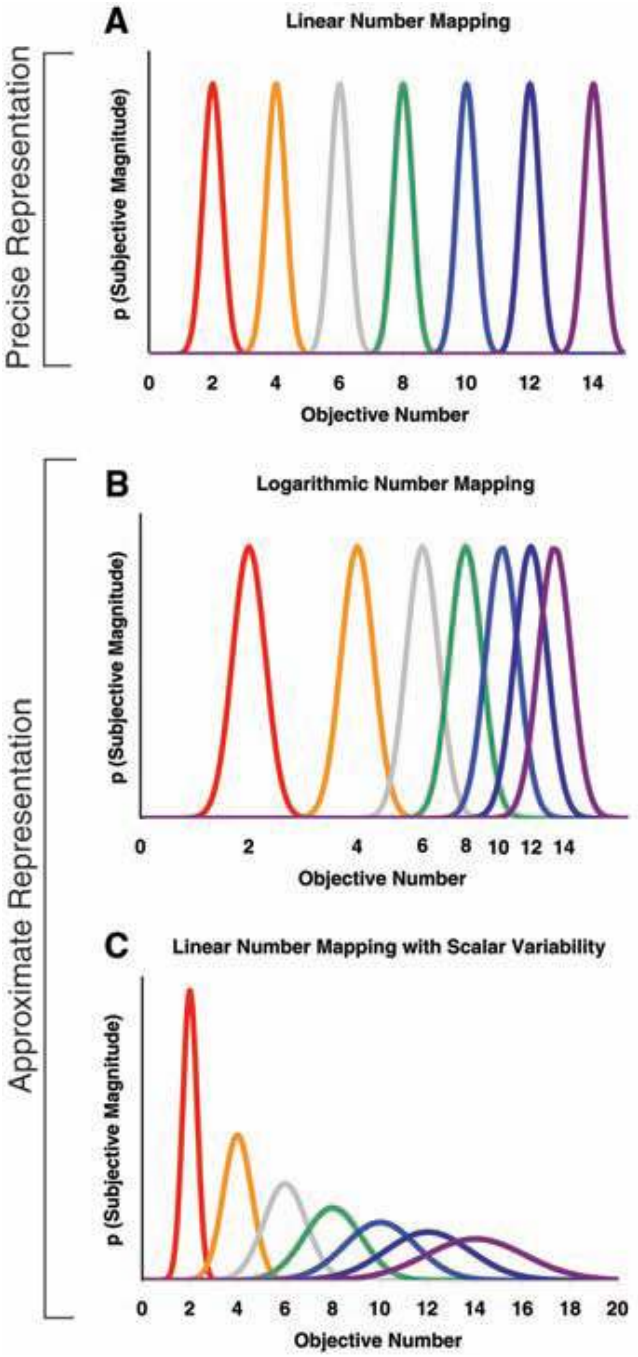

Figure 1.5. Proposed numerical representations of number. Panel A (top) shows a precise linear scale of number with constant variability between values, Panel B and C show approximate numerical representations; a logarithmic numerical scale with constant generalisation between values and a linear numerical scale with scalar generalisation between values, respectively. Figure from Cantlon, Cordes, Libertus and Brannon (2009).

1.4 Researching numerical competence in nonhuman animals

A variety of different procedures has been used in research on numerical competence in nonhuman animals. These can be separated into the different processes under investigation (e.g. relative or absolute number discriminations), which can be further distinguished in terms of the structure of stimulus presentation (simultaneous or sequential) and the sensory modality in which the stimuli are experienced (e.g. visual, auditory, tactile, etc.). A large proportion of research has involved stimulus-counting procedures, where subjects must respond to the numerical characteristics of external stimuli, such as the number of dots or items presented. Fewer studies have investigated the ability to discriminate and produce a certain number of responses in nonhumans (e.g. Mechner, 1958; Broadbent et al., 1993).

1.4.1 Isolating control by number

It is widely believed that numerical discriminations are difficult for nonhuman animals to perform (e.g. Davis & Perusse, 1988); number is not a salient environmental property and animals only respond on the basis of number as a "last resort", when no other valid cues are available (Davis & Memmott, 1982). The issue of confounding cues has plagued numerical research since Clever Hans, the infamous horse that appeared to count and do arithmetic, but who was later found by Pfungst to be using the subtle signals of the tester or spectators to determine when to stop responding (Pfungst, 1911). The fundamental covariation between number and other stimulus characteristics has made isolating numerical control in discrimination tasks difficult.

It appears that some competing cues are more salient than others; Cantlon and Brannon (2007) specifically tested the salience of number relative to other features of simultaneously presented visual arrays, and whether number was only represented and used as a last resort.

They trained and tested three rhesus macaques who been extensively trained in numerical tasks previously, and another who had no previous training history in numerical tasks. Subjects were trained in a match-to-sample task with numerical values 1-4, and tested with two types of trials. In standard trials, correct choice matched the sample stimulus with respect to number as well as one of three nonnumerical properties, e.g. shape, colour or surface area. Following training with these stimuli, subjects were tested in probe trials where the previously confounded dimensions were now incongruent, such that the correct answer could only be selected by responding on the basis of number and ignoring the shape, colour and surface area of the stimuli. Biases in matching responses based on number, shape, colour and surface area were compared between the number-experienced and inexperienced monkeys.

Results showed that for the number-experienced monkeys, number was represented and utilized in conditions even when it was a redundant cue. In the three conditions where number was tested against shape, colour and surface area, the three experienced monkeys performed significantly above chance on the standard trials, and performance for all subjects significantly increased as the numerical ratio between the choice stimuli, calculated as the small number divided by the large number, decreased. This was the case in all conditions, with the exception of one subject in the number vs. surface area condition. On probe trials, response times were longer, suggesting these discriminations were more difficult than in standard trials. As in the standard trials, the likelihood of subjects choosing a number match increased as numerical ratio decreased. Generally, subjects were more likely to match on the basis of number than other cues for the easier discriminations with smaller numerical ratios (ratios with a bigger difference between the small and large numbers); this was especially true for the number vs. surface area condition, where there was a significant overall bias towards numerical matching. However at the most difficult ratio, colour and shape were the strongest determinants of matching, and two of the three subjects were more likely to match on the basis of surface area rather than number (Cantlon & Brannon, 2007).

It can be concluded that in this experiment, number-experienced monkeys attended to number; numerical ratios affected accuracy even if numerical-based responses were not directly rewarded. At large numerical ratios and difficult numerical discriminations shape and colour, but not surface area, were more salient cues than number, while at small numerical ratios number was more salient than colour, and for one subject number was more salient than shape.

The performance of the number-naïve monkey resembled that of the number-experienced monkeys in several important respects. Firstly, performance on standard trials was significantly above chance for all three trial types, numbers vs. colour, number vs. surface area and number vs. shape. For the latter two conditions, a significant effect of numerical ratio on matching was obtained; performance increased with decreasing numerical ratios. This finding suggests the number-naïve subject was attending to number. On probe trials, however, the subject demonstrated a significant bias towards shape and colour over number across all numerical ratios, and a significant bias towards surface area over number at the largest two numerical ratios. For the smaller, easier numerical ratios in the number vs. surface area condition, the subject exhibited a significant number bias (Cantlon & Brannon, 2007). This shows that for the number naïve subject shape and colour were more salient cues than number, but number was generally more salient than surface area with the more easily discriminable smaller ratios.

The performances of the two groups of monkeys were directly compared to assess the effect of numerical training on biases towards particular cues. Results showed that experience with number generally resulted in a greater likelihood of choosing a match based on number across all conditions and numerical pairs. However, the bias towards choosing a numerical match for the number-naive monkey was only significantly different from the number-experienced monkeys in the number vs. colour condition, not the number vs. shape or surface area condition. Additional analyses of surface area manipulations revealed the number-naive monkey was significantly affeceted by surface area in the probe trials, whereas the number-experienced monkeys largely responded on the basis of number, rather than surface area.

Cantlon and Brannon (2007) conducted a second experiment with one number-experienced and -naïve monkey to examine whether similar findings would be obtained with a larger range of numbers, 1-8. Because the number-naïve monkey had not received any differential reinforcement for numerical discriminations, it was considered that he had still not experienced numerical training. The number-experienced monkey's performance generalized to the wider range of numerical values; a significant number-matching bias for the three shape, colour and surface area conditions at the easiest numerical ratio, and for surface area averaged across all numerical ratios. The performance of the number-naïve monkey showed strong effects of numerical ratio on accuracy, suggesting it was able to attend and represent the wider range of values. The number-naïve monkey was significantly more likely to choose the number match over the shape or surface area match, but not the colour match, at the easiest numerical ratios. Additionally, the number bias in the surface area condition was also significant across all numerical ratios.

This research shows that numerical training is not required for number to influence responding in a match-to-sample procedure; both experienced and naïve subjects spontaneously represented and matched on the basis of number at small numerical ratios, even though it was not necessary. Not surprisingly, numerical control over responding was stronger for subjects that had exposure to previous training in numerical discrimination tasks. The use of numerical cues to determine responding was dependent on the numerical ratio of the values tested; number was more likely to be used to match more discriminable values with a smaller to larger numerical ratio. Subjects still appeared to use shape and colour to determine responding on a large number of trials, though it is not possible to quantify the amount of influence these cues had over responding relative to number. Nevertheless, number was a more salient cue than cumulative surface area, and this was consistent across all numerical ratios tested.

The findings of Cantlon and Brannon (2007) show that preferential responding to number over some visual cues can be obtained in a task involving simultaneously presented stimuli.

However it may be more difficult to establish numerical control when temporal cues are available, such as in tasks where stimuli are presented successively. Are nonhuman animals able to still respond on the basis of number in the presence of competing temporal cues?

A seminal paper that investigated the influence of temporal variables on responding in numerical tasks is Breukelaar and Dalrymple-Alford (1998), who reported experiments based on the procedure used by Meck and Church (1983). In their first experiment, rats were trained to discriminate between two possible sequences of two or eight sounds, where duration and number were confounded. Once responding had stabilized, control by time and number were tested separately in transfer tests; one of these two variables varied between the training values while the other was held constant at an intermediate value. Under these conditions, time was the only determinant of responding. In their second experiment, control by time was stronger and acquired more rapidly and when separate time- and number-relevant cues were used. Furthermore, the rats that displayed high accuracy in numerical and temporal discriminations, after extensive training with time and number separately, based responses solely on duration when stimuli involved conflicting cues, even if number was more accurate. Thus, they concluded, like Davis and Memmott (1982) that number was used as a "last resort" cue, when no other competing temporal cues were available.

Based on these studies, which have used symbolic stimuli and equal reinforcement of correct responses, it appears that nonhuman animals may be able to use number preferentially over the visual cues, such as surface area, and to a lesser extent, shape and colour, but temporal cues may have a stronger influence over responding that cannot be overcome with training. At this point, there is no unequivocal evidence showing that significant control by number over responding above and beyond other variables can be obtained, even in relatively simple numerical tasks.

1.5 Current Research

The current experiments examine the performance of pigeons and humans in a range of numerical discrimination tasks. Chapter 2 reviews the current literature on relative numerosity discriminations and describes an experiment (Experiment 1) that examines responding in a numerical bisection task. Pigeons were presented with series of light flashes, and trained to bisect three pairs of "large" and "small" numbers of flashes across three different experimental conditions. The number pairs were kept at a constant ratio of 1:3, while the absolute value of the numbers ranged from 2 to 32. After training with a set of anchor value pairs, subjects were then tested for transfer to novel values within and outside the range of values used in baseline. Of interest was whether subjects would be able to attend to and discriminate number when temporal characteristics of the stimuli were randomised. The representation of number developed in this procedure was also examined using psychometric functions calculated from transfer test data.

Chapter 3 reviews some of the existing research on absolute numerical discriminations, which is investigated in Experiments 2-4. These experiments introduce a new procedure, the numerical reproduction task, in which pigeons were required to discriminate the absolute number of flashes presented in sequence in a sample phase, and reproduce that number in keypecks, with an additional completion response, in a following production phase. Experiments 2 and 2A were conducted to determine whether pigeons would be able to learn to discriminate and reproduce the number of 2-, 4- and 6-flash sequences, and transfer this skill to novel values. Additionally, these experiments provide a test of the last-resort hypothesis (Davis & Memmott, 1982); the relative contributions of temporal and numerical cues in determining responding in this task was investigated when flash rate and sample phase duration were either correlated with number (Experiment 2), or randomised and only weakly related to number (Experiment 2A). These experiments, as well as the variability analyses discussed in Chapter 6 and the Prototype Response Class model examined in the General Discussion were previously published in 2007 in

the *Journal of Experimental Psychology: Animal Behaviour Processes*, *33*, in a paper titled "Numerical reproduction in pigeons" by Tan, Grace, Holland and McLean.

Experiment 3 examines acquisition of performance in the numerical reproduction task and provides a replication of Experiment 2A. The effect of delays on performance in discrimination procedures is reviewed in Chapter 5 and an experiment investigating retention interval (RI) manipulations on reproduction performance is presented. Retention intervals (RIs), i.e. the delays following the sample phase and preceding the production phase, were manipulated in Experiment 4, to examine effects of memory in the numerical reproduction task, and to test whether the usual effect of RIs could be obtained in a numerical reproduction, rather than bisection, procedure. Experiment 4 was previously published in *Behavioural Processes*, *78* in 2008, in a brief report titled "Effect of retention interval manipulation on performance in a numerical reproduction task".

Chapter 6 reports variability analyses of data collected from Experiments 2-3, with the aim of elucidating the numerical representation and response processes that is developed and used in the numerical reproduction procedure. Specifically, changes in relative response variability as a function of numerical magnitude and its implications for numerical processing and representations are examined. Data from Experiment 3 are also used in analyses of the subjective numerical scale; the fits of linear and logarithmic functions to response data were calculated and compared.

Finally, human performance in verbal and nonverbal numerical discrimination tasks is described in Experiment 5. Participants were required to discriminate the numbers 1-20 in tasks analogous to the pigeon bisection and reproduction procedures reported in the preceding chapters. The main aim of this experiment was to compare human and pigeon performance in similar tasks, and investigate whether responding was based on similar numerical processes and representations.

2 Chapter 2: Discrimination/bisection

*2.*1. Introduction

Relative numerosity discriminations have played a major part in the investigation of nonverbal numerical abilities and the representation of number in both humans and nonhuman animals. Judgments of relative numerosity are considered one of the most basic numerical discriminations, involving fairly imprecise, dichotomous judgments of "few" vs. "many". Successful performance in these tasks does not necessarily require a true concept of number, but this skill may very well serve as a basis for the development of more sophisticated numerical abilities, such as absolute number discriminations and a true concept of number.

It seems reasonable to assume that relative numerosity skills are within the capabilities of a range of nonhuman species; for example, individual organisms able to discriminate foraging patches that contained more food and fewer predators would be more likely to survive and reproduce than those that could not. Tasks involving relative numerosity discriminations might provide a useful tool for investigating a rudimentary numerical ability, as well as the representation of number that used to determine responding. Consequently, the investigation of this ability has received considerable attention from numerical researchers, with a significant proportion of studies dedicated to this type of discrimination.

2.1.1 Spontaneous relative numerosity discriminations

Evidence of spontaneous numerical abilities demonstrates a natural understanding of number, and provides evidence for evolutionary continuity in numerical understanding. Accordingly a large proportion of research has been dedicated to the testing of animals' natural abilities to represent and discriminate number, using procedures which require little to no previous training. These experiments generally test simple numerical abilities, requiring subjects

to discriminate relative numerosity, usually choosing the larger of two presented sets.

Hauser, MacNeilage and Ware (1996) utilised a paradigm used with human infants (Wynn, 1992) to investigate and compare basic numerical understanding of infants and wild rhesus monkey. They used a preferential looking procedure to test simple arithmetic operations. Wynn found that infants looked longer when expectations about number were violated. Infants saw one object placed on an empty stage, then a screen lowered to obscure the stage, and then another object placed behind the screen. It was found that looking times were longer when one or three objects were seen when the screen was removed, than if there were two.

Hauser et al. (1996) used this procedure to investigate rhesus monkeys' natural ability to subtract and add eggplants presented in a similar manner. Subjects were wild male and female rhesus monkeys living on an island, and each subject participated in a single trial obtained when subjects remained in one position long enough for stimulus presentation. They presented familiarization trials to ensure subjects were not merely looking longer due to different stimulus displays (e.g. two eggplants may be more interesting than one); subjects saw one or two eggplants placed in one or two compartments of the display box without the screen, or one or two eggplants revealed after the screen was removed. In the test trials, subjects saw either a possible result, when the same number of eggplants was placed behind the screen (1 or 2) and remained after the screen was removed, or an impossible result, when the number of eggplants changed before the removal of the screen. In the addition condition, subjects looked significantly longer at the final display in the impossible than the possible conditions, relative to the looking time for the familiarization conditions.

Hauser et al. also tested subjects in a 2-1 subtraction task, where on test trials subjects saw two eggplants placed in the display box, one was removed from behind the screen, and then saw either one or two eggplants when the screen was removed. In this task, 7 of 8 subjects looked longer in the impossible condition than possible condition. However due to the large amount of variability in looking times, no significant difference in looking times for the familiarization trial and impossible trial was found. These results show subjects will look longer at displays that violate simple arithmetic

laws than those that do not, consistent with the finding of Wynn (1992).

Hauser and Carey (2003) extended Hauser, et al.'s (1996) procedure to more closely examine the content and format of the numerical representations developed in the expectancy violation procedure. In one experiment, they showed that rhesus monkeys were able to differentiate between possible and impossible outcomes when presented with 1 small + 1 small = 2 small or 1 large eggplant. Their ability to differentiate these outcomes suggests subjects were not representing and discriminating on the basis of overall contour length or volume, although it is possible the longer looking time at the large eggplant were due to the mismatch in object size, rather than number. Additionally, subjects were able to differentiate between the possible and impossible outcomes of 2 + 1 = 3, or 2 or 4. Because each subject was exposed to familiarization trials so that the outcomes were all equally familiar, the longer looking times for the impossible outcomes were not due to a familiarity preference. Longer looking times for impossible outcomes could also not be due to a preference for larger sets, since the impossible outcomes included both numbers larger and smaller (2 and 4) than the possible outcome (3). Two following experiments showed that when comparisons involved 3 rather than 2 addition operations and required frequent representation updating, performance fell to chance; no significant difference in looking times at possible and impossible outcomes with a 2 + 1 + 1= 3, 4, or 5 comparison, or 1 + 1 + 1 = 2 or 3 were found. Thus, even though subjects could successfully differentiate between comparisons involving the same absolute numerical values (the previous 2 + 1 = 3 or 4 comparison), the more complex arrangement of the stimulus presentation impaired performance considerably.

Further investigation into the spontaneous representation of numbers in addition operations was conducted in a later study by Flombaum, Junge and Hauser (2005), with adult rhesus monkeys using the same habituation-discrimination paradigm. Each subject was exposed to only one experimental trial, consisting of two familiarization phases followed by a test trial. In the first familiarization phase (F1), the experimenter removed a screen placed in front of the stage to show a number of lemons, placed in a row, equal to the number to be presented at the end of that subjects' test

trial. This was to ensure that longer looking times during the test trial was not due to preference for novel displays, or a certain number of stimuli. In the second familiarization phase (F2), a number of lemons were placed on the stage equal to the number present at the beginning of the test trial. Each subject experienced either an impossible or possible test trial. During test trials, the same number of lemons as those presented in F2, was placed in a row on a stage. An occluder was then lowered onto the stage to hide the lemons, and the experimenter added an additional number of lemons to the stage behind the occluder, one by one, with the subject watching. The occluder was then removed to present the final quantity. During the impossible test trials, the experimenter covertly added or removed items from the stage before removing the occluder, thus resulting in an improbable final number of lemons. Any test trials in which the subject that looked away from the display stage during the addition events was not analysed; some subjects were also excluded from final coding and analysis of video records due to poor quality. Mean looking times for the impossible and possible outcomes were compared between and within groups.

Over a series of experiments, Flombaum et al. (2005) tested whether rhesus monkeys could discriminate large ratio differences between values that were larger than 3 (outside the small number range), and could discriminate between these numbers as the sums from several different addition operations. The first three experiments compared looking times for the addition operations 3+1, 2+2 and 4+4 respectively, with the possible outcome of 4 and the impossible outcome of 8 for the first two experiments, and vice versa for the third experiment. Significantly longer looking times were predicted for the impossible outcome in each experiment, and this was obtained for all addition operations. A fourth experiment tested the effect of different numerical ratio using possible and impossible values with a ratio of 2:3 instead of 1:2; subjects were tested with a 2+2 = 4 or 6 discrimination. If discrimination was dependent on the ratios of the two values, rather than the absolute difference, then performance would drop with the smaller 2:3 ratio. There was no statistically significant difference in looking time between the 2+2 = 4 and the 2+2 = 6, which Flombaum et al. (2005) interpreted as providing evidence for ratio-dependent discrimination. Note,

however that this could also be an effect of a decrease in the absolute difference between the impossible and possible values, which had halved relative to the outcome values of 4 and 8 in the first three experiments.

A fifth experiment was also conducted in which continuous dimensions that may have been confounded with number, such as total volume, or row length, was controlled by cutting the ends of lemons so that there were three sizes: large, medium and small. Subjects were tested with 3+1 medium lemons= 4 large lemons, or 8 small lemons. All subjects in the test trials in this experiment initially saw three medium-sized lemons placed on the stage. These were hidden behind the occluder, and another medium lemon was added to the set. The occluder was then removed and subjects saw either four large lemons or eight small lemons, depending on whether they were in the possible or impossible group, respectively. Mean looking times for this experiment were significantly longer for the numerically impossible eight-lemon outcome, but not the possible four–lemon outcome. This suggests subjects' looking times were based on the number of lemons, rather than the continuous variables such as length or volume.

The findings from this experiment demonstrate rhesus monkeys are able to represent numbers larger than 4 to discriminate possible and impossible sums of addition operations, without any prior training (Flombaum et al., 2005). The fifth experiment showed subjects were discriminating primarily on the basis of number and not other confounding dimensions, and the familiarization phases of the procedure ensured the differences in looking time was not due to a preference for novel events or larger numbers. Subjects looked significantly longer at impossible outcomes with a ratio of 1:2, but not a 2:3 ratio, which was interpreted as showing a dependence of discriminability on the ratio of the small and large values, decreasing as ratios increased. Thus, the skill seemed to be limited by the proportional variability in the representations, with larger ratios resulting in greater representation variability.

Nonhuman animals also appear to be able to choose the larger of two sets of items presented with little to no explicit training in the task. In these experiments, subjects are generally presented

with two sets, either using food items as both stimuli and reinforcement, or an abstract symbol representing the number of food items they would receive as reinforcement.

Rumbaugh, Savage-Rumbaugh, and Hegel (1987) used numerosity judgments to assess summation in chimpanzees. In their first experiment, subjects were allowed to choose and consume between two groups of 0-4 pieces of chocolate presented in varying rations on two trays, each containing a pair of food wells. A clear preference for the larger quantity emerged, with preference ranging from 84-100%. Subjects were then tested for whether their preference for the larger quantity would be maintained when samples consisted of two pairs of quantities. Half the trials consisted of randomly generated numbers of chocolates between 0-4, while the other half consisted of "meaningful" summation comparisons, which met the following conditions: 1) the tray that consisted of the greater *total* number did not include the greatest *individual* number in its pair; 2) pairs were never identical or of equal totals; and 3) any quantity from 1-4 was not used in both pairs in the same trial. These conditions ensured subjects were not making choices based on any single quantity in a food well.

Overall, subjects consistently preferred the larger sum. Performance improved over the 5 days of testing and the two subjects performed better on the randomly arranged trials than the meaningful comparisons, with preference for the larger sum reaching 100% on the last day for the random trials compared to 61% and 58% for the meaningful trials. Furthermore, performance was better when the difference between the ratio of the total sums was larger than when it was smaller, e.g. 2:3 vs. 5:6. To provide a more focused analysis of the meaningful trials, subjects were then exposed to seven days of testing with just meaningful comparisons. Performance over these tests continued to improve and the overall mean preferences for the larger sum for the two subjects were 85% and 88%; once again an effect of ratio on preference for the larger sum was obtained.

The different quantities were arranged in specific geometric patterns to promote discrimination during early training, and another experiment was conducted to remove this

confound. In this experiment, subjects were tested with meaningful comparisons, and the items were dropped unsystematically into the food wells in no specific formation. If the subjects had been responding to the spatial arrangement in the previous experiments, accuracy would decrease significantly once this cue had been removed, however this was not the case; performance of both subjects in fact increased, with choices for the larger sum reaching 92%.

In a final experiment, Rumbaugh et al. (1987) tested for generalization to novel, more complex judgments by including comparisons involving 5 items, as well as 0-4. They found that the strong preference for the larger sum was maintained with both the familiar and novel trials. Furthermore, performance did not seem to be based on subjects avoiding the pair that included zero, and better performance when presented with larger ratios did not seem to be a result of a greater ability to discriminate smaller absolute numbers; there was no difference in performance in sets involving different absolute numbers in the same ratio (e.g. 3:4 ratio made up of 3 vs. 4 or 6 vs. 8 chocolates. Rumbaugh et al. proposed that in their experiments, subjects subitized the number of items in each of the wells, and used a perceptual fusion process to amalgamate the contents of each pair of wells into one group. Subjects could have then compared the two groups and chosen the larger total. However, based on their results it was not possible to conclude that subjects were really counting the items; only that judgments of relative or absolute quantities can be based on abstract symbols representing numerosities – in this case, Arabic numerals.

Hauser, Carey and Hauser (2000) conducted an experiment with over 200 semi free-ranging rhesus monkeys in which subjects observed two researchers placing individual pieces of apple into two separate containers. The researchers then walked away from the containers, providing subjects with the opportunity to approach one of the boxes. The ability of the rhesus monkeys to discriminate 1) a rock from a slice of apple; 2) larger numbers of apple slices differing by one; and 3) numbers of apple slices differing by more than one, was tested across various conditions in two experiments. The researchers attempted to test each animal only once to ensure that spontaneous numerical abilities were truly being tested; some trials did include monkeys that had been involved in at least one

previous numerical procedure; however, their performance did not differ from subjects that had only been tested once.

All monkeys were able to discriminate and select the apple slice when it was presented with a rock. When presented with different quantities of apple slices, the rhesus monkeys preferentially selected the box containing the larger number of apple slices, up to the conditions of 3 vs. 4 and 5 vs. 3. However, performance did not exceed chance levels in all following conditions in which one or both numbers were greater than four, suggesting a set-size limit on discriminative ability.

Hauser et al. (2000) also tested the role of duration in the discriminations, to ensure subjects were responding on the basis of number and not time. In a second experiment, one slice of apple in one box was always replaced by a rock. This largely involved conditions where the number of objects was always the same, with the rock being placed in the box with the smaller number of apple slices. To test whether subjects were merely avoiding the box with the rock, they included a condition where the rock was placed in the box with the larger number of apple slices. An additional condition was also included where subjects were given a choice between half an apple, or three pieces of one-sixth of an apple, to test whether subjects were responding on the basis of volume rather than number.

Results showed that performance of subjects was consistent with that in Experiment 1. Monkeys preferentially selected the box with the greater number of apple slices if the numerical difference was the same as those they chose correctly in the earlier experiment. This performance was maintained even when the actual number of objects, the total duration and activity of placing objects into the box were the same. Subjects did not appear to make choices based on avoiding the box with the rock, making the correct selection when the rock was placed with the larger number of apple slices, and showing no significant preference with 5 slices of apple vs. 4 slices of apple and rock, consistent with Experiment 1. Furthermore, the rhesus monkeys chose the three 1/6 slices of apple over the equal volume choice of half an apple, suggesting choices were made on the basis of number, not volume.

The drop in performance for comparisons with more than four objects suggested a set-size

limit on discriminations. Hauser et al. (2000) interpreted this finding as being consistent with the object-file and subitizing models of numerical discrimination and representation, but inconsistent with a scalar analogue-magnitude model. This finding differs from previous research which has demonstrated that animals including macaques, are able to form and use analogue-magnitude representations in discriminations of numbers larger than 3 (e.g. Washburn & Rumbaugh, 1991; Brannon & Terrace, 1998), and suggests that perhaps animals are able to develop different forms of representations depending on the nature of the task and procedure. There are several factors that may explain the difference in performance; Hauser et al. (2000) used a "real-world" procedure with semi-wild monkeys who may have struggled to attend fully for the full duration of the item presentations, due to the other distractions and lack of experience in the task. In addition, subjects in Hauser et al's. experiment only experienced one trial each and were rewarded regardless of their response, while most laboratory numerical experiments involve thousands of trials of training and differential reinforcement of correct responding.

The finding of a set-size limit on discrimination differs from results obtained by Flombaum et al. (2005). One possible explanation for this may be the nature of stimulus presentation; stimuli were presented simultaneously in the Hauser et al. (2000) experiment, whereas stimuli were presented sequentially in the experiments by Flombaum et al., as well as the experiments discussed below. Set-size limits may be restricted to tasks in which stimuli are presented simultaneously; with sequentially presented stimuli, subjects may be better able to retain a numerical representation over an extended period of time and update their representations as new objects are added to the array.

Beran (2001) investigated relative numerosity judgments of sequentially presented food items with two chimpanzees, Sherman and Lana. Each item to be counted was only visible prior to placement in cup, so whole sets were never seen in their entirety. Thus, subjects were required to perform mental addition or subtraction with the objects, a more difficult task than the discrimination of sets presented simultaneously.

In the first experiment, subjects were allowed to choose between two sets of M&Ms

placed into two cups; it was expected for subjects to prefer (and consequently reliably select) the larger set. Beran manipulated the total number of items, the absolute difference and the ratio difference between the two quantities. Items were placed, one at a time, at irregular intervals, in the left cup first, followed by the right cup. To address concerns about subject attending to the time it took to dispense the whole set, rather than the number itself, control trials were conducted in which the experimenter's hand was purposely held over the left cup for longer than the right cup, independent of the number of items placed in each. This temporal manipulation would result in a response bias towards the left cup if subjects were timing, rather than responding on the basis of number.

Results showed there was no significant difference in performance between control and regular trials, suggesting subjects were not relying primarily on duration to determine responding. Across all trials, the accuracy of the two subjects was significantly correlated with both the ratio between the quantities and the total number of items placed in both cups; accuracy was higher with smaller ratios between the quantities and total number, and it was found that the larger quantity was selected significantly more often on the easier trials with ratios less than 0.70 than on the more difficult trials with ratios greater than 0.70. A significant positive correlation between accuracy and the absolute difference between the two quantities was only found for one subject,where larger differences in quantities resulted in better performance.

In Beran's (2001) second experiment, the chimpanzees were required to monitor, sum, compare and select the larger of two sets of items, which were placed into their respective cups in two different intervals. Part of each set was placed into the left, and then right cup by one experimenter, before the remainder was placed into the cups in the same order by a second experimenter. Both experimenters were blind to the number of items placed in the cups by the other experimenter. Control trials were also included to assess reliance on duration as a cue for responding, as in Experiment 1.

Results were similar to the previous experiment; no significant difference in performance

was found between control and regular trials. For both subjects, accuracy was significantly correlated with both ratio and difference between the quantities; larger differences and smaller number/total number ratios corresponded to a greater proportion of correct trials. Additionally, total quantity had a significant effect on performance for one subject; a significant negative correlation was found between total quantities and performance. The results of this experiment suggest subjects were able to sum two temporally-separate quantities of sequentially presented items that made up the different sets, and mentally represent and compare the sets to select the largest total quantity, with no significant decrement in performance relative to the simpler first experiment.

Beran's (2001) third experiment further tested the chimpanzees' ability to represent and sum items in a third experiment. Subjects were required to sum three sets of M&Ms placed into each cup in temporally separate sequences. In this experiment, control trials were not presented, and the maximum difference between the total quantities in the cups on any trial was only one. Under these conditions, only one subject maintained responding above chance; Sherman was able to select the cup containing the larger number of items reliably and his performance was significantly better than in Experiment 1. The performance of Lana did not differ significantly from chance, and this appeared to be due to a failure to attend to the presentation of the stimuli. Consequently, analyses for the third experiment involved only Sherman's data.

Beran (2001) investigated Sherman's performance as a function of the different trial types, in particular as a function of whether the last sequence placed into each cup was larger, smaller or equal for the cup with the larger total quantity relative to the cup with the smaller total quantity. If the subject had merely been attending to the last sets placed into each cups, then performance would be better when the last set was larger for the cup with the larger total quantity. However, there was no significant difference between the trial types, suggesting Sherman was not merely using the differences in the last sets presented to determine his choice.

Lastly, the effects of removing one item from the first cup after a single set of M&Ms had

been placed in each cup was investigated. This manipulation would not affect performance if subjects were attending to the absolute number of items between the cups. One to five items were placed in each cup in the same manner as in Experiment 1, and prior to selection, one M&M was removed from the cup on the left. Responding in three different scenarios were compared. When the right cup contained a larger number of items before and after the removal of an M&M from the left cup, both chimpanzees were able to select the correct cup at levels significantly better than chance. This finding is not surprising, as the removal of the item should not have affected the subjects' choice since the right cup contained more items initially. Subjects also selected the larger quantity on 75% of the trials in the slightly more difficult situation where both cups initially contained an equal number of M&Ms, so that the left cup had fewer M&Ms once one had been removed. The last scenario, where the left cup had the larger quantity before and after the removal of one M&M, was the most difficult as subjects had to select the left cup even though an item had been removed from it. Only Sherman was successful in this situation with performance at 78% correct. Thus one of the two subjects was able to select the cup that contained the larger quantity of M&Ms, even after one was removed from that cup. To do this, the ability to recognize the absolute difference between the two cups, and monitor the addition or subtraction of items would be required.

Beran (2001) suggested that subjects may have used elementary numerical operations, forming representations of the items in each set placed in each cup, as in Experiments 2 and 3, and summing and comparing these to determine their response. Alternatively, subjects may have used a "counting-on" process, updating their representation of the first set of items placed in each cup by adding or subtracting the different sets of items.

These experiments were later extended by Beran (2004) to investigate response processes further. In his first experiment, he trained Lana and Sherman to select one of two sets of 1-10 marshmallows contained in separate cups. Both quantities were placed into the cups, one item at a time, before the chimpanzee selected a cup by touching it with a finger. Accuracy was

examined as a function of the total number of presented items, the difference between the two sets and the ratio of the two sets. Both subjects selected the larger of the two non-visible, sequentially presented sets at a high rate. Performance improved as the size difference increased between the sets (a distance effect) and as the ratio of the smaller to the larger set decreased. There was little evidence supporting the prediction that only sets of less than four items can be remember and compared made by the object-file model (Feigenson et al., 2002), with both subjects performing significantly better than chance on trials with both sets containing at least four or five items. The results of Experiment 1 could not determine whether subjects were making judgments based on the relative difference between the two sets or the absolute sizes of each set, and so, Beran conducted a second experiment to test this.

In this experiment, a third wholly visible set was revealed after the first two sets had been placed sequentially into the opaque cups, requiring subjects to discriminate between two sequentially presented nonvisible sets, and one simultaneously presented visible set. Subjects would not just be comparing the relations between the sets if subjects consistently selected the set that contained a larger number of items regardless of whether it, or the other sets, was visible or nonvisible. Results supported this; the chimpanzees reliably selected the larger nonvisible set over a visible, immediately available smaller alternative and only selected the visible set if it contained the largest number of items or was close to the largest number. It was unlikely that temporal cues were being used to determine judgments because subjects were able to select correctly the largest set of both sequentially and simultaneously presented items.

Beran (2004) also conducted a third experiment, in which 1 to 3 items were removed from one of the two containers before selection. For subjects to select the larger set correctly, they would have to remember the original number of items in each container and recognize the effect of item removal on the final numerical comparison. The number of items removed from the containers had a significant effect on performance; both subjects were able to select the larger set after removal of 1 item at a rate significantly better than chance, but only Sherman was

able to perform better than chance when 2 items were removed, and this was only for trials in which the numerical relation between the two sets was reversed after removal of the items. Neither subject selected the larger set at more than chance rates when 3 items were removed. In fact, there was no significant effect of the initial or final ratio of the smaller to larger set on performance when more one item was removed. Beran suggested this may indicate subjects did not understand the effect of removing two or three items on the set sizes.

A recent study by Beran (2007) addressed three issues that often arise in numerical discriminations. He conducted four experiments in which he tested the role of nonnumerical cues in determining responding by manipulating inter-item rate of presentation, duration of set presentation, and visual surface area, and the effect of set size on performance. In the first experiment, two male rhesus monkeys, Murph and Lou, were tested in a computerized procedure that was similar to that used by Beran (2001; 2004), where subjects watched two sets of red items be dropped into two "containers" by a image of a human hand. Subjects were initially trained on a 1 vs. 0, then 1 vs. 4, then 2 vs. 5 comparisons, over 3 sessions, before being tested with sets of between 1 to 10 items. Performance on the training sessions was significantly above chance for all three discriminations, showing subjects understood the task requirements by the end of training. Performance from the test sessions was analysed as a function of percentage correct at each ratio (smallest set divided by larger set size). Performance was largest correlated with ratio, and also significantly better than chance for all ratios except the highest five. Beran was interested in comparing performance with sets containing less and more than 4 items, to test the predictions made by the object-file and analogue magnitude models of numerical representation. An object-file model predicts performance to drop to chance levels for sets of 4 or more items, whereas an analogue magnitude model predicts performance to decrease as a function of the numerical ratio. Data supported the latter; there was a significant correlation between ratio and percentage correct when both sets had at least 3 items, and performance was significantly above chance for all ratios, with the exception of the five highest ratios and the ratio

0.60 for both monkeys and 0.75 for Murph. The data from this experiment show rhesus monkeys are able to discriminate and choose the larger of two sets of items presented sequentially on a computer monitor for sets of less and more than 4 items are presented, if the ratio between the two sets is not large. Hauser et al. (2000) reported a set-size limit on performance, however this was not found here.

In a second experiment, the interitem interval, or rate of presentation was varied randomly in one of the sets so that it could not consistently be used to determine the correct response. The total duration for stimulus presentation in the first and second sets were held equal and constant. Performance for both subjects under these conditions did not differ significantly from that in the first experiment, suggesting that interitem interval or sample duration did not play a major role in determining subjects' choice.

Beran (2007) also tested whether subjects were responding to the amount of red presented on trials, rather than the actual number of items, by manipulating the size of items presented. For any given set, all items could be made up of small squares or large squares. There were four types of trials in this experiment: both sets of items could be made up of all large or all small squares, or the larger numerical set could be made up of large squares while the smaller numerical set was made up of small squares, or vice versa. For the first three trial types, correct responses could be based on either amount or number. However, for the last type of trial, with ratios of at least 0.286, the number-based and amount-based strategies would be in conflict, and thus these critical trials could test which property was determining responding. Only a relatively small number of trials were presented (approximately 21% of total trials) and these were presented amongst other trial types, to ensure a true test of the spontaneous use of numerical cues and to prevent the development of control by a nonnumerical area or amount cue. Beran compared performance on the critical trials with the trials in which both sets contained items of equal sizes (identical to Experiment 1). No significant difference in performance on these two trial types was found, suggesting amount was not a major influence on responding. Most

importantly, performance for both monkeys on the first 40 critical trials was significantly greater than chance, 30/40 and 28/40 for Murph and Lou respectively, and was significantly correlated with the numerical ratio.

Finally, Beran (2007) manipulated the amount of time each item was visible by changing the speed they fell into the containers and the duration of the total presentation set to assess their effect on performance in Experiment 4. There were four types of trials; the presentation of the larger set took longer in total, and the items fell relatively slowly or relatively quickly, or the presentation of the smaller set took longer in total and the items fell relatively slowly or quickly. Thus, trials could be categorized as either having: valid number, total duration and item duration cues; only valid number and total duration, but not item duration cues; valid number and item duration, but not total duration cues; or only valid number cues. Additionally, a small proportion of trials in all categories had durations that were of similar length, and these were categorised as not possessing valid total or item duration cues. Results suggested that timing did play some role in determining responding. Although there was a high correlation between numerical ratio and performance, and performance was better than chance when all cues available or when item duration was not a valid cue, when total duration was in conflict with number, performance fell to random levels for both subjects. A significant correlation was still obtained between ratio and percentage correct on these trials for Murph, but not Lou. Thus, it appears that total duration was not the only cue subjects were relying on; performance was still above chance levels on trials where the set with the smaller number took longer to present. Subjects most likely attended to both time and number during presentation and used both cues as a basis for their choice.

van Marle, McCrink and Santos (2006) tested the ability of capuchins to quantify objects and substances using a procedure similar to Beran (2001). Subjects watched the placement of 1-4 items into two opaque cups and were able to select and consume the contents of one of the cups. After initial training with one and zero items, subjects were tested with a 1 vs. 2 comparison, and then 1 vs. 4, 2 vs. 3 or 3 vs. 4. Results showed subjects were able to perform

significantly better than chance in the 1 vs. 4 and 2 vs. 3, but not in the 3 vs. 4 comparison. This difference in performance was significant, and performance was best for the largest ratio between the two quantities (1 vs. 4). The researchers concluded that performance appeared to be constrained by ratio, rather than absolute set size, consistent with Beran (2001, 2004), but contrary to Hauser et al. (2000) and Feigenson, Carey and Hauser (2002). However, van Marle et al. only provides a relatively weak test of the set-size hypothesis since their experiment only used values up to 4 which is the upper limit predicted by the object-file model; a stronger test would involve values greater than 4.

A second experiment was also conducted to test whether capuchins would also be able to determine and select the larger of two portions of a non-discrete substance, banana puree, measured and presented in scoops (van Marle et al., 2006). Training values and comparisons were the same as in the first experiment. Performance was significantly above chance for all comparisons of 1-4 scoops and appeared to increase as the ratio between quantities increased.

The researchers concluded that the similar ratio-dependent performance with both discrete objects and continuous substances showed that numerical discrimination was based on an analog magnitude process rather than an object-tracking process. Presumably, the latter would have resulted in set-size-dependent performance. However, it should be noted that the presentation of the continuous substance occurred discretely, in separate spoonfuls, so it was possible that subjects were counting the separate events. Van Marle et al. (2006) argued that their subjects' ability to discriminate the 3 vs. 4 comparison with the continuous substance, but not discrete objects provided evidence against this possibility, although because their manipulation was also confounded with training it is unclear whether this increase in performance was due to increased experience or the discrimination of amount rather than number. Additionally, the role of temporal cues in determining responding was not investigated – although presentation was irregular due to being contingent on the attention of the subject, rate and duration of presentation was not measured.

Hanus and Call (2007) tested great apes' abilities to estimate, compare and select the larger of two sets, presented either as whole sets sequentially or simultaneously, or as individual items placed into opaque cups in a procedure similar to that by Beran (2001). The sequential presentation of whole sets prevented direct perceptual comparison between quantities, which was possible with simultaneous stimuli. Bonobos, chimpanzees, gorillas and orangutans were tested with all three presentation types. Values from 0 to 10 were used to investigate and compare abilities to discriminate number, and effects of ratio (the small quantity divided by the large quantity), and difference between quantities and total quantity on performance.

In Experiment 1, all subjects saw simultaneously and sequentially presented sets of every combination of the lower quantities of 1 to 6, as well as the pair 0 and 1, before being tested with 18 selected pairs of up to 10 items. The percentage of trials on which subjects selected the larger quantity by touching it with their finger, or occasionally, their tongue, was recorded. Subjects received the contents of whichever dish they touched first. Performance was significantly above chance when sets were simultaneously available, or presented one at a time. Additionally, there was no significant difference in performance between species, or between the low- and high-quantity tests. Analyses of individual performance revealed all subjects except two gorillas and one orangutan performed above chance in the low and high quantity tests. Regression analyses investigated the relative contribution of ratio, difference and total quantities in explaining performance and found that ratio was able to explain the majority of the variance for each individual species, as well as for the pooled data. Accuracy increased as ratio decreased and differences between quantities increased; that is performance was worse with discriminations involving high ratios and small differences. These findings show that apes were able to discriminate between whole sets of quantities presented successively and simultaneously, even with high ratios and large quantities. Because subjects were able to compare sets even when they were not perceptually available, it is likely they were using numerical representations to determine responding.

Hanus and Call (2007) conducted a second experiment to investigate whether performance would remain accurate when items were presented item-by-item rather than as a whole set. Some subjects were excluded form this experiment as they were not available at the time of testing or failed to attend to the new procedure sufficiently. The procedure was identical to Experiment 1, except that items, rather than whole sets, were presented and placed into one of two cups individually before subjects made their choice. Performance analyses by species group for the low-quantity discriminations showed all species except the bonobos chose the larger set at levels significantly greater than chance. Performance in the high-quantity pairs did not differ significantly from chance for any of the individual species; individual analyses showed only one chimpanzee and one orangutan performed significantly above chance on these discriminations. Regression analyses showed ratio accounted for the greatest proportion of variance in responding compared to other variables, although much less than in Experiment 1 (38% vs. 81%).

Hanus and Call's (2007) results showed that apes are able to discriminate and select the larger of two whole sets presented either simultaneously or sequentially. Furthermore, some individuals were also able extend this performance to numerical sets presented item-by-item. No significant differences in species' abilities were found, although there were major individual differences in performance. These findings are largely consistent with previous research with chimpanzees and orangutans (Beran, 2001, 2004; Call 2000), although chimpanzee performance was considerably lower than that obtained by Beran (2001). This difference may be due to the much greater laboratory experience in language and quantity discrimination experiments of Beran's (2001) subjects, since those used by Hanus and Call's experiments were largely naïve. Finally, numerical ratio appeared to be the main determinant of choice in these experiments, although difference between quantities also influenced responding. This, along with the finding that subjects were able to discriminate quantities of up to 10, subjects provides support for an analogical magnitude system over an object-file model of numerical discrimination.

2.1.2 Symbolic relative numerosity discriminations

Some research has also examined the ability of nonhuman animals to make relative numerosity discriminations using symbolic stimuli and food reinforcers. Given the more complex task, more training is usually necessary for subjects to reach criterion performance, although the use of food reinforcers that varied with sample number appeared speed acquisition such that the required number of training trials is still relatively limited, compared to other laboratory tasks.

Washburn and Rumbaugh (1991) trained two rhesus monkeys to select one of two numerical symbols, presented on a computer screen, after which they received the corresponding number of food pellets. Subjects experienced 1000 trials in a task where they were differentially reinforced to select the larger of two different Arabic numerals, ranging from 0-5, using a joystick. The numbers 6, 7, 8 and 9 were then introduced into the task. Both subjects successfully selected the larger value with pairs of numerals from 0-5, buto only one subject consistently transferred this ability to trials with numerals 6-9. The experimenters found that when food pellets were presented arrythmically and temporal cues were controlled, performances of both subjects were more comparable, and both chose the larger numerical value significantly more often than chance. To further test generalization, subjects were presented with arrays of five numerals from 1-9, randomly selected and presented around the screen. Each numeral selection resulted in the delivery of food pellets of the same number. Results demonstrated that subjects were able to order the numerical values, reliably selecting the largest numeral, although performance appeared to depend on the number of possible choices and the relative difference between the numerals.

Washburn and Rumbaugh (1999) concluded that subjects had learnt to associate Arabic numerals with their corresponding number of pellets, and to arrange the numerals ordinally. They stated that performance did not seem to be based on temporal cues, and logical transitivity

alone could not explain the positive transfer to novel values 6-9 after initial training; knowledge about both the relative differences and the size of the relative differences between pairs of numerals would have been necessary for successful performance in this task. However, it was noted that their data do not distinguish between whether subjects had developed a "complex matrix of relative values" (p. 193), or had developed proper knowledge about absolute number, and it was not believed either subject was truly counting.

Olthof, Iden and Roberts (1997) showed that two squirrel monkeys were able to make summation and ordinal symbolic judgments, using numerical symbols that were associated with different numbers of food items. They adapted the procedure used by Washburn and Rumbaugh (1991), using food wells covered by a piece of wood displaying one of the following Arabic numerals: 0, 1, 3, 5, 7 and 9. Subjects were initially trained to choose the largest number of different pairs of simultaneously presented numbers. Although subjects required more intensive training in this task than in the more naturalistic tasks discussed above, the procedure still only involved a relatively few number of training sessions; on average both subjects received approximately 165 sessions of training and testing prior to the last phase of testing. Training began with values 0 and 7, and intermediate values were successively introduced in paired discriminations across 4 phases, with the values 1 and 9 introduced last. The pairs 1 and 3, 1 and 5, 1 and 7, 3 and 9, 5 and 9 and 7 and 9 were excluded for training and were used to test ordinality. These were selected, in the same manner as Washburn and Rumbaugh (1991), so that subjects could not use transitive inference to discriminate the larger number with these pairs. Performance in these tests was significantly above chance for one subject (100%), Jake, but not for the other, Elwood (83%). Performance on the number pairs that included 1, and the 3 and 9 pairs was 100%, and 90% on the 5 and 9 pair. Neither monkey was able to perform above 50% on the 7 and 9 pair.

In the last phase of testing, subjects were presented with four simultaneously presented numbers, and were required to choose the largest of the four successively, then the remaining

three, and finally two symbols on each trial. All 15 four-number combinations of the six numbers presented in training were presented once per session each day. The spatial position of the numbers in each combination was varied over a 4 day block, according to a Latin-square design. For both subjects, the percentage of correct choices for discriminations with four numbers was higher than, but did not significantly different from, chance (25%) in the first block. On the majority of the following blocks, however, performance was significantly above chance. Elwood's performance in discriminations with three numbers was below chance for the first 2 blocks and improved to a level significantly above chance for the later blocks. Jake performed significantly above chance in the first block with three numbers, but did not maintain this consistently over the remaining blocks. Both subjects were more accurate on the final choice between the two last numbers, and performed significantly above chance across all blocks, a finding that is not surprising given that this was most similar to their training conditions.

Olthof et al. (1997) concluded that the poor performance on the initial four-choice discrimination was due to a generalization decrement from the two-choice procedure in the previous phases. However, this could also be evidence for a failure to learn the ordinal relationships of the numerical stimuli and this suggestion is also supported by the additional failure of Jake to reach performance significantly better than chance in the three-choice discrimination.

In three subsequent experiments, Olthof et al. (1997) examined the ability of squirrel monkeys to sum single numbers, pairs of numbers or groups of three numbers and to select the greater total. Subjects were tested with discriminations of single number pairs first, and preference for the larger number was significantly greater than chance for all pairs, except for 0 vs. 1, 5 vs. 7 and 0 vs. 3. On the two-number pairs, the researchers had expected performance to start at chance and improve over sessions; however both subjects performed significantly above chance from the very first session, on all trials except for Jake on the 5 + 0 vs. 3 + 3

discrimination. This showed that subjects did not require additional training to be able to select the larger sum of two pairs of numbers, despite the totals sometimes equalling quantities that had not previously been encountered, both within and outside the range of values used in training. Additionally, these choices could not be based on rules about individual numbers such as "avoid the smallest number" or "choose the largest number", as performance was still high on trials where those rules would have resulted in an incorrect choice.

Subjects were later tested with new two-number pairs, including some that included a single number and a two-number pair; the single number could be less than or greater than the sum of the two numbers, so accurate responding could not be based solely on the consistent selection of just the single number or the two-number pair. Under these conditions, both subjects continued to choose the larger number or sum on between 75-87% of total trials. Performance did fail to reach significance on some trials, mostly discriminations involving 0, 5, 7 and 9. Poor performance may have been due to a lack of discriminability between 0, 5, 7 and 9; a modification to the symbols for 5 and 9 resulted in a distinct improvement in performance in these two pairs. Subjects were able to discriminate and select the larger quantity even when the largest presented number was both a single number and in a pair of numbers (e.g., 3 vs. 1 + 3). Preference was weaker when both values were large and had only a small difference between them, suggesting performance may have been affected by the absolute magnitude of the numbers presented. This, however, may be due to a strong bias to select the number 9 when presented, since this was always the correct choice in any pair of single numbers being discriminated, rather than any effect of numerical magnitude on performance. Additionally, performance was not influenced by experimenter cues, both subjects continued to choose the larger quantity at rates significantly higher than chance when the experimenter was blind to stimuli presented. There was no systematic trend in data across sessions, suggesting subjects were not learning new associations over testing, and the high performance at the beginning of each new test showed that preference for the larger quantity was present from the first session.

In their final experiment, Olthof et al. (1997) tested the monkeys' ability to select the larger quantity when each choice contained three numbers. Combinations of the numbers 0, 1, 3, 5, 7 and 9 were tested in different orders on the three-number cards, and were presented during sessions amongst single-number pairs. The apparatus and procedure was otherwise identical to preceding experiments. Performance on the three-number vs. three-number discriminations was significantly higher than chance over the first four sessions, as well as overall; both Jake and Elwood chose the larger sum on over 71% of the trials. Although performance fluctuated across sessions, no significant trend was observed. The overall preference for the larger sum was lower than those in the previous experiment with two-number vs. two-number pairs, and closer examination suggested that responding could be explained by subjects merely choosing the stimulus with the largest single number: preference for the largest quantity was greatest on the 1+1+3 vs. the 0 + 3 + 9, and the 0 + 1 + 5 vs. 0 + 3 + 9 pairs, and the lowest on 1 + 1 + 7 vs. 0 + 5 + 5 pair, and the 1 + 1 + 9 vs. 0 + 7 + 7. However, this would not explain the higher preference for the larger in the 1 + 1 + 5 vs. 1 + 3 + 5 pair, and lower preference on the 1 + 1 + 3 vs. 0 + 1 + 5. Thus, results provide partial support for the possibility that subjects were able to sum and select the larger sum when pairs of three numbers were presented, however performance was less accurate than with pairs of two numbers.

In general, the results of these experiments show that the two squirrel monkeys were able to respond to the ordinal level relationships between six number symbols, reliably selecting the larger number when presented with any combination of two symbols. There is some evidence that subjects may have developed representations of the absolute numbers associated with each symbol; monkeys were able maintain accurate performance when presented with novel pairs, even though these judgments could not have been based on ordinality only. Certain combinations appeared to be more difficult to discriminate than others, with weaker preferences obtained for sums that differed by smaller than larger amounts, consistent with Weber's law. Additionally, the subjects' immediate preference for the larger sum when presented with sets of

two or three number symbols could not be explained solely by the choice of the largest single number or avoidance of the smallest, and a blind testing procedure in one experiment showed responding was not influenced by inadvertent experimenter cues. Olthof et al. (1997) cautioned that their data do not suggest that subjects understood the summation relationships between symbols or that subjects conducted mental arithmetic to make their decisions. Rather, they propose subjects represented the quantity of peanuts presented on each trial, either perceptually or in a Meck et al. (1983) pacemaker-accumulator type mechanism, and associated the different representations of the amount of peanuts with their respective number symbol.

2.1.3 Food and number confounds

A pertinent issue relating to research using food as both stimuli and reinforcers is the possibility that subjects were responding to the conditioned appetitive responses associated with each symbol, rather than learning the actual quantities the symbols or combinations of symbols represented (Olthof, et al. 1997). This problem would be exacerbated in procedures where a extended training is required to reach reasonable performance. If a strong correlation between the mass of food used and number exists, subjects may exploit their natural tendency to rely on variability in mass rather than number to determine responding, similar to children's preferential use of cues such as surface area or contour over number (e.g. Clearfield & Mix, 2001; Mix, Huttenlocher & Devine, 2002). Their high sensitivity to variation in size and amount is consistent with this (Menzel, 1960, 1961). Although some experiments discussed previously introduced control conditions to test for these, not all did.

Olthof and Roberts (2000), in an adaptation of an experiment by Olthof et al. (1997), showed pigeons summed the values of visual symbols based on the mass of food, rather than the number of food items, represented by the respective symbols. In a first experiment, subjects were trained to choose between two food wells that were covered by a card showing a symbol

corresponding to one of the 0, 1, 3, 5, 7 or 9 grain pieces contained in the well. Subjects were expected to select the larger sum. After training with two single symbols, summation was tested by presenting subjects with two wells covered by cards, each containing a pair of number symbols. In summation tests, subjects were tested with a range of number pairs that had large and small differences between the sum of symbol pairs, and with which optimal performance could not have been achieved by merely choosing the largest single number, or avoiding the smallest single number. Subjects were rewarded with the number of pieces of grain equal to the sum of the symbols. Note that in this experiment, both number and mass were confounded.

Subjects' performance was similar to that obtained with squirrel monkeys by Olthof et al. (1997); an immediate preference for the larger sum was demonstrated, and the larger sum was selected for all pairs, including pairs that summed to novel values, except the pairs 1 + 3 vs. 3 + 3 and 0 + 5 vs. 3 + 3. Olthof and Roberts (2000) oncluded that pigeons had learnt the ordinal position of the different symbols and were responding on the basis of these representations. Errors could not have been explained by subjects merely selecting the symbol pair that contained either the larger or smallest single number, as subjects were able to perform significantly above chance on similar comparisons with different values (eg. 0 + 9 vs. 5 + 7 or 1 + 3 vs. 0 + 5). Due to the covariation between number and mass, two more experiments were conducted to tease out the influences of these two variables. They found that when different numbers of food items that had equal mass were treated as equivalent, subjects showed no preference for many small pellets over fewer large pellets, and only chose the larger number when the other choice was zero. Additionally, when number was held constant (at 1) and mass varied, pigeons showed a significant preference for the larger mass, and transferred to the summation pairs. Consequently, these results would suggest that number is not a primary cue for responding; subjects preferentially represented and responded by mass over number, although this may be dependent on the use of food items as stimuli.

All experiments discussed thus far have required subjects to select the larger of the

presented arrays, and showed largely positive results. However, these findings may not necessarily reflect an understanding of number or numerosity, but merely an adaptive perceptual strategy of selecting the larger food source, or the symbol associated with the larger food source. Boysen and Berntson (1995) investigated this question. They used a two-subject task in which two chimpanzees, with differing experience with numerical training, were presented with two arrays of candy of different numerosities. One subject, the "selector", was required to choose one of the arrays, which was then given to the other subject, the "observer". The researchers found neither subject was able to learn the optimal strategy of selecting the smaller array despite their extended training with numerical discriminations and their obvious distress at acquiring the smaller amount. In fact, performance was at chance for the less number-experienced subject, and was significantly lower than chance and became less optimal as the reward ratios increased for the more experienced subject. Performance of the more experienced subject differed when the food arrays were later replaced by numerical symbols; choices immediately and reliably reflected the more optimal strategy of selecting the smaller value. Unfortunately, the less-experienced subject was not tested with numerical symbols. These findings show that although subjects were able to discriminate and select the smaller quantity of food when represented symbolically, they were not able to overcome the competing natural tendency to choose the larger array when food was used as stimuli. The use of representational stimuli allowed subjects to demonstrate their understanding of the numerical task free of the natural constraints on foraging responses.

A stronger demonstration of sponftaneous numerical ability would be the successful discrimination of number using non-food stimuli in natural settings. Jordan, Brannon, Logothetis and Ghazanfar (2005) showed monkeys were able spontaneously to represent and discriminate number cross-modally; matching the number of conspecific voices heard to faces presented on a video display. Discrimination between the quantities of two vs. three was tested in a habituation-discrimination procedure (similar to Wynn, 1992; Hauser & Carey, 2003; Flombaum et al., 2005).

Auditory stimuli were presented via a speaker located between two monitors which each displayed 60s of a continuous loop of 1s video showing either 2 or 3 simultaneously vocalizing rhesus monkey faces. The videos were edited to make the mouth movements synchronous, and vocalizations were equated for duration and amplitude and were also synchronized with the videos, so that amodal cues could not be used to match.

Results showed that the majority (75%) of monkeys would look significantly longer at the display that matched in number of faces the number of voices heard than the display that did not, spending, on average, 60% of their total looking time looking at the numerical match. This difference held for monkeys that heard two calls as well as three calls, suggesting that subjects were spontaneously separating and enumerating the vocalizations and matching them to the visual displays that were presented. The fact that each subject only experienced one trial of either the two or three-composite stimulus emphasises that previous training in this discrimination was not needed for successful performance. The synchrony of the visual and auditory elements would have prevented subjects from using synchrony cues, or the temporal cues of rate or duration of vocalization to match stimuli. Thus, it appears that rhesus monkeys are able to represent the numerical correspondence of multi-modal cues spontaneously, despite overlap and synchronicity of presentation.

Similar research has showed that cotton-top tamarins are able to discriminate the relative numerosity of voice syllables spoken by humans in another habituation-discrimination procedure (Hauser, Tsao, Garcia & Spelke, 2003). Stimuli consisted of spoken Consonant-Vowel (CV) syllables spoken by an adult female, and two adult males, one with an average pitched voice and one with a low-pitched voice. These were selected based on previous research that had shown that CV syllables would elicit strong orienting responses in a habituation-discrimination procedure. Subjects were initially familiarized with eight exemplars of each target number, played ten times each in a randomized order. Speaker and syllables were varied in the familiarization stimulus set, while syllable durations and inter-stimulus intervals (ISIs) were held constant. The constant syllable durations and ISIs resulted in the total sequence durations being longer for larger target numbers, and

this was controlled for by holding total sequence duration constant while varying the syllable durations in the test sequences. These manipulations ensured that sequence or item duration, and rate or amount of acoustic energy was not correlated with changes in number.

Test sequences were presented immediately after the familiarization set. Six test trials in total were presented, alternating between three trials with the same target number as in familiarization, and three trials of a different number. The test subjects were divided into two groups, and across these groups, familiarization number and whether the first test trial was the same or different number to the target number was controlled. The recorded response was a head turn or orientation to the speaker playing the syllables.

The test trials in their first experiment examined contrasts between the numbers 4, 8, 6 and 5, and found subjects responded significantly more often to the different number than the number to which they were familiarised, for the values 4 vs. 8 and 4 vs. 6. However, no significant difference in responding was found for 4 vs.5 discrimination. The second experiment aimed to further test whether discrimination thresholds were based on the difference or ratio of the two numerosities. Test trials involved discriminations between 8 vs. 12 and 8 vs. 10, and found subjects responded significantly more to the former comparison, but not the latter. That is, the tamarins discriminated the numerosities at a 1.5, but not 1.25 ratio. Analyses across the 4 vs. 6 and 5, 8 vs. 10 and 12 conditions in Experiments 1 and 2, showed a significant effect of numerical ratio on mean proportion of responses on the test trials; subjects responded significantly more frequently with a ratio of 1.5 than 1.25. No significant effects or interactions for set size. Hauser et al. (2003) concluded that this shows discrimination was dependent on the ratio between the set sizes, and independent of the absolute magnitudes.

Jaakkola, Fellner, Erb, Rodriguez and Guarino (2005) conducted an experiment with dolphins which avoids some of the issues encountered by procedures which use food as stimuli, and in doing so, also demonstrated nonhumans are able to select the fewer set of two sets of items. In two separate experiments, subjects were trained to select one of two arrays that

contained a fewer amount of dots. The arrays of dots varied in terms of size and position, and inadvertent cuing was avoided by having a "blind" trainer carry out the training and testing. After initial training with 3 sets of comparisons of pairs between 1-7, and varying in ratio, testing was then carried out with all possible pairwise comparisons between the values 1-8. Results from both experiments showed subjects were able to select the "fewer" sample on more than 80% of the novel comparisons, and performance was not significantly related to surface area of dots. Error analyses showed a decrease in number of errors as the ratio between the two numerosities increased, consistent with Weber's law. Jaakkola et al. also found performance dropped to chance levels for the 7 vs. 8 comparison- this could be interpreted as a set-size limit, however this is larger than the normally predicted set-size limit of 4 and this result is also predicted by Weber's law. Their experiment shows that bottlenose dolphins were able to discriminate relative numerosity and make ordinal judgments about them, without having to learn each pair-wise comparison individually.

The foregoing review has shown that there is considerable evidence that nonhuman animals are able spontaneously to discriminate relative numerosity in natural settings. One limitation of these studies is the difficulty in distinguishing responding based on numerical characteristics from responding based on other stimulus characteristics. Results of Olthof et al. (2000) and Boysen and Berntson (1995) indicate that animals have a bias towards responding on the basis of food mass, instead of number. Given the natural tendency to select the larger of two sets of food, this raises questions about whether subjects in many of these studies were truly responding on the basis of number alone, or whether their responding was influenced by confounding cues. The often unavoidable covariation between number and a myriad of other characteristics, such as temporal variables like duration and rate of presentation for sequential stimuli, or visuo-spatial variables like area, location, or mass for simultaneously-presented stimuli makes strict testing of numerical abilities difficult in less-controlled environments outside of the laboratory.

2.1.4 Numerical bisection

As well as permitting greater control over non-numerical cues, another advantage of laboratory tasks is that they allow the testing of more complex behaviours. The studies discussed so far have only required subjects to make one type of response, selecting either the smaller or larger. A more difficult and stronger test of relative numerosity understanding would require subjects to categorise stimuli as either a "small" or "large" number. This sort of discrimination is tested in numerical bisection tasks. A typical bisection procedure involves baseline training with two exemplars or samples differing in number and controlling for other cues such as area or stimulus duration. Subjects are required to choose one key following the "small" sample, and another key following the "large" sample. Baseline training is followed by testing with novel intermediate values.

Early research focused on the ability of rats to discriminate between different numbers of sound events. A typical experiment was performed by Fernandes and Church (1982) who trained rats to discriminate successive sound events. In order to obtain reinforcement, subjects had to press the right lever ("few") after the presentation of 2 sounds, and the left lever ("many") after the presentation of 4 sounds. Subjects were initially trained with three different sequences to prevent classification based on the duration of the stimulus sequence. The three different sequences were 2 bursts of 0.2s of white noise of separated by either a short (0.8s) or long (2.8s) interval and 4 bursts of 0.2s white noise separated by short 0.8s interval. Thus, inter-sound intervals remained constant across all three sequences, and sequence duration of the 2-long and 4-long sequences were equal. Thus, temporal variables alone were not reliable predictors of the correct response. Under these conditions, subjects were able successfully to discriminate between these two values, and the sequence structures were such that temporal cues alone were not sufficient determinants of performance. Subjects were also trained and tested with novel sequences, which varied in terms of the number of sounds (2 or 4), and inter-sound intervals

(between 0.8s and 2.8s), as well as the length of the white noise bursts (0.2s or 0.4s). During testing with the novel sequences, performance gradually worsened as subjects increasingly responded to trials with 4 sounds with longer inter-sound intervals as "few", rather than "many". Fernandes and Church suggested that this may have been due to subjects stopping counting prematurely after a particular duration had passed, and responding on the basis of the shortened count. Further evidence for the influence of sequence duration was also found; performance was greatest for 2-short sequences than 2-long or 4 sequences, and the sequence most frequently misclassified in the last 5 days of training was the 4 sequence used in original training, which was also the only 4 sequence whose total duration overlapped 2 sound sequences.

Meck and Church (1983) investigated whether numerical and temporal discriminations can be made independently, and whether a single system might be responsible for both these discriminations. Ten rats were trained to discriminate between two sequences of white noise bursts; a response to the left lever was reinforced following a two cycle noise of 2s duration, and a response to the right lever was reinforced following an eight cycle noise of 8s duration. Thus, number and duration were perfectly correlated. Subjects were then tested for number and duration discrimination by respectively holding either duration or number at an intermediate value, while varying the other variable among values 2, 3, 4, 5, 6, and 8. Results showed that the probability of a "long" or "many" response increased as a sigmoidal function of signal value. Additionally, the point of subjective equality was close to 4, the geometric mean of the two extreme signal values. The psychometric response functions for time and number were similar when the ratio of the extremes was constant; suggesting that either two different mechanisms with the same sensitivity were used for processing time and number, or the same mechanism was used for both. As the latter was more parsimonious, this possibility was further investigated in a second experiment.

If both time and number are represented and processed by a single mechanism, then a manipulation that affects one should also affect the other. Meck and Church (1983) administered

1.5mg/kg of methamphetamine to rats in this task to determine whether the drug would exhibit a selective or global effect on numerical and temporal discrimination. Previous research had shown that administration of amphetamine in a similar temporal estimation task resulted in a leftward shift in the psychometric functions and consequently also the PSE, suggesting an increase in pacemaker speed. The procedure from Experiment 1 was modified to maintain performance; responses on the left key were reinforced after either a 2s or 2 cycle signal, and a right key was reinforced after an 8s or 8 cycle signal. For training, signals presented on trials were either time or number relevant. For time-relevant trials, the number of signals were held constant at 4 cycles and were either 2 or 8s in duration, while the opposite was true for the number-relevant trials. Subjects were exposed to three saline and methamphetamine test sessions each, on alternating days. Also during testing with saline and methamphetamine, half of trials in the session were unreinforced probe trials where either the number or duration was held constant at 4 cycles or seconds, while the other variable varied between 3, 4, 5, and 6 cycles or seconds, respectively.

The psychometric functions for both the temporal and numerical discriminations exhibited a significant, leftward shift of about 10% for the methamphetamine sessions relative to the saline sessions. Sensitivity was similar to that obtained in the previous experiment. This suggests that numerical and temporal processing may share a common mechanism based on an internal pacemaker, which is affected by changes in dopamine and consequently the administration of methamphetamine.

A final experiment by Meck and Church (1983) investigated the mapping of number onto duration; in particular how much of an increment in one numerical unit was equivalent to an increment in one unit of time. It was predicted that, if time and number were processed using the same internal pacemaker, one response rule applied to number could also be applied to time. For initial training, 6 rats were divided into two groups, one group was reinforced for pressing either a stationary lever following a 2s white noise signal, or a moving lever following an 8s white

noise signal, while the other group was reinforced for responding according to the opposite rule. Additional durations were then added in probe trials, replacing half of the training trials per session. These signal durations were values between the two training values, logarithmically spaced; 2.0, 2.2., 2.5, 2.8, 3.2, 3.6, 4.0s. Subjects were then tested with five different test signals of 10, 12, 14, 16 or 20 cycles of white noise. The noise signals could be categorised based on either the total duration or the number.

Psychometric functions for duration and number mapped directly onto each other, using a least squares method of analyses. Meck and Church (1983) found that the median signal value for time and number associated with 50% long and many responses (the PSE) was equal to 2.84s and 14.1 segments, both closer to the geometric and arithmetic mean, and providing a ratio of 200ms as an estimate of time equivalent to each count.. A model based on this assumption was able to account for 99% of the variance in the data, with no systematic deviations.

The results of Meck and Church's (1983) research show that rats are able to process and retain numerical and temporal information simultaneously. Psychophysical functions obtained for these discriminations were indistinguishable from each other, and were equally affected by methamphetamine administration, suggesting the same pacemaker was used to process both types of information.

Later research investigated the simultaneous processing of time and number, extending the research of Meck and Church (1983). Roberts and Mitchell (1994) adapted Meck and Church's discrimination procedure, requiring pigeons to discriminate between sequences of light flashes that varied in both duration and number. Their first experiment involved the discrimination between sequences of 2 and 8 flashes, which had total durations of 2 and 8s, respectively; time and number were confounded. Subjects were reinforced for pecking the right key following one sequence of flashes, and for pecking the left key following the other sequence of flashes. Once subjects had acquired this discrimination with 80% accuracy, probe tests were introduced to test for control by number and time. On timing tests, the number of flashes

presented was held constant at 4, while the interflash intervals were manipulated such that the duration of the flash sequence ranged from 2, 3, 4, 5, 6, 7, 8s. Conversely, on counting tests, the duration of the flash sequence was held constant at 4s, while the number of flashes were 2, 3, 4, 5, 6, 7, and 8. Fourteen probe trials were interspersed among 50 training trials. All responses on probe test trials were reinforced to maintain response rate, and differential reinforcement continued on regular training trials.

Findings replicated those of Meck and Church (1983); psychophysical functions showed that the proportion of large/long responses increased as time and number increased. Control by time appeared to be stronger than control by number; psychophysical curves for the former were steeper suggesting subjects were better able to discriminate 2s from 8s than 2 from 8 flashes.

To test whether control by numerical and temporal variables could be influenced by additional training, Roberts and Mitchell (1994) conducted two additional experiments. In the second experiment, subjects were specifically trained to discriminate the number of flashes; subjects were required to discriminate two from eight flashes by pecking either the left or right key, respectively after their presentation. The total duration of the flash sequences was 2s, 4s, or 8s. Roberts and Mitchell noted that two of the sequences, 8 flashes/2s duration, and with 2 flashes/8s duration were ambiguous, because the predicted response differed depending on whether time- or number-based responding was assumed. The pacemaker accumulator model predicts chance performance with these sequences as it assumes that timing and counting are processed identically beyond the accumulator stage. Therefore, if subjects are able to perform at levels significantly above chance with these sequences, temporal and numerical information must be processed separately in working memory.

There was a large, significant discrepancy in performance with different flash sequences early in training; percentage correct on sequences identical to those in previous training (with time and number confounded) was at least 75%, whereas percentage correct on ambiguous trials ranged from 0% to 40%. Additional training resulted in an improvement in discrimination

accuracy, with average percentage correct for the last 5 sessions not differing significantly and exceeding 75% on all trial types. These results show that pigeons are able to discriminate between two and eight flashes when sequence duration was not correlated with number. Subjects appeared to respond initially on the basis of time, resulting in the poor performance on the new trial types; however accuracy improved with subsequent training. This suggests that subjects learnt to respond purely on the basis of flash number.

Subjects were then placed in the time- and number-tests used in Experiment 1. Baseline trials with 2 flash/2s and 8 flash/8s sequences were presented with either time- or number-controlled trials, where one variable was held constant and the other varied between values 1-8. Sessions containing time or number-probe trials were presented on alternating days. Of particular interest was whether superior control by number would be obtained, or whether control by time and number would be equal.

Results suggested that control by time and number had equalized. Response curves for number showed numerical control had increased from the first experiment. Additionally, temporal control appeared to have been weakened by the additional explicit numerical discrimination training. Thus, the pigeons' numerical and temporal discrimination ability appears to be manipulable, and strengthening control by one dimension occurs at the detriment to the other.

Roberts and Mitchell (1994) conducted a third experiment to determine whether subjects could respond on the basis of either number or time when cued. Half of the trials within sessions were the familiar, time and number confounded 2 flash/2s and 8 flash/8s sequences, while the other half were the ambiguous trials, 2 flash/8s and 8 flash/2s sequences. Side keys were lighted red or green, to differentiate between trials where time-based responses or number-based responses were reinforced, respectively. So, for instance, if the red keys followed the presentation of an 8 flash/2s sequence, the left key was the correct choice, whereas if green keys followed then the right key would be correct. Roberts and Mitchell reasoned that if subjects

were able to successfully respond on the ambiguous trials using the postsequence cues, this would suggest that time and number are differentiated during working memory, unlike the original Meck and Church (1983) mode control model.

Results showed that, given extended training, pigeons were able to process both temporal and numerical information simultaneously, and respond accurately to either time or number based on postsequence cues. Thus, results were consistent with Roberts and Mitchell's (1994) claim that time and number are distinguished in working memory. However, it is also possible that subjects learnt a conditional discrimination based on another feature that covaried with duration or number in the ambiguous flash sequences. Flash rate, or inter-flash interval was another cue which differentiated between the 8 flash/2s and 2 flash/8s sequences, and subjects may have been responding to this, rather than duration or number.

A final experiment was conducted to test this possibility. Subjects were still required to make either time- or number-based responses, but in addition to the trial types used in the previous experiment, probe trials where either duration or number was held constant at 4s or flashes while the other varied from 2 to 8s or flashes were also included. This manipulation would vary flash rate differently on the time or number controlled sequences, and so responding should change accordingly if subjects were using flash rate as a cue.

Performance on the training and ambiguous trials was maintained and was significantly higher than chance. Response curves for the number and time-controlled probe trials revealed significant control by time and number, when their respective cues were presented after the flash sequence. Furthermore, when test key cues did not match the dimension tested, flat curves showing little control for that dimension was observed (e.g., when number control sequences were followed by the timing cue, or vice versa). This shows subjects were not using flash rate as a conditional cue, otherwise performance on these trials would have varied systematically.

Roberts and Mitchell (1994) reported several significant findings. Firstly, it appears pigeons are able to process both temporal and numerical information simultaneously, keeping

track of both the duration and number of flashes presented sequentially. This has specific implications for the mode control model, which does not differentiate between temporal and numerical representation beyond the accumulator stage (Meck & Church 1983). Consequently, Roberts and Mitchell proposed a modification to the model to allow the separate storage of information about time and number in working memory. Another noteworthy result was that equivalent control by both time and number was obtained, and the strength of their control could be influenced by training. Subjects could also be trained to respond on the basis of time or number on ambiguous trials, even though the correct time- and number-based responses were different, and the cue for which response was required was presented after the sequence.

Alsop and Honig (1991) examined relative numerosity judgments with pigeons with sequential visual stimuli. Subjects were presented with red and blue flashes of light, and had to peck either the left or right key if there were more blue or red flashes, respectively. Training was conducted with seven flashes, and testing involved five or nine flashes. Results from their three experiments showed that subjects were able to successfully discriminate "more" vs. "less" to a certain extent; performance was also influenced by some confounding factors. The order of the flashes affected discrimination; responses were biased towards the colour of later-occurring flashes. This finding is a reflection of the successive processing; with the relative numerosity judgments dependent on memorial processes, and the rapid decay of the stimulus elements in memory. Performance was also affected by the duration of the flash, with longer flashes appearing to have greater influence over responding. This may be a result of an increase in stimulus discriminability, or a decrease in stimulus decay, or a combination of both. Although subjects were able to respond accurately in this procedure, it is unclear whether subjects were responding on the basis of numerosity, or merely the total duration of the individual red and blue flashes. The authors suggested different methods for testing this possibility, but the confounding of number, duration and salience made determining the precise nature of discrimination in this procedure difficult.

A similar procedure was used by Keen and Machado (1999) to examine how relative

numerosity discriminations were affected by the difference in number of elements between the two samples, and the total number of elements. Pigeons were shown red and green light sequences on different side keys, and responses selecting the sequence containing fewer elements were reinforced. The researchers used a comparatively large range of values; across different conditions, the total number of elements ranged from 4 to 28 in multiples of 4, with differences ranging from 0 to 14. Generally, results showed that discriminative accuracy increased with a Weber-like function (difference between samples/total number of elements) when the difference was greater than 0. Accuracy tended to decrease as the total number of elements increased, a phenomena known as the "size effect" (Moyer & Landauer, 1967). Additionally, a "distance effect" was also found; when the total number of elements was held constant, accuracy increased as the difference increased. One critical procedural feature of this study was that stimulus delays were response-dependent, that is, a key-peck was required to end the presentation of each sample stimulus and continue the trial. This was to ensure subjects attended to the sample stimuli, but also increased the correlation between sample duration and frequency. Although temporal variables played some role in determining performance, regression analyses suggested stronger control by number than duration; the cumulative frequency of the two stimuli was a better predictor of performance than the cumulative duration.

2.1.4.1 Bisection of response number

Rilling and McDiarmid (1965) investigated pigeons' ability to discriminate between two fixed ratios. Two pigeons were trained to discriminate between one ratio that was held constant at FR50 (referred to as the "noise") and another ratio (referred to as the "signal"), which started at FR5 and was increased to FR35 once a criterion of 90% correct had been reached. Following that, the signal FR increased daily in increments of 2 until performance dropped below 60% correct. They found that performance decreased gradually as the difference between the signal and noise fixed ratios decreased, and fell below 60% when the signal ratio was a FR47 (difference of 3). It was concluded

that the gradual deterioration in performance suggested that the ability to discriminate response number was a continuous, rather than all-or-none process.

Fetterman (1993) conducted an experiment specifically to test the relative roles of number and duration in determining responding in a conditional discrimination of fixed ratios (FR). In his procedure, a response to one of two lighted keys was reinforced if the FR just completed was a relatively small number of key pecks, e.g. FR10, and a response to the other key was reinforced if the FR was large, e.g. FR30. Relative and absolute differences between ratios were varied across conditions, and within each condition subjects were trained with two FR values, and then tested with values intermediate to the training ratios in probe trials. Following this, subjects were then exposed to a final time-discrimination test, where subjects had to discriminate the times taken to complete the FRs in the last number discrimination condition. This transfer test involved the discrimination of the times for both the training and probe values.

Psychometric functions plotting the probability of a large response as a function of response number were calculated from the probe trial data for each subject. For all subjects, the probability of choosing "large" increased as relative ratio value increased, and functions were orderly and ogival. Responding was consistent with a scalar representation; functions for the different training ratios superimposed when plotted on the relative scale, and standard deviations were proportional to the magnitude of the stimuli.

The time from the first to last ratio response was used for the temporal analyses, as this was the temporal variable with the strongest correlation with choice. Psychometric functions based on the relative time to complete a ratio were of similar form to those of response number, although superposition was less evident. Scatterplots of ratio time plotted against ratio number show that there was some temporal influence on responding.. All birds were more likely to choose a large response when ratio time was long, rather than short, however the extent of this influence varied between subjects, with some showing stronger temporal control than others. Fetterman also calculated separate point biserial correlations between ratio time and choice for each probe FR. These were

significant or approaching significance for the majority of subjects and ratios, suggesting that ratio time exerted some control over responding, above and beyond response number.

To confirm this, multiple regression analyses were conducted to investigate the relative control of time and number over choice responses. Results showed that in 9 out of 11 cases, ratio time accounted for a unique proportion of variance after ratio value had already been included in the regression model, and in 8 of 11 cases, ratio value accounted for a unique proportion of variance above and beyond ratio time. Individual data showed variability in numerical and temporal control, with some birds showing stronger control by number than time, or vice versa. However, the high correlations between ratio value and time, some reaching 0.70, requires that caution must be taken when interpreting the results of the regression analyses, due to the strong multicollinearity between the predictor variables.

Results of the transfer tests to the time-based discrimination further corroborated other analyses. Psychometric functions of temporal discrimination performance were similar to that of numerical discrimination, with the probability of choosing the large alternative increasing as both time and number increased. Disruption of discrimination performance was more evident for some subjects than others, suggesting subjects relied on both temporal and numerical cues to differing degrees. This was confirmed by the obtained individual differences in beta weights.

The main finding from Fetterman's (1993) research was that responding was controlled by multiple cues, temporal and numerical, and the extent of this control was subject to individual variation; some subjects appeared to have a greater natural propensity towards attending to number rather than time, or vice versa. However, the regression analyses, which were central to his experiment, had the issue of multicollinearity due to the strong covariation between the temporal and numerical variables, so any conclusions must be hesitantly drawn. Further experiments that address the problem of covariation would be have to be conducted before any solid statements can be made about multiple stimulus control in numerical discriminations.

2.1.5 Differentiating numerical and temporal control

The possibility of confounding temporal and numerical variables can be reduced by presenting stimuli simultaneously, rather than sequentially. With this method, visual characteristics such as area and size of the array may vary with number, however, these seem to be easier to control for than temporal cues.

Honig and Stewart (1989) trained three pigeons to discriminate between arrays that consisted of a matrix of two elements that varied in colour, shape or size. Training was conducted with uniform arrays of red and blue dots; one element was the S+ and the first response to this stimulus array after 20s stimulus was reinforced with access to food, the other element was the S- and all trials with this stimulus ended after 20s with no reinforcement. After performance had reached high levels in discrimination training (95-99% responding to S+), testing was conducted in which arrays consisted of differing proportions of the S+ and S-. The ratios of S+: S- used in test arrays were 36 and 0, 25 and 11, and 21 and 15. The S+ and S- values for each pair of numbers were counterbalanced across participants and the locations of the S+ and S- stimuli within the matrix was randomised. No reinforcement was given on test trials. Results of tests showed that responses to the stimulus decreased as the proportion of positive elements decreased, which suggests that subjects were responding based on the relative numerosity of the positive elements presented in the arrays.

To assess whether responding was based on the actual number of red and blue elements or their relative proportions, Honig and Stewart (1989) conducted transfer tests with larger numbers, using matrices of 64 elements instead of 36. Although the number of elements had increased, the relative proportions remained the same. Additionally, the array was slightly larger, and dots were more closely spaced. Subjects were tested with these 64-dot arrays after several additional sessions of training with all red or all blue 36-dot arrays. After this, two of the three subjects also experienced a second set of tests following an additional 6 training sessions with 36-dot arrays. Discrimination prior to these tests was close to perfect, and response functions were similar to those obtained with the

36-dot arrays, although somewhat steeper. No consistent, systematic effect of the changes in stimuli on overall responding was found; for two subjects, responding declined on the first or the second transfer test, while the third subject exhibited a significant increase in response rate in its first and only set of transfer tests. This did not appear to affect the response gradients, and their similarity with responding obtained with the 36-dot arrays. The similar performance under these conditions suggests subjects were responding to the relative proportions of the red and blue dots presented in the arrays, rather than the absolute number of the elements.

In additional experiments, Honig and Stewart (1989) varied the form and size of elements to investigate whether subjects were using other characteristics to determine responding. They trained four new pigeons were trained to discriminate between 36-element arrays with Xs and Os instead of red and blue dots; these new elements would not allow subjects to respond based on the relative area of the S+ and S-. Training conditions were the same as the previous experiment, with half of the subjects trained with Xs as the S+ and Os as the S-, and the other half of the subjects were trained with the opposite. After 12 training sessions, subjects were placed in two consecutive test sessions where the proportions of elements were manipulated. Subject were then given an additional three training sessions before being placed in a two-session transfer test with 16-element arrays with differing proportions. Response gradients for the three subjects that acquired the discrimination showed the same linear decreasing pattern previously obtained; response gradients were also similar with the 36- and 16-element arrays. These results replicate and extend Experiment 1. Despite the manipulation that prevented the use of area as a cue for responding, subjects discriminated the relative, not absolute, number of elements in the array, and responding was not affected by a decrease in the number of elements contained in the array.

Honig and Stewart (1989) manipulated the size of elements in a third experiment as an additional test of area discrimination. They noted that by varying the size of elements, confounds were also introduced, such as the total area of the fixed number of elements and the distance between large or small items. They attempted to limit these confounds, and reported that subjects tended to

respond on the basis of the size of elements, rather than other covarying characteristics of the array. Six subjects were trained with arrays consisting of 36, 25, 16 and 9 small dot and 25, 16, 9 and 4 large dots. All arrays took up the entire display screen except for the four large dot arrays. These were concentrated in a quadrant of the screen, varying between trials, in order to maintain similar spacing between elements to the other arrays. Three sets of mixed arrays, consisting of 9, 16 and 25 elements, were used for testing. Six uniform training arrays were also used, consisting of proportions of 3 and 6, 4 and 12, 8 and 8, 20 and 5 and 10 and 15 either large or small dots. All patterns were made up of four different randomized arrangements of dots. Half of subjects were trained with large dots as S+, and half were trained with small dots as S+.

All subjects first received initial training with positive arrays only, before receiving discrimination training with both positive and negative uniform arrays. After performance had stabilised, transfer tests began. A first transfer test involved two sessions in which combinations of both uniform and mixed arrays containing 9 and 16 elements were presented. After additional training sessions, a second two-session transfer test with uniform and mixed arrays of 25 elements was then conducted, followed by five more additional training sessions and a final transfer test containing all 25 and 16 element arrays.

All subjects acquired the discrimination, reaching at least 90% accuracy, albeit acquisition for some occurred faster than others. Performance was also maintained in the training sessions presented between transfer tests. Response gradients were orderly and followed the same decreasing pattern previously obtained. Gradients became steeper with extended training and testing. Results suggested that pigeons were able to discriminate between the relative numerosities of elements that varied in size. If discriminations were based at least partially on the total area, subjects should have responded more to arrays containing a larger number of the S+ element (e.g. 25) than to arrays containing a fewer number (e.g. 4). Inspection of the total number of responses revealed no such trend, suggesting total area did not influence response rates. Responding was influenced by the total area covered by the elements. In the first and third transfer tests, arrays consisting of different numbers of elements

were presented concurrently. If subjects were influenced by the total areas of the elements then birds trained with large dots as S+ should respond less to the arrays containing fewer S+ elements than those containing more S+ elements, for any given proportion of S+ and S-, whereas birds trained with small dots as S+ should show the opposite trend. This pattern was not found for birds trained with large dots as S+, but was found for the birds trained with small dots as S-, suggesting area of elements within the presented matrices may have influenced responding. Honig and Stewart (1989) also tested for transfer to different sized dots, replacing the large dots in arrays to medium-sized dots. Subjects quickly learnt the discrimination under these conditions, achieving at least 95% accuracy after seven training sessions. Response gradients were similar to those obtained with large and small dots, suggesting pigeons were not affected by changes in element size.

Whether subjects could discriminate the relative numerosity of elements from different "natural" categories, using drawing of birds and plants was also investigated (Honig & Stewart, 1989). Individual elements differed in colour and shape, though overall the bird stimuli were more solid, with slightly bolder and more saturated colours. Subjects were trained with uniform arrays of 16 birds and flowers. These were also included in test trials with arrays in the same proportions as the 16 element arrays in the previous experiment; 16 and 0, 12 and 4, 8 and 8. Subjects acquired this discrimination quickly, reaching at least 90% performance in no more than three sessions. Response gradients showed an orderly linear function, consistent with previous findings, suggesting that pigeons are able to assess the relative numerosity of complex stimuli from natural categories. It is possible that pigeons were only responding to the specific elements in each category, since all bird and flower pictures appeared in the training arrays; and thus gradients may have been a result of the decreasing presence of stimulus characteristics in testing arrays in which those stimuli decreased in proportion.

To test this, an additional experiment was conducted where naïve pigeons were trained with arrays of 9 elements from two new conceptual categories, unicorns and flowers. Independent sets of elements, which were not presented in training, were used to test for transfer to novel elements.

Subjects showed positive transfer between training and transfer arrays suggesting that pigeons were not responding to individual instances and searching for their presence in the test arrays. The total number of responses between training and transfer were similar suggesting there was no generalisation decrement. Overall, the results of this experiment suggest that subjects were able to identify and categorise elements correctly and discriminate their relative numerosity.

Emmerton (1998) conducted two experiments where pigeons were required to discriminate between visual arrays of different numbers of dots. In her first experiment, subjects were presented with a pair of arrays of dots that varied in size and were arranged randomly. To obtain reinforcement, subjects were required to select the smaller numerosity of the two arrays. Training and testing used numerosity values ranging from 1 to 7; initial training was conducted with 40 slides of the pairs 1/2, 3/5, 2/6, 5/7 and transfer trials were then conducted with novel exemplars of the same numerosity pairs replacing ten of the initial training stimulus pairs. After training and transfer, subjects were tested with numerosity shifts; the pairs 1/3, 3/7, 2/4 and 5/6 replacing the ten transfer pairs. Performance was high overall, and results showed that accuracy was largely influenced by the numerical difference between the stimulus pairs; subjects performed better when there was a larger difference between the S+ and the S-, a numerical distance effect. However, Emmerton suspected that array density, which tended to covary with number, might have also influenced responding; her second experiment investigated this possibility.

The second experiment of Emmerton (1998) was the same as the first, with two exceptions. Firstly, the numerosity judgments in which subjects' accuracy was especially high was changed from 2/4 and 2/6 to 2/3 and 2/4. Thus, subjects were required to select the smaller of the following numerosity pairs: 1/2, 1/3, 2/3, 2/4, 3/5, 3/7, 5/6 and 5/7. For any given session, the smaller correct numerosity remained the same while the large, incorrect numerosity could vary between two values. In addition, array density was manipulated so that items were either "near" or "far" for each array, resulting in four different possible density arrangements for any pair. Similar to the first experiment, subjects' accuracy increased as the numerosity difference between the arrays increased for all pairs

except for when both numerosities were greater than 4 (pairs 5/6 and 5/7). The arrangement of the array also influenced performance, albeit opposite to that found in previous research; instead of overestimating numerosity when dots were far apart and underestimating numerosity when dots were close together, results suggested that pigeons were doing the opposite. Emmerton explained this by proposing that when the dots were closer together, subjects were more likely to detect dots and consequently produce a correct response, however when dots were further apart the reverse was true.

2.1.6 Investigating the representation of number

The representation of number can be investigated using relative numerosity discriminations. Accuracy and psychometric discrimination plots (proportion of the "large" key choices as a function of number) provide information about how responding may be determined. One concrete finding obtained with bisection procedures is the superpositioning of psychometric functions when plotted on a relative scale. That is, when normalized, functions overlap, regardless of the absolute numerical range with which they were originally obtained. Superpositioning is consistent with both Weber's law and scalar variability (i.e., response variability increases proportionally with numerical magnitude, resulting in constant relative variability).

Bisection points, also known as the point of subjective equality (PSE), can be calculated from the psychometric functions. The bisection point is the numerical value where the proportion of "large" responses is at chance (50%). The location of bisection points may provide some information about the nature of the underlying numerical scale. Gibbon (1981) discussed the interpretation of the location of bisection points in temporal discrimination procedures; his arguments would apply to numerical discrimination also, assuming similar underlying principles.

If representations of number were scaled logarithmically, with constant generalization between values and discriminations were based on the relative similarity of ratios of the current stimulus to

each training stimulus, the bisection point of the learned values should be located at the geometric mean. This is also predicted if the scaling of the numerical representation were linear with scalar response variability, and if the same similarity judgment rule was applied. A different bisection point location with a linear number scale exhibiting scalar variability is predicted if a "proximity" rather than "similarity" rule was used; if subjects responded based on which anchor value most likely generated the noisy test stimulus, then the PSE would occur at the harmonic mean. If numerical representations were linearly spaced, and there is constant rather than increasing variability across numerosities, then bisection points should occur at the arithmetic mean (Gibbon, 1981).

Research has obtained somewhat conflicting evidence with respect to bisection points. The majority of counting and timing experiments have obtained PSEs at the geometric mean (e.g. Fetterman, 1993, Roberts, 2005), although others also have obtained bisection points at the harmonic mean (e.g, Fetterman, Dreyfus & Stubbs, 1985) and sometimes in between both (Fetterman, Stubbs, Dreyfus, 1986). Some research with humans has found bisection points at the arithmetic mean (Droit-Volet, Clement, Fayol, 2003). Consequently, based on bisection point data alone, the nature of the underlying subjective numerical scale is still unclear.

There are alternative methods of investigating numerical representation in bisection procedures. The analyses used by Fetterman (1993) permitted the investigation of the scaling of subjective number. Subjects were trained on a variety of different scales, in which absolute and relative ratio sizes were manipulated and psychophysical analyses were conducted to determine the parametrics of the numerical scale used to determine responding. Data generally conformed to scalar principles. Psychometric functions for scales with a 2:1, 3:1 and 4:1 ratio all superposed when plotted on a relative scale, consistent with Weber's law and previous temporal discrimination experiments (e.g. Meck & Church, 1983, Fetterman & Killeen, 1992). Variability, measured as the standard deviation of the psychometric functions, increased proportionally with numerical magnitude; standard deviations increased as a linear function of the bisection point. Additionally, relative variability, measured as the coefficients of variation or the standard deviation divided by the bisection point value

(referred to as the Weber fraction), remained constant as number increased, again, consistent with Weber's law and scalar variability. Fetterman (1993) also examined the location of the bisection points and plotted these against the geometric, arithmetic and harmonic mean to determine which had the best fit. Fits were similar, but the harmonic mean was the worst predictor of bisection points of the three, accounting for 94.6% of total variance, compared to 97% and 96% for the geometric and harmonic mean respectively. The arithmetic mean was a better predictor for small FR values, whereas the geometric mean was a better predictor for large FR values. The slope for the geometric mean plot, but not the plots for the arithmetic or harmonic mean, did not differ significantly from 1. These results would suggest that the geometric mean was the best estimate of the subjective midpoint for the relative numerosity discrimination.

Emmerton and Renner (2006) extended this research further, replicating Fetterman's (1993) experiment with visual simultaneously presented stimuli consisting of combinations of 5, 10, 15, 20, 40, 60, and 80 dots, while controlling for brightness and density of the visual array. Another unique feature of their experiment was the testing of transfer to novel numerical values both outside and inside the training range; previously no experiments had tested extrapolation of numerical discrimination of visual arrays in pigeons. Of particular interest was whether the data revealed the same scaling effects as found by Fetterman. Results were consistent with Fetterman's findings and supported scalar variability. All psychometric functions for the 2:1, 3:1 and 4:1 ratios superimposed when the ratio of the anchor values were held constant. Analyses of the bisection points showed that these did not differ significantly from the geometric mean, but did differ significantly from the arithmetic and harmonic means of the training values. The difference limen, half the difference of numerosity values that correspond to 25 and 75% of the proportion of large choices, was calculated as a measure of response variability, and increased proportionally to the bisection points. Dividing the difference limen by the bisection point values provides a relative response variability measure, equivalent to the Weber fractions calculated by Fetterman (1993). When plotting these against the bisection points, the slope of

the linear regression was almost 0 (slope = .0005), suggesting relative response variability remained constant as subjective estimates of numerosity increased.

Brannon, Wusthoff, Gallistel and Gibbon (2001) adapted the time-left procedure used by Gibbon and Church (1981), in which subjects had to respond on the basis of the difference between two numerical values, to test for a linear relationship between subjective and objective number. Subjects had to compare a numerical difference, which varied between trials, with a constant numerical value and choose the smaller value. If the subjective numerical scale is logarithmic, then the size of the difference relative to the constant value will be dependent on the ratio between the two numbers, rather than the difference. Therefore, if the ratio between the difference and the constant numerical value does not change, then the subjective difference will not increase. Conversely, if the subjective number scale is linear, the subjective difference should increase if the absolute difference between the two numbers increased, regardless of their ratio.

In their procedure, subjects were required to make an initial number of pecks (*I*), to an illuminated center key, which in turn produced light flashes in the food hopper, according to a variable ratio schedule so that the number of pecks required to produce a hopper light flash varied around a mean value. This manipulation was to reduce the correlation between time and number. After a variable number of hopper flashes (*T)* had been generated by center-key pecks, the center key was darkened and the side keys illuminated. Subjects were then required to make a choice between the "standard" and a "number-left" side keys, and peck until a certain number of hopper flashes, specific to that key, had been produced. On the standard key, the required number of hopper flashes (*S*) that had to be generated was always constant, whereas on the number left key the number of flashes was equal to *I-T*. Consequently, the standard key would be the better choice when *I-T* is greater than *S*, whereas and the number-left key would be the better choice if *I-T* is less than *S*

After initial training, subjects were tested in four different conditions, in which the *S* and *I* values were varied in terms of their ratio and absolute difference. During testing, a small proportion of trials were forced-choice trials, where only the standard or the number left key was illuminated. In

the first testing condition, the required number of hopper flashes (*T*) varied between 1-7, and *S* = 4 and *I* = 8. On the standard key, 4 hopper flashes had to be generated before reinforcement was delivered, and on the number-left key, 8-*T* flashes were required. In the second testing condition the values were multiplied by a factor of 1.5, so *S*=6 and *I* = 12, thus increasing the distance between the numbers while keeping the ratio the same. In this condition, *T* varied between 1, 2, 3, 4, 6, 8 or 10 flashes. The values were then returned to those used in the original testing condition for 10 sessions, before subjects were placed in the fourth and final testing condition. In this condition, the *S* and *I* values were changed to 3, and 6, while *T* varied between 1-5 flashes.

Brannon et al. (2001) plotted the probability of choosing the number-left key as a function of the number of flashes produced by pecking the center key (*T*) for each of the four conditions, and calculated the indifference point, the value for *T* at which the probability for choosing the number-left key was equal to 0.5, for each function. Thus, *T* was equal to the subjective difference between *S* and *I*, since subjects should have chosen the number-left key when $I-T < S$, and conversely should have chosen the standard key when $I-T > S$. It was predicted that if the indifference point increased when *S* and *I* values increased, but maintained the same ratio, this would provide evidence for a linear subjective numerical scale, since if the numerical scale had logarithmic spacing, then the subjective difference should only be dependent on the ratio, and not the difference of the two values. The probability of choosing the number-left key should increase as *T* increases, and this was confirmed by the results. Analyses showed that the indifference point shifted towards larger values of *T*, as the *S* and *I* values increased to 6 and 12, and shifted towards smaller values of *T* as the *S* and *I* values decreased to 3 and 6. Analyses showed that all four functions superimposed when plotted against the *T* divided by the maximum *T* value for each condition, suggesting that increasing and decreasing the *S* and *I* values by the same factor resulted in a proportional increase or decrease, respectively, in the psychometric function and consequently also the indifference points.

To check whether subjects' responding was controlled by the number of hopper flashes and not the time spend producing them, Brannon et al. (2001) plotted the cumulative probability of

terminating keypecks on the center key as a function of time spent pecking that key. Results showed that there was much variability in the durations for each of the different flash numbers, and consequently a considerable amount of overlap. To examine temporal control, the probability of choosing the number-left key was plotted as a function of time spent pecking further the center key for each value of *T*. The probability of choosing the number-left key did not change as time increased, but the former did increase for each increase in *T*, suggesting subjects were responding based on the number of hopper flashes generated, rather than the time spent generating them.

Brannon et al. (2001) proposed that their experiment demonstrated that subjects were able to perform numerical subtraction, shown by their reliable preference for the key associated with the smaller keypeck requirement. Additionally, the shift in the indifference point in the first block of testing trials after the new *S* and *I* values were introduced suggested that generalization was immediate, and consequently that pigeons appeared to have an abstract understanding of the task. Brannon et al. also claimed that the failure to obtain constant indifference points for different *I* and *S* values pairs with a constant ratio provided clear evidence for a non-logarithmic numerical scale.

However, Dehaene (2001) believed the findings of Brannon et al. (2001) could be explained more parsimoniously by assuming the pigeons were merely representing the first number of hopper flashes, *T*, and associating each value of *T* with the key that resulted in the smallest delay to reinforcement. Thus, subjects could learn the associations between the number of hopper flashes, responses and rewards, without having to use subtraction. To test this, he ran simulated a simple neural network that was able to learn number-response associations, using a similar schedule to that used by Brannon et al.

The output of Dehaene's (2001) simulator replicated the findings of Brannon et al.(2001); psychometric functions increased systematically as a function of the value *T*, and increasing or decreasing *I* and *S* linearly affected the location of the indifference point. Dehaene noted that although Brannon et al. argued that the indifference point shifts was indicative of a linear numerical scale, the linear increase in the indifference point with *S* was obtained by the simulator with both

linear and logarithmic scales. Another feature of the data noted is the sub-optimal location of the indifference point, which systematically shifted towards numbers smaller than the ideal value when *T* = *I-S*. This is not easily explained by Brannon's hypothesis, but can easily be accounted for by associative learning hypothesis; the variability, and consequently the overlap between representational distributions for coded numbers increase as magnitudes increased, and consequently activate increasingly overlapping sets of neurons, resulting ultimately in asymmetrical generalization towards larger numbers. Brannon et al. stated that the immediate generalization in the first block of transfer trials demonstrated the abstract knowledge subjects had developed about the task, however Dehaene points out that because reinforcement was still available for those trials, learning may still have taken place. The simulation, when a fast learning-speed was assumed, exhibited a small indifference point shift similar to that obtained by Brannon et al., providing additional support for an associative process rather than actual subtraction. A generalization test that used unrewarded probe trials would be needed to provide a stronger test of Brannon et al.'s claim.

To further investigate whether numerical scales are logarithmically or linearly spaced, Roberts (2005) trained pigeons to bisect two number scales, using 1-16 and 2-32 keypecks. In a trial, subjects were divided into two groups and were required to make either 1 or 16 keypecks, or 2 or 32 keypecks to a center white keylight. For all trials, except those only requiring 1 keypeck, the keylight was darkened for an interval which randomly varied between 0 and 2s between each keylight illumination. Once the required number of keypecks had been made, the side keys were illuminated red and green. If subjects had previously made 1 or 2 keypecks, a response to the green (small) key resulted in reinforcement and a response to the red (large) key resulted in a black-out, whereas if subjects had previously made 16 or 32 keypecks, the consequences were reversed. The location of the red and green keys varied randomly between trials. If subjects made a wrong response, the same sample number was repeated after the inter-trial interval up to four times. After performance reached a high level on these discriminations, subjects were tested with intermediate even numbers in unreinforced probe trials that were randomly distributed amongst the training trials. Subjects were still required to

make a choice between the small and large side keys, but were only reinforced for correct responses on training trials. Following training and transfer testing with one discrimination, subjects were then switched to the other discrimination.

Roberts (2005) found that subjects were able to perform in this task at very high levels of accuracy, reaching levels of 95% and 87% correct for the first 1 vs. 16 and 2 vs. 32 discriminations respectively. Positive transfer to the second set discriminations was found, with performance obtained at similar levels. Psychophysical curves were calculated for the 1 vs. 16 scale and 2 vs. 32 scale by plotting the percentage choice of the large key against the number of responses, and superimposed when plotted on the same axis. Superimposition suggests the discrimination of both scales were based on the same process. Bisection points were calculated for the 1-16 scale and the 2-32 scale, and were not significantly different from the geometric means; 4.0 on the 1-16 scale and 8.0 on the 2-32 scale. To examine whether subjects were discriminating time, not number, Roberts examined differences in the average duration of keypecks made on the center key as a function of red and green key choices. If subjects were timing, more response errors on trials with the small number would have been made when the duration of pecking was long, rather than short, and conversely, more response errors on large number trials would have been made with a short pecking duration than a long duration. Results found only two significant differences between average durations for the red and green key choices, one for each of the two scales, and only one of those, the mean durations for two pecks, was in a direction consistent with a timing hypothesis. Based on this, Roberts concluded there was little evidence supporting the hypothesis that subjects were responding on the basis of time, rather than number.

Roberts (2005) then conducted an additional experiment, in which midpoints of the 1-16 and 2-32 scales were preset at the arithmetic mean. Thus, only responses to the red key were reinforced if the sample number was located below the arithmetic mean, and only responses to the green key were reinforced if the sample number was above the arithmetic mean. This was to test the hypothesis put forth by Dehaene (2001) in his review of Brannon, Wusthoff, Gallistel and Gibbon's (2001)

subtraction experiment. Subjects were initially trained and tested with either the 1-16 scale or the 2-32 scale before being transferred to the other scale. Training began with the extreme scale values, and intermediate values were added once subject had reached 80% accuracy.

Results showed that overall accuracy was significantly greater for the full 1-16 scale than the 2-32 scale, although the general pattern across numerical values was similar. Roberts (2005) suggested there may have been greater interference from extra numbers in memory with the larger number scale; odd numbers, rather than the usual even numbers may have been represented due to greater errors in enumeration with the 2-32 scale. For both scales, accuracy increased as a function of distance from the midpoints, as expected. Interestingly, however, the accuracy plots were not symmetrical; close to the center of the scale, accuracy was significantly higher for larger than smaller numbers, whereas at the extremes of the scale, accuracy was significantly higher for smaller than larger numbers. Counting times were also analysed in the same manner as in Experiment 1, and only found one significant difference in the 1-16 scale that would suggest that subjects might have been timing. However, given there were no other significant differences found, it did not appear subjects were using time instead of number.

Roberts (2005) then generated an associative model based on the results of Experiment 1 and 2, and generated predictions based on two different assumptions; whether number was represented on a linear scale with scalar generalization, or whether it was represented on a log scale with constant degree of generalization. In the model, it was assumed that responses to the red and green key would gain equal amounts of associative strength after 1, 2 or 16, 32 keypecks, respectively. The probability of choosing the large key would be equal to the amount of associative strength for that response divided by the total associative strength for both keys, for any given number value. These probabilities were calculated and plotted for the logarithmic and linear scales. For Experiment 1, the probability of a large response function for the log scale increased gradually as number increased, and indifference between the small and large choices occurred at the geometric mean of 4 pecks. The linear scale function increased more steeply, reaching the maximum proportion of 1 at 3 pecks. Data

from Experiment 2 were modelled using the basic same assumptions as Experiment 1. However, as the reinforcement conditions differed, with responses to the red key being reinforced for values below the arithmetic mean for both scales and responses to the green key being reinforced for values above the arithmetic mean, it was assumed that associative strength increased at every number and also spread to adjacent numbers. Proportion of correct choices were then calculated, assuming either a logarithmic or a linear scale, by dividing generalized associative strength for the correct key for numbers above or below the arithmetic mean, by the total generalized associative strength. Both the logarithmic and linear functions showed the overall pattern of weakest performance near the arithmetic mean, with increasing accuracy as the distance between the sample number and arithmetic mean increased. However, the logarithmic function predicted the asymmetry obtained in the results around the midpoint better than the linear function did. Both functions show the higher accuracy for larger numbers than lower numbers around the midpoint, but only the logarithmic function showed the asymmetry at the extreme values found in the data. In contrast, the linear scale predicted the opposite pattern (greater performance with the large extreme values than the small extreme values).

Roberts (2005) attributed the asymmetries obtained in the data and predicted by the models to differences in generalization. As numbers increase on a logarithmic scale, the scale compresses, such that the higher numbers are closer to the midpoint than lower numbers. This results in greater generalization across the midpoint from the high to low numbers than from the low to high numbers. However, since the amount of generalization affecting each value is constant, there are limits to the distance associative strength can generalize across; greater generalization of associative strength for the "small" response would extend towards the higher numbers, due to the scale compression, and weaken the strength of the "large" response. Conversely, the strength of the "large" response will not generalize to the values at the other end of the scale because of the greater distance between them. This results in the asymmetries around the midpoint, and at the extreme ends of the scale. The prediction of the linear scale - greater accuracy for values above the midpoint than below- is due to the greater degree of generalization for higher than lower numbers. Thus, Roberts concluded that an

associative model assuming a logarithmic numerical scale, with constant generalization or variability, provided a better fit of the bisection data than a linear numerical scale with scalar generalization.

2.1.7 Current experiment

The research discussed so far has largely used bisection procedures in which temporal and numerical characteristics of stimuli covaried. Roberts and Mitchell (1994) showed that control by these variables over responding is affected by training. Although many researchers have demonstrated control by number through the manipulation of temporal and numerical variables in transfer tests, or in statistical analyses, the existing covariation between number and time during training may have reduced numerical control over behaviour.

The current experiment examines the performance of pigeons in a numerical bisection procedure in which the covariation of numerical and temporal cues was limited. In this procedure, subjects are presented with a sequence of response-dependent keylight flashes, after which they are required to choose one of two responses. The relationship between number and time was degraded throughout baseline training and testing to increase control by number and to reduce the influence of temporal variables. Both sample phase duration and flash rate were randomized to reduce correlations with number. Subjects were initially trained and tested with relatively small values, 2 and 6. Following this training, they were trained to discriminate larger numbers, obtained by multiplying the original values by 2 and 4 respectively; subjects were required to discriminate between 4 and 12 flashes, and 8 and 24 flashes. This manipulation would also allow a better investigation of other aspects of performance that might distinguish between a logarithmic and linear scale, such as superimposition of bisection functions and coefficients of variation.

Additionally, transfer tests included values outside of the training range; the majority of experiments to date have only examined transfer to numbers within the baseline training values

(for the exception, see Emmerton & Renner, 2006); a stronger test of numerical competence would include numbers that require the extrapolation of learning.

This experiment examines the pigeons' ability to bisect pairs of numbers of different sizes, with a constant ratio. Of particular interest are the obtained response variability, and the location of the bisection point as these results may provide information about how number might be represented.

2.2 Method

2.2.1 Subjects

Subjects were four pigeons, *Columba livia*, numbered 181-184. All had experimental histories involving choice procedures, but no experience with either counting or timing tasks. Subjects were maintained at approximately 85% of their free-feeding weights by additional feeding, when necessary, after experimental sessions. Water and grit were continuously available in their home cages.

Apparatus

Four chambers containing a row of three keys were used. The chambers measured 40 cm by 40 cm by 32 cm, with the row of keys situated 21 cm above the floor. The center key was located 16 cm from each wall and the other two keys 8 cm on either side of the center key. Each key required a force of approximately 0.15 N for a response to be registered. The panel also contained a houselight situated 8 cm above the center key, and a food hopper, situated 13 cm below. During reinforcement, the houselight and response keys were dark, while the hopper was illuminated and raised to allow access to the wheat. Fans attached to each chamber provided ventilation and masking noise during experimental sessions. Sessions were controlled and recorded by a computer running MED-PC software, located in an adjacent room.

Procedure

Sessions were conducted at approximately the same time each day, seven days a week. Sessions ended when 72 trials had been completed or after 120 minutes, whichever came first. Each trial was preceded by a 12 s inter-trial-interval (ITI), during which the houselight and keylights were dark. Trials consisted of a sample phase and discrimination phase. At the start of the sample phase, the houselight was illuminated. After a variable delay, the center key was illuminated red and a single response to the center key extinguished it. This sequence (delay, illumination of the center key, response) will be referred to as a 'flash'. Baseline training involved two trial types, with sample phases consisting of *N* or *N* x 3 flashes. Trial types were randomly determined, subject to the constraint that there were three trials each with *N* or *N* x 3 flashes in every block of six trials.

To degrade the correlation between flash number and temporal cues, sample phase stimulus delays were determined using a double-randomization procedure. At the start of a trial, an expected sample phase duration was selected randomly without replacement from a list of durations (in seconds): {5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15}. The programmed average inter-flash interval for the particular trial was then calculated as expected sample phase duration divided by flash number. Finally, the individual inter-flash intervals were determined by multiplying the average flash interval by a delay sampled without replacement from a distribution of 12 delays with an average of 1 s generated by an exponential progression (Fleshler & Hoffman, 1962). This procedure was expected to decrease the correlation with number for both sample phase duration and flash rate.

After the last flash had been completed, the houselight and all keylights were darkened for a 2-s retention interval. Following the retention interval, the houselight, and left and right keys were illuminated to signal the start of the discrimination phase. For reinforcement to be

obtained, a response to the white key was required if 2 flashes occurred during the sample phase, whereas a response to the blue key was required if 6 flashes occurred. The locations of the blue and white keys were randomized on individual trials. Once a key had been pecked, the trial ended and both keylights were darkened.

If subjects made the correct response during the discrimination phase, reinforcement was presented. During reinforcement, the keylights and houselight were extinguished while the hopper was raised and illuminated for 4.5 s. If subjects made an incorrect response, a 5 s blackout occurred, followed by a correction trial. Correction trials were identical to the preceding regular trial, except that only the correct keylight was illuminated at the beginning of the discrimination phase. A single peck to the illuminated key resulted in 1.5 s hopper access.

Subjects received 127 or 128 sessions of baseline training with 2- and 6-flash trial types until performance had appeared to reach asymptote for all four birds. Transfer tests were then conducted with the novel values: 0, 1, 3, 4, 5, 7, and 8. These new numbers were presented in probe trials randomly arranged among the 2- and 6-flash baseline trials. Consequently, transfer test sessions consisted of 54 baseline trials and 18 probe trials. In any given session, there were 3 trials each for 4 of the novel values, and 2 trials each for 3 of the novel values. The probability of reinforcement on probe trials was 50%.

Following the completion of these transfer tests, subjects were returned to the original 2 and 6 baseline training conditions. Subjects were then divided into two groups, A and B for training with two new sets of baseline values. Group A was placed in training with 4 and 12 flashes, while Group B was placed in training with 8 and 24 flashes. After training under these conditions, subjects were then transfer tested with values 0, 2, 6, 8, 10, 12, 14, 16, and 0, 4, 12, 16, 20, 28, 32 for Groups A and B, respectively. Group A were then placed in baseline training with 8 and 24 flashes, and Group B placed in baseline training with the 4 and 12 discrimination. Due to a programming error, data from the first 27 sessions in this condition were unusable, and to ensure performance had not been compromised, subjects experienced a total of 60 sessions of

baseline training with these values, before being placed in their respective transfer tests. Finally, subjects were returned to the original 2 and 6 flash baseline discrimination, before being placed in a final transfer test with values 0, 1, 3, 4, 5, 7, 8. The arrangement of the experimental conditions, and the number of baseline training and transfer test sessions can be seen in Table 2.1

Table 2.1 The order of experimental conditions and number of sessions in the baseline training and transfer tests sessions of each condition for Group A (subjects 181 and 182) and Group B (subjects 183 and 184).

	Condition 1			Condition 2		Condition 3		Condition 4	
Group A (181 and 182)	2v6			4v12		8v24		2v6	
Group B (183 and 184)				8v24		4v12			
	BL	TT	BL	BL	TT	BL	TT	BL	TT
No. of Sessions	127/128	49	72	42	31	60	42	32	33

2.3 Results

2.3.1 Baseline Training

For all conditions, data were aggregated over the last 10 sessions of baseline prior to transfer tests, and the first 10 sessions of transfer testing. The additional sessions of transfer tests were conducted for purposes of quantitative modeling and are not reported here. Although the two groups experienced the 4 vs. 12 and 8 vs. 24 conditions in different orders, independent sample t-tests showed no significant difference in performance between Groups A and B, so data from all subjects were collated across those conditions.

Performance in the small and large number trials is plotted in Figure 2.1 below. To test for order effects on overall performance, a one-way repeated measures ANOVA was conducted and obtained a significant effect of order on performance averaged across small and large number trials, $F(3,9) = 5.74$, $p<.05$. Tukey post-hoc tests showed performance in the first 2 vs. 6 condition was significantly lower than performance in the second and third 4 vs. 12 and 8 vs. 24 conditions, $p<.05$. There was no significant difference between the last 2 vs. 6 condition and all

other conditions. Consequently, only results from the last 3 conditions are reported here.

Although the effect of flash number on responding was the main focus of analyses, the relationship between number and the temporal variables, cumulative sample phase duration (hereafter referred to as sample phase duration) and flash rate, was also examined. Flash rate was calculated by dividing flash number by the cumulative sample phase duration.

Overall, the sample phase duration increased with flash number in baseline training in all conditions. Averaged across subjects, sample phase durations were 11.12 s [SE= 0.21] and 12.36 s [SE= 0.80] for 2- and 6-flash trials, 18.91 s [SE = 0.93] and 22.95 s [SE = 1.57] for 4- and 12-flash trials, and 30.75 s [SE = 3.25] and 33.53 s [SE = 0.32] for the 8- and 24-flash trials. However, the difference in sample phase durations only reached significance for 4- and 12-flash trials, $t(3) = 6.17$, $p<.05$.

Flash rate was significantly lower for the smaller than larger value in each condition. Average flash rates were 0.30 flashes/s [SE = 0.01] and 0.60 flashes/s [SE=.03] for the 2-flash and 6-flash trials, 0.26 flashes/s [SE = 0.02] and $M = 0.59$, [SE = 0.05] for the 4-flash and 12-flash trials, and $M = 0.30$, [SE = 0.03] and $M = 0.81$, [SE = 0.01] for the 8-flash and 24-flash trials. Dependent sample t-tests found a significant difference in flash rate for all conditions, $t(3) = 16.78$, $p <.05$, for the 2- and 6-flash trials, $t(3) = 12.04$, $p<.05$ for the 4- and 12-flash trials, and $t(3) = 19.76$, $p<.05$ for the 8- and 24-flash trials.

To provide a more conservative test of numerical control, outliers (cases with sample phase durations greater than three standard deviations from the mean) were excluded from the calculation of the correlational and the logistic regression analyses reported below[1]. Correlations between sample phase duration and flash number were relatively small, but significant at $p < .05$ for all subjects except Pigeon 182 in the 2 vs. 6 condition ($p = .05$); averaged across the baseline and transfer tests for all birds, $r = 0.16$, SE= 0.04 for the 2 vs. 6 discrimination, $r = 0.26$, SE = 0.04 for the 4 vs. 12 discrimination, and $r = 0.14$, SE = 0.12 for the 8 vs. 24 discrimination.

[1] Using the criterion of sample phase duration >30s resulted in an average of 2.08% of total trials being excluded.

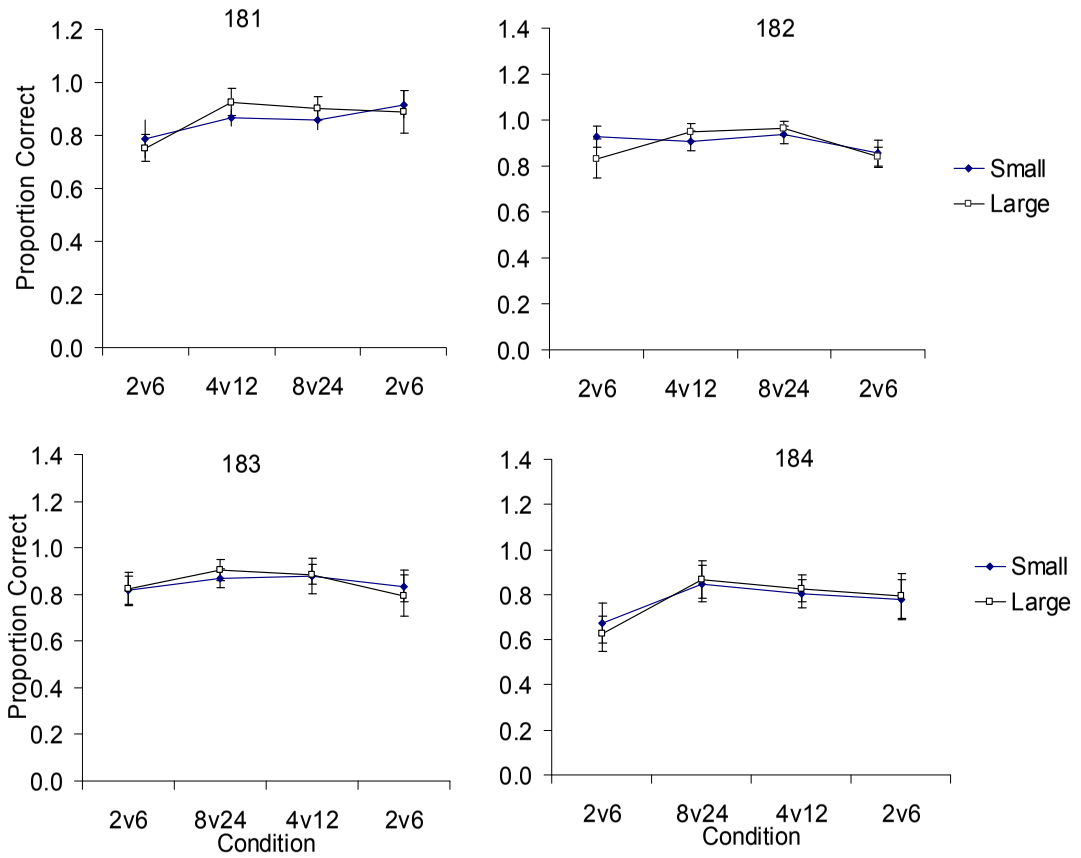

Figure 2.1. Individual average proportion of correct baseline training trials on the small and largenumber trials (e.g. 2 or 6 in the 2v6 discrimination, respectively) in the four training conditions. Error bars represent $\pm$ 1 standard deviation.

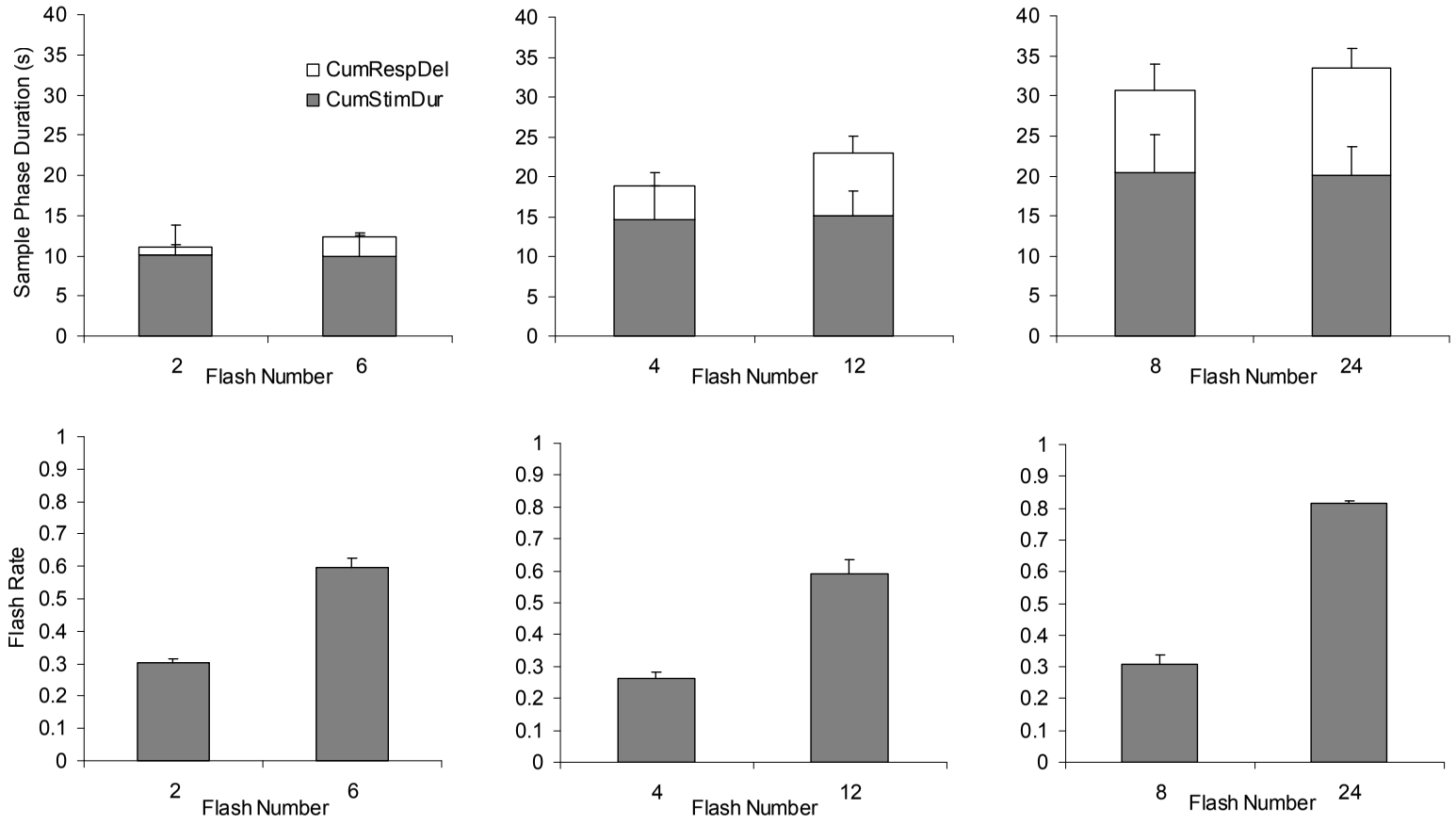

Figure 2.2. Plot of mean cumulative sample phase duration (upper panels) for the last 10 sessions of baseline training in all three conditions: grey bars show mean cumulative stimulus duration, white bars show cumulative response delays. Lower panels show mean flash rate, calculated as N/cumulative sample phase duration. Error bars show + 1 S.E.

A much stronger relationship was found between flash rate and flash number, likely due to the upper limit placed on the sample phase duration. Correlations between these two variables were significant for all subjects in all conditions; averaged across baseline and transfer tests, r =0.47, SE=0.03 for the 2 vs. 6 discrimination, r = 0.70, SE = 0.02 for the 4 vs. 12 discrimination, and r = 0.73, SE = 0.03 for the 8 vs. 24 discrimination.

These analyses suggest that the randomisation procedure degraded the relationship between temporal variables and flash number, but the response-dependent nature of the procedure resulted in some covariation between sample phase duration, flash rate and flash number. This can be seen in Figure 2.2.

Proportions of "large" responses were calculated for all trial types. Subjects were able to discriminate 2 and 6 flashes successfully. Performance, calculated as the overall proportion of correct responses, exceeded 60% correct for all four subjects, and was significantly different from chance, $t(7)= 2.49$, $p<.05$. Performance was equal for the two trial types, average M= 80.34% [SE= 5.11] for 2-flash trials and average M = 75.77% [SE = 4.70] for 6-flash trials.

Similar results were obtained for the 4- vs. 12-flash and the 8- vs. 24-flash discriminations. The average proportion of blue-key responses on the small and large trial types were significantly different in the two conditions; M = 0.09 [SE = 0.03] and M= 0.85 [SE = 0.02], $t(3)=15.78$, $p<.05$, for the 4- and 12-flash trials, respectively, and M = 0.09, [SE = 0.02] and M = 0.87 [SE = 0.02], $t(3) = 19.19$, $p<.05$, for the 8- and 24-flash trials, respectively. Performance in these conditions was higher than the previous 2 vs. 6 discrimination; M = 89.96% [SE = 2.52] and M = 85.39% [SE = 2.35] for the 4- and 12-flash trials, respectively, and M = 90.92% [SE = 1.98], and M = 87.04% [SE = 2.11] for the 8- and 24-flash trials.

Results of a repeated-measures ANOVA conducted on proportion correct values, with trial types (small or large number) and condition (2v6, 4v12 or 8v24) as factors, showed a significant effect of trial type on performance, $F(1,3) = 25.74$, $p<.05$, and no significant effect of condition, $F(2,6) = 3.57$, $p = 0.09$, or interaction, $F(2,6) = 2.74$, *n.s.* A Tukey post-hoc test

showed the significant effect of trial type was due to significantly more accurate performance on the small value trials on the 4 vs. 12 and 8 vs. 24 discriminations, $p<.01$ and $p<.05$, respectively.

Hierarchical logistic regression analyses were conducted to investigate the relative control by flash number, sample phase duration and flash rate over responding, in particular whether significant numerical control over responding could be obtained after controlling for the temporal variables. In these analyses, sample phase duration and flash rate were entered into a logistic regression model predicting the probability of a 'large' response at the first step, and flash number was entered in the second step. Logistic (B) coefficients and odds ratios were calculated for the full model between response and flash number, sample phase duration and flash rate. Nagelkerke R^2 values were calculated as an approximate measure of variance accounted for by the full model, and a chi square test (based on the differences in the -2 log likelihood ratios) was used to test for significant improvements in fit after the addition of flash number into the regression model. These values are reported in Table 2.2.

Table 2.2 shows that significant B coefficients for flash number were obtained for all four birds in all conditions with the exception of 181 in the 8 vs. 24 discrimination, which approached significance. Flash number was the only significant predictor of a "large" response for 181 in the 2 vs. 6 and 4 vs. 12 conditions, 182 in the 4 vs. 12 and 8 vs. 24 conditions, 183 in the 2 vs. 6 condition, and 184 in the 4 vs. 12 condition. Of the temporal variables, sample phase duration had the least influence over responding, with B coefficients significant only for Pigeon 182 in the 2 vs. 6 condition, and 183 in the 4 vs. 12 condition, and 184 in the 8 vs. 24 condition. Stronger control was exhibited by flash rate; it was a significant predictor for 184 in the 2 vs. 6 condition, 183 in the 4 vs. 12 condition, and 181, 183 and 184 in the 8 vs. 24 condition. For all birds, a significant increase in -2 log likelihood ratios was obtained when flash number was added to the model, with the exception of 181 ($p = 0.07$). The full model accounted for a range of variance for the four subjects, with Nagelkerke R^2 values varying from 41% to 84%.

Table 2.2. Hierarchical logistic regression results from last 10 sessions of baseline training for all three conditions

2 vs. 6	**181**		**182**		**183**		**184**	
	B	**B(Exp)**	**B**	**B(Exp)**	**B**	**B(Exp)**	**B**	**B(Exp)**
Sample Duration	-0.11	0.89	-0.10**	0.90	0.02	1.02	-0.04	0.97
Flash Rate	-2.27	0.10	-1.11 (p=0.06)	0.33	0.50	1.64	-1.56**	0.21
Flash Number	1.37***	3.92	1.01***	2.73	0.71***	2.04	0.80***	2.22
Nagelkerke Rsq	0.71		0.57		0.48		0.41	
Chi square	165.00***		217.17***		123.97***		188.76***	
4 vs. 12	**181**		**182**		**183**		**184**	
	B	**B(Exp)**	**B**	**B(Exp)**	**B**	**B(Exp)**	**B**	
Sample Duration	-0.07	0.94	0	1	0.08**	1.08	-0.41	0.96
Flash Rate	-0.38	0.69	2.92	18.58	6.97***	1066.78	-0.16	0.86
Flash Number	0.62***	1.87	0.56***	1.75	0.2**	1.22	0.396***	1.49
Nagelkerke Rsq	0.69		0.79		0.66		0.47	
Chi Square	40.95***		23.03***		7.22**		58.15***	
8 vs. 24	**181**		**182**		**183**		**184**	
	B	**B(Exp)**	**B**	**B(Exp)**	**B**	**B(Exp)**	**B**	**B(Exp)**
Sample Duration	0.003	1.00	0.03	1.03	-0.04	0.96	0.06**	1.07
Flash Rate	5.38**	217.77	3.63	37.77	3.93*	50.84	3.59**	36.35
Flash Number	0.10 (p=0.06)	1.10	0.26***	1.29	0.23**	1.25	0.10**	1.10
Nagelkerke Rsq	0.68		0.84		0.72		0.59	
Chi square	3.21 (p =0.07)		12.75***		10.83**		6.79**	

Note: * *p*<.05, ***p*<.01, * *p*<.001**

Response latencies during the choice phase were calculated and analysed to test for a size effect, which would predict significantly longer response times in trials with a larger number of flashes. Choice latencies in all conditions were longer in the large-number trials than the small-number trials, and were more variable. Mean choice latencies in the 2 vs. 6 discrimination were 0.93 s [SE = 0.075], and 1.11 s, [SE = 0.118]. Mean choice latencies in the 4-flash trials and 12-flash trials were 0.94 s [*SE* = 0.065], and 1.12 s [*SE* = 0.14], respectively, and 0.98 s [*SE* = 0.081] and 1.24 s [*SE* = 0.146] for the 8-flash trials and 24-flash trials, respectively. Results of a repeated measures ANOVA found a significant effect of small/large trial type on response

latency, $F(1,3) = 11.51$, $p<.05$, and an effect of condition that approached significance, $F(2,6,) = 3.96$, $p = .08$. No significant interaction was found. Results of a post-hoc Tukey HSD test revealed response latencies on the 24-flash trials of the 8 vs. 24 discrimination were significantly longer than the 8-flash trials. Differences in response latencies for the small and large number trials in the 2 vs. 6 and 4 vs. 12 discriminations approached significance, $p = .07$ and .06, respectively. These findings provide limited support for a size effect; choice latencies were significantly larger on the larger-number trials in only the 8 vs. 24 discrimination, the condition with the largest difference between numbers.

Transfer testing

Data from the first 10 sessions of transfer testing from each condition were aggregated and analysed in the same manner as baseline data.

Mean sample phase durations were calculated for the 2 vs. 6, 4 vs. 12 and 8 vs. 24 conditions (see upper panel, Figure 2.3). Repeated measures ANOVAs found significant effects of number on sample phase duration in the 4 vs. 12 condition, $F(8,24) = 10.15$, $p<.001$ and the 8 vs. 24 condition, $F(8,24) = 8.77$, $p<.001$. No significant effect of flash number was obtained in the 2 vs. 6 condition, $F(8,24) = 1.18$, *n.s*. Significant linear trends were obtained for in both the 4 vs. 12 and 8 vs. 24 conditions, $F(1,3)=33.18$, $p<.05$, and $F(1,3) = 46.39$, $p<.01$. Flash rates were calculated for all trial types except 0-flash trials (see lower panel, Figure 2.3). Flash rate increased more markedly with flash number and results of repeated measures ANOVAs showed significant effects in the 2 vs. 6 condition, $F(8,24) = 49.87$, $p<.001$, the 4 vs. 12 condition, $F(8,24) = 109.84$, $p<.001$, and the 8 vs. 24 condition, $F(8.24) = 102.94$, $p<.001$. Significant linear trends were also found: $F(1,3) = 485.67$, $p<.001$, $F(1,3) = 174.98$, $p<.05$, and $F(1,3) = 537.59$, $p<.05$ for the 2 vs. 6, 4 vs. 12 and 8 vs. 24 conditions, respectively.

Correlations between the temporal variables and flash number were calculated in the

same manner as for baseline trials[2]. Once again, correlations were relatively low for both sample phase duration, average r =0.20, SE = 0.02 for the 2 vs. 6 condition, average r = 0.24, SE = 0.01 for the 4 vs. 12 condition, and average r = 0.20, SE = 0.07 for the 8 vs. 24 condition. Correlations between sample phase and flash rate were also comparable to baseline, average r = 0.55, SE=0.03, average r = 0.74, SE = 0.02, average r = 0.75, SE = 0.02 for the 2 vs. 6, 4 vs. 12, and 8 vs. 24 conditions respectively.

Hierarchical logistic regression analyses were conducted to investigate the respective roles of sample phase duration, flash rate and flash number in predicting the probability of a large response. These analyses used data from the first 10 sessions of transfer testing from each condition. A summary is provided in Table 2.3. Flash number accounted for a significant amount of additional variance above and beyond the temporal variables, flash rate and sample phase duration, for all birds in all conditions with the exception of 184 in the 8 vs. 24 condition, which approached significance. In the full model, flash number was a significant and the strongest predictor of response for all subjects except 184 in the 8 vs. 24 condition and 183 in the 4 vs. 12 condition, who had greater control by both flash rate and sample phase duration. The amount of variance accounted for by flash also approached significance for 181 in the 2 vs. 6 condition. Control by sample phase duration was also significant for 181 and 182 in the 2 vs. 6 condition. The approximate variance accounted for by the full model ranged from 37 to 80% with the Nagelkerke R^2 values of 60% or greater obtained for all birds in all conditions, except 184. These results suggest that for the majority of subjects, significant control by flash number over responding was obtained, above and beyond flash rate and sample phase duration.

[2] Using the criterion of sample phase duration>30s resulted in an average of 2.21% of total trials being excluded.

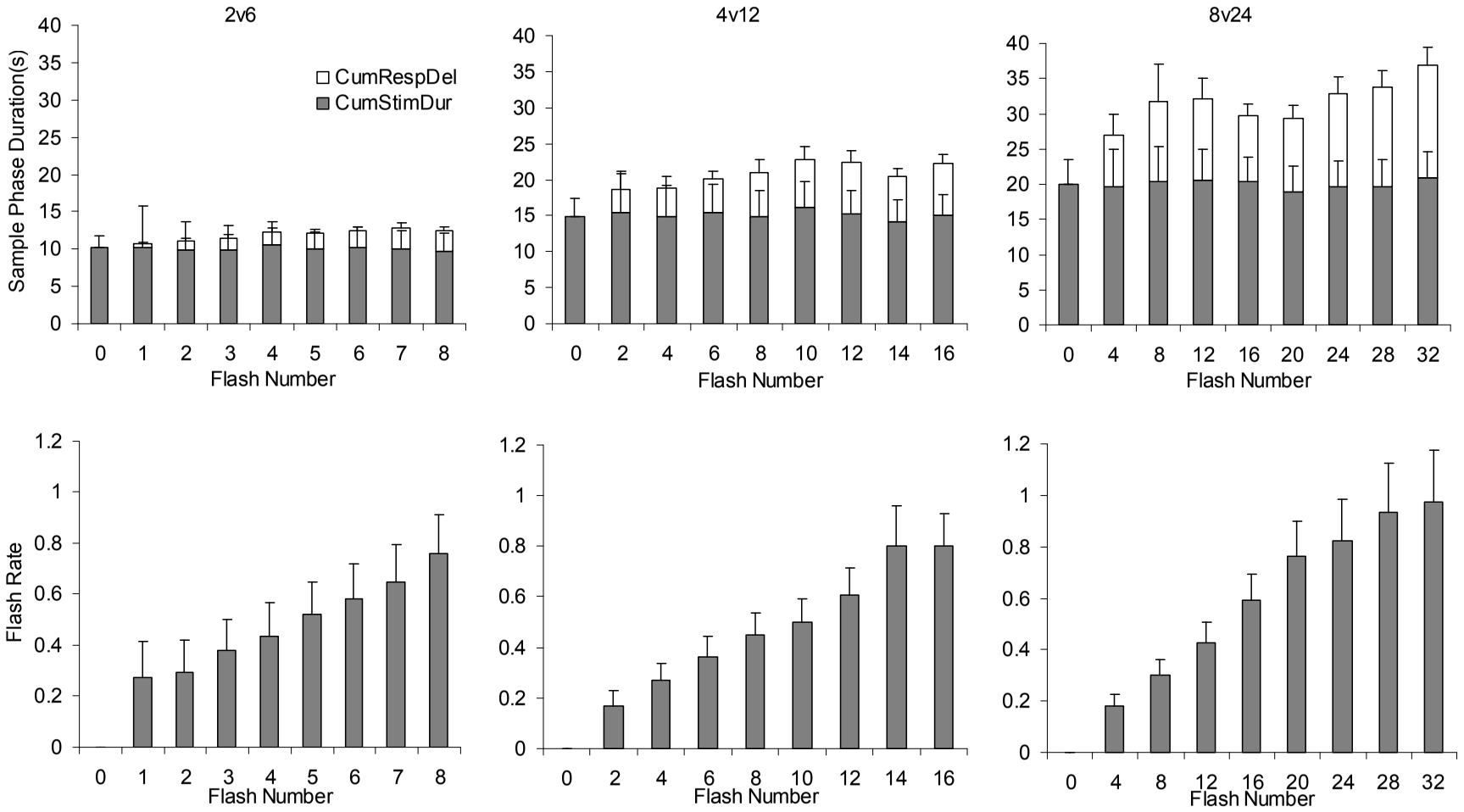

Figure 2.3. Plot of mean cumulative sample phase duration (upper panel): grey bars show mean cumulative stimulus duration, white bars show cumulative response delays. Lower panel shows plot of mean flash rate. Error bars show + 1 S.E.

Table 2.3. Hierarchical logistic regression results from first 10 sessions of transfer testing for all three experimental conditions.

2 vs. 6	**181**		**182**		**183**		**184**	
	B	**B(Exp)**	**B**	**B(Exp)**	**B**	**B(Exp)**	**B**	**B(Exp)**
Sample Duration	-0.09*	0.92	-0.09**	0.91	-0.02	0.98	-0.003	1
Flash Rate	-2.06 (p=0.06)	0.13	-0.8	0.45	-0.89	0.41	-0.68	0.51
Flash Number	1.19***	3.29	1.10***	3	0.98***	2.67	0.70***	2.01
Nagelkerke Rsq	0.64		0.65		0.60		0.42	
-2 log likelihood change	138.19***		210.70***		173.24***		135.60***	
4 vs. 12	**181**		**182**		**183**		**184**	
	B	**B(Exp)**	**B**	**B(Exp)**	**B**	**B(Exp)**	**B**	**B(Exp)**
Sample Duration	-0.01	0.99	0	1	0.07*	1.08	-0.002	1
Flash Rate	1.35	3.84	0.27	1.31	5.33***	206.95	0.49	1.64
Flash Number	0.46***	1.59	0.53***	1.69	0.22**	1.24	0.27***	1.32
Nagelkerke Rsq	0.65		42.99***		0.62		0.37	
-2 log likelihood change	35.48***		42.99***		11.86**		29.28***	
8 vs. 24	**181**		**182**		**183**		**184**	
	B	**B(Exp)**	**B**	**B(Exp)**	**B**	**B(Exp)**	**B**	**B(Exp)**
Sample Duration	-0.04	0.96	-0.05	0.96	-0.06	0.94	0.05**	1.05
Flash Rate	1.71	5.54	-0.23	0.79	0.82	2.26	3.10***	22.12
Flash Number	0.228***	1.255	0.380***	1.463	0.241***	1.272	0.052 (p=0.07)	1.054
Nagelkerke Rsq	0.70		0.80		0.62		0.44	
-2 log likelihood change	18.90***		102.95***		55.38***		3.06(p=0.08)	

* $p < 0.05$ ** $p < .01$ *** $p < .001$

Average choice latencies in the three conditions during transfer testing are shown in Figure 2.4. Choice latencies appeared to increase as flash number increased. A repeated-measures ANOVA revealed a significant difference in response latencies in all three conditions, $F(8,24) = 5.49$, $p<.001$, $F(8,24) = 6.31$, $p < .05$ and $F(8,24) = 3.09$, $p<.05$ for the 2 vs. 6, 4 vs. 12 and 8 vs. 24 conditions, respectively. Linear trend analyses found choice latencies increased

linearly with flash number in the 4 vs. 12 condition, $F(1,3) = 11.42$, $p<.01$, but not in the 8 vs. 24 condition, $F(1,3) = 5.10$, $p = .11$. The linear trend in choice latencies in the 2 vs. 6 condition approached significance, $F(1,3) = 8.79$, $p = .05$. Thus, unlike baseline training, significant size effects on choice latencies were found in the 4- vs. 12-flash, but not the 8- vs. 24-flash discrimination. Similar to baseline, the difference in choice latencies for the 2- vs. 6-flash discrimination approached significance. These results provide some support for a size effect and also suggest the size of this effect is not dependent on the relative magnitude of the flashes being discriminated.

The mean proportion or probability of a "large" response was calculated for both baseline and transfer trial types, and the psychometric plots for each bird for each condition are shown in the Figure 2.5. Positive transfer to novel values was obtained to the extent that probability of a large response was directly related to the number of flashes, and also extended to values outside of the training range (0, 1 and 7, 8). Repeated-measures ANOVAs found a significant effect of flash number on the probability of a large response for the 2 vs. 6 discrimination, $F(8,24) = 128.30$, $p<.001$, the 4 vs. 12 discrimination, $F(8,24) = 118.85$, $p<.001$, and the 8 vs. 24 discrimination, $F(8,24) = 78.85$, $p<.001$.

To test for extrapolation to values outside of the training range, planned comparisons were conducted to compare proportion of large responses to the lowest two transfer test values relative to the lower baseline training value, and the highest two transfer test values relative to the higher baseline training value. Planned comparisons were conducted to test whether subjects had developed any understanding of "0", shown by any significant difference between responding on the 0-flash trials to the two next lowest transfer test values. If the proportion of "large responses" were significantly lower on the 0-flash trials relative to the next lowest transfer test value and the lower anchor value used in training, this would suggest subjects had developed some understanding of the numerical value of 0, despite not receiving any explicit training with this value.

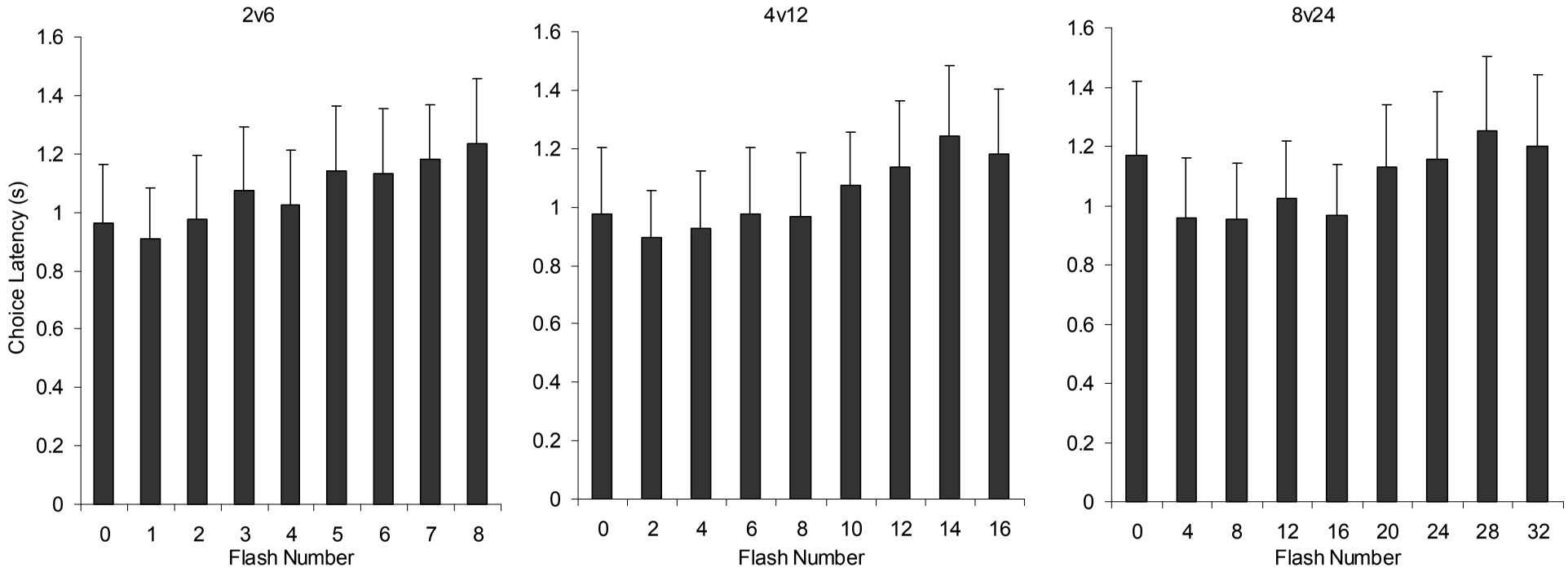

Figure 2.4. Average choice latencies in the first 10 sessions of transfer testing in the 2 vs. 6 (left panel), 4 vs. 12 (middle panel) and 8 vs. 24 (right panel) conditions. Error bars show $\pm$ 1 S.E

For the proportion of large responses in the 2 vs. 6 condition, planned comparisons found that the proportion of large responses was significantly greater on 7- and 8-flash trials than 6-flash trials, $F(1,3) = 13.32$, $p < .05$, and the difference between the proportion of large responses on 0- and 1-flash trials and 2-flash trials approached significance, $F(1,3)= 6.13$, $p = .089$. No significant difference between the 0- and 1-flash trials was found, $F(1,3) = 0.48$, *n.s*. In the 4 vs. 12 condition, a planned comparison found no significant difference between the proportion of large responses for the lowest two transfer values, 0 and 2, and the lower baseline training value 4; $F(1,3) = 5.43$, *n.s*. The proportion of large responses for the highest two transfer values, 14 and 16, was significantly higher than the higher baseline training value, 12; $F(1,3) = 13.31$, $p < .05$. No significant difference between the 0- and 1-flash trials was found, $F(1,3) = 2.37$, *n.s.* Responding in the 8 vs. 24 condition was similar to the 4 vs. 12 condition. Planned comparisons found no significant difference between the lowest transfer test values 0 and 4, and baseline value, 8, $F(1,3) = 3.68$, *n.s.*, and a significant difference between the highest transfer tests values, 28 and 32, and baseline value, 24, $F(1,3) = 13.01$, $p < .05$. No significant difference between the 0- and 4-flash trials was found, $F(1,3) = 0.01$, *n.s.* The failure to find significant differences among the lower extreme values may be due to floor effects in responding. Overall, these results confirm that transfer successfully occurred for flash numbers outside the range included in training, consistent with the hypothesis that pigeons were able to respond differentially to new numbers of flashes despite no previous exposure to these.

When the psychometric functions for the three different conditions were plotted along the same scale, the functions superimposed, suggesting constant Weber fractions; constant relative variability across the three different number scales. These can be seen in Figure 2.6. A repeated measures ANOVA was conducted with the individual data, using condition and N/S (sample number divided by the small anchor value) as predictors, and found a significant effect of N/S, $F(10,90) = 334.261$, $p<.001$, and no significant difference across conditions, $F(2,9) = 0.99$, *n.s.* and no significant interaction, $F(20,90) = 0.73$, *n.s.*

To provide estimates of the bisection points for individual subjects, a two-parameter logistic function was calculated and fitted to the baseline and transfer test data:

$$P(\text{large}) = \frac{1}{1 + e^{-A(n-B)}}$$

where n = number of flashes, A = slope of the middle part of the function, and B = bisection point for that function. These functions provided a good estimate of the data, accounting for between 97-99% of variance for each subject in each condition.

The bisection point estimates were tested to see whether they were located closer to the geometric mean or arithmetic means of each scale. The predicted bisection point values, including the geometric and arithmetic means of the baseline training values in each condition are shown in Table 2.4. All 16 bisection points were closer to the arithmetic mean than the geometric mean. Results of single sample t-tests found no significant difference between bisection points and the arithmetic mean for the 2 vs. 6 discrimination, $t(3) = 0.70$, $p = 0.53$, the 4 vs. 12 discrimination, $t(3) = 1.78$, $p = 0.17$, and the 8 vs. 24 discrimination, $t(3) = 1.12$, $p = 0.34$. However, bisection points were found to be significantly different from the geometric mean, $t(3) = 5.22$, $p<.05$ for the 2 vs. 6 discrimination, $t(3) = 8.05$, $p<.005$ for the 4 vs. 12 discrimination, and $t(3) = 4.37$, $p<.05$ for the 8 vs. 24 discrimination.

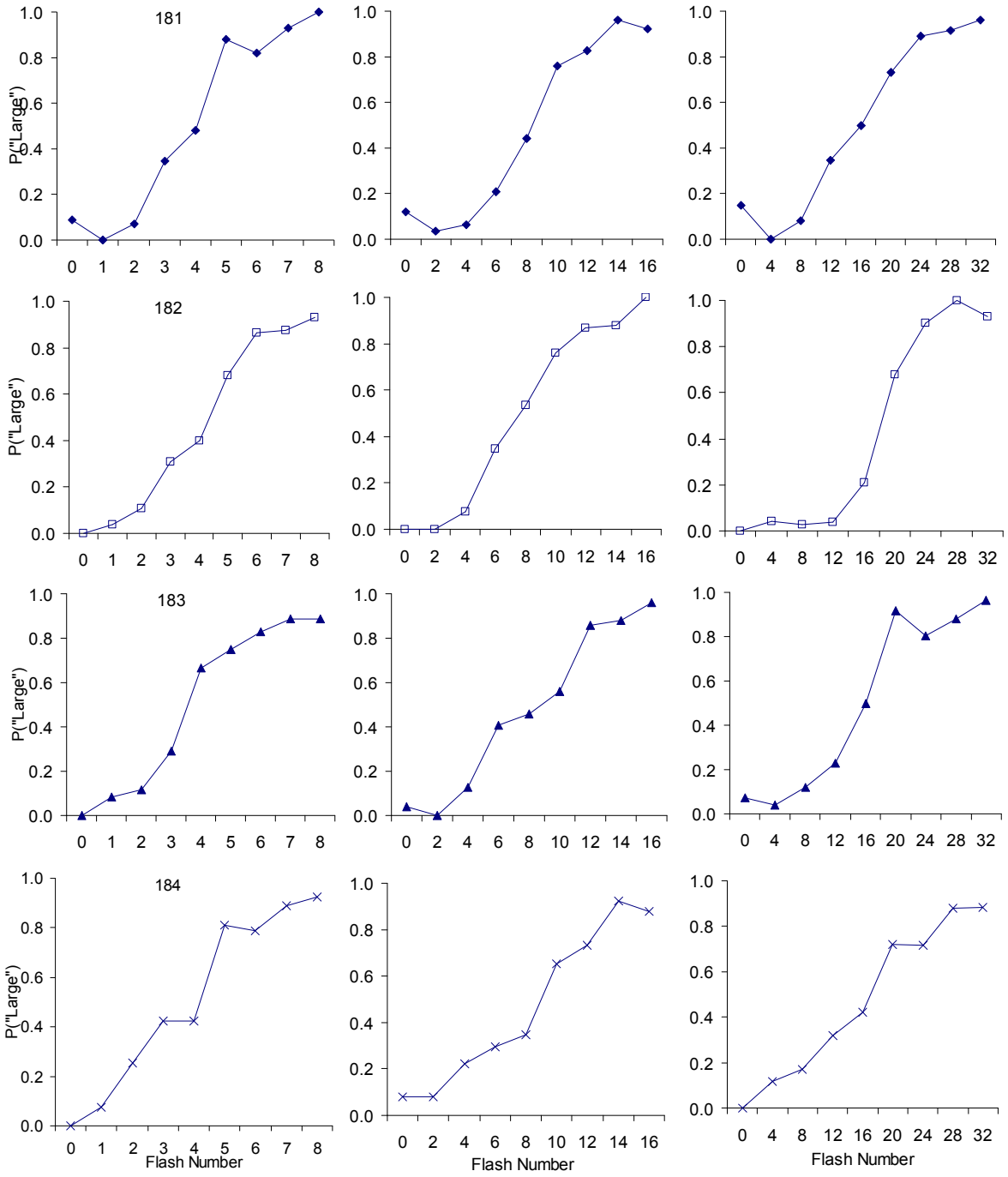

Figure 2.5. Probability of a large response plotted as a function of flash number for individual subjects for the 2 vs. 6 (left column), 4 vs. 12 (center column) and 8 vs. 24 (right column) discrimination conditions.

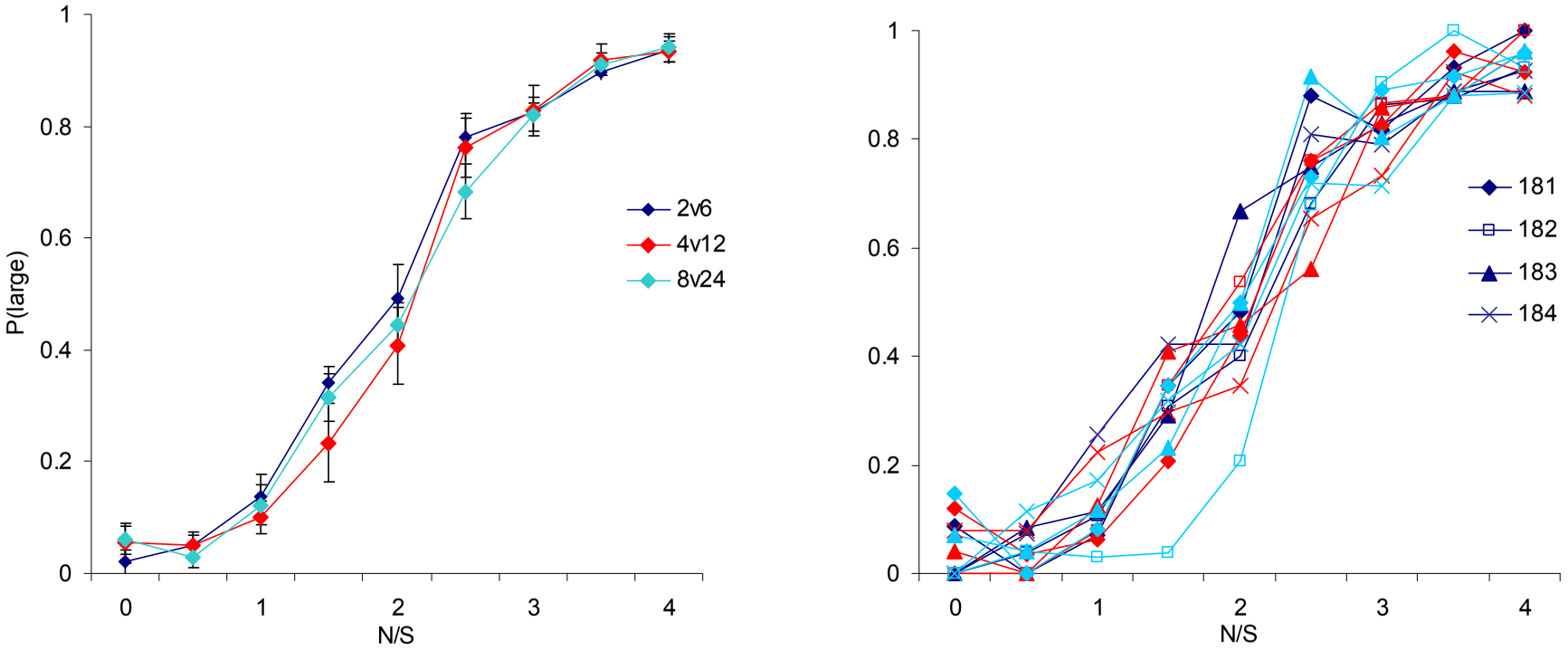

Figure 2.6. Group average (left panel) and individual (right panel) psychometric functions for the 2 vs. 6 (dark blue series), 4 vs. 12 (red series) and 8 vs. 24 (light blue series) discrimination conditions, plotted on a relative scale (sample number divided by the smaller anchor value).

Table 2.4. The arithmetic mean, geometric mean and bisection points for all subjects in all conditions

	Condition		
	2v6	**4v12**	**8v24**
181	3.87	8.40	15.90
182	4.21	7.84	18.60
183	3.84	8.32	15.71
184	3.83	8.65	16.77
Arithmetic Mean	4	8	16
Geometric Mean	3.46	6.93	13.86

Difference limens (DLs) were also calculated for each subject and each condition. The DL represents the smallest value required for a difference in numerosity to be discriminated, and was calculated by interpolating the numerical values at which the proportion of large responses made equaled 25% and 75%, and halving the difference between these values. These values are plotted in Figure 2.7. Two different patterns in DL values emerged; 182 and 184 showed an obvious increase in DL values as scales increased, whereas DL values for 181 and 183 increased from the 2 vs. 6 to the 4 vs. 12 discrimination, but did not increase for the 8 vs. 24 condition. A repeated measures ANOVA obtained a significant effect of condition on DL values, $F(2,6) = 8.89$, $p<.05$ For all subjects, the numerical difference necessary to achieve a change in responding increased as the numbers of flashes increased; a trend analysis confirmed a significant linear trend $F(1,3) = 13.62$, $p<.05$.

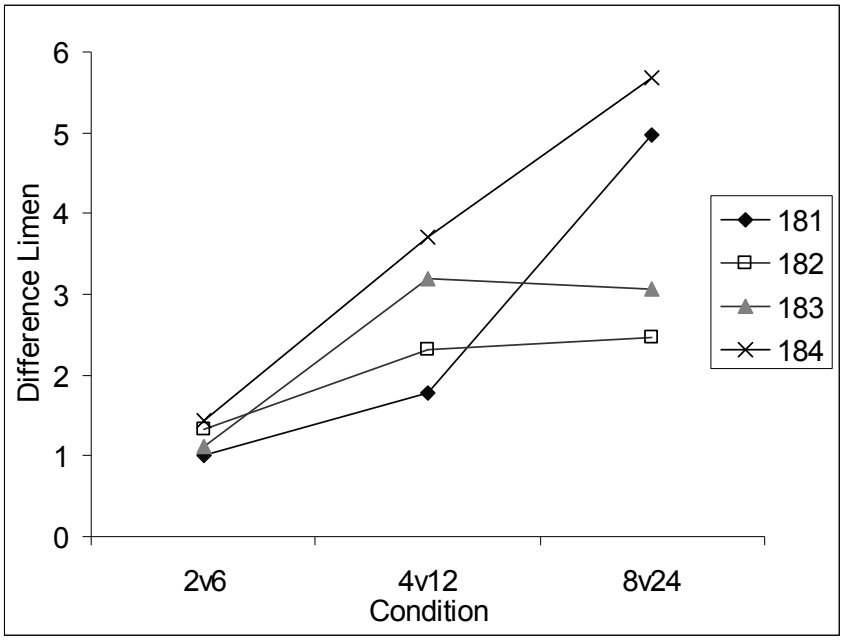

Figure 2.7. Individual difference limen (DL) values for the 2 vs. 6, 4 vs. 12 and 8 vs. 24 discriminations.

To assess changes in relative response variability as a function of flash number, Weber fractions identical to those calculated by Emmerton and Renner (2006), and similar to the coefficients of variation used by Fetterman (1993), were obtained by dividing the individual DL values by the PSE for each subject. These are plotted in Figure 2.8. Scalar variability would be characterized by a significant linear relationship between the DL and PSE values, and constant Weber fractions as PSEs increased, when plotted on a log-log scale.

Results were generally consistent with scalar variability. A bivariate regression found a significant positive linear relationship between DL and PSE values, $\beta = 0.76$, $p < .01$, $R^2 = 0.57$, variability increased as PSEs increased. Additionally, although a negative correlation was found between the log Weber fractions and log PSEs, the relationship did not differ significantly from zero, $\beta = -0.38$, $p = 0.21$, $R^2 = 0.15$.

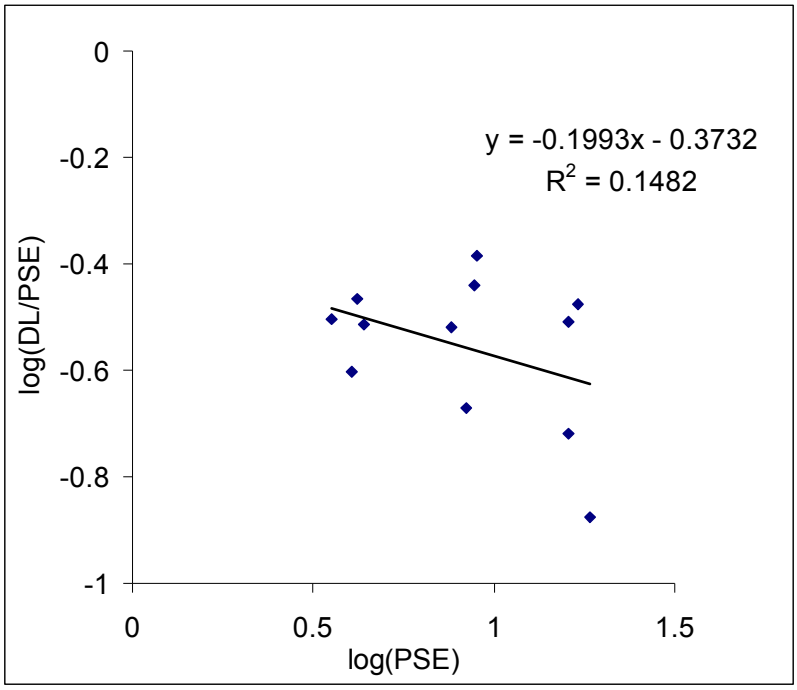

Figure 2.8. Obtained Weber fractions plotted as function of PSE on a log-log scale

2.4 Discussion

Results suggest subjects were able to successfully discriminate 2 and 6, 4 and 12 and 8 and 24 flashes. Although the response-dependent stimulus presentation resulted in some covariation between sample number and flash rate, results of hierarchical logistic regression analyses revealed responding was largely based on number, rather than the duration of the sample phase or flash rate. This provides further confirmation that pigeons are able to discriminate number, even in the presence of covarying, albeit somewhat unreliable, temporal cues. Performance was somewhat dependent on numerical magnitude; difference limens increased significantly as number increased suggesting the difference required for two numbers to be distinguished increased with magnitude. It is worth noting that in the examination of individual data, this pattern was only found in 2 of the 4 subjects; the other two subjects showed no increase in difference limen from the 4 vs. 12 to 8 vs. 24 condition. As these birds experienced these conditions in different orders, this difference cannot be due to an order or learning effect. Although implications of this finding are unclear, it indicates that for some pigeons there is no "size" effect in numerical discrimination- variability does not appear to always increase with numerical magnitude.

This experiment extends those of Emmerton and Renner (2006) who demonstrated extrapolation in relative numerosity judgments with visual stimuli presented simultaneously. These results show that pigeons are capable of transferring their relative numerosity discrimination ability of sequential visual stimuli to values up to 2 units outside of the baseline training, as well as values within the baseline training range. Some extrapolation to values outside of the lower extremes of the training ranges was found, with some subjects (e.g. 182) showing a lower proportion of large responses on 0-flash trials, relative to trials with the next two largest number of flashes. This is clearest on the 2 vs. 6 condition, but is also seen, to a lesser extent in the 4 vs. 12 and 8 vs. 24 conditions. However, the difference between

responding on the 0 and lowest anchor value trials only approached statistical significance on the 2 vs. 6 condition, and was not significant in the other two conditions. No significant difference was found between the second lowest numerical trial type and 0. This is most likely due to floor effects as often responding on the second lowest numerical trial type was equal or close to 0.

There has been a paucity of research investigating nonhuman's understanding of zero; although the results of studies that have been done are promising. Squirrel monkeys (Olthof et al., 1997) and chimpanzees (Boysen and Berntson, 1989) have been able to use zero in addition operations, using Arabic numerals. Merrit, Rugani and Brannon (2009) successfully demonstrated the understanding of some properties of zero in a match-to-sample task, and relative numerosity task with a "select smaller" rule. Their subjects treated all empty presented sets as equivalent, and treated empty sets as occupying a lower place in the numerical continuum than non-empty sets. Distance effects were also obtained; accuracy increased as the numerical difference between the empty set and distractor increased. Thus, their results suggest subjects had some grasp of equivalence of empty sets, and their location on the numerical continuum.

However, there does not appear to be a natural understanding of the ordinal properties of zero with symbolic stimuli. Biro and Matsuzawa (2001) found that although the chimpanzee, Ai, had considerable training in matching Arabic numerals and arrays of dots, and could also successfully select Arabic numerals in increasing order, she did not show any positive transfer when tested with zero. Additionally, studies with Alex the parrot have demonstrated a spontaneous use of the word "none" as a response when asked to name a characteristic for a numerosity not present (Pepperberg & Gordon, 2005), although further testing showed that Alex did not say "none" when asked to say how many items were inside two empty cups (Pepperberg, 2006). Pepperberg proposed that Alex used "none" to describe the absence of a particular characteristic of a set of items, but not to describe an absence of actual items. It is also possible that "none" was Alex's response to signal a failed search (Merrit, Rugani & Brannon, 2009). It should be noted that although he did not respond for 5 of 8 trials, on the remaining three trials,

Alex used the label "one" for the empty sets, suggesting knowledge that an empty belonged at the lower end of the numerical continuum.

A true zero concept, that includes the understanding of its cardinal and ordinal properties is highly complex and often only gradually develops after children have learnt its fundamental characteristics (Wellman & Miller, 1986). Consequently, although it seems highly unlikely that nonhuman animals would show spontaneous knowledge about the higher order properties of zero, it is reasonable to hypothesise that nonhumans may develop or possess the mechanisms and knowledge that may serve as a foundation for a zero concept. Results from the present study and previous research support this view.

Representation and Response Rules

Bisection points were closer to the arithmetic than the geometric mean for all three discriminations (2 vs. 6, 4 vs.12, 8 vs. 24). This finding is inconsistent with previous research in numerical bisection with nonhumans, which typically has produced bisection points at the geometric mean (e.g Fetterman, 1993, Jordan & Brannon, 2006, Meck & Church 1983, Roberts, 2005).

The only other study that has obtained strong evidence for bisection at the arithmetic mean is that of Droit-Volet et al. (2003), which found this in both verbal *and* nonverbal counting of children and adults. Note that PSEs at the arithmetic mean is predicted for verbal counting, as human verbal representations of number are generally linear with constant, not scalar variability. Jordan and Brannon (2006b) suggested that certain characteristics of the experiment might be responsible for the anomalous bisection point locations in the nonverbal condition of Droit-Volet et al. Since Droit-Volet et al.'s task involved the sequential presentation of stimuli; the individual items could have been enumerated by a parallel process, and consequently elicited a more linear format of numerical representation. This could also be applied to the current

experiment, which presented light flashes successively. This critique, although reasonable, is not very robust. A large number of bisection procedures have used stimuli presented successively not simultaneously, and have also obtained bisection points at the geometric, not arithmetic mean, e.g. Fetterman, (1993), Roberts, (2005), Meck and Church (1983), Roberts and Mitchell (1994). Thus sequential stimulus presentation does not appear to be solely responsible for the different bisection point location in this experiment, or Droit-Volet et al.'s study.

Jordan and Brannon (2006b) also suggested that subjects might have been able to count subvocally in the nonverbal counting condition of Droit-Volet et al. (2003) due to the long durations of stimuli presentation, despite the verbal distractor task. If subjects had been counting stimuli in the same manner as in the verbal counting condition, bisection points at the arithmetic mean are not surprising. However, this does not explain how bisection points were located at the arithmetic mean in the current experiment, as pigeons could not have been verbally counting.

Results of Fetterman (1993) suggested that the arithmetic mean may be a better predictor of bisection points than geometric means for small numerical FR values, e.g. less than 15; however, the difference between the arithmetic and geometric means with these values are small. The obtained bisection points at the arithmetic mean for the 2 vs. 6 and 4 vs. 12 conditions would be consistent with Fetterman's findings; however bisection points with the values 8 and 24 were also located closer to the arithmetic than geometric mean, suggesting this result is not only limited to small numerical magnitudes

Bisection point locations in temporal bisection procedures are dependent on a few different factors, which also apply to frequency or number bisection (Brown, McCormack, Smith & Stewart, 2005). Humans appear to be far more sensitive to these influences than nonhuman animals. One factor is whether the spacing of stimulus values is logarithmic or arithmetic (linear). Bisection points can be shifted leftwards, towards the geometric mean if items are logarithmically rather than linearly spaced (Wearden & Ferrara, 1995, Allan, 2002). This effect is further compounded by the ratio of the longest/largest to shortest/smallest stimuli; with large

long:short ratios, e.g. 9:1 the effect of stimulus distributions is much greater than with small long:short ratios, e.g. 4:1 (Wearden & Ferrara, 1996; Brown, et al., 2005). For humans, a long:short ratio of 2:1 or less will generate bisection at the geometric mean and no effect of stimulus distributions, while values greater than 2:1 will produce bisection at the arithmetic mean and effects of stimulus spacing. Animals, on the other hand, can still produce bisection points at the geometric mean with values of up to 4:1 (Wearden & Ferrara, 1996).

It is possible that the results in the current experiment may be explained by the effects of stimulus spacing described above (Wearden & Ferrera, 1996; Brown et al., 2005); stimulus spacing was arranged arithmetically, not logarithmically, which may have skewed bisection points towards the arithmetic mean. However, stimulus spacing effects with nonhuman animals have only been found in one temporal bisection study with rats (Raslear, 1985), which involved very large ratios of 100:1. With our large:small ratio of 3:1, which would still be considered a small ratio in terms of effects on bisection points, it seems rather unlikely that stimulus spacing, and large:small ratio would be solely responsible for bisection at the arithmetic mean.

An alternative explanation is that the obtained bisection point locations reflect the structure of the subjective numerical scale subjects developed in this procedure. Recall that bisection at the geometric mean is generally taken as evidence for either a logarithmic scale with constant generalization between values or a linear scale with increasing generalization between values. This is the general finding and the two main possibilities which have been suggested for nonverbal numerical representation. On the other hand, bisection at the arithmetic mean would suggest discriminations were based on a linear representation of number, with constant variability between numerical values. This is the type of number scale that normally would be associated with human verbal representation of number, where values along the continuum are equally spaced and the generalisation does not increase with magnitude.

Although location of bisection points suggested that variability in representations (and responding) did not increase proportionally with magnitude, combined analyses of data from the

2 vs. 6, 4 vs. 12 and 8 vs. 24 conditions suggested that response variability across the different ranges was scalar, increasing proportionally to number, consistent with Weber's law. Psychometric functions superimposed when plotted on a relative scale, and relative variability remained constant; the slope of Weber fractions (DL/PSE) did not differ significantly from zero when plotted against PSEs on a log-log scale. Additionally, a positive linear relationship between the PSEs and the difference limens was found. Thus variability across ranges, but not within ranges, is scalar and conforms to Weber's law. One possible interpretation of this is that subjects were able to limit response variability in their relative numerosity judgments for the transfer values within each condition, but as numerical magnitude increased across conditions, response variability increased proportionally.

What kind of response rule did subjects use to perform this discrimination? One possibility is a proximity or likelihood ratio rule, where responses are based on the independent assessment of which anchor value most likely generated the test signal (Gibbon, 1981). Subjects learn the absolute values of training stimuli and generalization around the two anchor values produces the typical bisection functions in transfer tests. This would predict bisection at the arithmetic mean if constant variance is assumed as well as transfer to novel values outside the training range, provided that the differences between the sample value and each of the anchor value is not equal. Alternatively, responding based on a similarity rule is determined by the relative similarity between the test value and two anchor values, based on their ratios. This rule, however, does not predict bisection at the arithmetic mean (Gibbon, 1981).

Are there other explanations for the obtained data? It is unlikely that subjects performed the numerical bisection task by calculating the arithmetic mean as a threshold value and assessing the relative difference between the test value and the arithmetic mean, responding "small" if it was lower, and responding "large" if it was higher. This response rule would predict response phase choice latencies that decreased as the difference between the test values and arithmetic mean increased. However, results did not support this prediction – if anything, choice

latencies increased linearly as numerical magnitude increased.

Alternatively, responding may have been determined by comparing test values to one of the anchor values (e.g. if X, respond "small", if not X, respond "large"). If this were the case, it would be likely subjects would have used the smaller anchor value as the comparator, since it would be easier and quicker to determine whether the number of flashes was equal to, say 2 rather than waiting for 6. This could possibly account for the pattern seen in the choice latencies, but it is unclear what implications this would have for the location of the bisection point.

Another possibility is that subjects may have developed a mental number line along which they placed numerical values during training and testing within each condition. Assuming a linear subjective scale with constant variability between numbers, this would account for the obtained arithmetic mean bisection, as well as the increasing response latencies.

At this stage, it is difficult to state decisively what implications the bisection point and variability results have for the representation of number and the process of relative numerosity discrimination, given the paucity of research on numerical bisection, and within the studies that have been conducted, a lack of results consistent with those obtained here.

What these results *do* show is that pigeons are able successfully discriminate and bisect different ranges of numerical values, with a ratio of 1:3. Number had significant control over responding, above and beyond the temporal variables, flash rate and sample phase duration, providing further evidence that number-based responding in nonhuman animals is possible. As stimuli were presented successively, subjects were required to monitor and remember the number of flashes seen during the sample phase, before being able to make their choice response. Because only one sample was presented at a time, a representation of the two anchor values, at least, must have developed for subjects to compare test values against and to allow correct responding. Furthermore, the finding that subjects were able to extrapolate to novel numbers both within and outside the training range suggests that an understanding of relative numerosity had been developed.

This experiment has produced unusual results that are worthy of further investigation; it would be interesting to investigate stimulus spacing effects, and the effect that discrimination with different numerical anchor value ratios would have on performance. The most significant outcome is the suggestion that perhaps it is possible for nonhuman animals to develop a linear representation of number with constant variability, similar to that obtained with human verbal counting. Relative numerosity procedures test a relatively simple numerical ability and responding is limited to one of two choices. Would similar findings be obtained with a more complex discrimination of absolute number, and in a procedure where greater response variability is afforded? The following experiments aim to answer these questions.

3 Chapter 3: Numerical reproduction

3.1 Notes on Experiment 2 and 2A

The current chapter presents two experiments that have previously been published in a paper titled "Numerical reproduction in pigeons", co-authored with my supervisors, Randolph Grace and Anthony McLean, and an earlier Master's student, Shasta Holland in Tan, Grace, Holland & McLean (2007) in *Journal of Experimental Psychology: Animal Behavior Processes, 33*. Experiment 2 includes data from Master's research conducted by Shasta Holland that has been reanalysed for the current research. These analyses differ from the original Master's research in three main ways; 10 instead of 5 sessions of baseline and transfer data were used in all analyses; hierarchical regression analyses were conducted to assess relative control by time and number; and additional variability analyses were conducted (these are reported in Chapter 6). Experiment 2A reports results of my Honours research, conducted in 2006, including additional variability analyses reported in Chapter 6.

3.2 Introduction

Research has shown that nonhuman animals are able to discriminate relative numerosity with relatively high accuracy under a variety of conditions. However, the ability to make accurate judgments of more or less is a relatively simple ability, which may not require a very sophisticated understanding of number. A stronger test of numerical competence would involve absolute number discriminations, where subjects have to be able to discriminate and respond to one particular numerical value.

In order to demonstrate true discriminations on absolute number, the following possibilities must be eliminated: 1) that the judgments are based on relative numerosity; 2) that

discrimination is stimulus specific; 3) that discriminations are based on another variable correlated with number, e.g. area or duration (Davis & Perusse, 1988). These issues are normally addressed by using tasks that require the discrimination of more than one number, using multiple stimuli in training and novel stimuli in testing, and controlling for stimulus characteristics that would normally covary with number.

Absolute number discriminations have not been as widely researched as relative number discriminations; this is most likely due to the high demands of the task and the extensive training that is required to obtain reasonable performance. The research that has been conducted has usually been performed in laboratory settings, and largely focussed on matching of responses to numerical stimuli (e.g. Boysen & Berntson, 1989; Xia, Emmerton, Siemann & Delius, 2001), as well as constructive tasks in which subjects have to produce a specific number of responses or select stimuli in numerical order (Beran & Rumbaugh, 2001; Beran, Rumbaugh, & Savage-Rumbaugh, 1998; Biro & Matsuzawa, 2001; Brannon & Terrace, 2000; Mechner, 1958; Xia, Siemann, & Delius, 2000).

A natural propensity to discriminate and represent absolute number may exist in nonhuman animals. Davis and Bradford (1986) investigated numerical discriminations in a task that incorporated features which simulated foraging behaviour in a rat's natural environment. Their procedure involved training rats to select a wooden tunnel in a certain ordinal position of a row of tunnels placed along the side of an enclosure filled with bedding material. Each rat was allowed to roam freely about the enclosure, but had to "count its way" (p. 267) to the correct tunnel to retrieve a food reward. In their first experiment, subjects had to choose the *n*th of six identical tunnels, by entering the tunnel through a hinged door to obtain its reward. Incorrect tunnels were also baited, but had inserts to prevent access to food following an incorrect choice. Probe trials with the inserts removed were also conducted to ensure performance was not based on the presence or absence of the inserts. Between each trial, bedding was raked and the "correct" tunnel was switched with one of the others to reduce olfactory or visual cues. The

spacing of tunnels along the enclosure wall was also randomized to minimize the influence of spatial cues. Subjects were trained for ten sessions to either select the third, fourth or fifth tunnel, and then tested.

All subjects learnt to select to the correct tunnel, often within the first block of test trials, and maintained performance at a level significantly above chance. There was no significant effect of ordinal number on performance, and almost all errors involved choices of tunnels either immediately preceding or following the target tunnel; 70% of these errors involved early entries. Davis and Bradford (1986) noted that all of the four subjects trained to select the 5th tunnel developed a "working down" strategy, where instead of counting up to the 5th tunnel from the start box, subjects merely ran to the 6th and last box and worked backwards. Thus, under these conditions, subjects resorted to a simpler approach where they only had to "count" to 2, rather than to 5. However, by the end of the first block of trials all but one subject had largely stopped using this response rule. Another noteworthy feature of this experiment is that subjects were not food-deprived during testing; that is numerical sensitivity was not food-dependent, and in fact subjects would often enter the correct tunnel but not eat the bait inside.

A second experiment was also conducted where the tunnels were arranged in an L-shape along two walls adjacent to each other, such that subjects had to turn a corner while running along the array of tunnels. This manipulation, which changed both the spatial arrangement and distances between boxes, did not appear to impair performance; all subjects responded correctly during the first trial in session 1 or during the second test session, and rates of correct responses remained high and above chance. Thus, subjects were able to transfer their ability to select the tunnel at a particular ordinal position to a different spatial configuration that altered the absolute location and distance between boxes.

Rats were also tested for long-term retention of performance, going through additional testing one year and 18 months after the first two experiments. Under the same conditions as in Experiment 1, 9 of the 10 subjects were still able to perform at levels significantly above chance

one year later. Both of the two remaining subjects were still able to select the correct tunnel 18 months after the initial experiments. These results show that the rats were able to remember and respond correctly in a task following an extended period of inactivity (Davis & Bradford, 1986). Overall, Davis and Bradford's findings showed that rats were able to learn to select the correct tunnel based on its ordinal position relatively quickly, and this ability was transferred to spatial changes in array and retained over an extended period.

Suzuki and Koboyashi (2000) extended Davis and Bradford's (1986) work in a series of experiments, controlling for possible alternative cues for responding and investigating performance with large target numbers. Additionally, they recorded behaviour to measure any indicating acts or tagging behaviour subjects may have been using. The general procedure was similar to Davis and Bradford's and required subjects to locate and enter a box positioned at a certain ordinal location to receive reinforcement.

Their first experiment was a partial replication of Davis and Bradford (1986); three rats were trained to obtain food by entering the fourth of 6 boxes. Each box had hinged doors at the opening, which rats had to push open to get to the food, and incorrect boxes had an additional stopper behind the door to prevent rats from entering the wrong box. Initially subjects were trained to enter the right box, with bait inside over a period of 4 days. Training trials began with all cues available and these were gradually removed. Training began with the bait only in Box 4 with the door removed, then with the door; by the end of the 4th day of training, the spacing of boxes varied pseudorandomly, all boxes were baited and had doors attached. Testing involved 10 days of 10 test trials each. To reduce the presence of olfactory cues during testing, the wood chips on the floor of the enclosure were raked and spread around between trials. The absolute location of the boxes varied randomly within the enclosure, and the individual boxes were also switched between trials so that the location and physical characteristics of the correct box was always physically different each trial, although the ordinal and relative location remained the same. Intentional or indicating acts, measured as any gestures made towards the boxes without

touching them, were also recorded.

Results showed that all rats performed significantly better than chance during even the first block of tests, although one rat, Subject A, showed a slight tendency to choose box 3 instead of 4. These results are largely consistent with Davis and Bradford (1986). Additionally, indicating acts were observed; on an average of 25% of correct trials, and 12% of incorrect trials, rats often stopped, bowed or turned their heads at box doors as they ran past them. The higher occurrence of indicating acts on a proportion of the correct trials suggest that these behaviours may have improved performance, although they were not necessary to produce a correct response.

In a second experiment, Suzuki and Koboyashi (2000) doubled the length of the enclosure and the total number of boxes to 12, to test whether subjects were using the relative position of the box to determine responding, or selecting the box located closed to the middle of the array. In this experiment, the location of box 4 ranged from 49 to 238cm from the corner closest to the start box. Performance under these conditions was at exactly the same levels as Experiment 1, with the exception of one rat (subject A) that persisted in choosing box 3 instead of box 4. The percentages of correct and incorrect trials in which indicating acts were observed were similar to Experiment 1; 22% and 10%, respectively. Thus, subjects did not seem to be basing responding on the relative location of the target box within the enclosure.

The extent of rats' numerical ability was tested in a third experiment, increasing the ordinal number of the target box gradually as performance allowed. Each subject began eight sessions, each consisting of 1 training day and 2 testing days, with target box 5. If subjects made at least eight correct choices out of 20 test trials, training began with the next ordinal number, up until target number 12. Subject A showed poor performance, failing to reach criterion for box 7. The main type of error made by this subject was largely the selection of the box immediately preceding the target box. Consequently Subject A was excluded from later testing. Conversely, Subject B reached box 11, and Subject C reached box 12 within the eight sessions. There did not

appear to be any increase in variability in responding as the target number increased and performance did not decline as number increased either. Indicating acts were found on average of 27% correct and 15% incorrect trials. The researchers noted that as the ordinal number increased, the indicating acts occurred later in the sequence, close to the target box.

Two additional experiments were conducted by Suzuki and Koyobashi (2000) to test for a "working down" strategy, where subjects used distance from the last box to determine choice, and to test for the possible use of stopper presence/absence as a cue. Results from both these experiments showed that subjects were not using either strategies and performance was maintained despite the added controls.

To test for an understanding of ordinality, Suzuki and Koyobashi (2000) trained 4 naïve rats to select the 3rd box out of a number of total boxes that varied randomly between 3 and 12. Testing showed three of four subjects were able to choose the correct box at rates significantly above chance and this performance was acquired within the first 20 trials and did not vary with the absolute number of boxes. This demonstrates that performance was not dependent on the relative position of the correct box. Indicating acts were relatively infrequent in this group, with occurring on an average of 7% on correct trials and 6% on incorrect trials. The researchers then increased the difficulty of the task by increasing the ordinal number of the target box in the same manner as in Experiment 3. Of the three subjects, one failed to reach criterion at box 6, while the other two reached box 7. The latter two subjects were then tested in another experiment where box sizes were also varied to determine whether distance cues were influence performance. Both subjects continued to choose the correct box at levels significantly above chance under these conditions.

Suzuki and Koyobashi's (2000) results provide convincing evidence for the discrimination of ordinal absolute number in rats. Even when location and total number of boxes, visual, positional and olfactory cues were varied, subjects were still able to discriminate ordinal locations of boxes up to 12. Additionally, subjects did not require extensive training in

order to reach criterion performance. This may be due to the relatively simple task demands; subjects merely had to discriminate a single ordinal number, or the sequential presentation of boxes, or it may be due to the resemblance of this task to a natural foraging situation. Indicating behaviours occurred on between 5-25% of trials, suggesting they were not critical in determining choice. However, the increased frequency of these behaviours on correct trials suggests they may have assisted discrimination. It should be noted that there was no increase in response variability with number, suggesting subjects were not merely estimating box number but were discriminating accurately. But, because performance was less than perfect, subjects' performance cannot be considered true counting; rather they appeared to be using some sort of protocounting process to determine responding.

3.2.1 Symbolic absolute number discriminations

Several studies have examined the association between abstract symbols and number in nonhuman animals. Beran, Rumbaugh and Savage-Rumbaugh (1998) investigated numerical discriminations in a constructive counting procedure. They trained a chimpanzee, Austin, in a computerized task to select sequentially items presented on a screen, using a joystick, equal to a presented target number. The target number was presented as an Arabic numeral located above a horizontal line that bisected the computer screen. Austin had to move the cursor to the target number to start the trial, and as this happened, several items were presented in the bottom half of the screen. The number of items varied on individual trials, but was always at least equal to the target number. These items had to be selected by moving the cursor to one and after a pause of half a second, that item moved to the top half of the screen and the cursor returned automatically to the center of the horizontal line. The trial was ended by moving the cursor back to the target numeral, but also was terminated automatically if the sequence of Arabic numerals was incorrect, or by ending the trial before selecting a few or larger number of items signaled by the target number. These errors were recorded, and the subject would have to repeat the same trial until the

correct response was made. Austin experienced 5 training conditions where he learnt to select the correct number of Arabic numerals or dots, with target numbers ranging from 1-9. Performance was then tested over 3 conditions. During testing, each stimulus, either numerals or dots, were placed in 1 of 11 locations in the bottom half of the screen. The first testing condition was similar to the initial training conditions; Austin was required to select Arabic numerals in the correct increasing sequence until the target number, which varied from 1-9, was reached. The following two testing conditions involved target numbers ranging from 1-4, with dots arranged in sequence, or dots arranged randomly or pseudo-randomly around the 11 positions, respectively. These conditions tested whether choices were being determined by position.

Performance on the sequential number and sequential dot trials was very high, however this alone does not indicate numerical understanding; the critical trials in the experiment for demonstrating an understanding of ordinality were the random and pseudo-random dot trials. A difference in performance on the random and pseudo-random dot trials would reveal a pattern in item selection. Results showed that Austin tended to use the existing sequential pattern for the smaller target numbers, 1 and 2, but never used the available sequential pattern for the larger target numbers, 3 and 4. Furthermore, he showed a distinct bias towards selection of the dots that were closest to the cursor when it was in the center of the screen, which resulted in actually avoiding a sequential pattern when selecting items, suggesting there was no advantage in arranging items that allowed a selection pattern other than proximity to the cursor.

Analyses across all random and pseudo-random trials showed that Austin's performance was significantly better than chance on all target numbers. Although performance was significantly better on the last 100 trials than the first 100 trials, performance was still significantly better than chance for the first 100 trials of each target number, suggesting additional training played only a partial role in determining correct responding. Errors were also examined, although only late exit errors, where more items than specified by the target number were selected, were explicitly discussed. Overall, Austin did not tend to repeat errors on

correction trials, suggesting his response strategy was based on number, as he would change the quantity of items selected on the correction trial. Performance on the correction trials tended to decrease as target number increased, from 94% on the target number 1 trials to 76% for the 4 trials. Additionally, it was found that Austin changed his selection sequence on 1/3 of the correction trials, and even more so on the 3 and 4 trials, while still selecting the correct number of items, suggesting successful responses on a correction trial was not a result of a non-numerical adjustment to an established selection strategy.

This experiment showed that a chimpanzee is capable of selecting a sequence of items equal to a presented target number. His responding met some of the criteria for the application of numerical tags. Behaviour was consistent with the one-to-one correspondence principle as only one item could be selected at a time, and followed the order-irrelevance principle- there did not appear to be any consistent selection pattern across target numbers and also for correction trials. Furthermore, Beran et al (1998) concluded that performance also adhered to the stable-order principle since dots were the tags to be used in each trial, and each constituted one "count". The performance of Austin in this task fulfills at least some of the requirements for counting; however, the adherence to some principles, namely one-to-one correspondence and stable-order principles seem to be largely determined by the use of identical dots in the procedure and its subsequent response limitations.

This study was further extended by Beran and Rumbaugh (2001), who trained two chimpanzees in a similar task. Subjects were presented with an Arabic numeral, ranging from 1-7, on the upper half of the screen. After touching the numeral with the joystick-controlled cursor, they were required to select, in sequence, the same number of white dots presented in randomly arranged arrays on the lower half of the screen. As the cursor made contact with each dot, the dot disappeared and reappeared on one of multiple shapes at the top of the screen; these served as feedback for number of dots selected. The position and number of shapes were randomly determined, with up to 30 shapes and at least as many shapes as specified by the target

numeral. Visual feedback was also excluded from some of the sessions, interspersed throughout the experiment. Subjects had to move the cursor back to the numeral to complete the trial. If the number of dots selected was less than the numeral presented or more dots were selected than the numeral, the trial was terminated. Subjects began the experiment with target numerals 1 and 2, and the target sets increased by an additional numeral after performance exceeded 70%. During early training, a correction procedure was included which incorrect trials were represented, however these were eliminated once subjects began working with the number 4.

Results showed that performance was significantly better than chance for the target numerals 1-7 and 1-6 for each subject, and the absence of feedback only appeared to affect performance for the largest number in each subject's target set. Thus, the chimpanzees appeared to be able to construct a numerical set successfully, based on the presentation of a target number. As a further test, Beran and Rumbaugh (2001) investigated whether subjects' responding was based on the duration, rather than number. Duration differed significantly as a function of number, however analyses of incorrect trials showed that although subjects tended to complete incorrect trials more quickly, this was not associated with a specific type of error. This finding, along with the large variability in durations on correct trials for all numbers, led Beran and Rumbaugh to conclude that differences in duration could not explain performance alone.

Analyses of response variability showed that the modal number of dots collected for each target was the number represented by the numeral, and errors of more than one dot were only obtained for the larger numbers 6 and 7. There was a significant negative correlation between proportion of correct trials and target number, and a significant positive correlation between the standard deviation of responding and the average number of dots collected. These results suggest responding exhibited scalar variability; however, it must be interpreted cautiously because of the response limit of n+1. It is unclear whether responding would show a similar pattern had there been no upper limit on the number of dots collected.

Xia, Emmerton, Siemann and Delius (2001) conducted a study by examining whether

numerosities could be associated with abstract symbols. Pigeons were trained to match sets of dots varying in number from 1-5 to corresponding letters, in a symbolic match-to-sample procedure. Subjects were required to peck and receive acoustic feedback for each element in the stimulus set, before the set was removed and the symbol array was presented with symbols arranged in a fixed X pattern. Correct responses were rewarded with access to food, and incorrect responses resulted in correction trials. Initial training only involved numerosities of 1 or 2, and their corresponding symbols. Subsequent numerosities were added individiually, as criterion performance was reached. Results of the first experiment showed only 5 of the 6 pigeons were able to complete the first training stage with numerosities of 1 and 2, and only 2 of the 5 were able to reach criterion in the final stage involving all 5 numerosities. Average response distributions calculated for the training stage with the values 1-4 showed the largest proportion of responses for any given stimulus set was for the corresponding letter symbol. Additionally errors were largely the selection of the symbol for the adjacent numerosity value of the correct symbol, suggesting that pigeons were responding on the basis of numerosity value, rather than spatial location.

In a second experiment, Xia et al. (2001) tested the possible role of various other cues in determining responding in the first experiment. They manipulated various aspects of the stimulus presentation; removing the acoustic feedback, the configurations of the numerosity arrays, changing the location of the elements in the stimulus set and of the symbols, including novel elements in the numerosity arrays, using heterogenous stimuli and reducing the response requirement during stimulus presentation to one peck. Performance was largely unaffected by the removal of the acoustic feedback, new stimulus array and response symbol configurations, and the use of heterogenous stimuli. The effect of including novel elements was mixed; there was positive transfer to square stimuli, but performance on trials with low, but not high, numerosities of triangles and butterflies was poor. Xia et al. attributed this decline in performance to weaker stimulus generalization due to the sudden change in area or shape.

Reducing the response requirement and duration of the stimulus presentation caused a drop in performance below chance levels for the higher numerosities, 3 and 4, but not the lower numerosities, 1 and 2. This suggests that responding to the stimuli, as well as visual perception, was being used to discriminate the numerosities. It was unlikely that subjects were using duration, or pecking rhythm to determine their response as they were still able to differentiate between 1, 2 and 3 despite only pecking once at the array.

Subjects' responding in this procedure met some of the criteria for counting, described by Davis and Perusse (1988). Pigeons were able to make judgments based on absolute number, using physical tagging of each individual item according to the one-to-one principle; additionally, keypecks to the stimulus elements did not occur in a fixed order, consistent with the order-irrelevance principle. The abstraction principle requires that absolute number discriminations be transferable to any type of items. Positive transfer from the original circular stimuli to novel stimuli occurred; however transfer appeared to be dependent on stimuli similarity, with poorer transfer occurring with fewer and more dissimilar items. Additional experience with these novel stimuli eliminated the drop in performance, such that responding was consistent with the abstraction principle. Unfortunately, the procedure was not able to test the principle of ordinality or stable-order principle by demonstrating that numerosities had been ordered along a continuum. Nevertheless, Xia et al (2001) have provided evidence for absolute numerical discriminations in nonhumans at a reasonably complex level.

Boysen and Berntson (1989) trained a chimpanzee to count food and objects using Arabic numbers, and tested her ability to use her representations of number to sum 0-4 food items or symbols placed in 2 or 3 sites. The chimpanzee, Sheba, had preliminary training in one-to-one correspondence in which she learnt to place one object per compartment in a divided tray, and select round cards that contained a number of metal disks that matched the number of food items, ranging from 1-3, presented on a tray. After reliable responding to the round cards with food items was obtained, they were replaced by plastic cards with Arabic numerals. She then was

required to match the Arabic numerals, presented on a monitor, to the original round cards with disks with the values 1-3, before the values 0 and 4 were introduced. It should be noted that it was at the introduction of the number 4 that Sheba's performance began to deteriorate, and she began to develop tagging behaviours similar to that her human teacher exhibited during the prior training. She began to touch, point to or move items in the tray before making a final decision and this tagging was reliably performed during all later numerical tasks. After training with 0-4 food items, Sheba's ability to match sets of items with an Arabic numeral was tested with a variety of inedible objects, using a double-blind procedure. Her overall performance under these conditions was 87% correct, suggesting she was able to transfer her skills developed in previous training with arrays of both homogenous and heterogeneous edible items, to heterogenous arrays of inedible objects.

Following these tests, Sheba's numerical ability was tested in two experiments. The first experiment involved a functional counting task, where one or two of three food sites were baited with oranges. These food sites were three distinct locations arranged such that the number of oranges was not visible from the start/choice platform, or any other food sites. The task required Sheba to move around all three sites and then select the correct Arabic symbol corresponding to the total number of oranges hidden at all three sites. For the first three trials, the experimenter walked Sheba around the three sites, and pointed and verbally indicated each orange before returning to the start platform, where Sheba made her choice. Following these trials, Sheba was left to move around the sites freely. Choices were indicated by physically touching, with her finger, one card of a row containing the values 1-3 arranged in order until the experimenter acknowledged her response.

Interestingly, Sheba's performance in the first session of testing was significantly above chance, suggesting no specific additional training was required for Sheba to generalize to the new requirements of the task; she had no previous experience with tasks that require her to move and track numbers in different locations, develop and remember a representation of the quantities

seen, and select the correct symbol corresponding to the sum of the two sets. Her high performance was maintained throughout the blind testing, in which the values 1-4 were used and the experimenter could not see or indicate Sheba's selection. Boysen and Berntson suggested that this could be explained by the extended application of numerical skills developed in previous numerical tasks, without having been trained explicitly in addition.

To further assess Sheba's ability to represent numbers, she was tested in a symbolic counting task with the numbers 0-4. This was identical to the previous task, except Sheba was now required to sum the numerical symbols displayed on a card at each of the two of three possible sites, instead of oranges. Also, testing trials included a blind condition, where the experimenter did not see Sheba's number selection, as well as a double-blind condition, where the experimenter did not see which numbers were placed at each site as well as Sheba's response. Sheba's performance was significantly better than chance for both the first session and blind tests, about 81% correct overall, and this level was maintained during subsequent blind and double-blind tests. More conservative probability test values were calculated using just response options that were not addends to ensure counting performance was not overestimated and Sheba was not merely just avoiding any of the addends when making her choice. Performance was still significantly better than chance under these conditions. Another possible strategy Sheba may have been using was merely selecting the next value higher than the largest addend (e.g. 1 + 3 = 4, select 4). To assess this, responding was examined on the critical trials where this strategy would fail; 2 + 2, or 0 + n, and results showed Sheba was not using this method to determine choice; performance was still significantly higher than chance at 94%. These results suggest that Sheba's ability to sum arrays of food items also applied to Arabic numbers, despite not having any additional training in summing combinations of numbers.

The performance of Sheba in this numerical discrimination provides strong evidence of counting, as defined by Davis and Perusse (1988). Her counting behaviour used tags that were organised in a consistent order and transferred to novel pairs of numbers, consistent with the

principles of cardinality and ordinality. Boysen and Berntson (1989) believed that serial learning of numbers played a large part in Sheba's ability to apply her numerical ability in different contexts, ensuring competence with smaller, simpler arrays before continuing with larger values, and different objects. They believed overtraining with labeling small arrays was necessary for flexibility in applying tags to develop. Through this process, Sheba developed an ability to apply counting principles to any collection of items, consistent with the abstraction principle proposed by Gelman and Gallistel (1978).

Tagging

One important process in the acquisition of true counting behaviour is the development of numerical tags and the understanding of their application. One of the defining characteristics of counting behavior is the use of serially-ordered number tags, or cardinality (Gelman & Gallistel, 1978). Although cardinality often involves the application of verbal labels, there is no requirement that tags be language-based. Any type of serially-learned markers may be used (Davis & Memmott, 1982); there is considerable evidence of nonverbal, behavioral tagging, in the use of fingers or other body parts during enumeration, in human cultures lacking a verbal counting system that supports this idea (Ifrah, 1985; Saxe, 1982).

Research shows that pointing and touching gestures, a form of behavioral tagging, improve counting performance in preschoolers (Alibali & DiRusso, 1999). The gesticular analog representation of number in finger symbol sets plays an integral part in the acquisition of counting by facilitating the development of one-to-one correspondence and the unification of elements being counted (Brissiaud & Greenbaum, 1992). Gesturing improves procedural competence by facilitating the application of children's knowledge of one-to-one correspondence, specifically aiding the partitioning of counted and uncounted items and the coordination of verbal and gesticular tags. Moreover, the use of gestures when counting appears to be most beneficial during, rather than after, the development of counting in children (Alibali

& DiRusso, 1999).

Given the importance of behavioral tagging for development of counting in humans, the investigation of analogues in nonhumans may provide some insight into their numerical abilities. However, there is only limited evidence for behavioral tagging in nonhumans. Early reports were anecdotal: Koehler (1950; cited in Davis & Memmott, 1982) described a crow that seemed to count the worms already eaten from a series of boxes by bowing the appropriate number of times before each empty box when returning to complete an interrupted trial. Mechner (1958) reported that one of his rats appeared to use the completion of a semicircle drawn with one paw to determine the criterion number of lever presses performed by the other. Additionally, Davis & Bradford (1986) reported rats performing tagging-type behaviours as they passed tunnel entrances on the way to the correct tunnel.

In the experiments conducted by Boysen and Berntson (1989), Sheba showed some evidence of behavioural tagging; pointing, touching or moving items that she was enumerating. Research with young children suggests that motor tags may facilitate the development of the counting principles, e.g. cardinality and abstraction (see Alibali & DiRusso, 1999 for review). If Sheba had been using tags consistently, rather than merely imitating the tagging behaviour of the experimenter during training, it would provide further support for the notion that Sheba had acquired some counting ability. Given the complicated nature of the task, it was unlikely Sheba had been using simpler numerical processes, such as subitising, to determine responding, and so the researchers concluded that Sheba was able to enumerate and sum the number of both food items and abstract symbols, and use them representationally.

Sheba's tagging behaviour was further investigated in a later study (Boysen, Berntson, Shreyer & Hannan, 1995). They reported that indicating behaviours were never reinforced directly; however Sheba was permitted a second recount when she made errors during initial training (a correction trial). If she still produced an incorrect response on the correction trial, the experimenter pointed to and usually touched each item while she counted them verbally, before

pointing to the correct numerical card as she repeated the correct number word. Every trial was recorded using a video camera and behaviour was scored by two external observers. After developing a behavioural glossary, the number of indicating acts made by Sheba in each session was scored. Sheba had extended contact with items while touching or moving the item around but also sometimes lost contact. An indicating act was scored if Sheba touched an item, then moved her hand away and touched the same item again within 1s, or if Sheba touched an item after moving it to a different area on the tray. Finally, the number of items to be counted and the final number chosen by Sheba were recorded.

Sheba's performance was significantly better than chance, even after results were adjusted to account for possible position cues. In terms of the indicating acts performed, Sheba used several behaviours, including pointing without making contact with the item, touching the candy, and moving the candy around the tray. Boysen et al. (1995) found that the number of indicating acts performed was significantly correlated with the number of items presented, as well as the Arabic numeral selected as Sheba's choice. Furthermore, the correlations between indicating acts and the number of items in the array were equal for both the correct and incorrect trials. This suggests that incorrect responses were not a result of errors in the performance of the indicating acts. Correlations between the number selected by Sheba and the number of indicating acts she performed were much higher on correct trials than on incorrect trials. Nevertheless, Sheba's responding on incorrect trials still exhibited systematic patterns; a significant correlation was found between the number of items presented and the number selected, and most errors only differed from the correct response by a single numerical unit. Additionally, Sheba exhibited a tendency to overestimate the actual number of items in terms of her final response and in her tagging. Regression analyses showed Sheba would generally produce twice as many tags as items presented on correction trials, and this tendency was exaggerated on incorrect trials. Based on this, Boysen et al. suggest Sheba may have used the indicating acts to physically proceed through the counting sequence, as well as for covert

tagging, and errors may have been a result of "recount errors", in which Sheba may have been counted items more than once. Given the reduced correlation between tags and response selection on incorrect trials, the indicating behaviours may have served to mediate correct performance.

A critical feature of tagging or indicating behaviors in counting is that the number of responses corresponds to the cardinality of the group of items being counted. Thus, the ability to discriminate and make different numbers of responses would serve as a precursor to behavioral tagging. This ability can be investigated in constructive counting procedures, where subjects are required to produce a particular number of responses.

In a pioneering study, Mechner (1958) investigated discrimination of number in response sequences in fixed-ratio and fixed consecutive number (FCN) schedules. In his procedure, rats were trained to respond on two levers in a Skinner box. Reinforcement was delivered after N consecutive responses on lever A had been completed, or after a minimum of N responses on lever B followed by another response on lever A. One of these schedules was randomly programmed for any given run. Consequently, subjects had to discriminate and produce N or N+1 responses in order to obtain reinforcement. One of Mechner's aims was to examine the effects of N, the response requirement on the probability of the termination of a response run as a function of the number of response already made, and the absolute probability of a response run of length n. Mechner found the response probability distributions varied as a function of the response requirement (N) with the maximum probability occurring at or near N and decreasing as N increased. The median and the variance of the run length also increased as N increased. The probability of the termination of a response run was always steepest around the value of N, and as N increased, the steepest portion of the function would flatten and the maximum probability value would also decrease. These results suggest that subjects were able to respond on the basis of the number of responses needed to obtain reinforcement, and the variability in responding increased as number increased – consistent with scalar principles.

The results of Mechner's experiment alone do not provide thoroughly convincing evidence for an ability to discriminate number; subjects may have been using the duration or rate of responding to determine when to terminate their response run. Wilkie, Webster and Leader (1979) adapted the procedure to unravel the relative contributions of time and number. Subjects still obtained reinforcement after a response on key A after at least N responses were made on key B, but they also introduced a blackout on both keys following each response on key B- this effectively disrupted the regularity of the temporal cues. Performance was virtually identical with both fixed and variable blackouts; subjects would still switch to key A after N key pecks on key B even though inter-response times on the B key, and consequently overall duration, varied greatly in the variable blackout condition. These findings suggest that the number of responses made were able to function as controlling cue for behaviour, independently of the duration. Requirements for reinforcement also appear to influence response number in this procedure: Platt and Johnson (1971) showed, by differentially reinforcing food-tray entries and error contingencies, that rats would only produce the minimum number of responses required for reinforcement.

Experiment 1 by Machado and Rodrigues (2007) extended Mechner's (1958) FCN schedule to examine performance while varying response criterions between the factors of 4 between 4 and 32, inclusive. In their procedure, subjects could obtain reinforcement by either pecking the left key exactly *n* times and switching to the right key, or if they failed to reach or exceeded this requirement, had to continue pecking the left key until a total number of at least 16 pecks on the left key had been made. The response criterion, *n*, was varied across conditions. Results were consistent with previous studies (Mechner, 1958; Platt & Johnson, 1971); subjects were able to learn to switch to response key B after making a certain number of *n* responses had been made on response key A over a wide range of *n* values. That is, subjects were able to learn response patterns differing in number. In a second experiment, Machado and Rodrigues also showed pigeons were able to learn and discriminate two different FCN schedules within the

same condition when subjects were either required to make exactly 4 responses and then switch response keys, or at least 16 times and switching.

Xia, Siemann and Delius (2000) trained pigeons to produce a specific number of responses when presented with one of several abstract symbols. Subjects were required to respond to a s_n stimulus, by making n pecks on that key followed by an additional peck to an enter key, in order to obtain reinforcement. If subjects made less than n pecks before pecking the enter key, or made n+1 pecks on the stimulus key, a time-out immediately followed. Subjects were initially trained with n = 1 to 4 (Experiment 1), before testing with values up to 6 (Experiment 2).

In Experiment 1, performance was above chance for all numerosities. The average modal number of responses to stimulus s_n was equal n, and the most common errors were response numbers of $n \pm 1$. Response distributions tended to flatten out as n increased, suggesting an increase in response variability as the response requirement increased. These results show that pigeons were able to discriminate between 4 different abstract symbols and their corresponding numerosities, and to produce different numbers of responses. In Experiment 2, n = 5 was added to the values for training and testing, before n = 6 was included in the stimuli set. Accuracy was above chance for all values, s_1-s_6 in testing. Although response distributions were steeper than in Experiment 1, they showed a general flattening with higher numbers.

As it was not clear how subjects were determining responding, Xia et al. (2000) analysed response duration-based errors to test whether responses were based on a timing strategy. The average duration between the first and last keypeck on the symbol key was calculated for the correct responses for each individual and numerosity for s_3-s_6. If subjects had been timing, the correct response times would have been memorized as a threshold time for switching to the enter key. Consequently, errors involving n-1 pecks may have been due to a slower than average response rate (and longer response duration), whereas errors involving n+1 pecks may have been due to a faster than average response rate (and shorter response duration). Results did not

support this, with the total number of cases not differing significantly from chance. These findings suggest that a timing strategy did not appear to be the sole determinant of responding in Xia et al.'s task.

There are some limitations to this study, however. It should be noted that not all subjects which began initial training continued onto training and testing with n=5, or n=6. Additionally, there was not enough data from the test stage alone for analysis and so data from their "consolidation" training trials, which did not have randomized stimulus presentations, were also included. The upper and lower limits placed on the possible number of responses made ($0 < n < n+1$) restricted response variability. This, and the exclusion of pigeons that failed to reach criterion performance in this task, as well as the lack of true test data with the values 5 and 6, makes it difficult to draw any strong conclusions that extend beyond that pigeons are able to respond constructively to abstract numerical stimuli.

3.2.3 A New Task: Numerical Reproduction

Here, we investigate a new procedure in which pigeons were trained to discriminate and reproduce the absolute number of response-dependent keylight flashes presented on a trial. The procedure combines aspects of both discrimination and production tasks and can therefore be described as *numerical reproduction* (cf. Zeiler & Hoyert, 1989, who studied a related temporal reproduction task).

The sample phase of each trial began with the onset of a houselight, and after a delay the center key was illuminated red. The length of the delays varied depending on experimental condition and is discussed in detail below. A single response extinguished the center key. The sequence of *center key on – response – center key off* is termed a 'flash'. During the sample phase, 2, 4, or 6 flashes occurred. After the flash sequence had been completed, all lights were extinguished for a 2s retention interval. The houselight was then illuminated, along with the center and right keys (red and green, respectively), to signal the production phase. To earn

reinforcement, subjects then had to make the same number of responses to the center key as flashes in the previous sample phase, followed by a single response to the right key. Incorrect responses were followed by a blackout and a correction trial. Thus, subjects were trained to reproduce the number of responses made to the lighted key in the sample phase.

The numerical reproduction procedure incorporates some features of Mechner's (1958) and Xia et al. (2000) counting procedures; subjects are required to produce a certain number of responses, plus an additional "completion" response following the presentation of a certain stimulus set. However, this procedure differs in one critical aspect; the permitted response variability. Typical FCN procedures (Machado & Rodrigues, 2008; Mechner, 1958; Platt & Johnson, 1971,) have no upper limit on response requirements; reinforcement follows any response run equal to or greater than the target number. Conversely, in the Xia et al (2000) experiment, a strict upper limit was imposed on response production; subjects immediately received a time-out if they exceeded the required number of responses. This severely restricted response variability and limited possible analyses. In the numerical reproduction procedure, there is no upper or lower limit on the number of responses emitted during the production phase, but only the correct number of responses are reinforced. Incorrect responses result in a correction trial. The stricter reinforcement contingencies may shape more accurate responding than obtained in the Mechner-type procedures, and the lack of limits on response number, unlike in Xia et al. (2000), will provide a complete picture of responding in the production phase.

The following two experiments use the numerical reproduction procedure in which the temporal organization of the sample phase was manipulated in various ways. Our basic goal was to determine whether pigeons could be taught to count in this procedure. An answer to this question requires a definition of counting. For the numerical reproduction procedure, true counting would require highly accurate performance in the production phase (i.e., a one-to-one correspondence between flash number and response number), accurate transfer to novel numbers, and evidence that responding was controlled by number independently of temporal

cues. It was not expected that the pigeons would be able to achieve these goals, but it may be possible to evaluate the extent to which they were able to approximate them.

One of the main aims of these experiments was to characterize the role of temporal variables and their influence on responding in the numerical reproduction procedure. Is it possible to obtain significant control by number above and beyond flash rate and sample phase duration when temporal and numerical variables are confounded, or only weakly correlated? What are the relative influences of time and number on responding? Also of interest was whether subjects would be able to learn to discriminate three different numbers simultaneously. Generally experiments introduce numerical values one at a time, starting from the lowest and gradually increasing requirements, but in this task subjects are exposed to 2-, 4- and 6-flash trials from the start of training; will subjects still learn to respond correctly to all three stimuli?

3.2.4 Experiment 2

In Experiment 2, the role of the temporal cues sample phase duration and flash rate were investigated through different temporal organization of the stimuli during the sample phase. There were two conditions, rate-controlled and time-controlled. In the rate-controlled condition, 2, 4, or 6 flashes were scheduled to occur at a constant rate during the sample phase, by keeping the inter-flash interval (IFI) constant at 2.5s. In the time-controlled condition, 2, 4, or 6 flashes were scheduled to occur within a 10s overall sample phase duration. Thus, in the rate-controlled condition, the IFI was constant while the sample phase duration covaried with number; whereas in the time-controlled condition, the sample phase duration was constant while the IFI covaried with number. Note that the obtained values of these temporal variables were expected to vary, to some degree, from those programmed depending on subjects' response latencies. However, it was anticipated that such variation would degrade the otherwise perfect correlation between temporal variables and flash number, and thus facilitate multiple regression analyses in which the relative control of responding by temporal and numerical cues could be assessed (cf. Fetterman,

1993). Transfer tests were also included in both conditions to examine performance with novel flash numbers (1, 3, 5, and 7). Separate transfer testing was carried out with temporal organizations that were consistent with baseline (rate-controlled with rate-controlled, and vice versa) and inconsistent with baseline (rate-controlled with time-controlled, and vice versa).

3.3 Method

3.3.1 Subjects

Subjects were four homing pigeons, numbered 175-178. They had previously served in an experiment on choice, but had no prior experience with timing or counting-related procedures. Subjects were maintained at approximately 85% of their free-feeding weights by additional feeding, when necessary, after experimental sessions. Water and grit were continuously available in their home cages.

3.3.2 Apparatus

Four identical chambers, measuring 40 cm by 40 cm by 33 cm, were used. Each chamber had three keys 21 cm above the floor arranged in a row. The center key was located 16 cm from each wall and the other two keys 8 cm on either side of the center key. Only the center and right keys were used during sessions. Each key required a force of approximately 0.15 N for a response to be registered. The panel also contained a houselight situated 8 cm above the center key, and a food hopper, situated 13 cm below. During reinforcement, the houselight and response keys were dark, and the hopper was illuminated and raised to allow access to the wheat. Fans attached to each chamber provided ventilation and masking noise during experimental sessions. Sessions were controlled and recorded by a computer running MED-PC software, located in an adjacent room.

3.3.3 Baseline Procedure

Sessions were conducted seven days per week, at approximately the same time of day. Sessions ended when 105 trials had been completed or after 120 minutes, whichever came first. Each trial was preceded by a 12s inter-trial-interval (ITI), during which the houselight and keylights were turned off. Trials were composed of a sample phase followed by a production phase. At the start of the sample phase, the houselight was illuminated. After a delay, which varied depending on the current trial type and condition, the center key was lighted red; a single response to the center key extinguished it. This sequence (delay, illumination of the center key, response) will be referred to as a 'flash'. Each sample phase consisted of 2, 4, or 6 flashes. After the last flash had been completed, the houselight and all keylights were darkened for a 2s retention interval. Trial type was determined pseudorandomly, subject to a constraint that each type would occur a certain number of times in each block of 9 trials. The relative frequencies of trial types were equal at the start of training (i.e., 2, 4, and 6 flashes each occurred 3 times in each block of 9 trials), but were occasionally adjusted for individual subjects, if necessary, to facilitate learning of the task. For example, if a subject began to make two responses during the production phase on most trials regardless of flash number; the relative frequency of trials with 2 flashes would be reduced.

Following the retention interval, the houselight was illuminated and the centre and right keys were illuminated red and green, respectively, to signal the start of the production phase. To obtain reinforcement, subjects were required to peck the centre key the same number of times as flashes presented in the sample phase, and then peck the right key once, completing the trial. For example, if four flashes had occurred during the sample phase, after the retention interval, subjects had to produce four pecks on the centre key before pecking the right key once to earn reinforcement. Once the right key had been pecked, the trial ended and both keylights were darkened.

If subjects had made the correct number of center key pecks, a single response to the

right key was reinforced. During reinforcement, the keylights and houselight were extinguished while the hopper was raised and illuminated for 4.5s. A response to the right key after an incorrect number of center key responses produced a 5s blackout, followed immediately by a correction trial. Correction trials were identical to the preceding regular trial, except that only the center key was lit red during the production phase. Once the correct number of center key pecks had been made, the centre key was darkened and the right key was illuminated green. A peck to the right key was reinforced by 1.5 s access to grain. Pilot testing found that performance improved if the reinforcer magnitude for correction trials was different from correct responses on regular trials (4.5 s).

The temporal distribution of flashes during the sample phase was varied in two conditions to examine the effects of controlling the flash rate or total sample phase duration on performance. In the time-controlled condition, all flashes were scheduled to be presented within a 10s interval, that is, two, four or six flashes were programmed to occur within 10 s. A 1s response latency was assumed, so that the delay preceding illumination of the center key for each flash was calculated as [10s – N*1s] / N, where N is the number of flashes presented on a particular trial. Thus, because the total duration of the sample phase was controlled, the flash rate covaried with respect to sample number. Conversely, in the rate-controlled condition, for each flash the center key was illuminated after a constant delay (1.5s), and so the length of the sample phase covaried with number. Assuming 1s response latencies on average, a two-flash trial lasted 5, a four-flash trial lasted 10s, and a six-flash trial lasted 15s. Thus, in each condition, one of the temporal variables (flash rate or sample phase duration) covaried with number while the other was controlled.

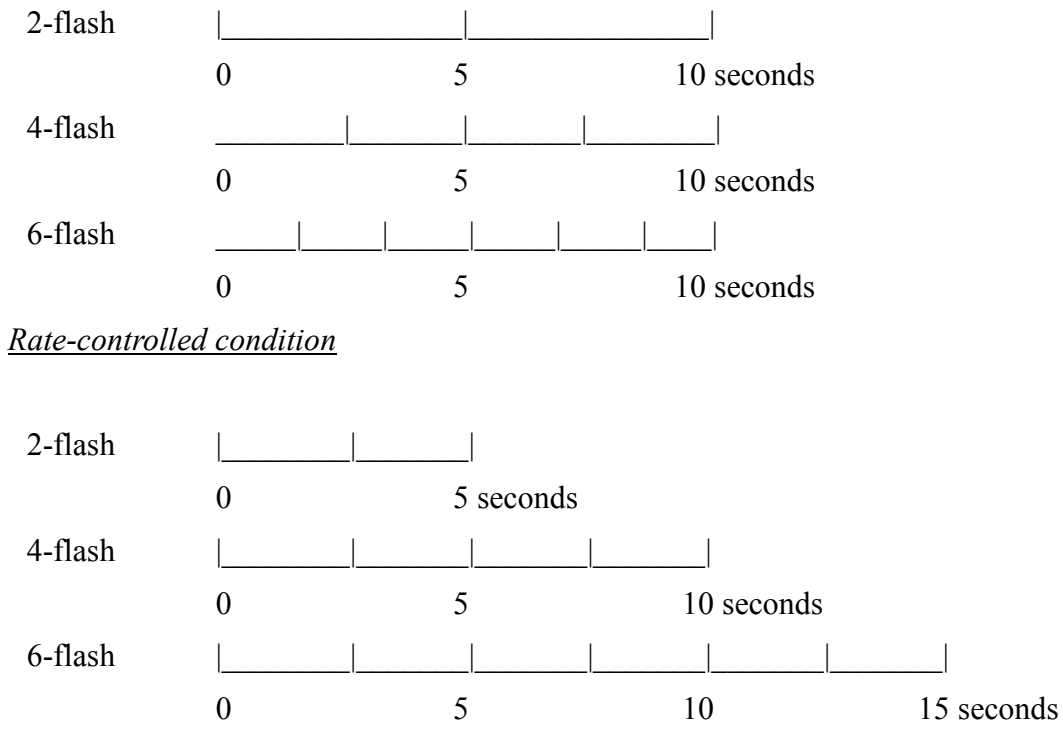

Figure 3.1. An illustration of flash presentation in the rate- and time- controlled tasks. In the time control procedure all flashes are presented in a 10-second interval, while in the rate control procedure one flash is presented every 2.5-seconds.

However, because the sample phase was response dependent, response latencies were expected to vary and thus it was not possible to ensure that the obtained flash rate (in the time-controlled condition) or the obtained sample phase duration (in the rate-controlled condition) was equal to the programmed value. Although the inability to control the temporal variables precisely may appear to be a weakness, it has the advantage of facilitating assessment of their relative control over responding. Multiple regression analyses were planned, in which control by flash number would be assessed after partialling out variance that could be accounted for by temporal variables (cf. Fetterman, 1993). These analyses would not be possible if the temporal variables were perfectly correlated with number.

3.3.4 Transfer Tests

After baseline training in each condition, two types of transfer test sessions were conducted. Test sessions included 4 types of probe trials in which the flash number was 1, 3, 5, or 7. There were 5 of each type, for a total of 20 probe trials in each test session. The remaining 85 trials were baseline trials, and the relative frequency of different flash numbers (2, 4, and 6) was similar to baseline.

There were two types of transfer test sessions, which differed in terms of whether the temporal structure of the sample phase on probe trials was the same or different as baseline trials. During consistent transfer test sessions, the sample phase during probe trials was arranged exactly as in baseline trials. For example, if the flash number was 3 for a probe trial, then for consistent transfer test sessions in the time-controlled condition, the delay preceding center-key illumination for each flash was [10s – 3s*1] / 3 = 2.33s. During inconsistent transfer test sessions, the sample phase was arranged as if the alternate condition were in effect. That is, probe trials during inconsistent transfer test sessions in the time-controlled condition were scheduled as rate-controlled trials, and vice versa. For example, for inconsistent transfer test sessions in the time-controlled condition, probe trials with flash number = 3 had a delay of 1.5s preceding each keylight illumination, and a programmed sample phase duration of 7.5s (assuming a 1s response latency). Reinforcement on probe trials was determined randomly, such that the probability of reinforcement was equal to the obtained reinforcement probability on regular trials averaged over the last 10 baseline sessions.

3.3.4.1 Training

Because all subjects had previous experimental histories (although none involving numerical or temporal discrimination tasks), training began immediately in the baseline procedure described above. Pigeons 175 and 176 were placed in the time-controlled condition; Pigeons 177

and 178 in the rate-controlled condition. Subjects often responded to the darkened center key between flashes, and so a contingency was arranged so that a flash could not occur unless at least 2s had elapsed since a response to the dark center key. However, this resulted in a dramatic decrease in accuracy for all subjects, so this was removed. Subjects then received the number of sessions listed in Table 3.1 for the first condition, before transfer tests were conducted. Subjects received 10 consistent transfer test sessions, followed by 10 additional baseline sessions, and finally 10 inconsistent transfer test sessions. All subjects then began baseline training in the second condition. Pigeons 175 and 176 were switched to the rate-controlled condition; Pigeons 177 and 178 to the time-controlled condition. After completing baseline training in the second condition, all subjects received transfer testing similar to the first condition.

The order of conditions, number of sessions of baseline training prior to transfer tests, and distribution of baseline trial types, are listed for all subjects in Table 1. All statistical tests used the .05 significance level.

Table 3.1. Number of sessions of baseline training in each condition, with distribution of trial types in parentheses. Note that a distribution of '4-2-3' would indicate that out of every nine baseline trials, there were four with flash number equal to two, two with flash number equal to four, and three with flash number equal to 6.

	Condition 1		**Condition 2**
Pigeon	**Time Controlled**		**Rate Controlled**
175	111 (4-1-4)		51 (4-2-3)
176	101 (1-4-4)		65 (2-3-4)
	Rate Controlled		**Time Controlled**
177	101 (1-4-4)		61 (2-3-4)
178	111 (3-3-3)		41 (2-3-4)

3.4 Results and Discussion

3.4.1 Baseline Training

Data were aggregated across the last 10 sessions of training prior to transfer testing in each condition. The primary dependent variable was the number of responses during the production phase, but we also analyzed two temporal variables from the sample phase: duration and flash rate. Sample phase duration was the cumulative duration of the sample phase, including all ISI and response latencies. Flash rate was the reciprocal of the average ISI, and was calculated by dividing the flash number by sample phase duration (excluding response latencies).

Planned contrasts found a significant linear trend for flash rate in the time-controlled condition, $F(1,3) = 49.69$, $p < .01$, and sample phase duration in the rate-controlled condition, $F(1,3) = 66.75$, $p < .005$. Sample phase duration increased with flash number in the rate-controlled condition, $M = 6.52$s [$SE = 0.90$], $M = 11.18$s [SE=1.31], and $M = 16.93$s [$SE = 2.04$], for 2-, 4-, and 6-flash trials, respectively, while flash rate remained approximately constant, $M =$ 0.39 flash/sec [$SE = 0.05$], $M = 0.42$ flash/sec [$SE = 0.05$], and $M = 0.42$ flash/sec [$SE = 0.04$]. Conversely, in the time-controlled condition, flash rate increased with number, $M = 0.17$ flash/sec [$SE = 0.02$], $M = 0.38$ flash/sec [$SE = 0.02$], and $M = 0.75$ flash/sec [$SE = 0.08$], for the 2-, 4- and 6-flash trials, respectively, while sample phase duration remained approximately constant, $M = 15.05$s [$SE = 2.42$], $M = 12.59$s [$SE = 1.03$], and $M = 10.52$s [$SE = 1.45$]. Planned contrasts for sample phase duration in the time-controlled condition, and for flash rate in the rate-controlled condition were not significant. This confirms that the procedure was effective in arranging different temporal organizations during the sample phase in the two conditions.

Correlations were calculated between the temporal variables and flash number for both conditions. Outliers (defined as sample phase duration > 30s) were omitted from the calculation

of these correlations, as well as the multiple regression analyses reported below[3]. As expected, correlations between flash rate and number were high in the time-controlled condition (Mean r = 0.86; SE = 0.02), as were those between sample duration and number in the rate-controlled condition (Mean r = 0.83; SE = 0.03).

Accuracy of responding was assessed by calculating the proportion of correct responses made during the production phase. As shown in the left panel of Figure 3.2, accuracy decreased as a function of flash number in both the time- and rate-controlled conditions. Accuracy was moderate for 2-flash trials (M = 0.41, SE = 0.24) but low for 6-flash trials (M = 0.18, SE = 0.11). A two-way repeated measures ANOVA with number and condition as factors found a significant main effect of number, $F(2,6) = 9.81$, $p < .05$, but no main effect of condition and no significant interaction. Thus, overall accuracy was moderate to low and decreased as a function of flash number, but there were no systematic differences between the rate- and time-controlled conditions.

The average numbers of responses made during the production phase are shown in the right panel of Figure 3.2. Overall, response number increased as a function of flash number, and there appeared to be little difference between the conditions. A two-way repeated-measures ANOVA with flash number and condition as factors found a significant main effect of flash number, $F(2,6) = 125.15$, $p < .001$, but the effect of condition and the interaction were not significant. Planned contrasts on flash number found significant linear trends for both the time and rate-controlled conditions, $F(1,3) = 41.27$, $p < .01$, and $F(1,3) = 43.37$, $p < .001$,

[3] Using a criterion of sample phase duration > 30 s resulted in an average of 2.07% and 1.17% of trials being omitted in the time- and rate-controlled conditions, respectively. Inclusion of these outliers would have reduced the correlations between temporal variables and flash number, and increased the incremental variance accounted for by flash number in the regression analyses. Excluding the outliers thus resulted in a more conservative test of control by flash number.

respectively. This demonstrates that response number increased as a linear function of flash number in both conditions. To quantify this relationship, average response number was regressed on flash number for each subjects' data. Averaged across subjects, regression slopes were 0.50 (*SE* = 0.26) for the time controlled condition, and 0.49 (*SE* = 0.23) for the rate controlled condition. Thus, although response number increased linearly with flash number, the slope was less than unity.

Response number distributions provide a more detailed picture of responding during the production phase, and are shown, averaged across subjects, for both conditions in the bottom panels of Figure 3.2. The averages are representative of individual data. Responding was differentiated across the three trial types, but was similar for both conditions. The modal response was correct for 2-flash and 4-flash trials in the rate-controlled condition, and for 2-flash trials in the time-controlled condition. The relative frequency of modal responding was greatest for 2-flash trials.

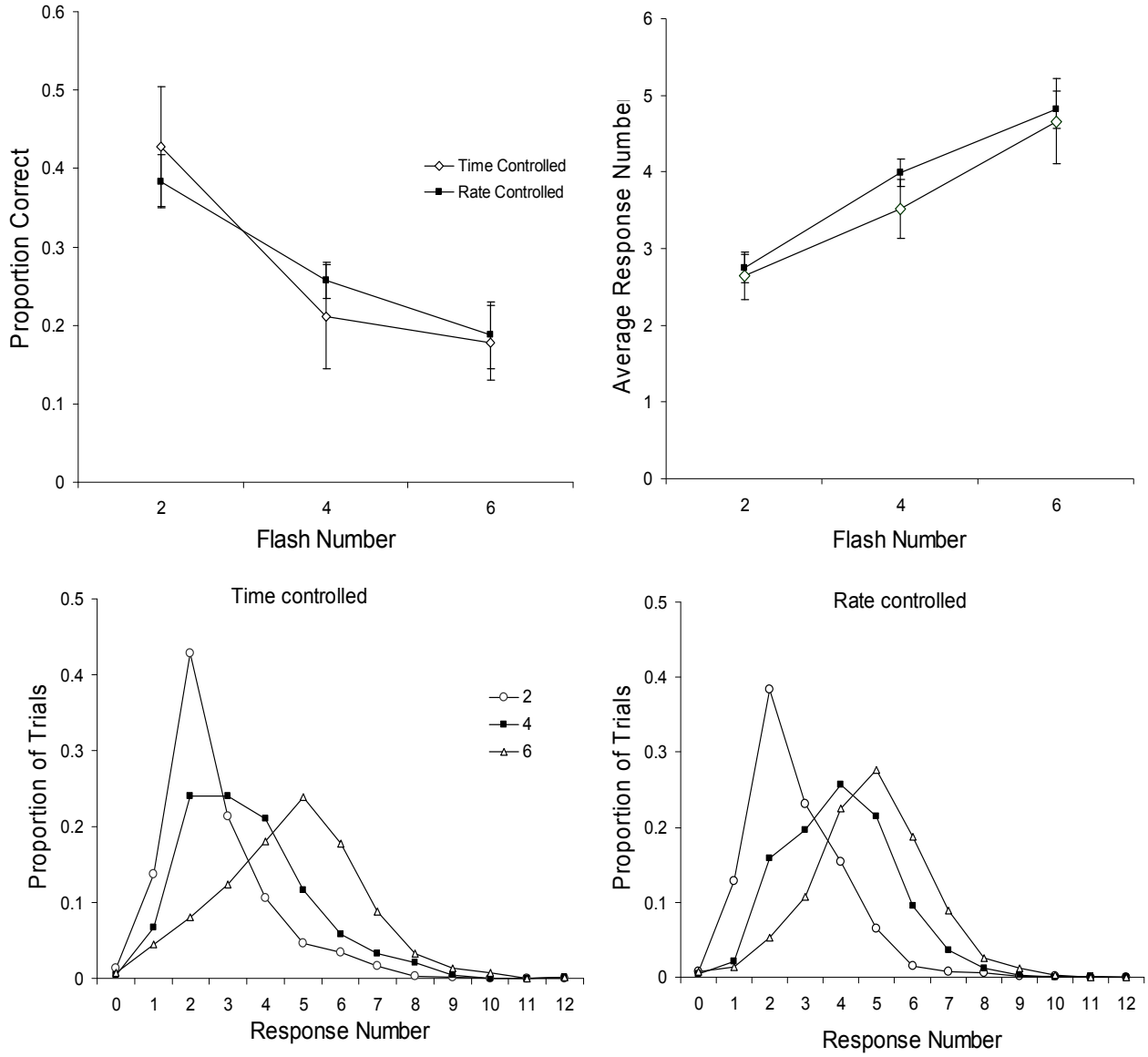

Figure 3.2. Average proportion of correct production phase responses (upper left panel) and average number of responses made during the production phase (upper right panel) as a function of flash number during the sample phase for both conditions in Experiment 2. Bars represent ± 1 S.E. The lower panels show distributions of numbers of responses made during the production phase for each baseline trial type (2, 4, 6 flashes) in the time-controlled (left) and rate-controlled (right) conditions of Experiment 2. Data are averaged across subjects.

3.4.2 Transfer Testing

Transfer tests were conducted to assess performance with novel flash numbers, in which the temporal organization of the sample phase was either consistent or inconsistent with baseline training. To the extent that subjects had acquired a rule-governed counting ability, equivalent positive transfer to novel numbers would be expected for both the consistent and inconsistent transfer tests. In the absence of such learning, it was anticipated that analysis of transfer responding would isolate the temporal cues that were most strongly related to responding in the production phase.

Data were aggregated over the 10 sessions of consistent and inconsistent transfer testing in each condition. The average numbers of responses during the production phase for transfer trials are shown in Figure 3.3. In the consistent transfer tests for both conditions, response number increased with the number of flashes presented in the sample phase, similar to that observed in baseline. Two-way repeated measures ANOVAs of probe trial data revealed significant effects. For the time-controlled condition, there was a significant main effect of number, $F(3,9) = 43.77$, $p < .001$, and a significant interaction $F(3,9) = 5.34$, $p < .05$, but no main effect of transfer test (i.e., consistent or inconsistent). Similar results were obtained for the rate-controlled condition; no effect of transfer test, but a significant effect of number $F(3,9) = 7.39$, $p < .01$, and interaction $F(3,9)= 33.29$, $p < .001$. Planned linear contrasts were also significant, $F(1,3) = 58.81$, $p < .01$ and $F(1) = 101.93$, $p < .005$, for both the time and rate-controlled conditions in the consistent transfer tests, respectively. These data show that average response number increased linearly with flash number in the consistent transfer tests, suggesting subjects were able to respond differentially to novel numbers. Tukey post-hoc tests were performed to investigate differences in the extreme trial types and found significant differences between the average response number for 1- and 2-flash trials in the rate-controlled condition ($p < .05$), but no significant difference between the average response number for 6- and 7-flash

trials in both the time- and rate-controlled conditions.

By contrast, responding during the inconsistent tests showed little or no evidence of positive transfer. Although average response number increased slightly with flash number in the time-controlled condition, the linear trend did not reach significance, $F(1,3) = 4.67$, $p = .12$. In the rate-controlled condition, average response number tended to decrease with flash number, and the linear trend approached significance, $F(1,3) = 8.94$, $p < .06$.

Overall, these results suggest that temporal cues were an important determinant of transfer responding, positive transfer to novel numbers was only obtained when the relationship between temporal cues and number was the same as baseline training. There was no evidence of differential responding between the conditions for the consistent tests, suggesting that both temporal cues, sample phase duration and flash rate, were capable of supporting transfer.

Average response distributions were also calculated for each of the transfer test trial types and are shown in Figure 3.4. In the time- and rate-controlled consistent transfer tests, clear differentiation between these trial types can be seen; with distinct shifts in response distributions to the right as flash number increased. Discrimination appeared to be better in the rate- than time-controlled condition, with response distributions in the latter for the 1- and 3-flash trials, as well as the 5- and 7-flash, trials showing a lot of overlap, suggesting subjects were responding to these two pairs of trial types as if they were two separate response categories. Response distributions for the inconsistent tests in both conditions show very little differentiation; the distributions superimpose a lot, and the modes are all located at approximately 4, regardless of trial type. This provides further evidence for deterioration in discrimination in the inconsistent transfer tests.

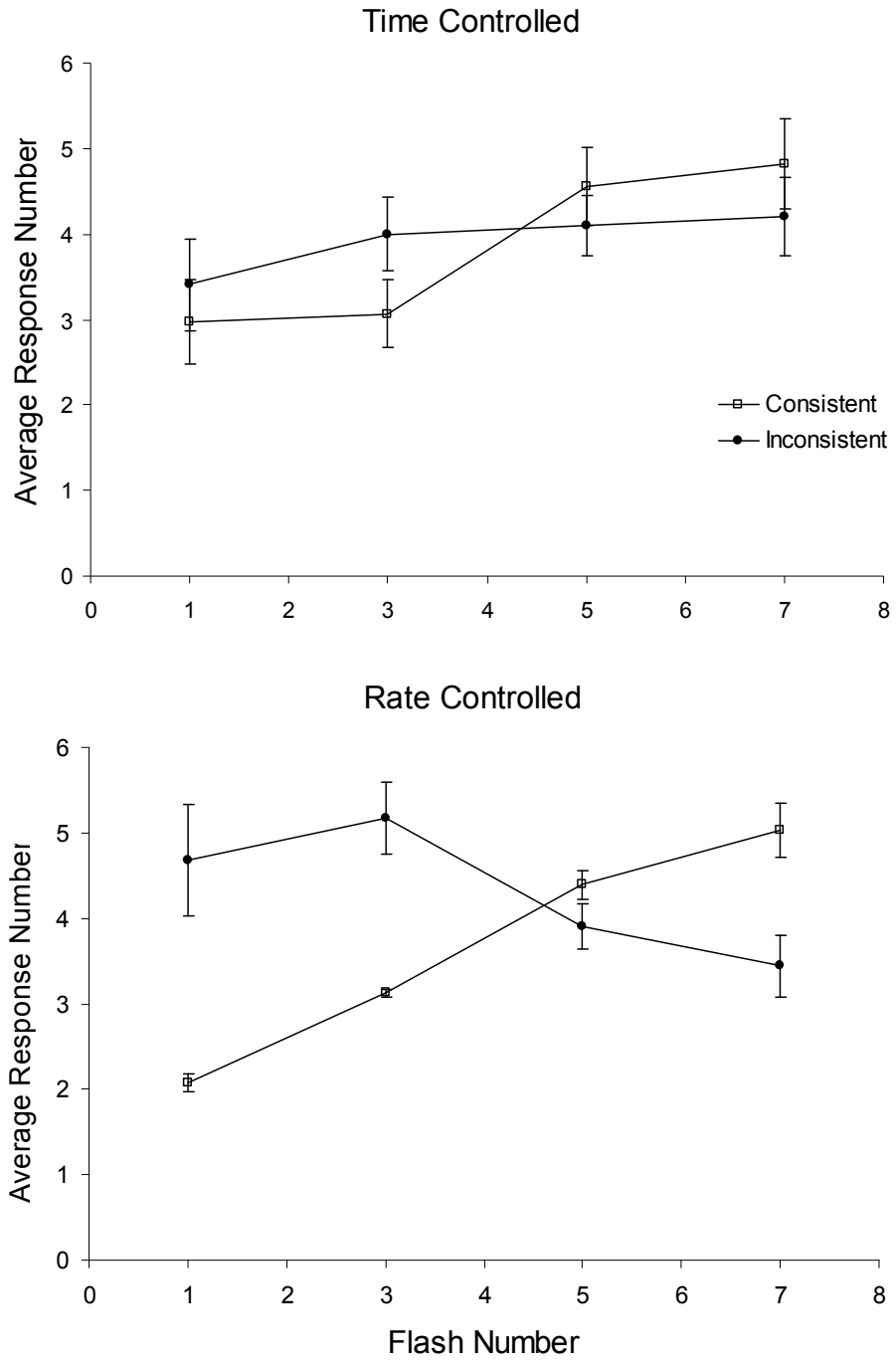

Figure 3.3. Average number of responses made during the production phase for transfer test trials in the time-controlled (upper panel) and rate-controlled (lower panel) conditions of Experiment 1. Bars represent +1 S.E.

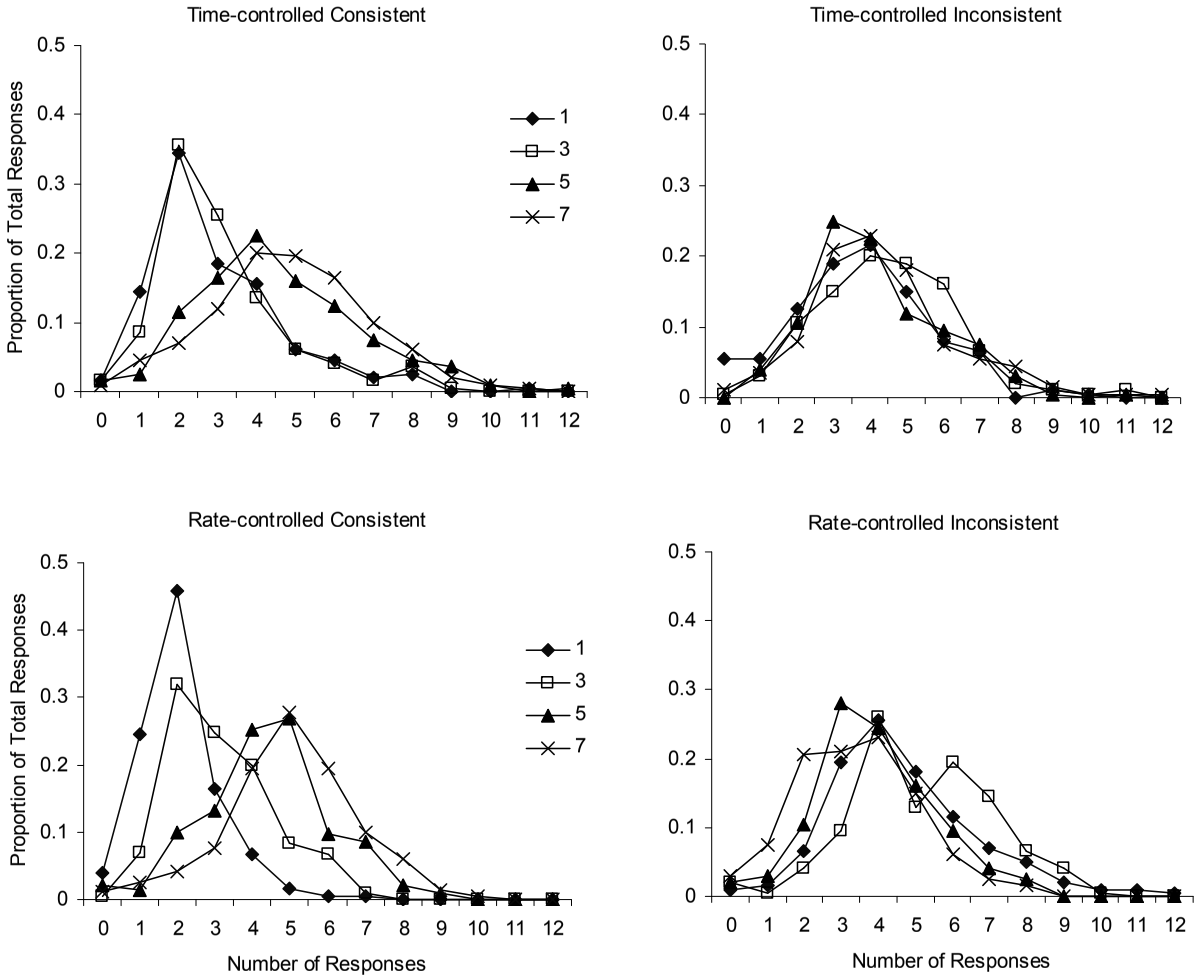

Figure 3.4 Average response distributions calculated for the transfer test trial types, 1-, 3-, 5- and 7-flash trials for the time-controlled consistent and inconsistent transfer tests (upper left and right panels, respectively) and the rate-controlled consistent and inconsistent transfer tests (lower left and right panels, respectively).

3.4.3 Multiple Regression Analyses

Hierarchical multiple regression analyses were conducted to investigate the relative influence of temporal variables and flash number on responding during the production phase. The major question was whether evidence could be obtained for control of responding during the production phase by flash number, after controlling for temporal cues. To answer this, sample phase duration and flash rate were entered into a regression model predicting response number at the first step, and then assessed for whether a significant increase in variance accounted for was produced when flash number was entered at the second step (cf. Fetterman, 1993). The increase in variance accounted for when the temporal cues were entered after flash number was also determined. Results are shown in Table 3.2.

Although temporal variables were highly correlated with flash number (*r*'s > .80), there was evidence that number acquired significant control over responding. For baseline data, flash number contributed significant additional variance beyond the temporal variables and had the largest beta weight in 6 out of 8 cases (exceptions were Pigeons 175 and 176, rate-controlled condition). For consistent transfer trials, flash number had the largest beta weight in 8 out of 8 cases, and contributed significant unique variance in 6 out of 8 cases (exceptions were Pigeon 178 time-controlled, and 175 rate-controlled). Significant control by flash number was also obtained in the inconsistent transfer trials for two subjects in Condition 2 (Pigeons 175 and 176, rate-controlled). The finding that similar results were not significant during Condition 1 suggests that when a particular temporal variable was correlated with number in the inconsistent transfer tests, prior exposure to that correlation was necessary for number to demonstrate significant control.

Table 3.2. Hierarchical multiple regression results from Experiment 2. Listed are beta weights and multiple R^2 values for regressions with cumulative sample duration, flash rate and sample number as predictor variables for response number for the last 10 baseline sessions and all transfer test sessions in each condition.

BASELINE									
Time controlled					*Rate controlled*				
Pigeon	**175**	**176**	**177**	**178**		**175**	**176**	**177**	**178**
Sample Duration	-0.12	-0.16***	0.06	-0.14*	**Sample Duration**	0.43***	0.41***	0.07	-0.20
Flash Rate	-0.03	0.10	0.03	-0.36***	**Flash Rate**	-0.03	-0.19	-0.05	-0.20***
Flash Number	0.54***	0.42***	0.48***	0.51***	**Flash Number**	-0.07	0.13	0.47***	0.76***
Multiple R^2	.36***	.38***	.25***	.07***	**Multiple R^2**	.13***	.30***	.28***	.37***
Number R^2_{inc}	.04***	.05***	.02***	.03***	**Number R^2_{inc}**	.00	.01***	.01***	.04***
Temporal R^2_{inc}	.01*	.03***	.00	.01**	**Temporal R^2_{inc}**	.04***	.07***	.01**	.02***
CONSISTENT TRANSFER									
Sample Duration	0.12	-0.11	0.00	0.24**	**Sample Duration**	0.16	0.23	-0.28	0.07
Flash Rate	0.05	-0.02	-0.18	0.14	**Flash Rate**	-0.04	-0.10	-0.26*	0.01
Flash Number	0.31	0.65***	0.68**	0.18	**Flash Number**	0.40	0.45*	0.81**	0.68*
Multiple R^2	.16***	.47***	.29***	.08***	**Multiple R^2**	.31***	.50***	.30***	.54***
Number R^2_{inc}	.01***	.08***	.04**	.01	**Number R^2_{inc}**	.01	.01*	.03**	.01*
Temporal R^2_{inc}	.00	.01	.01	.04*	**Temporal R^2_{inc}**	.01	.03**	.03*	.00
INCONSISTENT TRANSFER									
Sample Duration	0.06	0.18	-0.06	-0.24	**Sample Duration**	0.14	0.45**	-0.02	0.33**
Flash Rate	0.21	0.26	-0.12	-0.04	**Flash Rate**	-0.75**	-0.28	-0.69**	-.02
Flash Number	-0.06	-0.08	0.25	0.51	**Flash Number**	0.55**	0.52*	0.29	-0.18
Multiple R^2	.03	.07**	.10***	.08***	**Multiple R^2**	.18***	.11***	.17***	.24***
Number R^2_{inc}	.00	.00	.01	.01	**Number R^2_{inc}**	.04**	.02*	.01	.00
Temporal R^2_{inc}	.03*	.03*	.02	.00	**Temporal R^2_{inc}**	.11***	.10***	.05**	.05**

* $p < 0.05$ **$p < .01$ *** $p < 0.001$

Overall, results from Experiment 2 show that the numerical reproduction procedure is useful in studying nonhuman numerical competence. Subjects appeared to be able to respond on the basis of number: the number of responses made during the production phase increased linearly as a function of flash number; positive transfer to novel stimuli was obtained; and number accounted for significant additional variance in responding beyond temporal cues in a majority of cases. However, overall levels of accuracy were only low-to-moderate, and positive transfer was only obtained when the temporal organization of the sample phase was consistent with baseline (see Figure 3.3 and 3.4). This suggests that subjects were not truly discriminating number alone but responded primarily on the basis of whichever temporal cue was the most valid predictor of the correct response.

Although significant, unique control by number over responding was obtained, discrimination was still largely based on temporal cues. Sample phase duration and flash rate were strongly correlated with number in the rate- and time-controlled conditions, respectively. Regression analyses indicated that responding was significantly related to flash number after controlling for temporal cues, suggesting that acquisition of responding in the task might have depended largely on temporal cues. Note that similar to Fetterman, (1993), some caution must be taken when drawing conclusions based on the results of the regression analyses due to the strong covariation between the predictors used in the model. The role of temporal cues is further suggested by the disruption in performance in inconsistent transfer tests. When the temporal organization of the sample phase was changed during the inconsistent transfer tests, responding no longer increased with sample number. This finding supports the results of Breukelaar and Dalrymple-Alford (1998), and the "last resort" hypothesis (Davis & Memmott, 1982); subjects responded preferentially to temporal cues over number. Would the bias for time-based responding still persist when temporal cues became less reliable predictors of the correct response? To test this, a second experiment was conducted, in which the relationship between

number and the temporal variables, sample phase duration and flash rate was degraded. The specific aim of this experiment was to test whether significant control by number, and accurate responding could be obtained in the absence of reliable temporal cues. Additionally, the reduced correlation between the numerical and temporal variables would permit stronger quantitative analysis of their respective abilities to predict response number.

3.5 Experiment 2A Method

3.5.1 Subjects

Subjects were four homing pigeons, numbered 191-194. All subjects had experimental histories involving choice procedures, but no experience with counting or timing tasks. Subjects were housed and maintained according to the same conditions as Experiment 2.

3.5.2 Apparatus

The apparatus used in this experiment was the same as Experiment 2.

3.5.3 Procedure

All details of the procedure were the same as Experiment 2 with the following exceptions. At the start of a trial, an expected sample phase duration was selected randomly without replacement from a list of durations (in seconds): {5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15}. The programmed average inter-flash interval for the particular trial was then calculated as expected sample phase duration divided by flash number. Finally, the individual inter-flash intervals were determined by multiplying the average flash interval by a delay sampled without replacement from a distribution of 12 delays with an average of 1s generated by an exponential progression (Fleshler & Hoffman, 1962). This double randomization procedure was expected to degrade the correlations between both temporal variables and number.

All subjects received 129 sessions of baseline training, except for Pigeon 192, who received 125 sessions. The percentages of baseline trial types were sometimes changed for individual pigeons, but for the final training procedure, trial types were determined pseudorandomly subject to the constraint that out of every 9 trials, there were 2, 3, and 4 trials of 2-flash, 4-flash, and 6-flash trials, respectively. All pigeons received at least 20 sessions of training with the final procedure.

After baseline training, ten transfer test sessions were conducted with novel probe trials. During transfer test sessions, trials with 1, 3, 5 and 7 flashes during the sample phase were intermixed with regular (2, 4, 6) baseline trials. Each session of testing consisted of 85 baseline trials (with 19, 28, and 38 of type 2, 4, and 6 respectively) and 20 probe trials (5 each with 1-, 3-, 5-, and 7-flashes), with the identity of each trial determined pseudorandomly. Reinforcement on probe trials was randomly determined; the probability of reinforcement was equal to the probability of obtained reinforcement on regular trials averaged over the preceding ten baseline sessions. In addition to the variables studied in Experiment 2, we also recorded the individual response latencies during the production phase for each trial.

3.6 Results and Discussion

3.6.1 Baseline Training

Data were aggregated over the last 10 sessions of baseline training for analysis. Although both total sample phase duration and inter-stimulus-intervals were pseudo-randomized in Experiment 2A, both sample phase duration and flash rate increased with flash number; the means and standard errors for sample phase durations were M =12.79 s [SE = 0.81], M = 13.90 s [SE = 0.74], and M = 14.35 s [SE=0.26], for 2-, 4- and 6-flash trials, respectively. The means and standard errors for flash rate were M = 0.43 flash/sec [SE = 0.004], M = 0.57 flash/sec [SE = 0.004], and M = 0.78 flash/sec [SE=0.004] for 2-, 4- and 6-flash trials, respectively. One-way repeated measures ANOVAs found significant effects of number on flash rate, $F(2,6) = 1077.90$,

$p<.001$, but not sample phase duration, indicating that flash rate increased with number. Correlations between the temporal variables and flash number were calculated for individual subjects. Outliers were excluded from these correlations as well as from the multiple regression analyses using the same criterion as in Experiment 2[4]. Correlations were considerably lower than those in Experiment 2. Averaged across subjects, the correlations with flash number for sample duration and flash rate were $r = .14$ and $r = .28$, respectively. This shows that the randomization procedure for determining the temporal organization of the sample phase was successful at reducing the validity of temporal cues as predictors of flash number.

The average numbers of responses during the production phase and proportion of correct trials are shown in the left and center panels of Figure 3.3. Response number increased linearly with flash number while accuracy decreased, $F(1,3) = 27.07$, $p<.05$, and $F(2,6) = 15.46$, $p<.005$, respectively. Levels of accuracy were low-to-moderate, comparable to Experiment 1, $M = 0.35$ [$SE = 0.49$], $M = 0.22$ [$SE = 0.04$], and $M = 0.15$ [$SE = 0.05$] for 2-, 4-, and 6-flash trials, respectively. Average response distributions for each trial type are displayed in the right panel of Figure 3.5. Responding was well differentiated between 2- and 4-flash trials, for which the modal responses were correct, but there was greater overlap between 4- and 6-flash trials. Overall, data are similar to Experiment 2.

[4]An average of 3.35% of baseline trials were omitted using a criterion of > 30s.

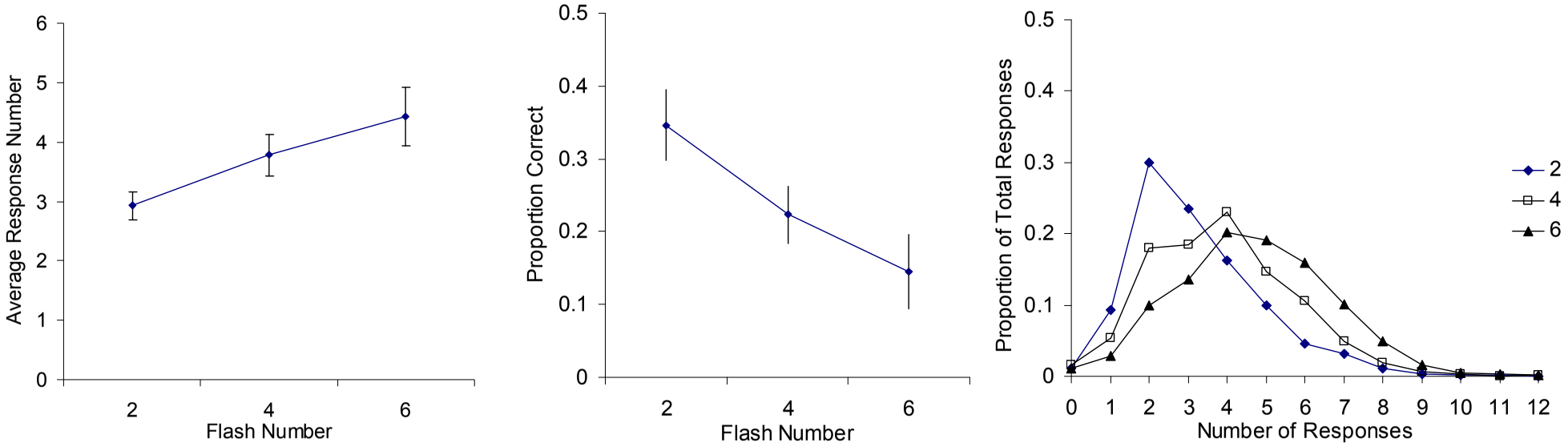

Figure 3.5. Average number of responses made during the production phase (left panel), proportion of correct responses (center panel), and distributions of response numbers during the production phase (right panel), for all trial-types in the last 10 baseline sessions in Experiment 2A. Error bars represent ± 1 S.E.

3.6.2 Transfer Testing

Average response number plotted as a function of flash number for the transfer and baseline trials in transfer-test sessions are shown in the left panel of Figure 3.6. Average response number appeared to increase as flash number increased on both the transfer and baseline trials. A repeated measures ANOVA conducted on data from the transfer test trials showed a significant effect of flash number on average response number, $F(1,3) = 19.58$, $p<.001$. A planned linear contrast revealed that response number increased linearly with flash number, $F(1,3) = 27.46$, $p<.01$. Tukey post-hoc tests comparing average response number between trials with 1 and 2 flashes, and 6 and 7 flashes, found results similar to the rate-controlled consistent transfer test in Experiment 2; fewer responses were made on trials with 1 compared to 2 flashes, $p<.001$, but there was no significant difference between trials with 6 and 7 flashes.

Response distributions were also calculated for the transfer test trials and are plotted in the right panel of Figure 3.6. Clear differentiation in response distributions for the 1-, 3- 5- and 7-flash trials can be seen; modes increase as flash number increases, and a corresponding increase in variability in the distributions can also be seen.

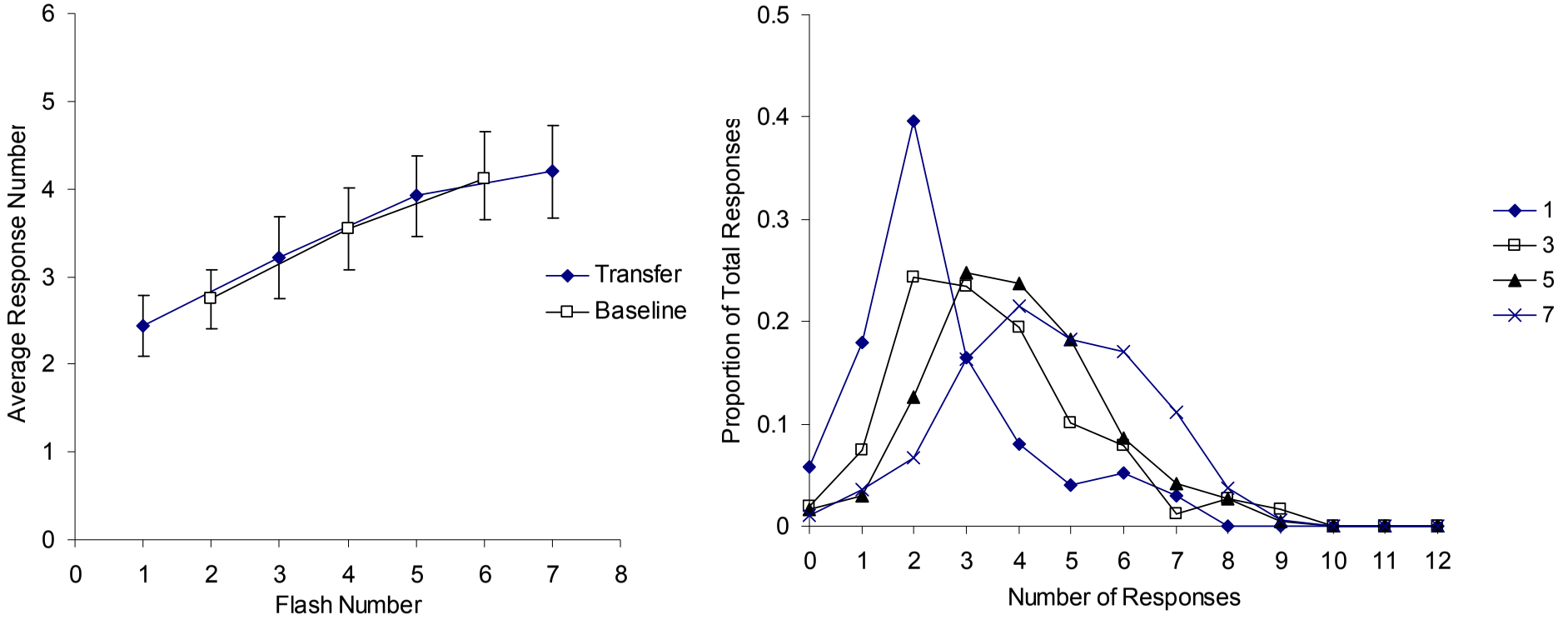

Figure 3.6. Average number of responses made during the production phase during baseline and transfer test trials (left panel) and distributions of response numbers for transfer test trials in Experiment 2A. Error bars indicate ±1 S.E.

3.6.3 Multiple Regression Analyses

Hierarchical multiple regressions were conducted to investigate the relative control by temporal and numerical cues over responding during the production phase. As in Experiment 2, the increase in variance accounted for was calculated when the temporal variables, sample duration and flash rate, were entered into a regression model after flash number, and vice versa. Results for baseline and transfer tests are reported in Table 3.3.

For baseline, number contributed significant additional variance beyond that explained by the temporal variables and had the largest beta weight for all four birds. R^2_{inc} values for number were also considerably larger than in Experiment 2 (M = .10 in Experiment 2A, compared to .04 and .02 for the time- and rate-controlled conditions in Experiment 2, respectively). However the total variance accounted for by the complete model (R^2) was only .14, substantially less than Experiment 2 (mean R^2 = .27 and .24, in the time- and rate-controlled conditions, respectively).

Results from the transfer tests also showed number accounted for a significant amount of variance above and beyond the temporal variables, and R^2_{inc} for number (M = .17) was greater than comparable results from Experiment 2 (M's = .04 and .02 for the time- and rate-controlled consistent transfer tests, respectively). Additionally, beta weights for flash number were greater than those for sample phase duration and flash rate for three of four subjects (with the exception of Pigeon 191). The total variance accounted for in transfer tests was greater than in baseline, and was comparable to values obtained in the time-controlled consistent transfer condition in Experiment 2 (mean R^2 = .24 and .25, respectively). Overall, these analyses show that randomizing the inter-flash intervals in Experiment 2A resulted in greater control by flash number over responding in the production phase independently of temporal cues.

Table 3.3. Hierarchical multiple regression results from Experiment 2A. Listed are beta weights and multiple R^2 values for regressions with cumulative sample duration, flash rate and sample number as predictor variables for response number for the last 10 baseline and transfer sessions.

BASELINE				
	Pigeon			
	191	**192**	**193**	**194**
Sample Duration	0.16***	-0.02	0.02	0.01
Flash Rate	-0.17***	-0.05	-0.13***	-0.06
Flash Number	0.28***	0.42***	0.44***	0.22***
Multiple R^2	.15***	.17***	.19***	.04***
Number R^2_{inc}	0.06***	0.16***	0.14***	0.04***
Temporal R^2_{inc}	0.09***	0.00	0.02***	0.00
TRANSFER				
	Pigeon			
	191	**192**	**193**	**194**
Sample Duration	0.24***	-0.12	0.16*	-0.04
Flash Rate	-0.01	-0.13	-0.05	-0.05
Flash Number	0.20*	0.59***	0.59***	0.36***
Multiple R^2	.13***	.31***	.39***	.12***
Number R^2_{inc}	0.03*	0.28***	0.26***	0.11***
Temporal R^2_{inc}	0.06*	0.02	0.04**	0.00

Note: * $p<.05$, ** $p<.01$, *** $p<.001$.

3.6.4 Inter-Response Time Analyses

Whether pigeons responded at a constant tempo within runs during the production phase was investigated by analyses of inter-response times. If so, the latency to make the first response (initial pause) and the latency to peck the green key (report latency) might be longer, but the within-run inter-response times (IRTs) should be constant.

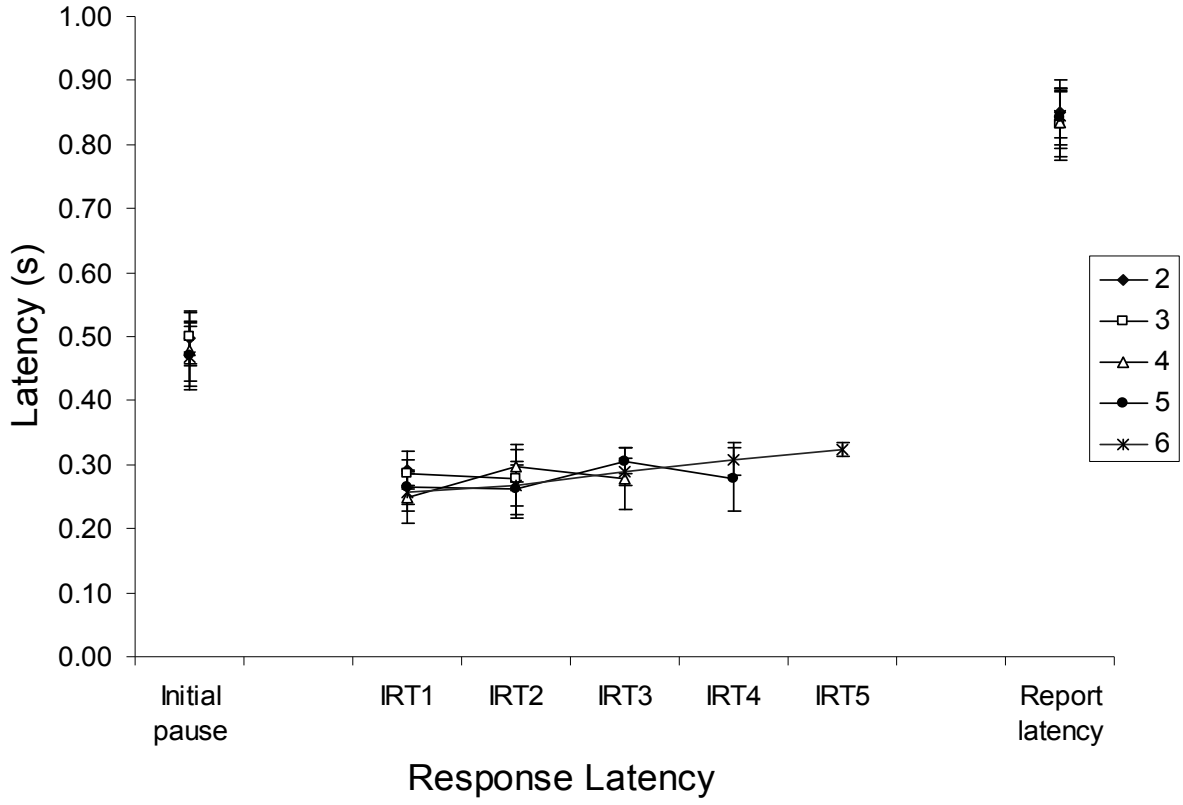

Figure 3.7. Average response latencies during the production phase for Experiment 2. For runs of two through six responses, the latency prior to the first response (initial pause latency), the successive inter-response times for responding on the red key (IRT1 through IRT5), and the latency to peck the green key (report latency) are shown. Error bars indicate ± 1 S.E.

Figure 3.7 shows the initial pause, report latency, and IRTs for runs between 2 and 6 responses. Data were pooled for individual subjects for the last 10 sessions of baseline, and averaged across subjects. For all runs, within-run IRTs were constant and relatively short (~0.3s), whereas initial pauses and report latencies were longer. One-way repeated measures ANOVAs on the latencies in Figure 3.5 were significant: $F(2,6) = 78.02$, $p<.001$, $F(3,9) = 53.65$,

$p<.001$, $F(4,12) = 50.14$, $p< .001$, $F(5,15) = 53.95$, $p<.001$, $F(6,18) = 65.13$, $p < .001$, for response numbers 2 through 6, respectively. Tukey post-hoc tests were conducted to investigate the differences between the initial latency, report latency and within-run IRTs. In all cases, the initial pause and report latencies were greater than within-run IRTs, but there were no significant differences among the within-run IRTs. This confirms that pigeons responded at a constant tempo during runs.

The within-run IRT results are similar to those of Xia, Siemann and Delius (2000). They trained pigeons to make from one to six responses to different stimuli projected on a key, and found that they responded at a high rate that was constant across number requirements. Xia et al. also observed that for each number, the final within-run IRT was slightly longer than those preceding; however, we found no evidence of this effect in our data. In Experiment 2A, responding at a high, constant tempo may have served effectively as a subdivision strategy, limiting variability as larger response numbers were produced (cf. Killeen & Weiss, 1987).

Results of Experiment 2A demonstrate that control by flash number, in the absence of reliable temporal cues, can be acquired in the numerical reproduction procedure. Moreover, degrading the validity of temporal cues appears to have increased relative sensitivity to number: Beta weights for number in the multiple regression analysis of baseline performance were always greater than those for temporal cues, and the unique variance associated with number was substantially greater compared with Experiment 2.

3.7. General Discussion

The goal of these two experiments was to investigate numerical competence in nonhumans using a novel task in which pigeons were trained to match the number of responses made during a production phase to the number of keylight flashes seen in a prior sample phase. In Experiment 2, the effects of temporal patterning in the sample phase were studied in two conditions in which flashes were either programmed to occur at a constant rate, or within an

overall constant duration. Thus, in the rate-controlled condition, sample phase duration covaried with flash number whereas in the time-controlled condition, flash rate covaried with number. In Experiment 2A, the correlation between temporal cues and flash number was degraded by generating inter-flash intervals and sample phase durations pseudo-randomly. Results showed that in both experiments, responding during the production phase increased as a linear function of flash number in baseline training (2, 4, 6), and positive transfer was obtained to novel numbers (1, 3, 5, 7). Transfer only occurred in conditions where the sample phase organization was similar to baseline. If subjects had learned to count the number of flashes during the sample phase according to an abstract rule (Gelman & Gallistel, 1978), temporal cues would have been irrelevant and positive transfer would have been obtained in the inconsistent transfer tests in Experiment 2. Failure of such transfer rules out human-like counting behavior, and shows that temporal cues were an important part of what subjects learned. Training in Experiment 2 did not produce control of subjects' responding by numerical cues independently of temporal cues.

However, multiple regression analyses showed that flash number was associated with unique variance in production phase responding after controlling for temporal cues. Flash number was a significant predictor of responding in both conditions of Experiment 2, but the unique variance associated with number was overall greater in Experiment 2A when the predictive validity of temporal cues was degraded. Thus, although subjects appeared to use temporal cues when they were available as a reliable indicator for responding, some numerical control over behaviour was still present and this numerical control was enhanced by reducing the covariation between time and number.

Our findings replicate and extend that of previous research that examined temporal and numerical control in numerical bisection procedures. Results of regression analyses were similar to those reported by Fetterman (1993). In his study, subjects were trained to discriminate between a larger and smaller fixed-ratio (FR) response requirement by subsequent choice of a red or green key. Fetterman conducted multiple regression analyses in which both the time taken

to complete each ratio and the ratio value were used to predict choice. Results showed that both ratio time and ratio value were associated with unique variance in choice responding, indicating that performance was controlled by both temporal and numerical cues. Roberts and Mitchell (1994) also drew similar conclusions from their research, which also required the discrimination of two flash sequences. Birds processed both temporal and numerical information simultaneously, and the extent of control by the respective variables was affected by the conditions of training the subjects experienced.

Although subjects were clearly not truly counting in this procedure, subjects were able to discriminate number and respond differentially to three numerical values trained simultaneously, and transferred performance to novel numbers both within and outside the training range. Performance in this procedure meets the three criteria for demonstrating absolute number discrimination. As already discussed, significant control by number was obtained above temporal variables in both experiments, and in Experiment 2A, temporal cues were randomized so they were not reliable indicators of response requirement; thus responding could not have been based on non-numerical cues.

Discriminations could not have been stimulus specific as stimuli were uniform keylight flashes, which varied only in terms of number, sample phase duration and flash rate. Subjects may have been responding to the specific stimulus patterns in the time- or rate-controlled conditions in Experiment 2, as they remained unchanged throughout training, however this would not have been possible in Experiment 2A as both temporal variables were randomized every trial. Additionally, accurate performance during transfer testing to novel numbers and stimuli is also evidence that responses were not based on specific stimulus patterns.

It is highly unlikely that subjects were merely discriminating relative numerosity, due to the use of three training values, and the task requirement of both discriminating the number of flashes and reproducing that number in key pecks. Although merely generating “more” or “less” responses on trials with larger or smaller numbers respectively may have resulted in some

covariation between flash number and response number, it would not have been sufficient to produce the observed response accuracy, especially considering the single-unit difference in test and baseline values.

Thus, subjects were able to learn absolute number discriminations in the numerical reproduction procedure, successfully responding differentially to three different numbers of flashes, and transferring performance to novel stimuli if conditions were the same as training. However, it is unclear how performance develops over training; how does the relative control of the temporal and numerical variables emerge and change as subjects gain experience in this task? In addition, the relative distributions of trial types were manipulated during baseline training in Experiments 2 and 2A to facilitate acquisition in this task- is performance able to reach similar levels if these distributions remain constant? The following experiment was conducted to attempt to answer these questions.

4 Chapter 4: Acquisition in the numerical reproduction procedure

4.1 Introduction

The previous experiments have demonstrated that pigeons were able to discriminate and reproduce number in a numerical reproduction procedure, making increasing numbers of responses in the production phase as the number of flashes presented in the sample phase increased. Regression analyses showed significant control by number in both Experiment 2, in which either sample phase duration or flash rate covaried with number, and more strongly in Experiment 2A, when the relationship between temporal and numerical variables was degraded. In Experiment 2, transfer to novel numbers could only be obtained if temporal organization of the sample phase was the same as training and some temporal control by sample phase duration and flash rate was still found in Experiment 2A, suggesting temporal variables still had at least a partial influence on responding.

To date there do not appear to be any published studies that have examined the acquisition of performance in numerical discrimination procedures, so it is unclear how numerical and temporal control develop over training in numerical tasks. For example, does number influence responding at the beginning of training or is extended training required for subjects to begin attending to numerical cues? How does response differentiation develop?

The experiment reported in this chapter was conducted to replicate the results of Experiment 2A and to characterize acquisition of performance in the numerical reproduction procedure. This experiment differs from the Experiment 2A in that distribution of trial types were kept more or less constant during training to investigate whether performance would still reach similar levels. Additionally, temporal variable data were not recorded from the beginning of training in Experiment 2A, and so this experiment also examined changes in temporal and numerical control over responding during acquisition in the numerical reproduction task..

4.2. Method

4.2.1 Subjects

Subjects were 7 homing pigeons, numbered 195-197, and 185-188. All had experimental histories involving choice procedures, but no experience with either counting or timing tasks. Subjects were maintained at approximately 85% of their free-feeding weights by additional feeding, when necessary, after experimental sessions. Water and grit were continuously available in their home cages.

4.2.2 Apparatus

The apparatus used was the same as in Experiment 2A.

4.2.3 Procedure

The procedure was the same as in Experiment 2A, with some exceptions. Subjects 195-197 received 210 sessions of baseline training, whereas subjects 185-188 received 195 sessions of baseline training. The relative frequencies of the baseline trial types were held constant throughout training, with the exception of 16 sessions (sessions 101-139) in 185-188's training where the distributions were adjusted to improve the poor discrimination. For 195-197, trial types were determined pseudorandomly subject to the constraint that out of every 9 trials, there were 2, 3, and 4 trials of 2 2–flash trials , 3 4–flash trials, and 4 6-flash trials. Trial types for 185-188 were determined in the same manner, with the exception that for every 9 trials, there were 3 of each trial type. For sessions 101-139, 4-flash trials were excluded and for every 9 trials there were 4 2-flash trials, and 5 6-flash trials.

After baseline training, transfer test sessions were conducted with novel probe trials. During transfer test sessions, trials with 1, 3, 5 and 7 flashes during the sample phase were

intermixed with regular (2, 4, 6) baseline trials. Each transfer test session consisted of 85 baseline trials (with 19, 28, and 38 of type 2, 4, and 6 respectively) and 20 probe trials (5 each of type 1, 3, 5, and 7), with the identity of each trial determined pseudorandomly. Reinforcement on probe trials was randomly determined; the probability of reinforcement was equal to the probability of obtaining reinforcement on regular trials, averaged over the subject's preceding ten baseline sessions.

Data from the final 10 sessions of baseline training and first 10 sessions of transfer tests were used to assess final performance in the procedure. For pigeon 195, 5 sessions in the middle of the transfer tests were excluded due to non-completion; consequently this pigeon only had 60 sessions of transfer test data for analyses, compared to 196 and 197, who had 65 sessions. To assess acquisition, data were collapsed into 10 session blocks. Average response number, proportion correct and correlations between average response number and numerical and temporal variables were analysed across training and transfer test sessions in blocks of 10 sessions.

4.3 Results

4.3.1 Acquisition

Data from all baseline training sessions were divided into 10 session blocks and analysed to assess changes in responding and performance during acquisition, Data from subjects 185-188 and 195-197 were analysed separately, due to procedural differences in the distribution of trial types.

Correlations between response number and 1) flash number; 2) sample phase duration; and 3) flash rate were calculated for each 10 session block and are shown in Figure 4.1. As can be seen in the first two blocks of training sessions, correlations between response number and the temporal variables were relatively similar to correlations for flash number. However as training progressed, the disparity between correlations with temporal and numerical variables increased,

such that the correlations between flash number and response number were considerably higher than with sample phase duration and flash rate. There were some differences in correlations between groups. For subjects 195-197, response number was clearly more highly correlated with sample phase duration than flash rate, whereas for subjects 185-188 there was less differentiation, though flash rate tended to be more highly correlated. Also, correlations between flash number and response number tended to plateau for 195-197 at about 0.3 after about 80 sessions, while for 185-188 it continued to increase until after 160 sessions, reaching a level of approximately 0.45 at the end of baseline training.

The average number of responses made on each trial type were also calculated and are plotted in Figure 4.2 for both groups. Average response numbers were generally higher for subjects 195-197 than 185-188, although they did reach similar values on the 4-flash and 6-flash trials, approximately 4 and 5 responses respectively, by the end of baseline training. There is evidence of a gradual separation of the number of responses made on different trial types. This is most obvious in the data of subjects 185-188, but also can be seen to a lesser degree in the first 50 sessions of subjects 195-197. There also appears to be a factor that influences overall response output across all trial types, which varies across session blocks, as evidenced by the covariation in average response number across all trial types.

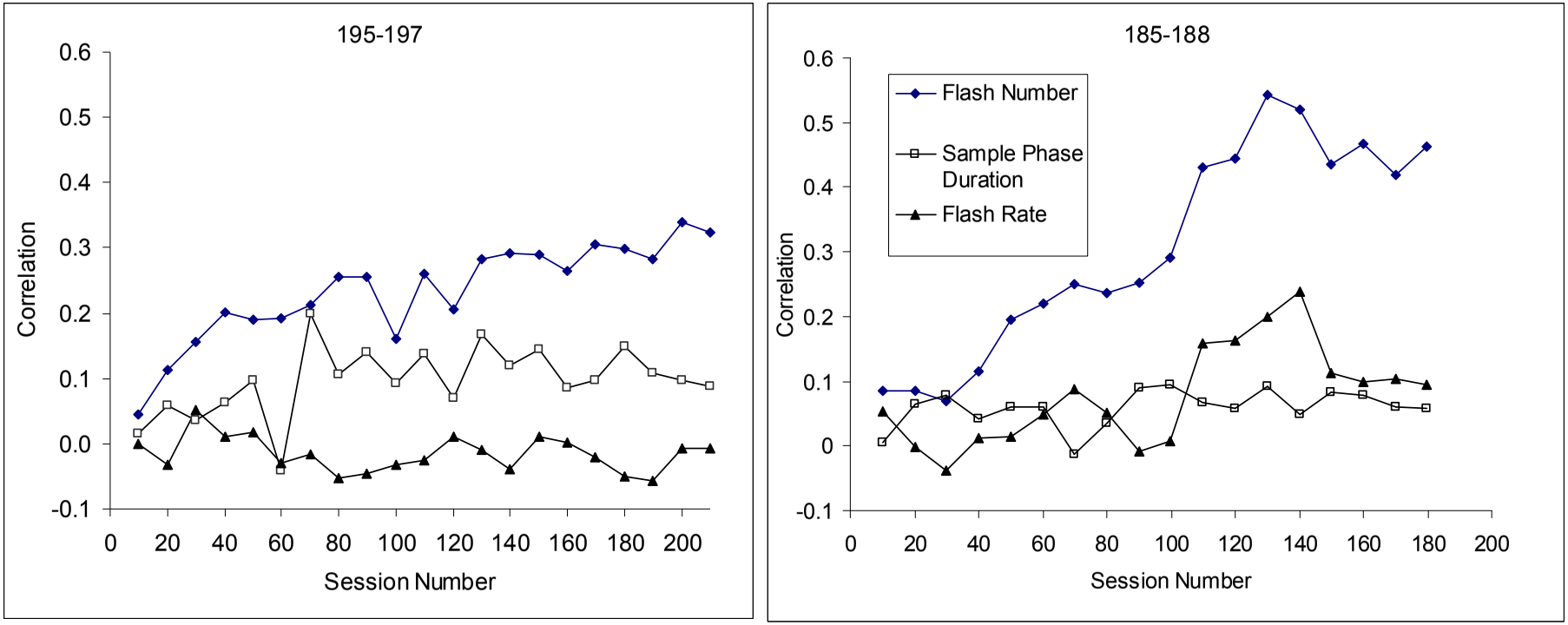

Figure 4.1. Average correlations with response number for subjects 195-197 (left panel), and 185-188 (right panel) across blocks of 10 sessions

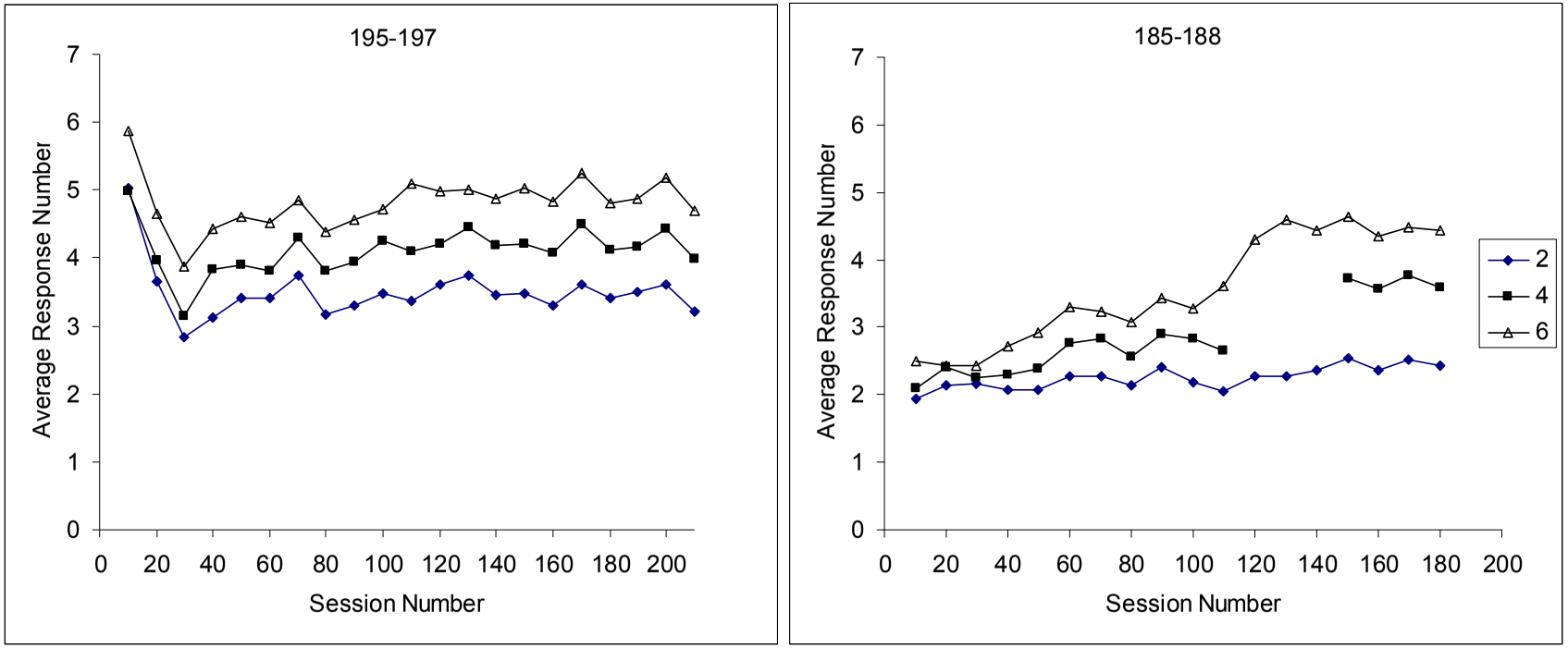

Figure 4.2. Average response number for subjects 195-197 (left panel) and 185-188 (right panel) for 2-, 4-, and 6-flash trials across 10-sessions blocks of baseline training.

The average proportion of correct trials per session for each block were calculated and plotted for all trial types in Figure 4.3. The plots for the different sets of birds differ considerably, and this may be due to the distribution of trials types in baseline training; subjects 195-196 had a ratio of 2:3:4 for 2-flash, 4-flash and 6-flash trials respectively, whereas subjects 185-188 had equal proportions of each trial type (3:3:3). The greater proportion of trial types with larger flash numbers may have skewed overall responding towards larger numbers for subjects 195-197 and resulted in the lower performance on the 2-flash trials. Conversely, 185-188 showed a strong bias towards a smaller number of responses, which resulted in a larger proportion of correct 2-flash trials, than compared to 4- or 6-flash trials. This bias persisted throughout all of baseline training, although performance also improved on the 4- and 6-flash trials for these birds after 4-flash trials were excluded for 30 sessions.

Hierarchical regression analyses were conducted, in the same manner as the preceding experiments to examine the changes in relative control of the temporal and numerical variables over responding at different stages in acquisition. Outliers, defined as sample phase durations > 30s, were excluded[1]. Individual data from the first ten sessions (S1-10), sessions 61-70 and 121-130 were used in these analyses. Control by the temporal variables, flash rate and sample phase duration, was assessed by adding them to a model in which flash number predicted response number. Conversely, numerical control was assessed by entering flash number into a regression model in which flash rate and sample phase duration predicted response number. These results can be seen in Table 4.1.

The performance of the full regression model in predicting response number improved over training; the average multiple R^2 increased from approximately 0 to 10% from the first block of baseline training to sessions 61-70, and in sessions 121-130, the average total variance accounted was approximately 25%. This provides evidence that numerical and temporal control over responding increased with training.

[1] The exclusion of outliers resulted in an average of 7.4%, 5.0% and 2.6% of total trials being excluded for sessions 1-10, 61-70 and 121-130, respectively.

Relative control by the temporal and numerical variables differed, and changed across session blocks. For sessions 1-10, significant beta weights were only obtained for flash rate and flash number for subjects 185 (flash number only) and 186 and 188 (flash rate and flash number). Also, for these three subjects, the unique variance associated with the numerical and temporal variables (R^2 inc.) were small, significant and approximately equal.

By sessions 61-70, beta weights for number were significant and large (> 0.25) for 6 of the 7 subjects, and approached significance for the remaining subjects. No significant beta weights for flash rate were obtained, and sample phase duration was a significant predictor of response number for two subjects. Beta weights for the temporal variables were all considerably smaller than the obtained beta weights for flash number. Furthermore, the increase in R^2 associated with number was significant or approaching significance for all subjects and greater than the increase in R^2 associated with the temporal variables.

Regression results for sessions 121-130 revealed an increase in temporal control by the temporal variables, in particular sample phase duration. The majority of subjects had significant beta weights for sample phase duration in this block of sessions, despite showing little control by this variable in previous sessions. Nevertheless, the beta weights for flash number were still the greatest of the three variables for all subjects; the group average for flash number was approximately 0.50, compared to 0.01 and -0.08 for sample phase duration and flash rate. Once again, the increase in R^2 associated with number was significant for all subjects and greater than the increase in R^2 for the temporal variables.

Together, these results provide evidence for early, superior development of numerical control, relative to temporal control over responding in this task, which emerges by sessions 61-70 and improves with further training.

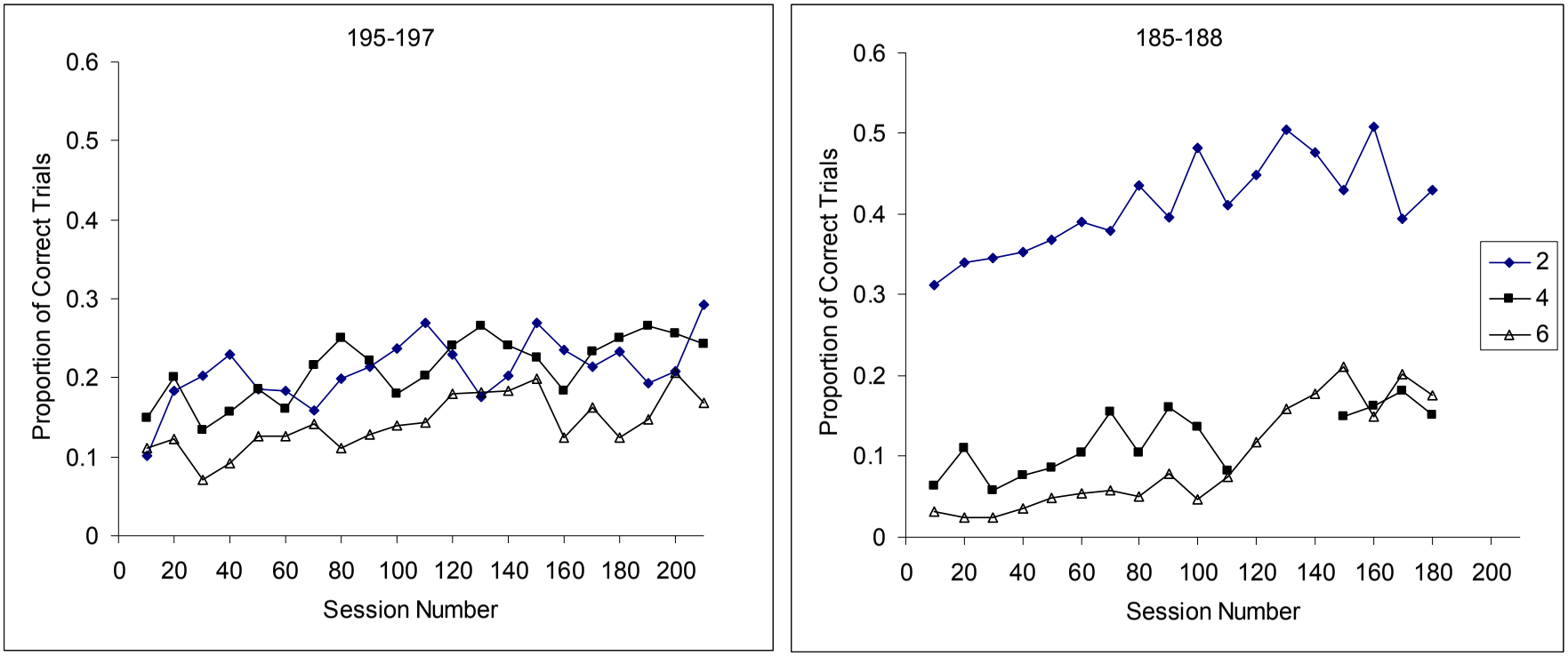

Figure 4.3. Average proportion of correct trials for subjects 195-197 and 185-188 for 2-, 4- and 6-flash trials across 10-session blocks of baseline training.

Table 4.1. Results of hierarchical multiple regression analyses for sessions 1-10, 61-70 and 121-130 of baseline training. Table shows beta weights for sample phase duration, flash rate and flash number and multiple R2 for the full models 1 and 2, and increase in variance accounted for when numerical and temporal variables are added to model 1 and 2, respectively.

Sesssions 1- 10							
Pigeon	**195**	**196**	**197**	**185**	**186**	**187**	**188**
Sample Duration	0.06	0.04, $p = .08$	-0.02	0.07	-0.03	-0.08	-0.08
Flash Rate	0.03	0.00	-0.01	-0.01	-0.13*	-0.01	0.32***
Flash Number	0.03	0.04	0.06	0.15**	0.14**	0.08	-0.19*
Multiple R^2	.00	.00	.00	.03***	.01**	.01	.03***
Number R^2_{inc}	0.00	0.00	0.00	0.001**	0.01**	0.00	0.01*
Temporal R^2_{inc}	0.00	0.00	0.00	0.005, $p = 0.08$	0.01*	0.01	0.04***
Sesssions 61-70							
Sample Duration	-0.01	0.10*	0.07	-0.13*	-0.08	-0.08	0.08, $p = .09$
Flash Rate	-0.02	-0.08, $p = .05$	-0.03	-0.08	-0.20*	-0.08	0.02
Flash Number	0.07, $p = .06$	0.25***	0.37***	0.24***	0.29***	0.32**	0.47***
Multiple R^2	.00	.09**	.15***	.04***	.09***	.06***	.25***
Number R^2_{inc}	0.00, $p = .06$	0.05***	0.09***	0.03***	0.02***	0.01**	0.10***
Temporal R^2_{inc}	0.00	0.02***	0.01**	0.01, $p = .07$	0.05***	0.00	0.00*
Sessions 121-130							
Sample Duration	-0.08, $p = .06$	0.23***	0.16***	0.002	-0.14**	-0.001	-0.09*
Flash Rate	0.00	0.03	0.02	-0.12*	-0.32***	-0.13*	-0.05
Flash Number	0.18***	0.30***	0.35***	0.54***	0.69***	0.64***	0.71***
Multiple R^2	.03***	.15***	.17***	.25***	.27***	.33**	.44***
Number R^2_{inc}	0.03***	0.07***	0.08***	0.12***	0.15***	0.10***	0.19***
Temporal R^2_{inc}	0.01, $p = .05$	0.05***	0.02***	0.01***	0.03***	0.01***	0.00, $p = .07$

Note: * $p<.05$, ** $p<.01$, *** $p<.001$.

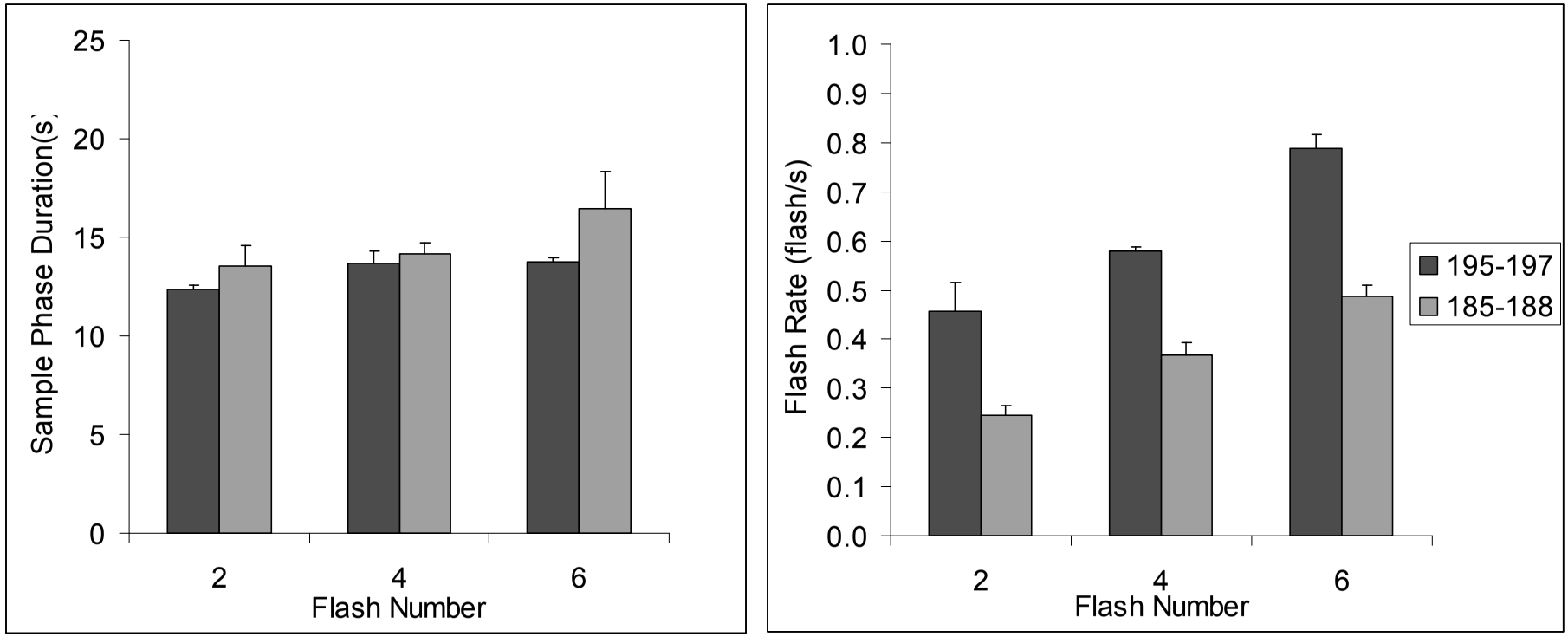

Figure 4.4. Average sample phase duration and flash rate for subjects 185-188 and 195-197 for the last 10 sessions of baseline training. Error bars show $\pm$ 1 S.E

4.3.2 Baseline training

Average sample phase duration and flash rate for each of the trial types were calculated for the two groups. Plots of means are shown in Figure 4.4. As in Experiment 2, the sample phase duration and flash rate were randomised to degrade their relationship with flash number, and thereby increase numerical control over responding, relative to temporal control. Consequently, any covariation between the temporal variables and flash number would be most likely due to the response dependent nature of the stimulus presentation.

Repeated measures ANOVAs were conducted to assess the effect of flash number on sample phase duration and flash rate, as well as any differences between groups. A repeated measures ANOVA found significant effects of flash number on both sample phase duration, $F(2,10) = 5.567$, $p<.05$, and flash rate, $F(2,10) = 54.65$, $p<.001$. Trend analyses showed that the increases in both these variables as flash number increased were linear, $F(1,5) = 18.77$, $p<.01$ for sample phase duration, and $F(1,5) = 62.44$, $p<.001$. A significant effect of group was obtained for flash rate, $F(1,5) = 80.97$, $p<.001$, but not sample phase duration. No significant interactions were found. These results are similar to the results of Experiment 2A.

To examine the relationships between the temporal and numerical variables further, correlations between flash number and sample phase duration and flash rate were calculated for each subject excluding outliers, defined as sample phase durations greater than 30 s, similar to Experiment 2[2].

Correlations between flash number and both temporal variables were relatively small, but still significant for all 8 subjects. The average correlation between flash number and sample phase duration was $r = 0.13$ for subjects 195-197, $r = 0.18$ for subjects 185-188. Correlations between flash number and flash rate were somewhat higher, $r = 0.25$ for subjects 195-197 and $r = 0.47$ for subjects 185-188. As in Experiment 2A, the double-randomisation procedure appeared to be effective in degrading the relationship between the temporal variables and flash

[2] The exclusion of outliers resulted in an average of 3,1% of trials being excluded from analysis.

number, though there was still some covariation present.

In baseline training, significant correlations with response number for both flash number and sample phase duration were obtained for all subjects. Correlations between response number and flash rate were significant for three of seven subjects, 185, 187 and 188 (average r = 0.11). Additionally, correlations between response number and flash number (average r = 0.41) were greater than between response number and both sample phase duration (average r = 0.18) and flash rate (average r = 0.05). This would suggest that, despite existing correlations between the temporal variables and flash number, the correlation between flash number and response number was still greater than those between the temporal variables and response number.

The proportion of correct trials per session were calculated for each subject and the group means are plotted in Figure 4.5. Performance was highest on the two-flash trials, and decreased as flash number increased. Results of a repeated measures ANOVA revealed a significant effect of flash number, $F(2,10) = 28.90$, $p < .001$. No significant effect of group on proportion correct was obtained, however, a significant interaction was found; Tukey HSD tests showed the significantly greater performance of 185-188 on two-flash trials than all other trials was largely responsible for this result, $p < .01$.

Average response numbers and response distributions for each of the three baseline trial types were also calculated and shown in the left and right panels of Figure 4.6, respectively. Average response number increased as a linear function of flash number, and was overall higher for subjects 195-197 than 185-188. Note that the response number scale is compressed with respect to flash number. A repeated measures ANOVA found a significant effect of flash number on average response number, $F(2,10) = 57.06$, $p < .001$, and the linear trend was significant, $F(1,5) = 66.67$, $p < .001$. There was also a significant effect of group, $F(1,5) = 10.69$, $p < .05$, and Tukey HSD tests showed that average response number for subjects 195-197 was significantly higher than for 185-188, $p < .05$. No significant interaction was obtained, indicating that response number increased with flash number similarly for the two groups.

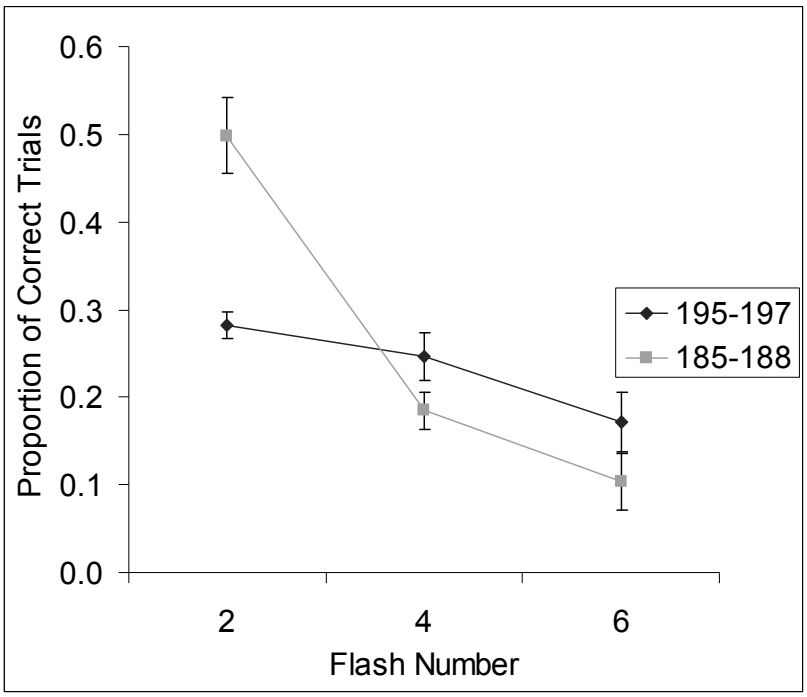

Figure 4.5. Average proportion of correct trials for last 10 sessions of baseline training. Bars show ± 1 S.E.

Average response distributions for the 2-, 4- and 6-flash trials for 195-197 and 185-188 are shown in the top and bottom right panels of Figure 4.6. The peaks (modes) of these distributions increased as flash number increased. However, the shifts in the modes were smaller than the respective changes in flash number. For subjects 195-197, peaks for the 2-, 4- and 6-flash response distributions were located at approximately 3, 4 and 4 and 5, respectively. For subjects 185-188, peaks were obtained at 2, 2 and 4. Another feature of the response distributions is the gradual flattening of the response distributions (increase in variability) as flash number increases; this is more obvious in the response distributions for 185-188.

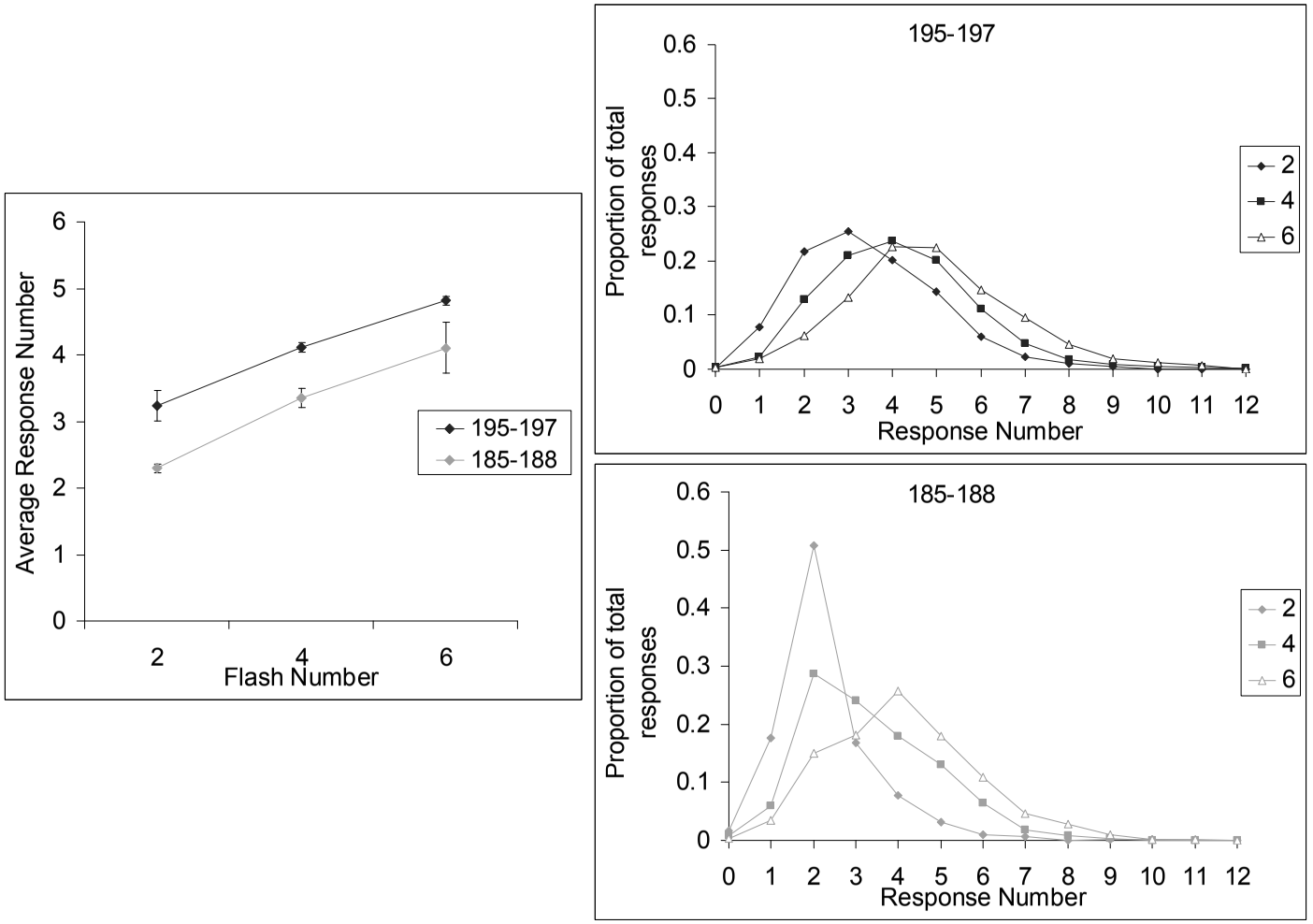

Figure 4.6. Average response number plotted as a function of flash number (left panel) and average response distributions (right panel)for subjects 195-197 and 185-188 in the last 10 sessions of baseline training. Error bars represent ± 1 SE.

4.3.3 Hierarchical regression analyses

Due to the covariation between temporal and numerical variables, hierarchical regression analyses were performed to assess the relative contribution of number and the temporal variables in determining the number of responses made by subjects, i.e. whether number or the temporal variables accounted for a significant amount of variance above and beyond the other. Flash number and the temporal variables (flash rate and sample phase duration) were entered into two different models predicting response number. In the first model, the temporal variables were entered in the first step and flash numbers entered in a second step. In the second model, these steps were reversed. Results of these can be seen in Table 4.1.

Significant control by number was obtained for all subjects in baseline training; flash number accounted for a significant amount of variance after the temporal variables had been added to the regression models, and in some cases, the unique variance associated with flash number was a significant proportion or all of the total variance accounted for by the full model. Additionally, beta weights for flash number were significant and larger than beta weights for both sample phase duration and flash rate for all subjects. Of the two temporal variables, greater control was exerted by flash rate than by sample phase duration; significant beta weights were obtained for flash rate in 4 subjects, whereas significant beta weights were obtained for sample phase duration for only 1 subject. The temporal variables accounted for a significant amount of variance over and above the variance accounted for by flash number for all but 1 subject, however the proportion of unique variance never exceeded 3%.

Table 4.2. Results of hierarchical multiple regression analyses of last 10 sessions of baseline training. Table shows beta weights for sample phase duration, flash rate, flash number and multiple R^2 for the full models 1 and 2, and increase in variance accounted for when numerical and temporal variables are added to model 1 and 2, respectively.

	Baseline						
	Subject						
	195	**196**	**197**	**185**	**186**	**187**	**188**
Sample Phase Duration	-0.04	0.05	0.08*	-0.10	-0.01	0.10	0.05
Flash Rate	-0.14***	-0.02	-0.11**	-0.22***	-0.22**	-0.09	-0.03
Flash Number	0.36***	0.31***	0.43***	0.54***	0.53***	0.48***	0.55***
Multiple R^2	0.12	0.10	0.19	0.19	0.21	0.23	0.30
Number R^2_{inc} (Model 1)	0.11***	0.10***	0.13***	0.12***	0.09***	0.07***	0.16***
Temporal R^2inc (Model 2)	0.01***	0.00	0.03***	0.02***	0.03***	0.03***	0.01*

* $p < 0.05$ **$p < .01$ *** $p < 0.001$

Overall, results show that during acquisition of performance in the numerical reproduction procedure, superior control by flash number over responding, relative to the temporal variables (sample phase duration and flash rate) is developed relatively early on, seen in the correlations and also the differentiation of response number across the different trial types. Differences in acquisition between subjects 195-197 and 185-188 may be due to the distribution of trial types experienced during training; birds 195-197 produced larger response numbers overall, whereas responding for birds 185-188 was biased towards 2 responses.

Analyses of stable baseline training data suggest subjects were able to successfully discriminate the different trial types on the basis of number; average response number increased linearly as flash number increased and response distributions shifted with flash number also. As in Experiment 2A, although sample phase duration and flash rate were randomised procedurally, some covariation was still obtained between the temporal variables and flash number. These enabled the use of hierarchical multiple regression analyses to determine the relative influence of these variables over responding, and revealed significant control by number above and beyond that associated with temporal variables.

4.3.4 Transfer tests

Data from the first 10 sessions of transfer tests were used in the analysis of transfer test performance.

Averages of the temporal variables were calculated and plotted as a function of number of flashes, in the same manner as the baseline analyses. Plots of average sample phase duration and flash rate are shown in the left and right panels of Figure 4.7, respectively. Both appeared to increase as a function of number, which was confirmed by repeated measures ANOVAs: A significant effect of number on sample phase duration, $F(6,30) = 7.35$, $p<.001$, and flash rate, $F(6,30) = 108.26$, $p<.001$, as well as significant linear trends for each, $F(1,5) = 152.94$, $p<.001$, and $F(1,5) = 432.96$, $p<.001$, respectively. As in baseline, a significant effect of group was obtained for flash rate, $F(1,5) = 105.26$, $p<.001$, but not sample phase duration. Tukey HSD tests showed that average flash rate for 195-197 was significantly higher than for 185-188, $p < .001$. Also, a significant interaction between group and flash number was obtained; $F(6,30) = 8.56$, $p < .001$. This appears to be due to the deviation from linearity in flash rate for 195-197, but not 185-188.

Correlations between the numerical and temporal variables were also calculated, again excluding outliers with a cumulative sample phase duration > 30 s[3]. Values were comparable to those in baseline, with the average correlation of flash number with sample phase duration and flash rate equal to 0.19, and 0.40, respectively. All individual correlations between the temporal and numerical variables for each subject were significant.

[3] The exclusion of outliers resulted in an average of 3.2% of trials being excluded from analysis.

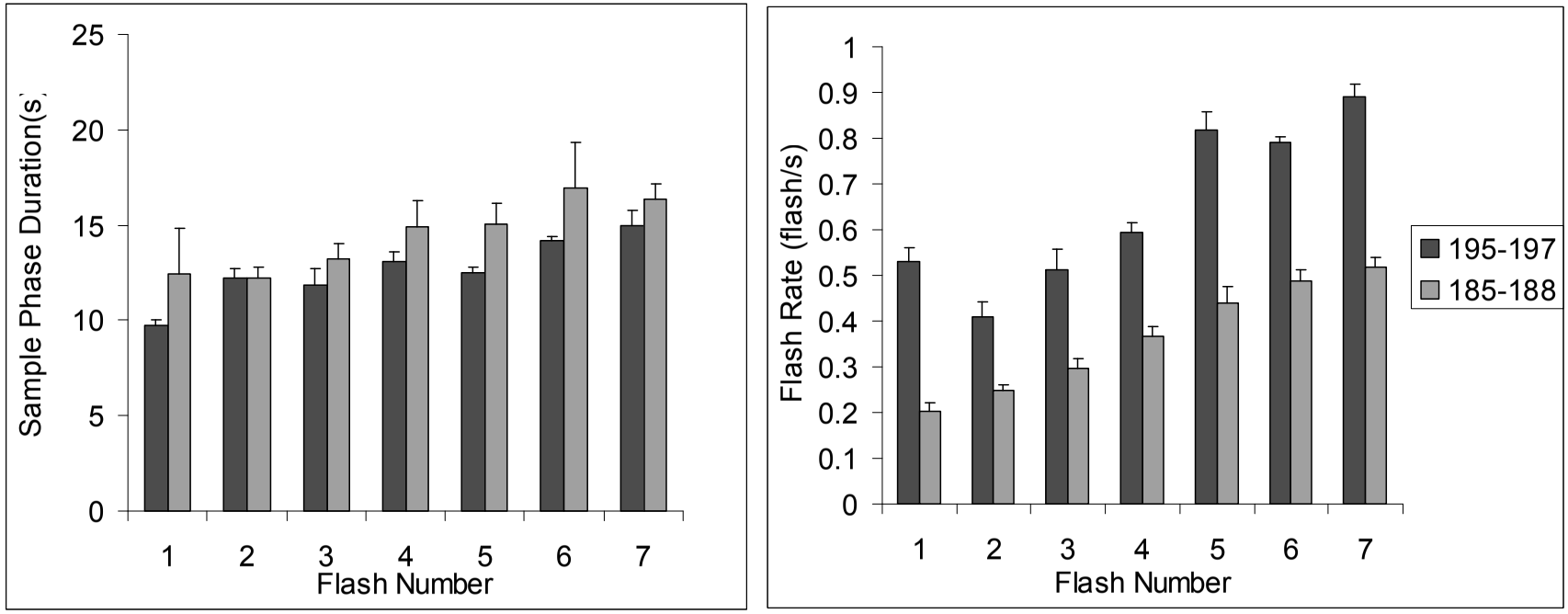

Figure 4.7. Average sample phase duration and flash rate for first 10 sessions of transfer tests.

Transfer test data were similar to those obtained in baseline; correlations between response number and flash number were all significant and largest of all three variables, average $r = 0.41$ and $r = 0.44$ for subjects 195-197 and 185-188 respectively. Correlations between response number and sample phase duration were smaller (but still significant), average $r = 0.24$ and $r = 0.17$ for 195-197 and 185-188 respectively. Correlations with flash rate were significant for subjects 185-188, average $r = 0.17$ and larger than those obtained for subjects 195-197, average $r = -0.003$; of the second group, only subject 197 had a significant correlation between response number and flash rate, $r = -0.08$, $p < .01$.

The average proportion of correct trials per session for both groups are plotted in Figure 4.8. A repeated measures ANOVA was conducted to compare performance in baseline and transfer testing on the 2-, 4- and 6-flash trials. There were no significant effects or interactions involving baseline/transfer testing, suggesting performance on baseline trials remained unchanged. Proportion of correct trials tended to decrease as flash number increased, as in baseline training. A repeated measures ANOVA found a significant effect of flash number, $F(6,30) = 7.99$, $p<.001$, and the linear trend was significant, $F(1,5) = 13.44$, $p < .05$. Thus, accuracy decreased linearly as a function of flash number, although this was less apparent for subjects 195-197. A significant interaction between flash number and group was also obtained, $F(6,30) = 5.32$, $p<.001$, and Tukey HSD tests showed this was due to the significantly better performance of 185-188 on the 1 and 2 trials.

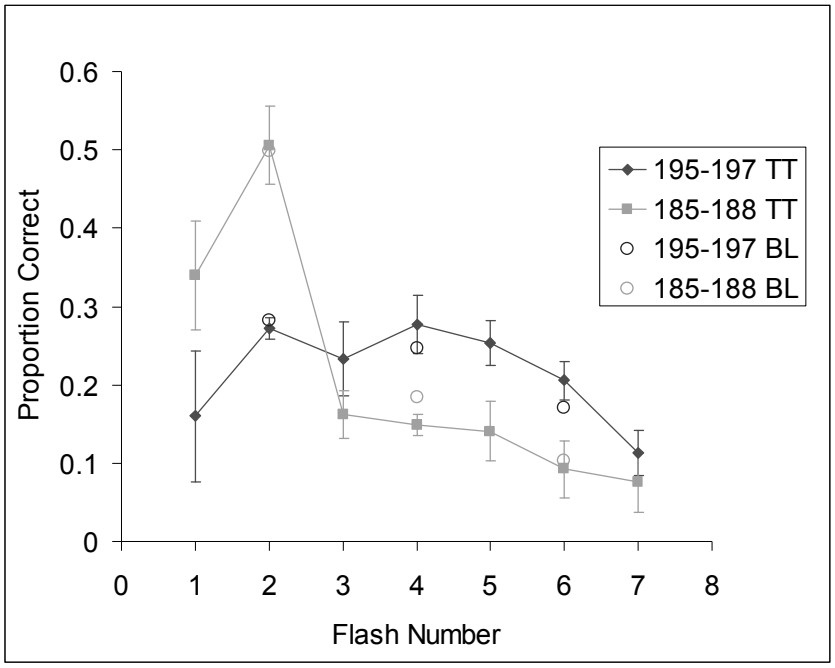

Figure 4.8 Average proportion of correct trials per session for first 10 sessions of transfer tests (filled symbols), and last 10 sessions of baseline (unfilled symbols). Bars show ± 1 S.E.

To assess transfer of performance to novel values outside the training range, planned comparisons were conducted for each of the two groups separately, comparing the average response numbers on 1- and 2-flash trials, as well as 6- and 7-flash trials. For subjects 195-197, no significant differences were found between the lower and higher number values (p= 0.09, and p = 0.49 for 1 vs. 2 and 6 vs. 7, respectively). For subjects 185-188, a significant difference in average response number on the 1- and 2-flash trials was found, $F(1,3) = 61.29$, $p<.01$, though no significant difference was found for the 6- and 7-flash trials. The failure to obtain significant differences in average response number on the 6- and 7-flash trials may be due to the increase in variability in responding corresponding to the increase in flash number.

Response distributions for transfer test data followed a similar pattern to baseline. The gradual flattening and increase in variability in distributions was clearly present, although the shifts in mode location were not so evident; only the response distributions for trials with flash numbers greater than 4 exhibited a mode that was not located at 2.

Hierarchical regression analyses were conducted to assess the relative control of the

temporal and numerical variables in predicting response number in the first 10 sessions of transfer tests. Results were similar to baseline and can be seen in Table 4.3; beta weights for flash number were significant and larger than those for either temporal variable, and in 5 out of 7 cases was the only significant predictor of response number. Significant beta weights were obtained for sample phase duration for 196, and both sample phase duration and flash rate for 197 and 188, suggesting there may have been a greater reliance on temporal cues. For all subjects, number accounted for a significant amount of unique variance in the full model, and in all cases this was at least half of the total variance. The variance uniquely attributable to the temporal variables was, once again, small, and less than 0.02 for over half of the subjects. Thus regression results demonstrate that the significant control by number, above and beyond the temporal variables, was maintained in the transfer tests.

Table 4.3. Results of hierarchical regression analyses of transfer data.

	Transfer						
	Subject						
	195	**196**	**197**	**185**	**186**	**187**	**188**
Sample Phase Duration	0.05	0.20***	0.14***	0.08	-0.06	0.09	-0.09*
Flash Rate	-0.06	-0.03	-0.09**	-0.02	-0.11	-0.07	-0.11*
Flash Number	0.42**	0.39***	0.40***	0.32***	0.54***	0.42***	0.64***
Multiple R^2	0.18***	0.21***	0.20***	0.12***	0.22***	0.19***	0.33
Number R^2_{inc}	0.13***	0.12***	0.14***	0.04***	0.08***	0.05***	0.16***
Temporal R^2_{inc}	0.01**	0.05***	0.04***	0.01**	0.00	0.02***	0.00

* $p < 0.05$ **$p < .01$ *** $p < 0.001$

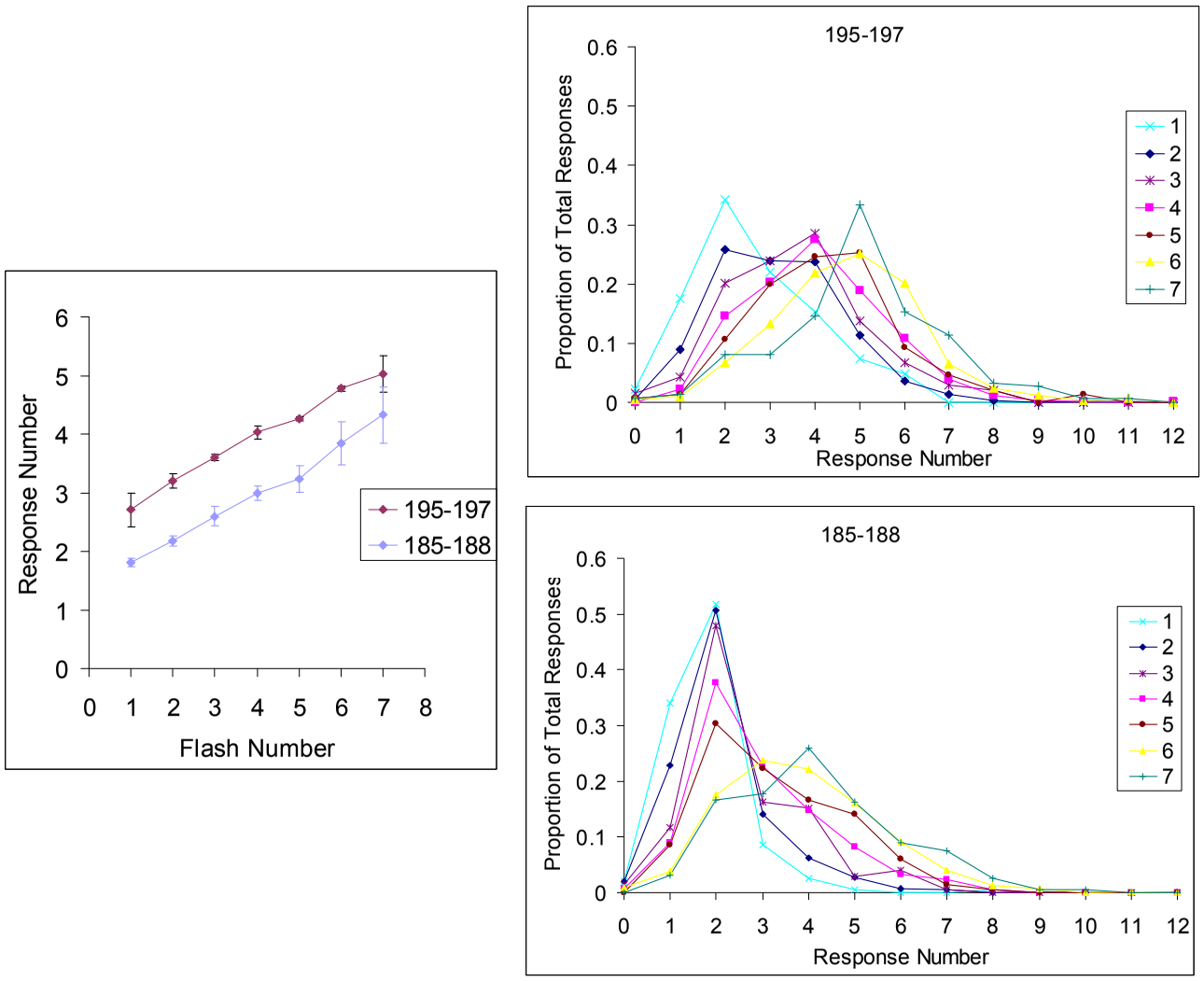

Figure 4.9. Average response number and response distributions for first 10 sessions of transfer tests.

4.4. Discussion

This experiment had two main aims: to replicate the results of Experiment 2A and to investigate the development of performance in the numerical reproduction procedure. The findings of the previous experiment were mostly replicated, although there were some notable differences in terms of performance.

Generally, performances of both groups of subjects were consistent with the previous experiment using the numerical reproduction procedure with randomised temporal variables. Regression analyses showed significant control by number over responding had developed, suggesting that the randomisation of flash rate and sample phase duration was sufficient to decrease their influence over responding. Interestingly, unlike in Experiment 2A, flash rate appeared to have a larger influence on response number than sample phase duration during baseline training, although this did not seem to persist through to transfer testing. This difference is most likely due to the greater covariation between flash rate and flash number during baseline training than in Experiment 2A, and supports the notion that subjects will use the most reliable cues to determine responding.

Response number increased linearly with flash number, and positive transfer was obtained to novel numbers, although there was limited evidence for transfer to numbers outside of the training range. A significant difference between average response number was only found for the 1- and 2-flash for subjects 185-188. This may be due to the increased variability in responding for higher flash numbers.

Response distributions shifted with changes in flash number, with the modal number of responses made increasing as flash number increased. Response distributions also tended to flatten with flash number, indicating that response variability tended to increase with flash number.

4.4.1 Acquisition performance

Analyses of performance throughout baseline training were informative, and suggested that subjects began attending to numerical cues relatively early on in training. Correlations between flash number and response number were generally larger than correlations between the temporal variables and response number from the first block of training for both groups of subjects, and this difference increased as subjects developed greater experience in the procedure, more so for subjects 185-188 than 195-197.

Analyses of average response number confirmed the early development of numerical control; differentiation between response numbers on the 2-, 4- and 6-flash trials began to emerge after approximately 40 to 50 sessions. For subjects 195-197, there was not much change in average response number after this point, suggesting asymptotic performance was reached early on. A similar pattern was found for 185-188 initially, but when 4-flash trials were excluded to improve performance, the average number of responses on 6-flash trials increased considerably. This also appeared to influence responding on the 4-flash trials when they were reintroduced, with average response number approaching 4, as opposed to 3. Interestingly, responding on the 2 trials was not affected at all, suggesting subjects clearly differentiated between responding on at least the 2-flash and higher flash number trials.

One interesting finding was the changes in overall number of responses made across trial types within the session. Within both groups, there appeared to be a general tendency to make more or less responses for all trial types which varied between blocks of ten sessions; average response number for the 2-, 4- and 6-flash trials would shift vertically in unison across different trial blocks. It is unclear what is responsible for this pattern; the mode-control model and neural network models of counting do not appear to be able to explain this. However, the Prototype Response Class (PRC) model proposed by Tan et al. (2007) includes a parameter that is able to account for these changes in responding (see the General Discussion for full description of the model). In the PRC model,

there are two parameters that are used in the calculation of the response functions; λ and δ. The parameter λ is essentially a measure of sensitivity in the difference in the number of responses generated and the response number associated with the prototype and affects the separation of the response distributions for the 2-, 4- and 6-flash trials. The parameter δ determines the overall probability of stopping responding on any given trial. Thus, the global shifts in overall response number can be predicted by the PRC model by increasing or decreasing δ across different blocks, while holding λ more or less constant.

Overall response number was also greater in subjects 195-197 than 185-188; the latter group appeared to have a very strong, persistent response bias towards smaller response numbers, which was reflected in their accuracy (see Figure 4.3). Although this may have resulted from individual idiosyncracies, the most likely explanation is the differences in trial-type distributions experienced in the two groups; the ratio of 2-, 4- and 6-flash trials was 2:3:4 for 195-197, whereas for 185-188 the trial types were distributed equally (3:3:3). The greater exposure to the 2-flash trials for birds 185-188 may have resulted in an overall bias towards making a smaller number of responses compared to 195-197. This small-response bias was partially corrected by manipulating the response distributions halfway through training- by improving the discrimination of 2-flashes from 6-flashes, responding on the 4- and 6-flash trials improved.

Proportion correct reflected the patterns in average response number. For subjects 195-197, there was no obvious difference in accuracy across trial types, although performance was worst on the 6-flash trials. For 185-188, performance on the 2-flash trials was considerably better than the 4- and 6-flash trials throughout baseline training. Proportion correct on the 6-flash trials was close to zero, but increased, once 4-flash trials were removed, to levels similar to performance on 4-flash trials (approximately 0.2).

In conclusion, this experiment successfully replicated the main features of Experiment 2A: 1) Response number increased linearly as a function of flash number; 2) Significant control by number, above and beyond the temporal variables flash rate and sample phase duration, was

obtained over responding; 3) Subjects transferred responding to novel values within the training range, and to a limited extent, outside of the training range. Acquisition was characterised by stronger correlations between response number and flash number than the temporal variables flash rate and sample phase duration, and early differentiation and maintenance of response number for the three different trial types. Responding also seemed to depend on the relative distributions of trial types during training.

5 Chapter 5: Effects of retention interval on performance

5.1 Notes on Experiment 4

Experiment 4 was published as a short communication in *Behavioural Processes*, *78* (2) in 2008, in a paper titled "Effects of retention interval on performance in a numerical reproduction task". It is the first investigation into the effects of retention interval (RI) delay on discrimination in a response production procedure; four pigeons were trained in the numerical reproduction procedure with a 2s RI, and RIs were then increased to 8s and decreased to 0.5s, or vice versa to assess effects on performance. Of particular interest was whether a significant "produce-small" or "produce-large" effect would be produced when RIs were increased or decreased, respectively, analogous to the findings obtained in discrimination procedures.

5.2 Introduction

Experiments 2-3 have shown that pigeons are able to discriminate number in a numerical reproduction task. In this procedure, stimuli were presented over an extended period of time, and a delay separated the stimulus and response phases. Thus, in order to respond correctly, subjects were required to remember the number of flashes seen during the sample phase during the sequence as well as the retention interval preceding the response phase. In order to determine whether response production in this task was affected by memory decay, retention interval was varied in this experiment.

The delayed-matching-to-sample (DMTS) task is often used to assess counting and timing behaviour, particularly how they might interact with memorial processes. A typical trial in this procedure consists of the presentation of sample stimulus (e.g. duration of houselight illumination, or a fixed-ratio requirement on a lighted response key), followed by a retention interval (RI), and then the presentation of two "choice" stimuli (e.g. red and green keylights). One choice response is

reinforced if the sample stimulus was short/small, and the other response is reinforced if the sample stimulus was long/large.

An interesting and robust finding has emerged from research with timing DMTS procedures. For pigeons and humans, increasing delays between the sample and choice phases, relative to baseline, results in a "choose-short" effect that lasts several sessions Specifically, subjects exhibit a bias towards the shorter choice alternative, such that accuracy on the short-samples remains high as RI delays increase but accuracy on the long-samples drops significantly below chance (e.g., Gaitan & Wixted, 2000; Lieving, Lane Cherek & Tcheremissine, 2006; Spetch & Rusak, 1989; Spetch & Wilkie, 1983; Wearden, Parry & Stamp, 2002).

A "choose-long" effect has also been found in timing procedures, although this effect appears to be less extreme than its counterpart. Accuracy remains high on long-samples, but decreases disproportionately for short-samples when RIs are reduced relative to baseline delays (Roberts, Macuda & Brodbeck, 1995, Santi & Hope, 2001; Spetch & Rusak, 1989, 1992; Zentall, Klein, & Singer, 2004).

Furthermore, analogous "choose-small" and "choose-large" effects have also been found in numerical procedures (Fetterman & MacEwen, 1989, Roberts et al., 1995, Santi and Hope, 2001, Santi, Lellwitz & Gagne, 2006). These findings suggest that temporal and numerical aspects of stimuli may share a common representation or response mechanism, and the majority of explanations for the choose-short effect have also been applied to similar results obtained with numerical discriminations.

A number of explanations have been offered to account for these effects. One of the earlier theories was the *subjective shortening* hypothesis (Spetch & Wilkie, 1983), which proposed that memory of the sample duration decays during the RI, in such a way that long durations are increasingly perceived as short durations as the delay increases. This explanation has also been applied to the choose-small effect found with numerical discriminations (Fetterman & MacEwen, 1989), but does not apply well to choose-large effects (Santi & Hope, 2001).

Spetch and Rusak (1989, 1992) later developed the relative-duration hypothesis, which suggested that subject's decisions were dependent on the "temporal background"; sample durations were judged relative to the trial duration, the summation all delays, including the RIs and inter-trial-intervals (ITIs). Therefore, total duration increases when RIs are lengthened, making the sample duration relatively smaller and resulting in the "choose-short" effect. Conversely, shortening RIs decreases the total duration, and increasing the relative length of the sample duration, leading to a "choose-long" effect. However, the relative duration hypothesis overestimates the "choose-long" effect, predicting an effect of equal magnitude to the "choose-short" effect (Spetch & Rusak, 1992). Results would also suggest that RI manipulations result in larger biases than ITI manipulations, most likely due to foreshortening effects and because RIs are closer to the choice time (Spetch & Rusak, 1992). The subjective shortening and relative-duration explanations are not mutually exclusive; in fact research findings support their interaction. The qualitatively similar and additive, but quantitatively dissimilar effects of manipulating ITIs and RIs suggests that subjective shortening can occur in a relative temporal context (Spetch & Rusak, 1992).

More recently, Zentall (1999, 2007) proposed that the choose-short effect can be accounted for by instructional ambiguity. Specifically, increasing the RI makes this interval more similar to the inter-trial interval, such that when making a choice response, subjects respond as if the RI was the ITI and consequently respond as if a 0s sample duration had just been presented. Indeed, when the RI and ITI are made more discriminable, the choose-short effect is decreased or eliminated (Spetch & Rusak, 1992; Zentall et al, 2004). If the instructional failure involves lack of differentiation between RI and sample delays, a similar failure might account for the choose-small effects found in numerical discrimination tasks (Fetterman & MacEwen, 1989), although again it is less successful at explaining choose-large effects (Santi et al., 2006). It is worth noting that both the relative duration and instructional ambiguity hypotheses are similar in that the temporal context of the discrimination is important, and both predict no choose-short effect when the RI and ITIs are differentiated. The former explanation attributes this to the RI and ITI no longer being part of the

same "background stream" (Spetch & Rusak, 1992, p57), whereas the latter suggests stimulus generalization between the two intervals is reduced.

Despite the reliability of the choose-short effect in discrimination studies that require a choice response, whether an analogous finding can be obtained in numerical tasks that involve the *production* of number is yet to be examined. Previous experiments (Experiments 2, 2A and 3) using the numerical reproduction procedure have demonstrated significant control by flash number over production-phase responding, above and beyond that exerted by temporal variables. Specifically, the number of production-phase responses increases, approximately linearly, with the number of flashes in the prior sample phase. However, some control over responding by temporal variables was also found, suggesting subjects were representing and responding on the basis of number and time in this procedure.

It is unclear how changes in RI delay between trial phases will affect performance in the numerical reproduction task. The assumption of a common representation of time and number implies that responding in a numerical task is based on a continuous, analogical scale of magnitude and predicts a "produce-small" effect in the numerical reproduction procedure, consistent with previous research, when the RI is increased. However, if responding is based on a categorical representation of number and determined by similarity between the current trial and the prototypes then RI manipulations may disrupt this discrimination, resulting in reduced or no differentiation between the different trial types (see Tan et al. 2007 for an example of a category-learning model). This would result in an increase in variability and a shallower or flat response number function; response number should increase on 2-flash trials, and decrease on 6-flash trials. The greater variability afforded by the less-restricted response variable in the reproduction procedure allows finer assessment of RI-dependent changes in response characteristics, such as average response number and response variation. This may provide further information about the nature of these effects that cannot readily be discerned from procedures using categorical choice responses.

5.3 Method

5.3.1 Subjects

Subjects were four homing pigeons, numbered 191-194. Subjects had extended training in the numerical reproduction procedure before exposure to the RI manipulations[1]. Subjects were maintained at approximately 85% of their free-feeding weights by additional feeding, when necessary, after experimental sessions. Water and grit were continuously available in their home cages.

5.3.2 Apparatus

The apparatus used in this experiment was the same as in Experiment 2A.

5.3.3 Procedure

Procedure was identical to baseline training conditions as Experiment 2A. The proportions of the 2-flash, 4-flash and 6-flash trials were equal and held constant throughout the experiment. The main difference in procedure was the manipulation of RI delays. The effect of delay on numerical reproduction performance was assessed by changing RI to 8s or 0.5s, from a baseline RI of 2s[1]. Manipulations were counterbalanced. Birds 191 and 192 experienced 10 sessions with a 0.5s RI, followed by another 10 sessions with an 8s RI before returning to baseline. Birds 193 and 194 experienced the opposite. Data from the first 10 sessions of all conditions were used for analysis.

5.3.4. Results

Separate one-way repeated measures ANOVAs for each RI condition were used to test for

[1] Pigeon 192 had experienced 472 sessions of training with a 2-s RI, 193 had experienced 483 sessions, and two of the birds, 191 and 194 had experienced 476 sessions of training in the procedure with a 2-s RI.

possible relationships between flash number and the temporal variables. No significant effect of flash number on cumulative sample phase duration was found, although the effect approached significance in the 0.5s and 8s RI conditions, $p = 0.07$ for both. For flash rate, a significant effect of flash number was found in the 2s RI condition, $F(2,6) = 17.72$, $p < .005$, but not the 0.5s and 8s RI conditions. Correlations between flash number and the two temporal variables were also calculated for each of the RI conditions. Averaged across birds, the correlations with sample phase duration were 0.05, 0.05 and 0.08 for the 2-s, 0.5s and 8s RI conditions respectively. The correlations with flash rate were 0.33, 0.48 and 0.40, for the 2s, 0.5s and 8s RI conditions, respectively. Together, these results suggest that the double randomisation procedure was effective in degrading the relationship between flash number and the temporal variables, duration and flash rate.

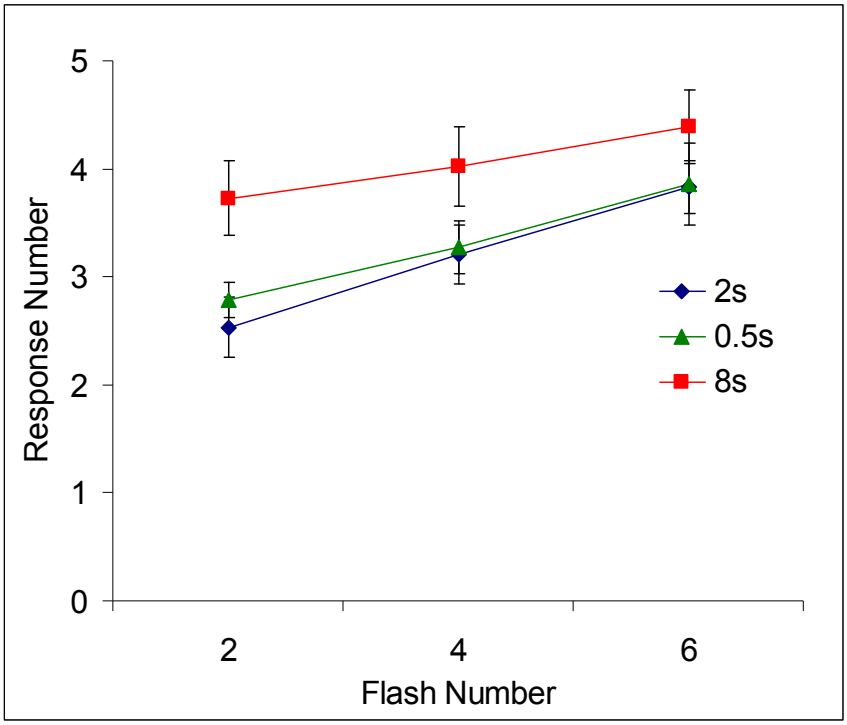

Figure 5.1. Average response number plotted as a function of flash number and retention interval.

Figure 5.1 shows the average response numbers for the 2-flash, 4-flash and 6-flash trials as a function of retention interval. Overall, average response number increased as a function of flash number, and increased relative to baseline (RI = 2s) when RIs were 8s and 0.5s. A two-way

repeated measures ANOVA found a significant effect of number, $F(2,6) = 73.17$, $p<.001$, and a significant linear trend, $F(1,3) = 81.09$, $p<.002$. No significant effect of RI was found, but there was a significant interaction, $F(4,12) = 4.98$, $p<.05$. Tukey post-hoc tests found significant differences in average response number between the 2-s and 8-s RI conditions for the 2-flash trials, $p<.001$, 4-flash trials, $p<.001$ and 6-flash trials, $p<.005$, but no significant differences between average response number for the three trial types in the 2-s and 0.5-s RIs conditions, $p = 0.33$, $p = 1.0$, and $p = 1.0$, respectively. These results suggest both RI and flash number influenced average response number, creating a significant "produce-large" effect that was largest for the 2-flash trials when the RI was changed to 8s.

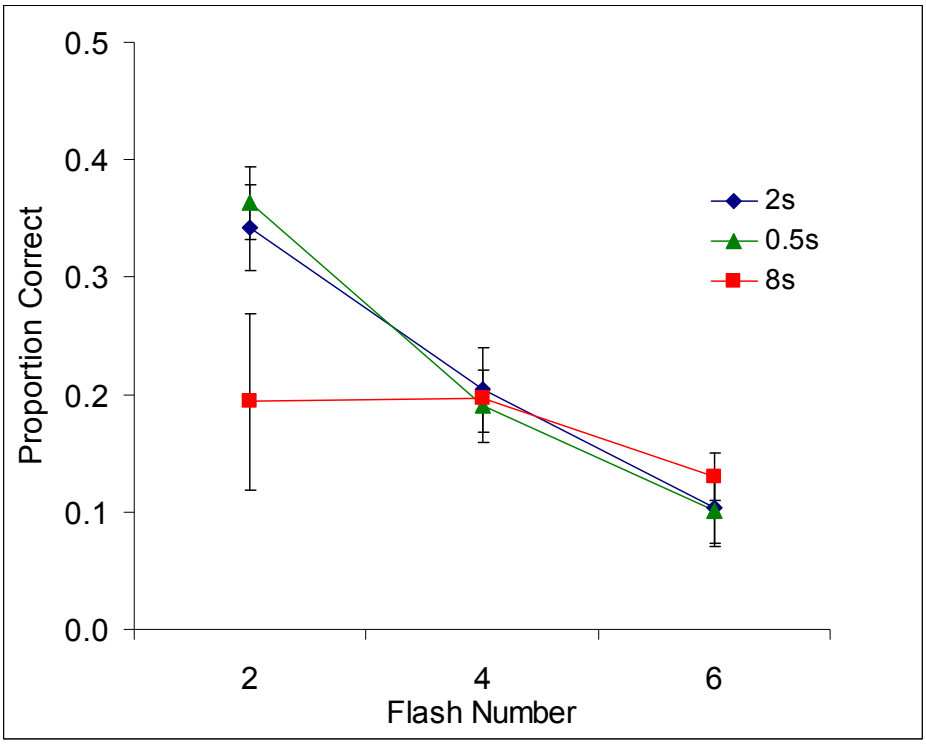

Figure 5.2. Proportion of correct trials per session for 2-flash, 4-flash and 6-flash trials plotted as a function of retention interval delay.

Accuracy, measured as the average proportion of correct trials is shown as a function of retention interval for each flash number in Figure 5.2. A repeated measures ANOVA showed that the proportion of correct trials differed significantly as a function of both flash number and RI,

$F(2,6) = 22.72$, $p<.005$ and $F(2,6) = 5.94$, $p<.05$, respectively. Significant linear trends were also obtained for flash number, $F(1,3) = 35.75$, $p<.01$, and RI, $F(1,3) = 15.36$, $p<.05$. Tukey post-hoc tests showed overall differences in proportion correct between 2-s and the 8-s RI, and 2s and 0.5s RI approached significance, $p = .05$ and $p = .06$, respectively. Although proportion correct decreased substantially as RI increased for 2-flash trials, but not 4- or 6-flash trials, the interaction between RI and flash number was not significant, $F(4,12) = 1.75$, $p = 0.20$. Generally, accuracy tended to decrease as flash number increased. The significant main effect of RI may be attributed to the decrease in accuracy on the 2-flash trials at the 8-s RI, but the relatively large variation in accuracy in this condition seems to be responsible for the non-significant interaction.

Response distributions for each of the trial types for each RI condition were plotted; these are in the upper right and lower panels of Figure 5.3. Response distributions for the 2-, 4-, and 6-flash trials became increasingly more similar as RI differed increasingly from baseline conditions; the modes of the distributions converged towards 3 and increasing overlap between individual trial-type distributions can be seen.

Effect sizes for each pair of trial types were calculated to analyse response discrimination. Effect size was calculated as the difference between the mean response numbers, divided by the pooled standard deviation. These are shown in the upper left panel of Figure 5.3. Discrimination decreased as retention interval increased, and was greatest between the 2- and 6-flash trials, intermediate for the 2- and 4-flash trials, and lowest for the 4- and 6-flash trials. Results of a two-way repeated measures ANOVA revealed a significant main effect of RI, $F(2,6) = 11.99$, $p<.01$, and trial type pair, $F(2,6)=28.16$, $p<.001$, on discrimination. Trend analyses revealed significant linear trends for both RI and trial-type pair, $F(1,3) =18.40$, $p<.05$ and $F(1,3) = 46.00$, $p<.01$. The interaction approached significance, $p = .06$. Tukey post-hoc tests revealed a significant difference in effect size between the 2s RI condition and the 8s RI condition for all three trial type pairs, $p<.001$, but no significant difference between the 2s and 0.5s RI conditions, although it did approach significance for the 2- and 6-flash trials, p =.06.

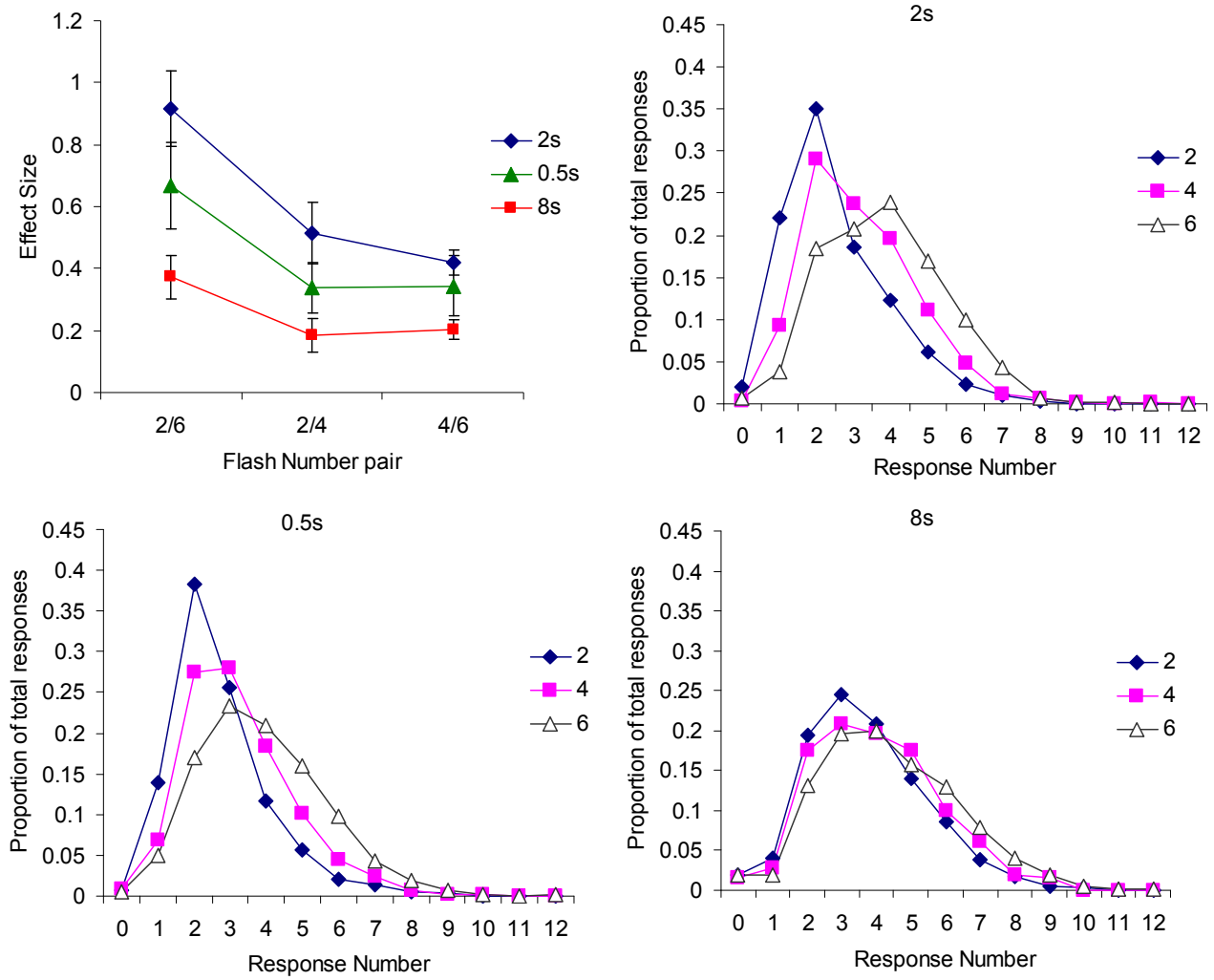

Figure 5.3. Upper left panel shows effect size plotted as a function of retention interval for the three pairs of trial types- 2 and 6-flash trials, 2 and 4-flash trials, and 4 and 6-flash trials. Upper left and lower left and right panels show response distributions for 2-s RI, and 0.5-s and 8-s RI respectively.

5.4 Discussion

The major goal of the present study was to determine the effect of retention interval on performance in a numerical reproduction procedure; in particular, whether a "produce-small" effect would be obtained with increased RIs. Interestingly, an opposite "produce-large" effect was obtained when RI delays were either decreased or increased; however, this effect was dependent on both number and RI. An increase in RI significantly increased average response numbers on 2-flash trials for but not 4- or 6-flash trials, and this effect increased with the difference in RI from baseline; changes in response number, average proportion correct and effect size were greater for the 8s RI, than the 0.5s RI.

The "produce-large" effect contrasts with the usual research finding of a bias towards the smaller number choice when RIs are increased in both numerical and temporal procedures (Fetterman & MacEwen, 1989; Roberts et al. 1995; Santi & Hope, 2001). At present a complete explanation for the "choose-small/large" effects is yet to be offered; current accounts only seem to be able to explain subjects' tendencies to produce large numbers of responses when the RI was shortened to 0.5 s, but not 8 s.

Were results consistent with a categorical representation of number, proposed by Tan et al. (2007)? It was predicted that changes to RI would degrade numerical control over responding, increasing variability as well as increasing response number on the 2-flash trials and decreasing response number on the 6-flash trials, creating a shallower function. Our results partially support these predictions. Response number increased for *all* trial types; average response number for the 2-, 4- and 6-flash trials approached 4, the average of the three trial types. However, a significant interaction suggested a change in slope when RI was manipulated; specifically a larger change in response number was found for the 2-flash trials than the 4- or 6-flash trials. Other analyses showed that discrimination between trial types decreased as RI differed increasingly from baseline, shown in the decrease in effect sizes, and the increase in overlap in the response distributions. Accordingly, the "produce-large" effect is consistent with a

categorical representation of number and may be understood as the result of deterioration in stimulus control; procedural differences disrupted numerical control over behavior, resulting in an increase in variability and decreased trial-type differentiation.

These results are also consistent with the "direct remembering" view of performance in DMTS (White & Wixted, 1999), which proposes that increasing RI duration can increase the variability and decrease the difference between the distributions of the stimulus effect related to each response type. The larger increase in response number with an 8-s RI than the 0.5-s RI can be understood as generalization of performance from the training delay to novel delays (Sargisson & White, 2001); the larger the difference in RI from baseline, the greater the "produce-large" effect.

It is also possible that the procedural disruption resulted in greater reliance on temporal cues, and if subjects were timing through the RI and using this total duration to determine responding, then an increase in average response number would be expected. Additionally, the differential effect of the change in RI on the 2-, 4- and 6-flash trials can be explained by a response-based ceiling effect. If there was an upper limit on the number of responses subjects could or would make in the production phase before switching to responding on the right "completion" key, then the RI-dependent increase in average response numbers should decrease as flash number, and consequently average response number, increased.

Procedural differences may also have affected results. The numerical reproduction task differs from that used in most studies of RI in two major ways. First and most importantly, this procedure requires the discrimination of absolute, rather than relative numerosity; instead of choosing "more" or "less" subjects must discern the absolute number of flashes presented in the sample phase and key-pecks made in the production phase. The production phase also differs; subjects are required to reproduce the sample number instead of just making a single choice response. The more difficult nature of the reproduction task results in performance and behaviour that differs from relative numerosity discriminations- accuracy is not constant across

sample values, and is significantly lower than in typical DMTS tasks- this may also contribute towards the conflicting findings.

In conclusion, the failure to obtain a produce-small effect and the result of a significant, number-dependent, produce-large effect in the numerical reproduction procedure are not readily accounted for by existing explanations of the original choose-short effect. This finding suggests that responding in the production phase is not susceptible to memorial decay; an increase in the duration of the interval between the sample and production phase did not result in a decrease in response number. An increasing deterioration of stimulus control with different RIs may be responsible for these effects, however further investigation with this procedure, using a greater range of delays and numbers is warranted to confirm this.

6 Chapter 6: Response variability and the representation of number

6.1 Notes on Chapter 6

The following chapter collates variability analyses of data from the numerical reproduction experiments reported in Chapters 3 and 4 (Experiments 2, 2A and 3). The analyses reported in the current chapter parallel the variability analyses conducted in Experiment 1, and these all relate to the same question, the underlying structure of the subjective numerical scale on which responses in this procedure is based. Portions of these chapters, namely the results of Experiments 2 and 2A, have previously been published in the paper "Numerical reproduction in pigeons", authored by myself and my supervisors, Randolph Grace and Anthony McLean, and an earlier Master's student, Shasta Holland in Tan, Grace, Holland & McLean (2007) in *Journal of Experimental Psychology: Animal Behavior Processes, 33*. This chapter also includes unpublished variability data and scaling analyses from the acquisition experiment (Experiment 3).

6.2 Introduction

Experiments 2, 2A and 3 have shown that pigeons are able to discriminate number in a numerical reproduction procedure successfully. Average response number increased linearly as the number of presented flashes increased, and significant control by number, above and beyond control by the temporal cues flash rate and sample phase duration were obtained. Subjects were also able to transfer discriminations to novel values both inside and outside the training range. Additionally, in contrast to previous research, responding in the production phase was not affected by memorial decay when retention intervals were lengthened.

For subjects to respond accurately in the numerical reproduction procedure, some mechanism would have been necessary for the retention of the number of external stimuli experienced during the sample phase and delay preceding the beginning of the production phase,

as well as for the number of stimuli held in memory to be translated into the corresponding number of responses generated during production phase. Assuming that the external stimulus value is converted into a perceived or psychological value, which is then later used to determine the number of responses required for reinforcement, an important question arises; what is the nature of the internal psychological scale that underlies the mapping of stimulus to response number? Alternatively, stated more concisely, how is number represented?

There has been much discussion about the structure and form of the subjective numerical scale used by nonhumans and humans in numerical tasks. Although it is not possible to observe the mental processes involved directly, behavioural measures can be used to test predictions made by models assuming different numerical scales.

Given the strong connection between numerical and temporal processing (Meck & Church, 1983; Roberts & Boisvert, 1998; Roberts & Mitchell 1994), theories and principles of timing and counting often mirror each other, with much overlap. Gibbon (1977) outlined distinguishing characteristics of four possible response processes that may operate in timing procedures, which can also be applied to numerical discriminations. In particular, Gibbon discusses differences in the mean, standard deviation and coefficient of variation as an indicator of different operating processes. The coefficient of variation is a measure of relative response variability, and is calculated as the standard deviation divided by the mean. I will describe these in terms of the discrimination of number, rather than time.

If no counting is occurring, then there should be no change in the mean number of responses, standard deviation, and consequently the coefficient of variation as number increases. The three remaining processes constitute absolute, poisson and scalar counting. In absolute counting, the variability in responding remains constant and it is only the mean that increases as number increases; the linear relationship between the response mean and constant standard deviation results in a negative linear relationship between the coefficient of variation and number. Poisson counting would be generated if responses were based on a count of events

generated by a Poisson process with a constant average rate; both the mean responses and standard deviations increase with number, however there is not an exact mapping between response and stimulus number and standard deviations increase according to the square root of the mean (binomial variability). This results in a similar, albeit less strong, decrease in coefficients of variation as number increases – relative accuracy increases as number increases. In scalar counting, response distributions for different numbers are scale transforms of the basic unit, such that the mean and standard deviation increase proportionally to number, resulting in constant coefficients of variation.

It is generally believed that both temporal and numerical discriminations are best explained by scalar processing. Scalar variability, seen in the superpositioning of bisection functions when plotted on a relative scale and constant coefficients of variation, is a reliable finding in responding in human nonverbal and nonhuman numerical discrimination tasks (Humans: Cordes, Gelman, Gallistel & Whalen, 2001; Whalen, Gallistel, Gelman 1999; Nonhumans: Beran & Rumbaugh, 2001, Cantlon & Brannon, 2007, Emmerton & Renner, 2006; Jordan & Brannon 2006; Roberts, 2005; Roberts, 2006; Both: Beran, Johnson-Pynn & Ready, 2008; Huntley-Fenner, 2001, Jordan & Brannon, 2006b). Additionally, there are two effects that suggest that response variability increases proportionally to magnitude. One is the distance effect, where discrimination improves as the distance between the two values being compared increases. For instance, greater accuracy and lower response times for a discrimination of 2 vs. 9 than 2 vs. 5 would be consistent with scalar variability. The other is a size or magnitude effect, where, if distance is held constant, discrimination accuracy is reduced and response time increases as numerical magnitude increases; for example, 12 vs. 15 should be a more difficult discrimination than 2 vs. 5 (Brannon, 2006; Brannon & Terrace, 2000, Moyer & Landauer, 1967, Olthof, Iden & Roberts, 1997; Olthof & Roberts, 2000; Roberts, 2005; Rumbaugh, Savage-Rumbaugh, & Hegel, 1987).

Patterns in response variability may reflect how number is represented. Assuming that

the structure of the numerical scale contributes to the variability in responding, two possible numerical scales have been proposed that are consistent with the obtained findings. The first is a logarithmically-spaced scale, such that the distance between numbers becomes increasingly compressed as magnitude increases. If variability is constant across numbers, then greater generalization should occur with higher than lower numbers due to the smaller distance between them. The second is a linearly or arithmetically-spaced scale, with an equal distance separating numbers, and with scalar generalization that increases proportionally to numerical magnitude. This predicts size and distance effects as generalization would increase proportionally with number, resulting in greater confusion between larger numbers.

There is some debate over which scale provides a better description of subjective numerical scales developed in discrimination tasks, with research providing support for both logarithmic and linear number scales. Distinguishing between the two is made all the more difficult as the two scales generally make identical predictions with respect to responding and response variability.

Recent neuroscientific research has obtained evidence supporting a logarithmic numerical scale (Dehaene, 2003; Nieder & Miller, 2003). Nieder and Miller (2003) specifically compared the possibility of logarithmic or linear coding of number, using response measures and neuronal recordings obtained from monkeys doing a delayed match-to-numerosity task. Monkeys were presented with a sample array of pseudorandomly-arranged dots, and after a 1s delay, were required to match to one of two test stimuli, presented successively by releasing a lever. Control stimuli were also used to ensure that subjects were responding on the basis of number, and not alternative cues. The possible numerical representational structures were compared by comparing changes to response distributions when plotted against either a logarithmic or a linear number scale. If numbers were being represented linearly, then, when plotted on a logarithmic scale, response distributions should become asymmetric with slopes becoming steeper as number increased. Conversely, if numbers were represented logarithmically with constant variability,

then, when plotted on a linear scale, distributions should remain symmetric but with slopes becoming steeper as number *decreased.*

Analyses of performance revealed both numerical size and distance effects, with accuracy improving with smaller numbers and increasing distance between the sample and correct match. Gaussian (normal) distributions were fitted to each of the data to test for symmetry, and linear, power and logarithmic functions were tested by fitting data along a linear scale, power function with exponents of 0.5 and 0.33, or logarithmic scale. Response distributions were calculated and were asymmetric when plotted on a linear scale, with shallower slopes for numerosities greater than the sample. When plotted on a logarithmic scale, distributions were more symmetric, suggesting the subjective numerical scale was logarithmic with constant variability. The linear scale provided the worst fit of data, while the fits of the logarithmic scale and power functions with both exponents were approximately equal and was close to 1 (0.98). Additionally, variance of distributions for each numerosity was constant when plotted on a power function scale with 0.33, and logarithmic scale, but increased linearly when plotted on a linear scale. Nieder and Miller concluded that these results provide support for a logarithmic scale..

Similar analyses were conducted with the neurophysiological data for both the sample and retention phases of the trials. Results mirrored those obtained with the behavioural data; size and magnitude effects were obtained, and neural activity filter functions were asymmetric when plotted on a linear scale and more symmetric when plotted on a logarithmic scale. For both the sample and retention phases, goodness of fit was worst for the linear scale, and best for the logarithmic scale. The variance of the neural functions also showed the same pattern as the behavioural data; variance was constant when plotted on a logarithmic scale, but increased with numerosity when plotted on a linear scale (Nieder & Miller, 2003).

Overall, these results suggest that a logarithmic scale provides a better description of the behavioural and neurophysiological data than a linear scale. Nieder and Miller also point out that a compressed numerical representation would increase the possible coding space and

consequently the potential numerical range that can be processed in perception and neurons. Evidence for logarithmic coding was present in both the acquisition and retention phases, with no change in the amount of compression seen in the data seen in the later retention phase. Based on this, Nieder and Miller proposed that in numerical processing, information is encoded logarithmically, such that the information used during the retention period and for production is also based on the compressed scale. Additionally, they hypothesised that the variability in numerical processing is introduced at the encoding stage, consistent with the Dehaene and Changeux' (1993) neural network model, rather than at the memory stage as described in the accumulator model (Gallistel & Gelman, 2000).

Bisection procedures have also been used to investigate the question of numerical representation, but the evidence is more equivocal (see Chapter 2, 2.1.5 for full discussion). Brannon, Wusthoff, Gallistel and Gibbon (2001) adapted the time-left procedure of Gibbon and Church (1981) for number; subjects had to respond on the basis of the difference between two numerical values. In subtraction procedure, subjects had to compare a numerical difference, which varied between trials, with a constant numerical value and choose the smaller value. If the subjective numerical scale is logarithmic, then the size of the difference (or the indifference point) relative to the constant value will be dependent on the ratio between the two numbers, rather than the difference. Thus, a logarithmic scale would predict no change in the subjective difference if the ratio between the difference and the constant numerical value remains constant, regardless of the absolute difference, whereas a linear subjective scale would predict an increase in the subjective difference when the two numerical values being compared increased while maintaining the same ratio. Brannon et al. (2001) proposed that their experiment demonstrated that subjects were able to perform numerical subtraction; subjects showed a reliable preference for the key associated with the smaller keypeck requirement, and indifference points increased and decreased as the absolute difference between the different *S* and *I* value pairs increased or decreased, respectively, with a constant ratio. That is, the indifference point shifted towards larger values when S and I values increased to 6 and 12, and shifted towards smaller

values when S and I values decreased to 3 and 6.

However, Dehaene (2001) argued that Brannon et al.'s (2001) results could also be explained by subjects only representing the first number of flashes (*T*) and learning which *T* value results in the shortest delay to reward. To test this, Dehaene ran some simulations of a neural network model which was capable of learning number-response associations (see Dehaene & Changeux, 1993). The output of Dehaene's (2001) simulator replicated the findings of Brannon et al. (2001); psychometric functions increased systematically as a function of the value *T*, and increasing or decreasing *I* and *S* linearly affected the location of the indifference point. Dehaene noted that a linear increase in the indifference point with *S* was obtained by the simulator with both linear and logarithmic scales, not just a linear scale as posited by Brannon et al. (2001). Another feature of the data noted was the sub-optimal location of the indifference point, which systematically shifted towards numbers smaller than the ideal value when $T = I - S$. This is not easily explained by Brannon's hypothesis, but is a logical consequence of associative learning; the asymmetrical generalisation towards larger numbers is caused by the increasing variability, and corresponding increasing overlap, of the representational distributions as numerical magnitude increased. Brannon et al. stated that the immediate generalization of performance in the first block of transfer trials demonstrated the abstract knowledge subjects had developed about the task, however Dehaene points out that because reinforcement was still available for those trials, this may possibly be attributed to continued learning in these transfer tests. The simulation, when a fast learning-speed was assumed, exhibited a small indifference point shift similar to that obtained by Brannon et al., providing additional support for an associative process rather than actual subtraction. Dehaene's (2001) analysis shows that some caution should be taken in interpreting Brannon et al's results: Although it is possible a linear subjective numerical scale was responsible for the obtained findings, the associative learning hypothesis proposed by Dehaene also provides a good explanation of the data.

Evidence for a logarithmic, rather than linear scale was obtained in a bisection procedure by Roberts (2005), as discussed in Chapter 2. Roberts developed an associative model that

generated different predictions assuming either a linear scale with scalar generalization, or a logarithmic scale with constant generalization. Both the logarithmic and linear functions showed a pattern of weakest performance near the arithmetic mean, with increasing accuracy as the distance between the sample number and arithmetic mean increased. Both models predicted higher accuracy for larger numbers than lower numbers around the midpoint, but only the logarithmic function showed the asymmetry at the extreme values found in the data. The linear function actually predicted the opposite pattern (i.e., greater performance with the large extreme values than the small extreme values). Based on his findings, Roberts concluded that an associative model assuming a logarithmic numerical scale, with constant generalization or variability, provided a better fit of the obtained bisection data than a linear numerical scale with scalar generalization.

Whereas the linear and logarithmic numerical scales that exhibit scalar variability are associated with nonverbal numerical discriminations in humans and nonhumans, a different number scale is thought to underlie verbal numerical discriminations in humans. The structure of this number scale is a linear scale with constant variability. This type of scale would correspond to a 'true' understanding of number, i.e.knowledge possessed by a normal adult human proficient in numerical tasks. A scale with this structure allows exact mapping between objective and subjective number and consequently, instead of scalar variability, would predict that variability does not increase proportionally with magnitude. Responding based on such a scale should have an error rate that remains approximately constant as number increases. Research investigating adult human verbal and nonverbal production of number has shown that verbal enumeration processes exhibited binomial rather than scalar variability, where variability increases proportionally to the square root of the average magnitude (Cordes, Gelman, Gallistel & Whalen, 2001). Thus, with this process and number scale, *relative* error rate should decrease as number increases, whereas with scalar variability, relative error rate should remain constant as number increases. Recall that this is consistent with a Poisson counting process described by

Gibbon (1977).

This pattern in responding also appears to be occasionally found in nonhuman animals, mainly limited to studies involving variations of Mechner's (1958) Fixed Consecutive Number/Fixed-ratio (FCN/FR) procedure. Hobson and Newman (1981) reviewed and examined performance with ratio schedules, noting that in ratio discrimination and counting/production procedures, but not timing procedures, relative discriminability (measured as the Weber fraction, $\Delta I/I$) improved as ratio number increased (discrimination: e.g. Rilling & McDiarmid, 1965; counting: e.g. Mechner 1958). They extended the previous research using FCN procedures, using pigeons instead of rats and larger ratio sizes to allow for better comparison between ratio counting and discrimination procedures. One of their key aims was to determine whether a single process can account for responding in fixed interval and ratio-counting schedules. Pigeons were trained in a procedure identical to Mechner (1958); on half of the trials, the fulfillment of an FR requirement on the center key produced reinforcement, whereas on the other half the first side-key response following a minimum number of responses on the center key had been made produced reinforcement. Subjects were tested with a range of ratio sizes.

Obtained response distributions for the ratio-counting task showed unimodal response distributions with run lengths just exceeding the minimum requirements for reinforcement. However, of greatest interest were the variability analyses. Coefficients of variation were calculated and plotted with data from previous studies (Laties, 1972; Mechner, 1958; Platt & Senkowski, 1970) and showed a clear decreasing pattern from FR 3-50, although slopes did begin to level off with larger ratios.

However, because Hobson and Newman (1981) used a mixed FCN/FR schedule, the different reinforcement contingencies, namely the direct reinforcement of center key responses not presenting pure FCN schedules, may have affected responding. Average run length would have been increased with an FR schedule than if center key responses were only reinforced indirectly after right key report responses on a FCN schedule. Response variability may also

have been affected, although it is unclear whether this manipulation would have increased or decreased response variability. This was investigated by Machado and Rodrigues (2007), who modified reinforcement contingencies so that subjects could obtain reinforcement by one of two ways. Subjects could either peck the left key exactly *n* times before responding once on the right key (i.e., a typical FCN schedule), or if they failed to reach or exceeded this requirement, subjects had to continue pecking the left key until at least 16 pecks on the left key had been made. Their variability analyses, as well as their reanalysis of Hobson and Newman's (1981) data, showed that coefficients of variation decreased hyperbolically as *n* increased, approaching asymptote for values greater than about 10 or 12. They also noted that Mechner's (1958) and Platt and Johnson's (1971) data showed a similar trend.

Some inconsistency in the results obtained in these procedures is apparent; for example, Platt and Johnson (1971) has been cited as demonstrating scalar variability, with constant coefficients of variation as number increases from 4 to 24 (see Gallistel & Gelman, 2000; Brannon, 2005). The reason for the mixed results is unclear; only some individual data was presented in Platt and Johnson's original study and Machado and Rodrigues (2007) noted that in this study, and other previous studies including Mechner (1958), do not report data that allows the direct assessment of scalar/nonscalar response variability.

Overall, responding in FCN-type schedules appear to violate Weber's law, with small numbers; relative response variability for values less than approximately 12 decreases hyperbolically. For values larger than 12, the scalar property still holds, with coefficients of variation remaining more or less constant. This does not appear to be affected by the relatively minor procecdural variations between different studies, and appears to be replicable. These results contradict the notion that number is represented and processed by a system that solely conforms to scalar principles, and suggest that in these types of procedures, at least, counting and timing are not based on the same processes.

Unlike relative numerosity discrminations, there is limited research investigating

response variability in numerical production procedures. These findings provide some information about how number may be processed and represented in numerical production procedures. However, in the FCN procedure there is no upper limit on the number of responses allowed; any number of responses that were equal or greater than the target number are reinforced. This leniency in reinforcement contingencies may have affected response variability such that it resulted in non-scalar responding to smaller numbers.

A different kind of production task was used by Xia et al (2001), where pigeons were required to produce a particular number of keypecks for each of 6 different stimuli. However, a strict upper limit was imposed on responding in this task; a time-out began as soon as subjects exceeded the target requirement. This prevented any in-depth analysis of response variability and its relationship with number.

The data obtained from the numerical reproduction procedure is potentially useful in determining how number may be processed and represented (Chapters 3 and 4). Subjects were trained with three different numerical values and tested for transfer with four additional novel values both within and outside the training range, providing sufficient numerical variability for response analysis. Additionally, there was no upper or lower limit on the number of responses that could be produced during the response phase, but only the correct number of responses was ever reinforced. Thus the procedure affords sufficient response variability to allow detailed analyses, yet reinforcement contingencies are strict enough to assume subjects were responding on the basis of number. Previous analyses (see Experiments 2, 2A and 3) have already demonstrated numerical control over responding, above and beyond control by temporal cues, so it is likely that numerical, rather than timing processes are being assessed here.

In the following analyses, response variability is assessed using data from the three numerical reproduction experiments. Of interest is the relationship between the coefficients of variation, a measure of relative variability calculated by dividing the standard deviation by the average response number, and flash number. Constant coefficients of variation would suggest

scalar processing and either a logarithmic number scale with constant variability or a linear number scale with increasing variability, whereas decreasing coefficients of variation would suggest Poisson processing and a linear number scale with constant variability.

Additionally, data from acquisition experiment (Experiment 3) is of interest. One question that has not been addressed is whether binomial variability develops after extended training or is present from the very beginning. One possibility is that responding initially conforms to scalar variability, but becomes more precise as subjects gain more experience in the task. A significant change in response variability over training has significant implications for the interpretation of the coefficients of variation as a reflection of a numerical response process or numerical representation; do changes in variability merely reflect a change in the nature of responding or a change in the underlying subjective numerical scale? Is it possible to distinguish between these two possibilities?

Data from the extended transfer testing in Experiment 3 will also be analysed to investigate the structure of the subjective numerical scale; in particular assessing whether responding was consistent with either a logarithmic or linear number scale.

6.3 Results

6.3.1 Experiment 2

Coefficients of variation (CVs) were calculated for individual subjects and conditions by dividing the standard deviation of response number during the production phase by the average number of responses. CVs were calculated for each trial type (2, 4, 6) over the last 10 sessions of baseline training, and are shown in Figure 6.1. A two-way repeated measures ANOVA with condition and flash number found a significant main effect of number, $F(2,6) = 31.88$, $p<.001$ but no significant effect of condition or interaction. A planned contrast on number revealed a significant linear trend, $F(1,3) = 59.72$, $p<.005$. As Figure 6.1 shows, coefficients of variation decreased linearly with increases in flash number, with no significant difference between the time-and rate-controlled conditions.

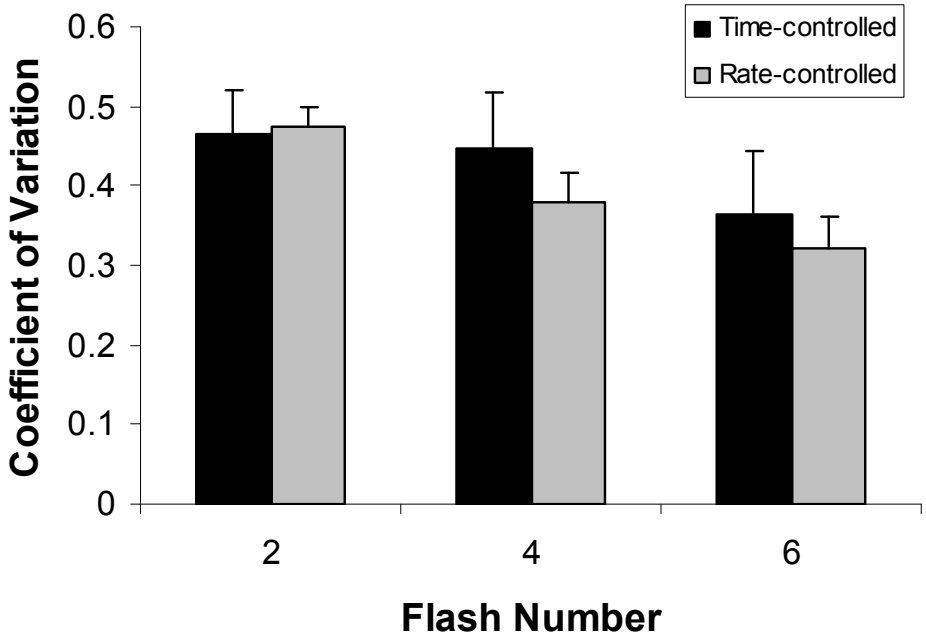

Figure 6.1. Mean coefficients of variation for the 2-, 4- and 6-flash trials averaged across all subjects in the time- and rate- controlled trials.

To provide a finer-detailed analysis of relative variability, log CVs were plotted against log response number, pooling data across subjects for both baseline and transfer tests (including baseline trials) in each condition. The slope of the function when CVs are plotted against average response number on a log-log scale is informative, and may allow us to differentiate between two possible numerical processes. A slope of 0, where CVs increase proportionally to average response number, would suggest scalar variability, consistent with previous research in human and nonhuman nonverbal numerical discriminations, whereas a slope of -0.5, where the standard deviation increases as a function of $\sqrt{n}$ (i.e., Poisson variability) would suggest binomial variability, a finding normally associated with human verbal discriminations.

Because there were 3 trial types in baseline sessions and 7 in test sessions, there were a total of 3 + 7 + 7 = 17 data points per subject in each condition. Results are shown in Figure 6.2. For both conditions, log coefficient of variation decreased with log response number, b = -0.25, p<.05 and b = -0.50, p<.001, for time- and rate-controlled conditions, respectively. The magnitude of the slope for the rate-controlled condition was significantly greater than in the

time-controlled condition, $t(198) = -2.00$, $p<.05$. Thus, responding in the rate-controlled condition exemplified binomial (Poisson) variability, whereas variability in the time-controlled condition was approximately midway between binomial and scalar.

6.3.2 Experiment 2A

CVs were calculated in the same manner as for Experiment 2, using the last 10 sessions of baseline and transfer data. The left panel of Figure 6.3 shows the average coefficients of variation for response number for each trial type during baseline. A repeated-measures ANOVA found that coefficients of variation decreased significantly as flash number increased, $F(2,6) = 34.43$, $p<.001$. A trend analysis revealed a significant linear trend, $F(1,3) = 155.52$, $p<.01$. The right panel of Figure 6.3 displays logs CVs for individual baseline and transfer data as a function of log response number. The regression slope was -0.54, $p<.001$. Thus, similar to the rate-controlled condition in Experiment 2, the standard deviation of response number increased as a negative function of the square root of the mean, exemplifying Poisson variability.

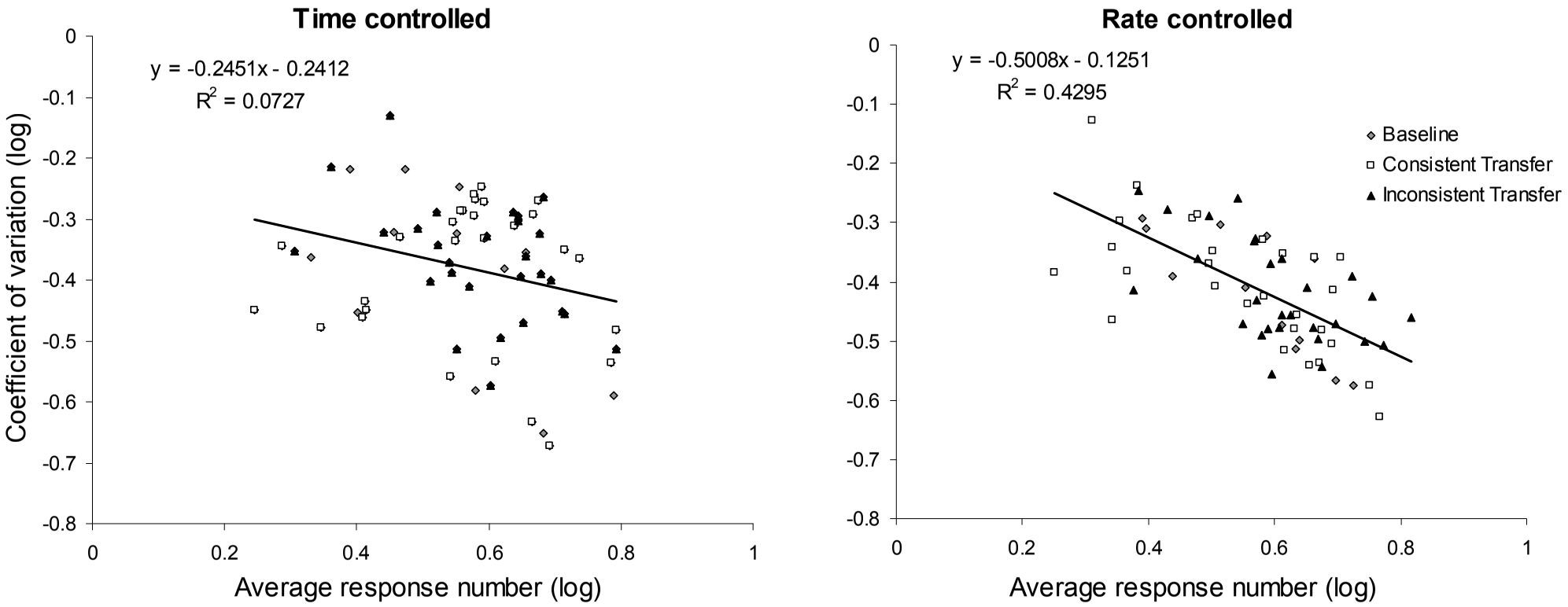

Figure 6.2. Scatterplot of log coefficient of variation for response number and log average response number for baseline and transfer tests in the time-controlled (left panel) and rate-controlled (right panel) conditions of Experiment 2. Each data point shows results for an individual subject, and regression lines are included.

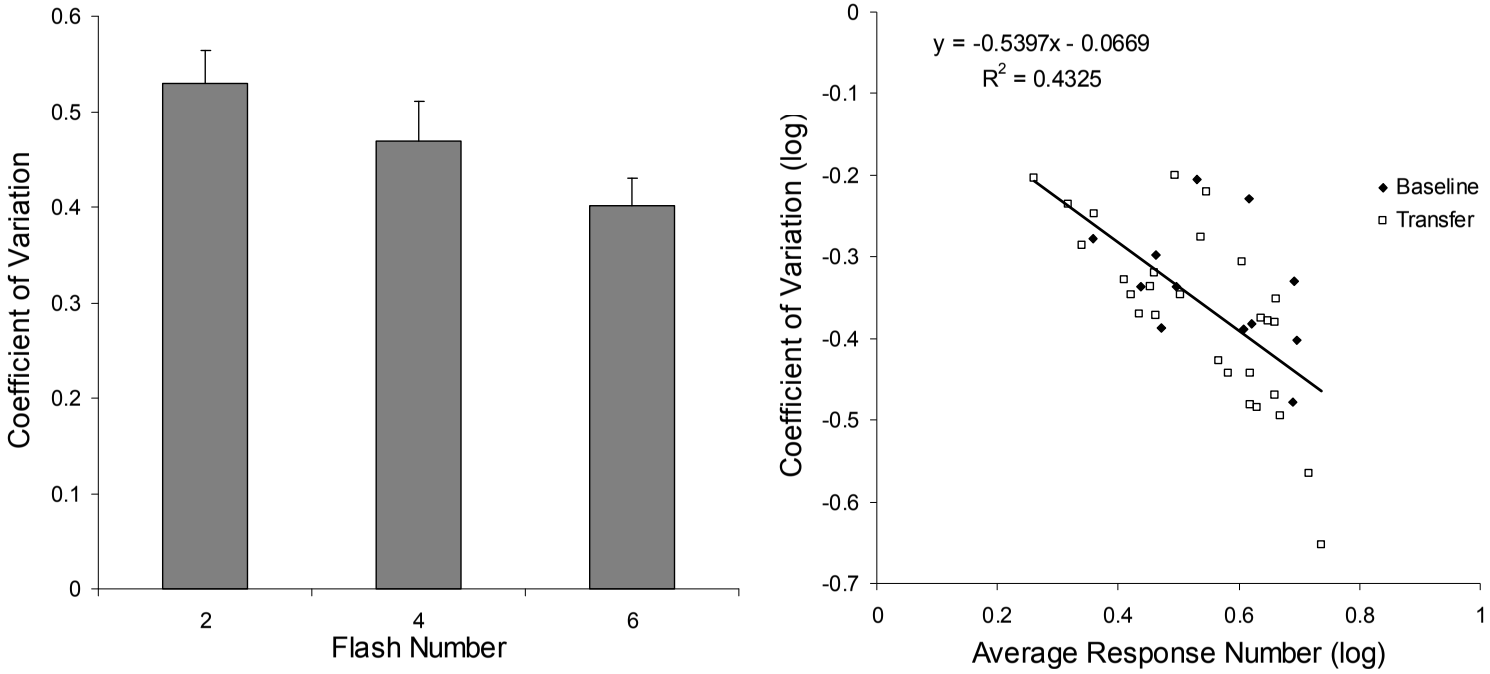

Figure 6.3. The left panel shows coefficients of variation for number of responses during the production phase for each trial type (2, 4, 6 flashes) during baseline training in Experiment 2A, averaged across subjects. Bars represent ±1 standard error. The right panel shows a scatterplot of log coefficient of variation for response number and log average response number for baseline and transfer tests in Experiment 2A. Each data point shows results for an individual subject, and the regression line is included.

6.3.3 Experiment 3

Individual CVs were calculated for the 2-flash, 4-flash and 6-flash trials for each 10-session block in baseline training. These were averaged for each group, subjects 185-188 and subjects 195-197, and are plotted in Figure 6.4. Scalar variability would represented by equal and constant CVs across all three trials types, whereas binomial variability would be represented by a decrease in CV as flash number increased. Repeated measures ANOVAs were conducted on the CVs obtained for sessions 1-10, 100-110 and the last 10 sessions of baseline training for subjects 185-188 and 195-197 separately.

For subjects 185-188, a clear change in relative response variability over time can be seen. Generally, a decreasing trend in CVs over time can be seen. CVs are essentially constant across trial types or increasing with flash number for approximately the first 100-120 sessions of training. At this point, 4-flash trials were excluded for 30 sessions and when they were reintroduced, the pattern in CVs is distinctly different and remains so for the rest of baseline training. CVs clearly decrease as flash number increased, consistent with binomial variability Results of the repeated measures ANOVA found a significant effect of session number on CV, $F(2,6) = 21.39$, $p<.005$, no significant effect of trial type on CV, $F(2,6) = 0.66$, *n.s.* and a significant session number and trial type interaction, $F(4,12) = 4.39$, $p<.05$. This confirms the general decrease in overall CVs and the change in CV patterns across trial types from increasing to decreasing, as subjects became more experienced in this task.

A similar, albeit less clear pattern is also shown in the data for subjects 195-197. CVs show a clear decreasing pattern across time, but variability patterns across trial types are a lot more erratic. At the beginning of baseline training larger CVs were generally found on the 4- and 6-flash trials than on 2-flash trials; by about 120 sessions, the data are more orderly, generally with the greatest CV values being obtained on the 2-flash trials. CVs for 4- and 6-flash trials are lower than that on 2-flash trials, and appear to be approximately equal. Results of the repeated measures ANOVA found a significant effect of session on CV, $F(2,4) = 9.45$, $p<.05$, but

no significant effect of trial type, $F(2,4) = 0.18$, *n.s.,* and no significant interaction $F(4,8) = 0.36$, *n.s.* These results suggest that a change in relative response variability occurred after approximately 100-120 session in both sets of subjects; CV patterns changed from being highly variable, scalar or increasing with flash number to less variable and decreasing with flash number.

Average CVs for the last 10 session of baseline were calculated and are plotted separately for 185-188 and 195-197 in Figure 6.5. A repeated measures ANOVA found a significant effect of number on CV, $F(2,10) = 27.68$, $p<.001$. No significant effect of group on CV, $F(1,5) = 0.73$, *n.s* and no significant interaction, $F(2,10) = 0.67$, *n.s* was found. A trend analysis revealed a significant linear trend, $F(1,5) = 37.40$, $p<.005$. For both sets of subjects, CVs decreased linearly as a function of flash number, consistent with binomial variability.

To quantify the relationship between relative variability and number, log CVs were regressed onto log average response number for both sets of subjects. For the baseline data, significant negative relationships between CV and response number were found for both 185-188, $b = -0.39$, $p<.05$, and 195-197, $b = -0.78$, $p<.05$. A scatterplot of baseline log CVs and log average response number can be seen in the left panel of Figure 6.6.

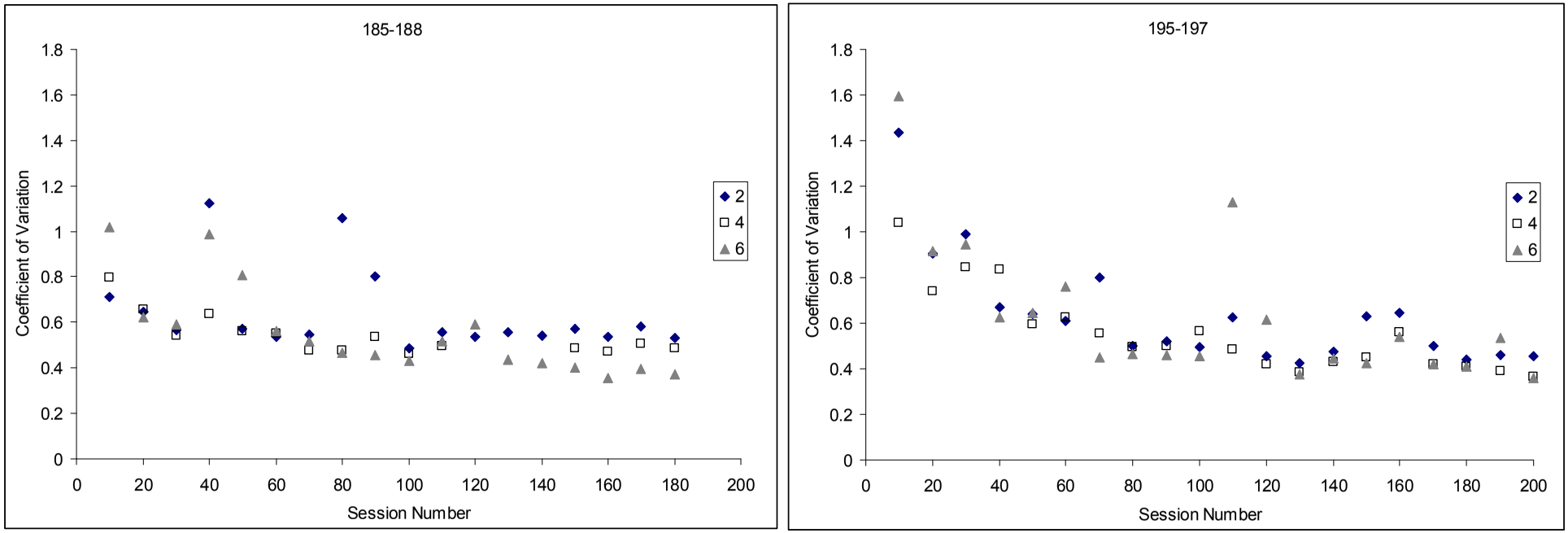

Figure 6.4. Average coefficients of variation for the 2-, 4- and 6-flash trials calculated across 10 session blocks for subjects 185-188 (left panel) and 195-197 (right panel) in Experiment 3.

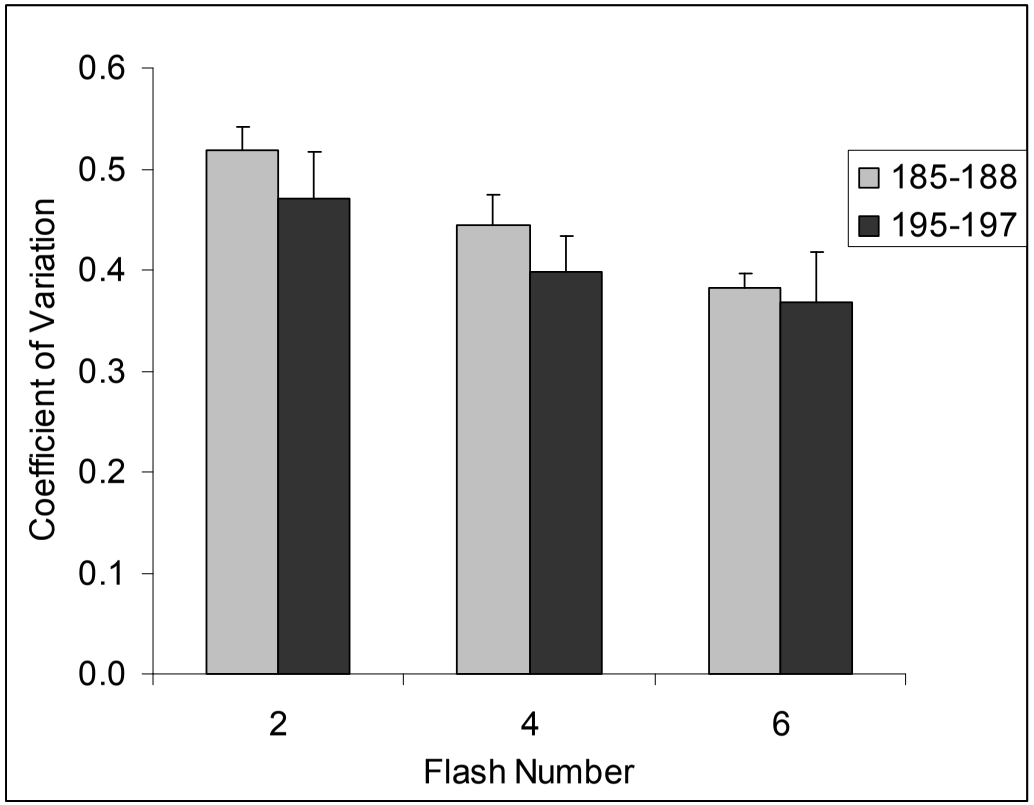

Figure 6.5. Average coefficients of variation for subjects 185-188 (light bars) and195-197 (dark bars) for the last 10 sessions of baseline training. Error bars represent ± 1 S.E.

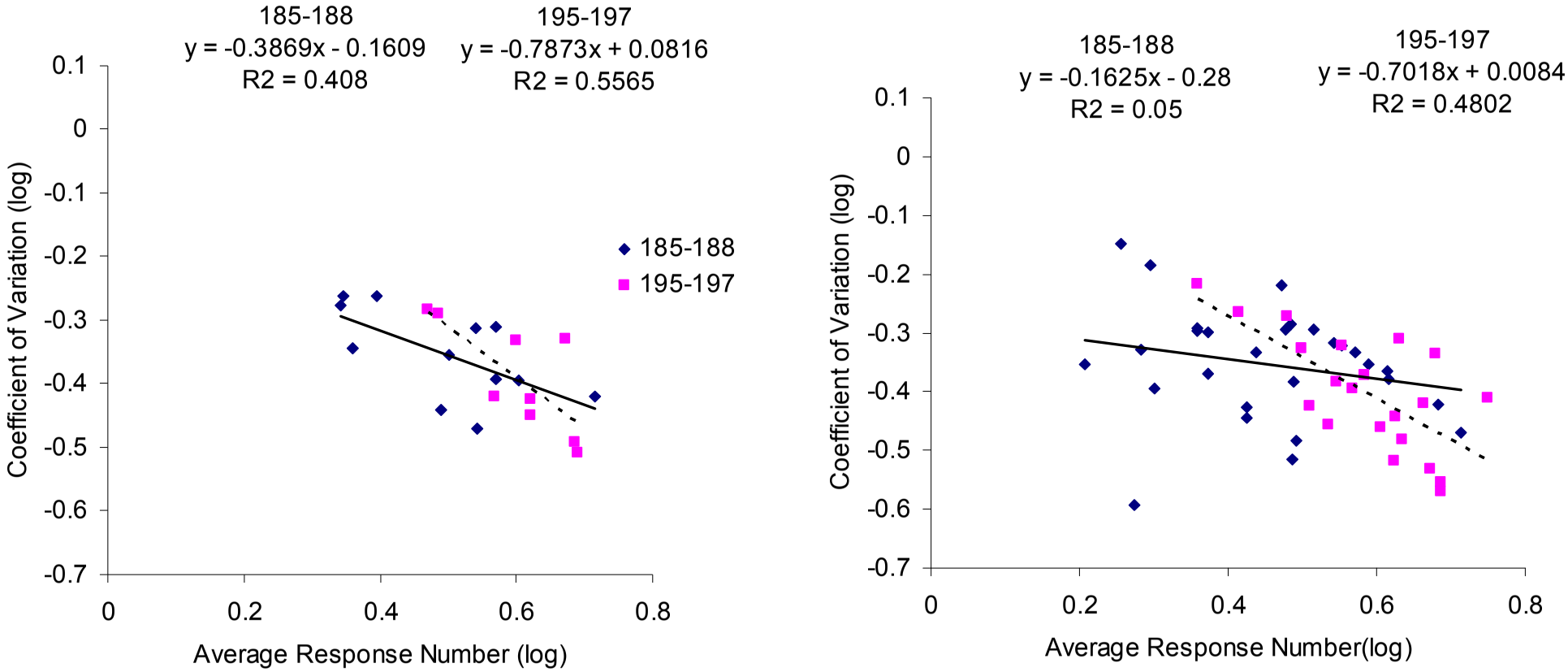

Figure 6.6. Average coefficients of variation plotted against average response number on a log-log scale for subjects 185-188 (filled diamonds) and 195-197 (unfilled squares) for the last 10 sessions of baseline training (left panel), and the first 10 sessions of transfer tests (right panel). Fitted linear regression lines are shown for 185-188 (solid lines) and 195-297 (dotted lines).

Relative response variability in the first 10 sessions of transfer tests was also investigated using regression analyses in the same manner as baseline. The average log CVs and log average response number for the two groups of subjects can be seen in the right panel of Figure 6.6. The regression slope for subjects 185-188 did not differ significantly from 0, $b = 0.16$, $p = 0.25$, whereas the regression slope for subjects 195-197 was much steeper, $b = -0.71$, $p < .001$. However, a between-groups t-test found no significant difference between the average slopes for 185-188 and 195-197, $t(5) = 1.52$, $p = 0.19$, and so data were collated for the following slope analyses.

Single sample t-tests were conducted with the obtained slope values for each subject's CVs and average response numbers to determine whether slopes differed significantly from 0 or -0.50. A significant difference was found between obtained slopes and 0, $t(6) = 2.52$, $p < .05$, and no significant difference between slopes and -0.5, $t(6) = 0.27$, *n.s.*

6.3.4 Summary

The results of these analyses show that in all numerical reproduction experiments, relative response variability more closely approximated binomial variability than scalar; the only exception was the responding in the time-controlled condition of Experiment 2, which showed variability approximately midway between binomial and scalar variability (slope of -0.25). This provides further evidence subjects were not timing in this task, as timing processes are strictly associated with scalar variability. Subjects were able to limit response variability such that relative variability *decreased* with number. Additionally, it appears that this characteristic is not present at the outset of training; it appears to develop as subjects gain more experience in the procedure.

The finding of binomial variability has implications for the representation of number. Generally scalar variability has been the signature of nonhuman nonverbal counting

performance, and consequently a logarithmic number scale with constant generalisation between values or a linear number scale with increasing generalisation between values have been proposed as the two main hypotheses for numerical representation, as both predict that relative variability remains constant with changes in number. However, neither of these scales is able to account for the binomial variability and so other possible scale structures must be considered.

Binomial variability is a signature of human verbal counting (Cordes et al., 2001) and the typical number scale associated with this level of numerical understanding is a linear number scale with constant generalisation between numbers. This number scale predicts binomial variability, as the error rates should remain constant as numbers increase; resulting in a decrease in relative response variability.

It should be possible to distinguish between the numerical scales predicting scalar variability and binomial variability by examining the "spacing" of responding, in terms of differences in average response number, along the numerical continuum. The two scales predicting scalar variability should effectively both predict differences in average response number that are logarithmically spaced ,whereas the scale producing binomial variability should predict differences in average response number that are linearly spaced..

6.4 Logarithmic or Linear? Subjective numerical scaling

The following analyses aim to elucidate the possible structure of the subjective numerical scale subjects developed in the numerical reproduction procedure using modelling techniques. Transfer test data from all the subjects used in the acquisition experiment (Experiment 3) were used to test whether the subjective numerical scales more closely resembled a scale with logarithmic or linear spacing. Each of these two scales was used to predict the obtained differences in average response number for every combination of trial-type pairs from the first 60 sessions of transfer tests. Their ability to account for the data was assessed and compared.

Functions were generated by multiplying a parameter by either the difference between any two given flash numbers for the linear model (k *[$n1 - n2$]), or the difference between the logarithm of the two flash numbers (k * [log($n1$) – log($n2$)]). Parameters were solved using Excel solver to maximise variance accounted for by each of the functions.

Obtained and predicted difference values for the logarithmic and linear models can be seen in Figure 6.7. All equations for the linear model fits had a slope of 1, whereas slopes for the logarithmic model fits were less than 1, suggesting the logarithmic model tended to underestimate the obtained differences between average response numbers on the different trial types. Generally, the plots show the linear model was able to predict differences in average response number much more successfully than the logarithmic model.

Accordingly, for all subjects, variance accounted for (VAC) by the linear model was greater than the logarithmic model. Individual VAC and k values can be seen in Table 6.1 below. A dependent-samples t-test showed that the average VAC for the linear model, 92.17%, was significantly higher than the average VAC for the logarithmic model, 73.21%; $t(6) = 3.08$, $p<.05$.

Table 6.1. Variance accounted for and *k* parameter values for logarithmic and linear scaling models for individual subjects.

	VAC (%)			*k*	
Subject	**Logarithmic**	**Linear**		**Logarithmic**	**Linear**
195	64.03	97.20		2.58	0.37
196	61.90	95.60		2.73	0.39
197	90.73	93.10		3.54	0.49
185	74.41	86.80		2.41	0.34
186	89.15	94.44		1.56	0.22
187	79.40	84.53		2.17	0.30
188	52.88	93.54		3.13	0.46

These findings are consistent with the hypothesis that subjects' subjective numerical scale has linear, rather than logarithmic spacing, and taken with the results of the CV analyses, would suggest subjects may have developed a linear representation of number with constant variability, rather than a logarithmic scale with constant variability or a linear scale with scalar variability.

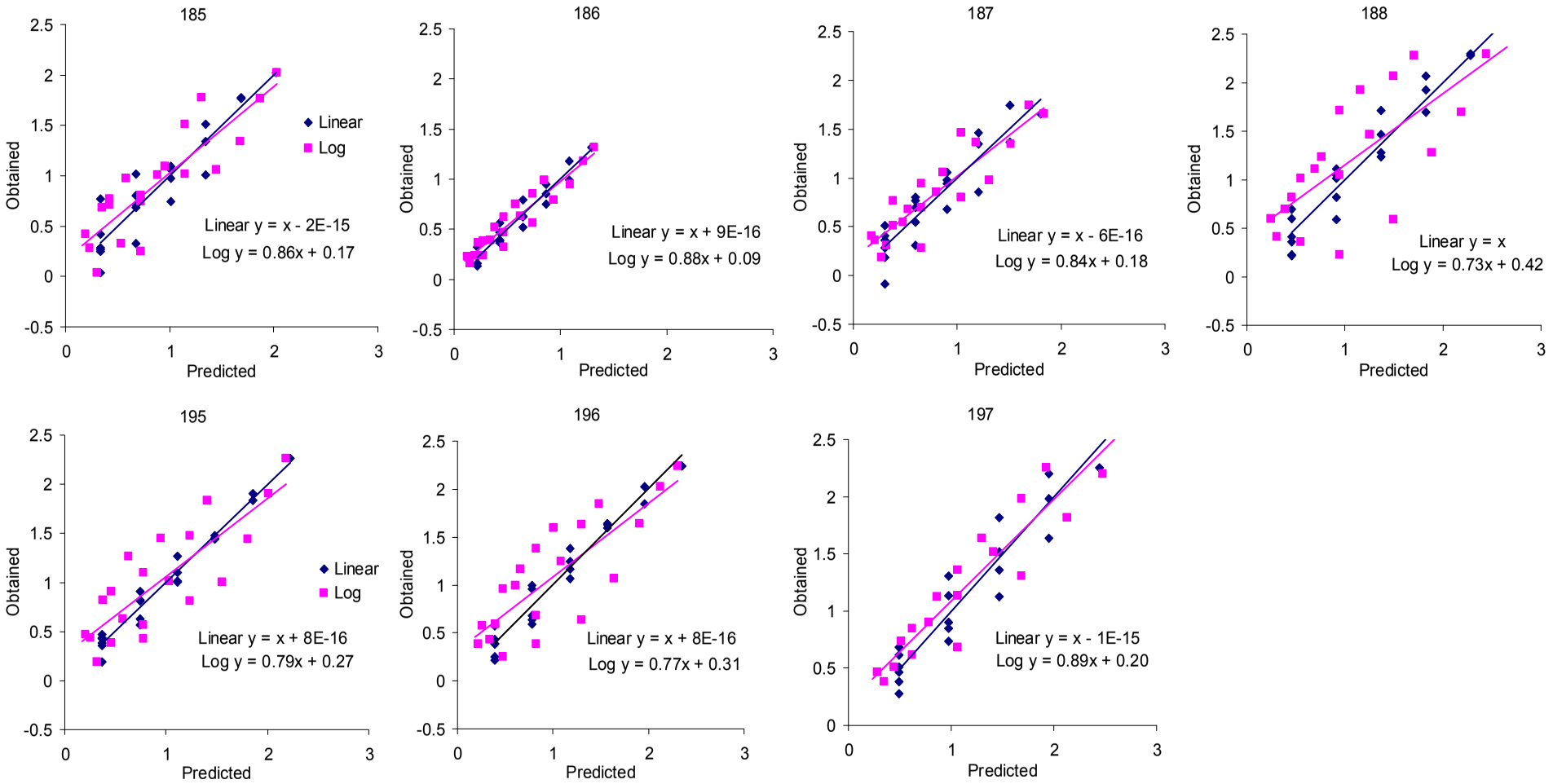

Figure 6.7. Fits of logarithmic (pink squares) and linear (blue diamonds) models to differences in average response number calculated for individual subjects in Experiment 3.

6.5 Discussion

Variability analyses of the Experiments 2, 2A and 3 have demonstrated that responding in numerical reproduction procedure does not conform to scalar principles: Although variability in responding in this procedure increased with number, the increase was not proportional. For all three data sets, the coefficients of variation, a measure of relative response variability, decreased as number increased, and when plotted against average response number on a log-log scale, slopes of the function generally did not differ significantly from -0.5. A slope of -0.5 characterises binomial variability, previously obtained with human verbal counting (Cordes et al. 2001), and is found when variability increases with the square root of number. This suggests discrimination and response processes used by subjects in this procedure enabled them to limit their response variability such that their relative accuracy in their reproduction of flash number increased as flash number increased.

Binomial variability has implications for the structure of the subjective numerical scale. This type of variability is typically associated with a linear number scale with constant generalisation between values, whereas scalar variability is normally associated with a logarithmic number scale with constant generalisation or linear with scalar generalisation. Results of distance analyses were consistent with the CVs; fits of a linearly scaled model of number to obtained transfer test data from Experiment 3 were considerably better than a logarithmic model for all subjects.

These variability results are similar to those obtained by researchers using Mechner FCN procedures (see Machado & Rodrigues, 2007), from which the numerical reproduction procedure was originally adapted. Responding in FCN procedures also tends to show decreasing variability, although it is unclear whether the effect was as strong as in the current experiments. None of the previous studies have examined the slope of the CVs or plotted them on a log-log scale to test this. Thus it is unclear whether variability in the FCN procedure is binomial, scalar,

or in between.

Consequently, individual data provided in Machado and Rodrigues (2007) were collated for variability analyses identical to those reported above. The individual CVs plotted as a function of average response number on a log-log scale are shown in Figure 6.8 with a trendline fitted to the average data. A linear regression analysis revealed a slope of b = -0.18, p<.005. The average slope of the individual CVs = -0.20 and differed significantly from 0, $t(5)$ = -3.76, p<.05, and -0.5, $t(5)$ = 5.34, p<.01. Thus, response variability in Machado and Rodrigues FCN procedure lies somewhere between scalar and binomial variability, similar to that seen in the time-controlled condition of Experiment 2.

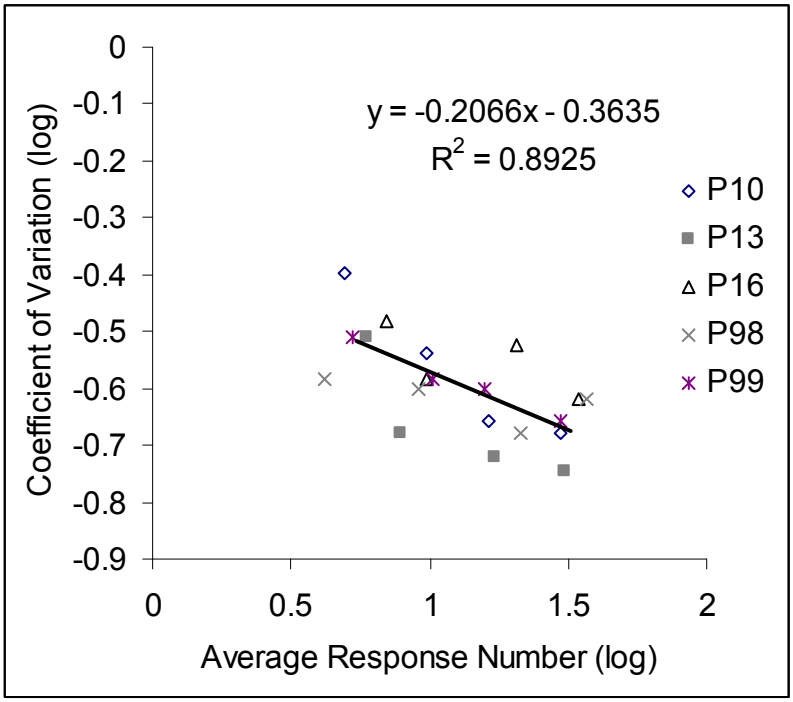

Figure 6.8. Individual coefficients of variation plotted against average response number on log-log scales, obtained from Machado & Rodrigues (2007)

The decrease in relative variability in responding in the numerical reproduction procedure was slightly greater than in a typical FCN procedure. This is somewhat surprising, given that subjects were required to discriminate both the number of flashes presented in the sample phase as well as the number of responses generated in the production phase in order to obtain

reinforcement. Thus, even though responding had to overcome two sources of error within the trial, relative variability still decreased as number increased. One possible explanation for the steeper decreasing slope obtained, relative to those obtained by Machado and Rodrigues, is the inclusion of correction trials during training in the numerical reproduction procedure. Requiring subjects to redo and correct responding on incorrect trials may have increased sensitivity to number and reduced variability in responding, a feature that is not present in the FCN procedures described. At this point the exact processes behind responding in this procedure are unclear.

Binomial variability is characteristic of human verbal counting (Cordes et al., 2001) in a numerical production procedure, and suggests variability patterns resemble that of responding based on a linear numerical scale with constant generalisation between numbers. This finding quite strongly contradicts previous research in nonhuman numerical discrimination (with the exception of responding in the FCN procedures), which has consistently obtained scalar variability in responding. Is there another, more parsimonious explanation for these findings?

It is possible that this anomalous finding is an artefact of a Mechner-type procedure; in particular, one that requires subjects to produce a target number on one key followed by a report response on another key. As decreasing CVs appear to be limited to procedures with this feature, perhaps there is something about this response process that is responsible for the restricted response variability.

This possibility will be at least partially addressed in the following experiment using human participants. Participants were required to enumerate, either verbally or nonverbally, a sequence of stimuli, and required responses include a production response condition analogous to that used in the numerical reproduction procedure as well as an additional report condition, which does not involve a sequence of keypresses followed by a response condition. If similar variability patterns were found in both the verbal and nonverbal counting condition with both these response types, it would suggest that perhaps the variability results are not solely attributable to procedural characteristics.

7 Chapter 7: Human verbal and nonverbal numerical discriminations

7.1 Introduction

The previous chapters have examined numerical discrimination in nonhuman animals, and demonstrated pigeons' ability to discriminate both relative and absolute numerosity in bisection and reproduction procedures, respectively. One particularly noteworthy result is the violation of Weber's law in responding in both tasks; obtained bisection points were located at the arithmetic, not geometric mean, and coefficients of variation for responding in the numerical reproduction procedure decreased as a function of the square root of number, rather than remaining constant. These results would suggest that, in terms of variability at least, numerical processing of subjects in these experiments was more akin to human verbal counting than nonhuman or human nonverbal counting (e.g. Cordes, et al. 2001).

If there is an evolutionary basis for numerical understanding, phylogenetic and ontogenetic continuity in numerical processing and representation would follow logically; animals, humans and infants should share the same or similar systems for discriminating number. This view has received considerable support over the last 10 years from research spanning the fields of developmental psychology, psychophysics, comparative cognition and neuroscience.

Researchers have examined a wide variety of numerical discriminations in humans, often adapting procedures used with nonhuman animals to allow for a more direct comparison of performance. I will selectively review some of this research, discussing experiments with adults and children that involve tasks that range from simple numerosity discriminations to complex discriminations of absolute number.

7.1.1 Relative numerosity discriminations

The original bisection experiment by Meck and Church (1983) has been applied to

humans of various ages, with similar results to that of nonhumans. Droit-Volet, Clement and Fayol (2003) investigated the role of number and time in a bisection task with 5- and 8-year old children, and adults. Participants were presented with sequences in which duration and number were confounded, and were asked to respond based on one or the other. Participants were also divided into counting and non-counting groups, to investigate the role of verbal counting in these discriminations. In the counting group, participants were instructed to count aloud and to adopt a comfortable counting rhythm, while in the non-counting group, participants were required to say repetitive speech aloud as fast as possible to suppress vocal or subvocal counting. The experimenter monitored vocalisations in the non-counting group to ensure compliance.

A first experiment investigated temporal bisection; participants were asked to respond based on total signal duration, and to ignore the number of stimuli. Participants initially experienced pretraining, in which they were presented with the two anchor values; 2 blue circles, presented in sequence and lasting 2s in duration, or 8 circles lasting 8s. These were presented in alternation 5 times. Participants were then trained to press one button on a response box following the short/few stimulus sequence, and a different button following the long/many stimulus sequence. Post-response feedback was provided in the form of a smiling or frowning clown, presented after a correct or incorrect response, respectively. After participants had completed a block of 8 trials with 100% accuracy, they moved onto the testing phase.

The conditions of the testing phase were much the same as in training, except no feedback was given on any trials. Also, during testing, participants were presented with two types of test sequences; on time-varying sequences, number was held constant at 4 stimuli, while varying the total sequence duration from 2, 3, 4, 5, 6, 7, and 8s, and on number-varying sequences, duration was held constant at 4s, while number varied from 2, 3, 4, 5, 6, 7, and 8s. Participants experienced 8 blocks of 14 trials each, presented in a random order.

Although no significant difference in acquisition was found between age groups in the counting group, the 5-year olds in the non-counting group took significantly more sessions to

reach criterion than the 8-year olds and adults. Overall, the slopes of the psychometric functions varied significantly as a function of time; the proportion of long responses increased as duration increased. Additionally, slopes for the time-varying sequences were steeper in the counting than non-counting group, whereas slopes for the number-varying sequences were flatter in the counting than in the non-counting group. This suggests that allowing a counting strategy improved sensitivity to number and reduced the interference of number in temporal bisection.

Sensitivity to time appeared to vary as a function of age; the 5-year olds were not able to process time and number independently without using a verbal counting strategy; in the non-counting group, bisection functions for both the time- and number-varying sequences were similar and superimposed. This was not seen in participants older than 8 years, suggesting a greater resistance to number interference was present.

In a second experiment, Droit-Volet et al. (2003) tested whether opposite effects would be found in a numerical bisection task. The procedure was identical to the first experiment, except participants were discriminating number, rather than duration. Unlike the temporal bisection task, acquisition of performance in the numerical bisection task did not differ between age groups, suggesting it was relatively easier for the 5-year old children to discriminate number than duration. The proportion of "many" responses increased as number increased for the number-varying sequences, but did not change when number was held constant and duration varied. Additionally, the slopes of the bisection functions were steeper in the counting than non-counting group for number-varying sequences, but did not differ between groups for time-varying sequences, suggesting a counting strategy improved number discrimination. There was also an effect of age on the slopes of the bisection functions for the number-varying sequences; these increased between 5 and 8-years of age. There was no difference between age groups on the time-varying sequences, which is not surprising, given that number was held constant.

Weber ratios, a measure of sensitivity calculated by dividing the difference limen (half the difference between the stimulus value resulting in 75% long/many and 25% long many

responses) by the bisection point value, were obtained and used as a measure of sensitivity. Weber ratios were smaller, showing greater sensitivity, with counting than without counting in both the temporal and numerical bisection tasks, and within the non-counting groups, greater numerical than temporal sensitivity was found in 5- and 8-year olds, but not adults.

This finding shows that in these tasks, number was a greater interference on time discriminations than vice versa. Droit-Volet et al. (2003) proposed that temporal processing required greater attentional resources than number, and was associated with a greater amount of error in its representation, due to the information being continuous, rather than discrete. This is further compounded by the inherent difficulty for younger children to ignore number as this appears to be a more automatic process, and consequently would require greater inhibition processes. Note that this finding is opposite to what is normally found in nonhuman animals (e.g. Breukelaar & Dalrymple-Alford, 1998), where temporal processing is generally the more automatic process with greater influence over responding than numerical processing.

Bisection points were similar for both the time and number bisection tasks, and were closer to the arithmetic than the geometric mean. Droit-Volet et al (2003) report that these are classically found in humans (e.g. Wearden & Ferrara, 1995), however bisection points at the arithmetic mean violates Weber's law and consequently goes against much previous research on temporal and numerical bisection in both humans and nonhumans (e.g. Allan & Gibbon, 1991; Beran, Johnson-Pynn, & Ready, 2008; Meck & Church, 1983; Roberts, 2006). It has been suggested that this discrepant finding may have been due to the procedural manipulation in the non-counting condition not being sufficient to prevent covert or overt counting, or the successive stimulus presentation eliciting a serial enumeration strategy, which may have resulted in a more linear representation of number (Jordan & Brannon, 2006). However, this would not account for the differences in performance between counting and non-counting conditions found by Droit-Volet et al. (2003), or the finding of geometric mean bisection in other studies that have used sequential stimulus presentation (e.g. Fetterman, 1993; Meck & Church, 1983).

Roitman, Brannon, Andrews and Platt (2007) conducted a similar study, adapting the Meck and Church (1983) procedure to allow for direct comparison between human performance and performance of monkeys conducted in a previous study. They trained 12 human volunteers to discriminate visual stimuli that varied in both duration and number. Participants were not given any explicit verbal instructions about how to respond. Each participant experienced 50 training trials, in which one of the two compound stimuli were randomly presented. After touching the fixation point on a touch screen to begin the trial, one red and one green response target were presented on the screen, followed by the presentation of the stimulus sequence. The two compound stimuli were either 4 or 16 flashes, with each flash lasting 50ms with a 150ms inter-stimulus interval, thus the stimuli were either 4 flashes lasting 0.8s or 16 flashes lasting 3.2s. Overall duration for the two stimuli was equated by including an additional delay of 2.4s between the onset of the targets and the presentation of the stimuli. This was to prevent participants from responding to trial characteristics other than the stimulus sequences. Following the presentation of the stimuli, participants were required to press one of the two response targets. Feedback was given after every response; if participants selected the green response target after the 4/short stimulus or red after the 16/long stimulus, the word "correct" was presented on the screen. Incorrect choices were followed by the presentation of "wrong" on the screen.

Following training, participants were tested with number- or time-varying stimuli. For number-varying stimuli, duration was held constant at 1.6s and the number of flashes consisted of even numbers between and including 4-16. Conversely, for the time-varying stimuli, number was held constant at 8 flashes, and durations ranged from 0.8s to 3.2s (with 0.4s intervals). A relatively large number of flashes and relatively short sequence durations were used to prevent the use of any verbal counting strategies. Trials containing the 13 novel stimulus types were randomly arranged between the regular compound stimuli from training, and were not reinforced, and did not include feedback. Regular training trials made up 48% of the session;

40% were reinforced, and 8% were unreinforced. Participants were told that half of the trials would be the same as training trials, although a small proportion would not be reinforced. They were also instructed to continue to use the same decision rules for the other trials, and that they would not receive feedback because there were no correct or incorrect answers.

Only four of the 12 subjects were sensitive to both the number and duration of the novel stimuli. The majority of subjects (7) were only sensitive to number, and 1 subject was only sensitive to duration. Generally, most classified novel stimuli according to number, with the proportion of many/long choices increased with number for the number-varying stimuli. There was little change in responding for the time-varying stimuli, however; only 5 of the 12 subjects classified time-varying stimuli according to differences in duration.

Bisection points for the number-varying stimuli (group average = 7.41) were closer to the geometric mean of 8 than the arithmetic mean. For those subjects sensitive to duration, the average bisection point was 2.30, which was located closer to the arithmetic mean of 1.75. However given the small difference between the arithmetic and geometric mean (1.6), this finding is difficult to interpret.

In a second experiment, participants were given explicit instructions to attend to either number or duration in testing trials. Participants were provided with 50 training trials, as in the previous experiment, before experiencing a block of either number or-time varying trials and were instructed to continue with the same decision rule they were previously using. Testing involved the 7 novel trial types, presented randomly and unreinforced. Additionally, reinforced and unreinforced training trials were also presented on 18% and 8% of trials respectively. After the completion of the first block of testing trials, participants were then instructed to attend to the other stimulus dimension, given a brief set of practice trials, followed by a second block of testing with novel stimuli, arranged identically to the first (Roitman et al., 2007).

Results showed that 7 of the 9 subjects successfully discriminated the novel stimuli based on the relevant, varying dimension. The remaining two subjects responded on the basis of

number, but not duration. When instructed to attend to number, all subjects successfully discriminated the number of flashes, and the average bisection point, 7.69, was located close to the geometric mean. Greater sensitivity to number was shown in responding than in the first experiment. Similarly, when told to attend to time, 7 of the 9 subjects successfully discriminated duration in the time-varying stimuli, and the average bisection point, 1.63 was located close to the geometric mean. Thus, temporal discrimination had improved relative to the first experiment, when duration was made more salient through explicit instructions. Additionally, a positive relationship between sensitivity to number and sensitivity to time was found; when plotted against each other, the slope of the function approximated 1, suggesting that participants were equally sensitive to number and time when instructed which dimension to attend to.

Roitman et al. (2007) tested whether participants may have been verbally counting in a third experiment; the large numbers and short durations may not have been a sufficient manipulation to restrict counting, and consequently a verbal distractor task was introduced. The flashing circle stimulus was replaced with flashes of A's and B's presented in sequence, in a random order. As well as being required to classify the stimulus sequence in terms of time and number, participants were also asked to report the last letter of the sequence after responding and received feedback. It was anticipated that directing attention to the verbal task would disrupt any verbal counting. Participants were told that the configuration of flashes matched the response targets and were not related to letters, and were instructed to deduce the choice rule by trial and error. Testing apparatus and procedure were exactly the same as Experiment 1, with the exception of the modification to the stimuli and the additional task. After 50 to 60 training trials, participants were placed in the testing phase, arranged the same as the first experiment; both time- and number-varying stimuli were interspersed among familiar training stimuli.

Accuracy on the letter identification task was significantly better than chance, suggesting subjects were successfully dividing attention between the discrimination and distractor task. An inverse relationship between accuracy and number of flashes suggested this task was a difficult

task, and anecdotal reports suggested that viewing the flashing letters prevented verbal counting. Discrimination performance under these conditions was similar to the first experiment; 10 of the 11 subjects were able to discriminate number and classify number-varying stimuli successfully. Bisection points were located close to the geometric mean, average = 8.69. Time-based responding was also similar to Experiment 1; only 4 of the 11 subjects successfully classified the time-varying stimuli. The bisection point was obtained at 2.5, once again, closer to the arithmetic than geometric mean. Sensitivity to time and number were both similar to that obtained in Experiment 1.

Because the presence of the verbal distractor did not appear to affect performance on the time and number discrimination task, Roitman et al. (2007) concluded that participants were not using verbal counting in the first two experiments, and that responding in these tasks were based on nonverbal processing of number and time. Their findings partially corroborated those of Droit-Volet et al. (2003); temporal processing was generally inferior to numerical processing, with subjects only successfully attending to and discriminating duration when provided with explicit instructions to do so.

Jordan and Brannon (2006b) also investigated numerical dicrimination in a delayed match-to-sample bisection task. Their two main aims were to investigate whether psychometric functions for children and monkeys (Jordan & Brannon, 2006) would superimpose, suggesting a common numerical representation, and whether performance would be consistent with Weber's law, whether bisection points would be obtained at the geometric or arithmetic mean, and whether functions with the same ratio but different absolute difference would superimpose.

To this end, Jordan and Brannon (2006b) trained 16 6-year olds to match novel exemplars of two sets of anchor values (2 vs. 8 and 3 vs. 12, presented in counterbalanced order) on a touch screen. Stimuli were yellow rectangles that contained an array of dots that varied in diameter and colour between stimuli. Sample stimuli and choice stimuli always differed in terms of colour, size, cumulative surface area, perimeter and orientation of elements to promote number-based

discrimination. Additionally, on half of trials the cumulative surface area and perimeter of sample stimuli were closer to the larger than smaller choice stimuli, and on the other half of trials the opposite was true. The size of the elements was also manipulated in the same way.

Each trial was initiated by a response to a picture presented on the touchscreen. This was then followed by the presentation of the sample stimulus, which was replaced by the choice stimuli after another response. Correct choices were reinforced by a sticker, as well as visual and auditory feedback, whereas incorrect responses were followed by a black screen and a short time-out, as well as negative auditory feedback.

Participants were instructed to choose the picture with the same number of items as the sample picture, and to respond as quickly as possible, to prevent verbal counting. They were shown examples of a correct and incorrect trial, before being presented with two practise trials. They then continued with training until they had reached a performance criterion of 80% correct and a minimum of 8 training trials completed. Participants were then placed in a bisection test, where intermediate values (3 to 7 or 4 to 11) were presented in probe trials that made up 30% of the total number. The remaining trials were identical to training. Choice responses on probe trials were nondifferentially reinforced with a sticker.

Post-tests were also conducted to assess verbal counting proficiency; participants were asked to provide a verbal count of a number of stickers, and to give a certain number of stickers to the experimenter. These showed all children understood the fundamentals of a verbal counting system. Results showed that the probability of choosing the larger choice stimulus increased as a function of number, and that functions for the two different scales (2 v. 8 and 4 v. 12) superimposed when plotted on a relative scale. Bisection points for the psychometric functions were located at the geometric, not arithmetic mean. Thus, responding exhibited scalar variability and conformed to Weber's law. Jordan and Brannon (2006b) noted that their requirement that subjects respond as rapidly as possible to the sample stimulus was sufficient to prevent verbal counting; analyses of response times showed that there was no increase in response latencies as the number of items increased, which would be

expected if subjects were verbally enumerating stimuli. This finding also suggests subjects were not using a serial enumeration process, which may be the primary response used with a sequential stimulus presentation (e.g. Droit-Volet et al., 2003), but a parallel enumeration strategy.

Responding of participants closely resembled that previously obtained with monkeys in a previous study (Jordan & Brannon, 2006). Psychometric functions obtained from both these studies superimposed when plotted together, suggesting both children and monkeys were using a similar nonverbal representation of number to respond in these tasks.

Cantlon and Brannon (2006) also compared ordering processes in rhesus monkeys and humans, in a task where they were required to select the smaller of two presented visual arrays. They initially trained 2 rhesus monkeys to order all possible pairs of values from 1-9 in ascending order, using a touchscreen. Correct responses were followed by positive auditory and visual feedback and a juice reward, while incorrect responses were followed by negative auditory feedback and a black screen for 3s. After approximately 100 sessions, the subjects were tested for 10 sessions with all new stimuli, including the novel numerosities 10, 15, 20 and 30. 75% of test trials were new exemplars of the numerosities 1-9, followed by feedback. 17% of trials consisted of one novel and one familiar value, and 8% of trials contained only novel values. There was no differential reinforcement on trials with novel values, so subjects would have to extend their knowledge developed with training values to the novel numbers. Stimulus characteristics were varied as in Jordan and Brannon (2006b), so that the density, perimeter or cumulative surface area of the elements or background, were not reliable cues.

Subjects rapidly learnt to order numerosities 1-9, reaching an overall accuracy of 82%. Performance was not affected by the density, surface area and perimeter controls, suggesting they were responding primarily based on number. Additionally, responding to novel pairs was at accuracy levels significantly above chance; subjects had successfully extrapolated their knowledge acquired in baseline training to novel, larger values. Accuracy tended to decrease, and response latency tended to increase as the ratio between numerical values (smaller/larger) increased; discrimination ability was

limited by the ratio of the values being compared, rather than the absolute numerical value (Cantlon & Brannon, 2006).

A second experiment was conducted to test whether ratio-dependent performance was also seen in a wide range of numerical values and to compare responding in monkeys and humans. In this experiment the same subjects from Experiment 1, as well as 11 university students were tested in the same task as in Experiment 1; human subjects were tested in a computer-based version of the touch-screen task. The monkeys received no further training, and were just exposed to extra testing trials. Human participants were tested in a single 40-minute session, and were verbally instructed to select the stimulus with the smaller number of elements on each trial, and were told to respond as quickly as possible without counting. Correct responses for humans were followed by positive visual and auditory feedback, while incorrect responses were followed by a 3-5s timeout. Both monkeys and humans were tested with all possible numerical pairs of the even values from 2-30, presented with equal frequency.

Results showed that, similar to Experiment 1, accuracy decreased and latency increased as numerical ratio (smaller/larger) increased. There was no effect of absolute set size, if ratio was held constant. Accuracy was similar for both monkeys (80%) and humans (87%); the difference in accuracy between species was actually smaller than the difference between the least and most accurate human participants. Asymptotic performance was reached relatively early during testing, suggesting any differences in performance were not due to differences in familiarity with the task or stimuli. Internal Weber fraction estimates (see Cantlon & Brannon, 2006 for calculation) revealed sensitivity to number was also similar for monkeys and humans. Cantlon and Brannon (2006) suggested that extensive laboratory training may have resulted in the similar sensitivity values. Generally, these results provide evidence for both qualitatively and quantitatively similar responding in monkeys and humans in this numerical discrimination task.

Beran, Johnson-Pynn and Ready (2008) compared performance of 19 4- and 5-yr old children and 7 rhesus monkeys using identical computerised bisection tasks. Their procedure involved

matching stimuli to symbolic, rather than stimuli that resembled the sample array, such as those used by Jordan and Brannon (2006), this was to prevent responding based on perceptual matches, or other specific properties of the response stimuli.

Each human participant was tested over four separate sessions. At the beginning of each session, subjects learnt the two anchor values for that session, which were assigned to each participant in a random order; 1 and 9, 1 and 6, 2 and 18, or 3 and 12. For the children, the experimenter initiated trials; this was to help control for variations in trial duration and to ensure participants were attending to stimulus presentation. Participants were presented with an array of white dots presented in the top centre of a black screen. The illumination of the dots varied; some were fully white, while others were partially filled. This degraded the relationship between illumination and quantity and would prevent subjects from responding accurately based on illumination alone. Response stimuli were located at the bottom of the screen, a white letter L and M on the left and right sides, respectively. Participants were required to select the L stimulus following the presentation of the small anchor value or the M stimulus following the presentation of the large anchor value by either pressing one of two keys on the left and right side of the keyboard (children), or by moving a cursor to one of the two response stimuli using a joystick (monkey). For children, a correct response was followed by a happy cartoon face and happy noises, while an incorrect response was followed by an unhappy cartoon face and an annoying sound. For monkeys,a food pellet and ascending tones followed a correct response, and a buzz, blank screen and a time-out followed an incorrect response. The children were partially instructed for the task; they were told to pay attention to what was on the screen and to decide which key went with each set. However they were not explicitly told the anchor values, or that the task was numerical. To prevent overt counting responses, they were told to respond as quickly as possible, and trials were cleared from the screen and not scored if response delay was greater than 5s; this resulted in less than 5% of total trials being excluded, as response latencies were generally very short.

Children and monkeys had different amounts of exposure to the procedure. Children only experienced 100-150 trials a session for 4 sessions, whereas monkeys on average received

approximately 5600 trials. However, the same performance criterion, 7/10 correct, had to be reached before testing started. The majority of trials during test phases were identical to those in training, but on one quarter of the trials, a quantity intermediate to the anchor values was randomly selected and presented. No response feedback was given on these trials.

Results showed some similarities and differences between species. The bisection points obtained for both species were closer to the geometric mean for 3 of the 4 anchor value pairs. The exception was the 1 vs. 9 discrimination, where bisection points were closer to the arithmetic mean. Additionally, for the 1 vs. 9 anchor values, the best fitting functions were linear rather than logarithmic, as found for the other anchor values. The results for this discrimination are consistent with the data reported by Droit-Volet et al (2003). It is possible that as this discrimination involved the smallest range of values, a more counting-like strategy emerged, resulting in a more linear representation of number.

Functions for the different anchor value ranges superimposed when plotted on a relative scale, although less well for children than for monkeys. The bisection functions for the monkey and children also superimposed when plotted together, providing additional evidence for a common representational system.

Generally, results of the research discussed so far show that responding for children, adults and monkeys in bisection procedures is quantitatively and qualitatively similar; within species, bisection functions for different ranges superimpose when plotted on a relative scale, and functions for different species superimpose when plotted on the same scale. Additionally, similar numerical sensitivity was found for monkeys and humans, despite differences in experience. Together, these findings suggest that the nonverbal representation of number developed in these relative numerositty discrimination tasks is similar, if not the same, in both monkeys and humans.

7.1.2 Numerical production and report

It appears, for the simpler relative numerosity discriminations at least, nonhuman and

human nonverbal numerical processing is similar; are similarities maintained for absolute number discriminations? Researchers have examined the performance of humans in verbal and nonverbal absolute number discriminations, requiring participants to either produce or report an absolute numerical value.

Whalen, Gallistel and Gelman (1999) examined processing in response production task, resembling that used by Mechner (1958). Participants were required to produce target number of key presses at fast rates that prevented vocal or subvocal counting. Their experiment investigated nonverbal counting processes, in particular whether participants could map a numerical value from a numerical symbol to mental magnitude, and to test for whether responding conformed to scalar variability.

Seven human participants completed 8 hourly sessions, and made 40 sets of keypresses for each odd number between 7 and 25. Participants were initially presented with an Arabic numeral that represented the required number joystick keypresses for that trial. The first keypress removed the numeral from the screen, and participants signalled completion by pressing a second joystick button. Unlike the nonhuman experiments, no prior or repeated training with differential reinforcement was necessary or given. Additionally, subjects were instructed to respond as quickly as possible, in order to prevent overt or covert counting strategies; the fastest possible keypress rate for humans is approximately twice as fast as the rates of typical counting behaviour. Participants were explicitly told not to verbally count the number of presses made but to determine the number of presses "by feel". No feedback about response accuracy was provided.

Results were similar to previous animal data. The average number of keypresses increased linearly with the target number. Additionally, variability in responding increased proportionally to number, standard deviations increased with number such that the coefficients of variation remained constant across numerical values. However, no quantitative analyses were conducted to ascertain the precise nature of the relationship between response number and

response variability and target number.

Inter-response intervals were analysed to test whether participants were subvocally counting. Average inter-response times ranged between 115-127 ms, considerably lower than the reported rates for subvocal counting (240ms). Additionally, total response times for silent counting were measured explicitly in the first and last sessions; participants were presented with Arabic numerals in the same range as testing, and pressed after they had silently counted as fast as possible to the presented number. These latencies were much greater than the total response times obtained for the same target number in testing, and also showed a significant increase for numbers in the teens and 20s, due to the extra syllables in the words. Response intervals were virtually identical with no systematic deviations, suggesting subjects were not subvocally counting, or "chunking" or dividing the number of responses into smaller amounts.

Subjects may have used total response time instead of number of key presses to determine when to stop responding. Whalen et al. (1999) tested this by asking participants to reproduce the duration of a presented tone by controlling the start and stop of second tone with keypresses. They found that the CVs for these timing judgments were much greater than for CVs for numerical keyhpress judgments; if participants had been timing then should have been no difference in variability. However the larger CVs for timing judgements suggest that this was not the case, and also that participants were much less sensitive to time. Participants were also asked to estimate amount of time it took for them to complete 7 to 25 keypresses, by producing a temporal duration equal to total duration of keypress response. Almost all participants greatly overestimated duration (e.g. 200%), suggesting that they were not able to time their response rate with any accuracy.

Whalen et al. (1999) also conducted a second experiment to investigate whether the repetitive motor movements influenced response variability in this procedure. Participants were presented with a black dot that flashes on and off in one location. They were then required to report how many flashes that had been seen, without verbally counting. The duration and rate of

the individual dot presentations were varied randomly to preclude timing behaviour. Performance was very similar to production performance; mean estimates were approximately correct and increased with number. Standard deviations also increased in direct proportion to the number of flashes presented, and coefficients of variation were constant. Although the dot presentation rates were generally significantly faster than covert verbal counting rates, regression analyses were conducted to test whether number or time was a better predictor of reported number. Dot number was a significant predictor of responding, whereas total duration was not.

This experiment of Whalen et al. (1999) shows that when they are prevented from verbally counting, humans' responding in a Mechner-type (1958) procedure closely resembles that obtained with nonhuman animals (e.g. Mechner, 1958, Platt & Johnson, 1971). Average response number, and response standard deviations increased proportionally to target number, consistent with Weber's law. Performance was also similar when subjects were merely required to report the number of flashes presented sequentially, suggesting response characteristics were not a result of the repetitive motor responding required in the Mechner procedure.

Numerosity estimation was investigated in children by Huntley-Fenner (2001), with a particular focus on the variability in responding, namely the coefficients of variation. Fifteen 5- and 7-year olds were presented with and required to judge the numerosity of an array of black squares on a white background, by pointing to a numeral along a horizontal number line. Differences in stimulus area and average brightness were held constant so they were not reliable cues for responding. The size of the squares covaried with number, though Huntley-Fenner argued that number was generally a more salient cue than object size and so it was unlikely participants used this as a cue. Additionally an inverse relationship existed between density and number, arrays containing greater numbers were less dense than those with smaller numbers, so it is possible, albeit doubtful, given the relationship is opposite to that normally found, that responding was based on this rather than number.

Participants were pretested to ensure they were able to recognise and identify numerals

up to 20; those that were not able to meet this requirement were excluded from further testing. Each participant experienced 9 practise trials, with arrays that consisted of 1, 2 or 3 squares, each presented 3 times. Accuracy feedback was only presented on the first three trials. Participants successfully completed the practise trials without having to count and were instructed to try not to count during the test trials.

Test trials involved arrays containing 5, 7, 9 and 11 squares, presented 40 times each. Trials were initiated by the experimenter only when the participant was looking at the monitor. Stimuli were presented rapidly (250ms) and followed immediately by a mask display to prevent verbal counting. Participants were then required to estimate the number of squares by pointing to one of the numerals from 1-20 situated along a number line. No feedback was given following responses.

To ensure the large number of response alternatives did not hinder performance, an alphabetic analogoue of the numerical task was used in a posttest- subjects were presented with the 5^{th}, 7^{th}, 9^{th} and 11^{th} letters of the alphabet and had to place it along a line of letters from the alphabet. Performance on post-tests was close to perfect, suggesting participants were able to respond accurately along a number/alphabet line, despite the large number of possible responses.

Performance on this task was relatively low, with overall accuracy only averaging 28%. However, over half (55%) of responses were within 1 number of the correct answer, demonstrating that responding was not random. The mean numerosity estimates increased in direct proportion to the presented numerosity, with slopes for all three age groups equalling 1 with an intercept of 0. Accuracy significantly decreased from 65% to 11% as numerosity increased from 5- 11, and accuracy also increased significantly with participants' age. As no response feedback was provided in this task, learning is unlikely to have occurred during testing. Accordingly, no significant effect of session number was found on accuracy, nor was any interaction between session number, age and accuracy. No significant change in response standard deviations across sessions was found, showing response precision did not increase with

experience in the procedure. The calculated coefficients of variation were constant across the different numerosities, suggesting scalar variability and consistent with previous research. Mean CVs were negatively correlated with participants' age (in days); estimates made by older children were less variable than those made by younger children, suggesting sensitivity to number may increase with age.

Generally, the results of Huntley-Fenner (2001) are consistent with the adult estimation data obtained by Whalen et al (1999). Children and adults are able to nonverbally discriminate the absolute number of keypresses, number of flashes presented sequentially, or number of items presented simultaneously in an array with response number increasing proportionally to the stimulus number and with scalar variability.

Boisvert, Abroms and Roberts (2003) examined human nonverbal counting in another computer-based task, requiring participants to estimate the number of geometric shapes presented sequentially by producing the number in keypresses or by verbal report. Participants were required to name properties of the simple geometric patterns as they appeared; a cognitively demanding task intended to prevent active subvocal counting.

In a first experiment, participants were trained and tested with the three numbers, 8, 16 and 32 tested individually over three sessions on three different days. A purple circle was presented at the beginning of each trial, before the stimulus sequence began. The stimulus sequence was then presented sequentially, and consistent of different coloured geometric shapes (red, green, blue, yellow or white squares, rectangles, circles or ellipses). After the target pattern had been presented, the end-of-trial (EOT) stimulus, a purple circle, was presented. Each pattern appeared on screen for 0.5 seconds, and inter-item intervals and inter-trial intervals varied randomly within the range of 0.3-2.4 seconds and 3 and 5 seconds, respectively. During training, participants were instructed to say the colour and shape of each pattern in the stimulus sequence and to "get a feel" for how many patterns appeared on each trial and to avoid actively counting. Every trial consisted of exactly the target number (8, 16 or 32) of patterns. For trials 6-10,

participants were asked to repeatedly press the space bar at least 5 times per trial, when they were confident the target pattern would appear and to keep pressing until the EOT stimulus appeared. After 10 training trials, participants were placed in 30 trials of testing. Trials consisted of two types, arranged pseudorandomly. Fifteen of these trials were empty probe trials in which the EOT stimulus did not appear and patterns continued to appear after the target number stimulus. Subjects were instructed to begin pressing the space bar in the same manner as in training trials, but then to also press the End key when they were confident the target number had been passed. Following an End keypress, the trial terminated and the ITI began. The other half of trials consisted of question trials in which four questions were presented after the EOT stimulus, asking participants to estimate four properties of the previous sequence; the number of patterns belonged to a particular colour or shape category (selected at random), the duration of sequence presentation, and the total number of patterns in the set. No feedback about participants' accuracy was provided.

Rates of responding were collected from the empty peak trials and placed in successive number bins, corresponding to the intervals during which each geometric pattern was presented. These were averaged across trials for each subject, and response curves were computed for each target number. Peak rates of responding for the 8, 16 and 32 target numbers were located at 9, 17 and 32, respectively, with the midpoints of each response run also located at similar values; 8.29, 16.40, and 30.66 respectively. Coefficients of variation were calculated to examine variability and found that the CVs for the target numbers 8, 16 and 32 did not differ significantly, suggesting scalar variability. Verbal responses were also analysed; average verbal estimates for each of the target numbers were calculated and equaled 8.48, 15.84 and 28.39, for the 8, 16 and 32 trials, respectively.

Boisvert et al. (2003) also examined correlations between the estimated and actual stimulus durations and estimates of target number. Although a significant correlation between the verbal estimates duration and number were significant, the correlations between estimates of

number and actual duration did not differ significantly from zero, suggesting subjects were not basing their responses on the duration of stimulus presentation.

A second experiment compared nonverbal and verbal estimates of number, using the values 8, 11, 14, 16 and 20. As well as providing a verbal estimate of each target number at the end of the sample presentation, participants were also required to provide a manual estimate, in the same manner as Experiment 1, with the exception that participants were also simultaneously required to say aloud the colour and shape of each object. Peak number curves of the manual response data showed that participants were able to accurately nonverbally estimate the target number with peak rates of keypressing occurring at 8, 12, 14, 17 and 20. The numerical midpoints of participants' response runs were also calculated, and did not differ significantly from the target numbers. Coefficients of variation did not differ significantly across target numbers. The same results were obtained for the verbal response data. Difference scores were computed by subtracting the target number from each of the verbal and manual estimates and divided by the target number. These provide a measure of the direction and amount of error in these two types of estimates. Positive significant correlations between the verbal and manual estimates were obtained for 11 out of 12 participants, which the authors purported showed that verbal and manual estimates of number were based on a common magnitude and numerical process.

The effect of performing a separate, confounding verbal counting task on the verbal and nonverbal estimates of target number was investigated in a third experiment (Boisvert et al. 2003). Each sample stimulus now consisted of a pattern made up of 1 to 10 identically coloured shapes and participants were required to say aloud the number of shapes and their colour, as well as estimate the number of patterns presented in total. Peak rates of keypressing for the target numbers 8, 11, 14, 17 and 20 occurred at 6, 10, 14, 14, and 18 respectively. Mean midpoints did not differ significantly from their target numbers for 8 and 11, but were significantly below target number for 14, 17 and 20. Additionally, mean coefficients of variation appeared to decrease as

target number increased, however this difference was not significant. For the verbal response, verbal estimates were all significantly lower than the target number, however despite poor performance, coefficients of variation were still constant across target numbers. Correlations between the verbal and manual estimate were calculated as in Experiment 2 and were found to be significant and positive.

These results showed that human participants were able to accurately estimate the numerical location of a target presented within a sequence in a peak number procedure, and that the variability of responding was scalar, consistent with a nonverbal counting process. Verbal and manual estimates of number were positively correlated. Interestingly, when participants had to perform an additional numerical task of reporting the number and colour of shapes in each stimulus pattern, verbal estimates of all the target numbers and manual estimates of the higher target numbers were significantly lower than the actual number. This finding suggests that this distraction task interfered somewhat with the main number estimation task, and Boisvert et al. (2003) propose that this may be due to interactions between multiple accumulators that monitored the number of stimuli in the distractor and main counting task.

Verbal and nonverbal counting processes are able to be differentiated using analyses of variability. Cordes, Gelman, Gallistel and Whalen (2001) tested human participants in verbal and nonverbal counting tasks and compared performance and variability in these procedures. Scalar variability, where errors increase proportionally to numerosity according to Weber's law, is a common finding in numerical discrimination procedures and is consistent with a mental magnitude accumulator model of numerical representation. However, there is another source of error in numerical discrimination and estimation tasks that does not predict scalar variability. Miscounts, counting an item twice or skipping an item, are equally likely with every count and therefore these errors should result in binomial variability, where errors increase in proportion to the square root of the numerosity. Scalar variability results in constant coefficients of variation (the ratio of the standard deviation to the mean) whereas binomial variability results in

decreasing coefficients of variation as numerosities increase. Additionally, Cordes et al. investigated variability around the threshold where processes change from subitising and nonverbal counting; if mental representations of numbers less than or equal to five were discrete, rather than continuous magnitudes, responding to values in this range should not exhibit scalar variability.

In their experiment, eight participants were presented an Arabic numeral on a computer screen which they had to reproduce in keypresses, made as fast as they could. Arabic numerals varied from 3 to 32. There were three conditions: A nonverbal counting condition where subjects had to repeat the word "the" at every keypress, a full counting condition, where subjects counted the number of keypresses aloud using full number words and also a tens count condition where they counted their keypresses in sets of tens. All participants took part in the nonverbal condition, and 6 took part in the full count and 7 took part in the tens count condition.

Cordes et al. (2001) plotted inter-response intervals as a function of number of presses to determine whether subjects used chunking strategies or response patterns when responding. Inter-response time functions were low and flat with a slope of near zero in the nonverbal and tens count condition, but increased with number in the full count condition. However this was expected as subjects had to count aloud using full number words, which consequently would cause inter-response times to increase with each decade.

The mean number of keypresses increased proportionally to number in all condtions, and in the verbal counting conditions the slope of these plots was almost equal to one. Standard deviations of responding also increased with number in all conditions, and this was most marked in the nonverbal counting condition. Cordes et al. then calculated coefficients of variation, the ratio of the standard deviation to the mean, and plotted these on a log-log scale. A slope of 0 would support the scalar variability hypothesis and was a predicted result for the nonverbal counting conditions, whereas a slope of -0.5 would support binomial variability and was predicted for the verbal counting conditions. Coefficients of variation for the nonverbal counting

condition did not differ significantly from zero for the majority of participants, and differed significantly from -0.5 for all but one participant. For the majority of subjects in the verbal full and tens count conditions, slopes differed significantly from zero, and only one subject in each condition had a slope that differed significantly from -0.5. These findings support the predictions of Cordes et al., and show that verbal and nonverbal counting can be differentiated by the relationship between response variability and the mean discriminated numerosity. Additionally, in the nonverbal counting condition, scalar variability was obtained both within and outside the typical subitising range (2-5), and coefficients of variation were also constant within and outside that range. This suggests that only one numerical process was operating over these values, and the finding contradicts the hypothesis that subjects used subitising to discriminate smaller numbers, and a nonverbal magnitude counting mechanism to discriminate larger numbers.

Multiple numerical representations?

Some researchers believe that multiple representations of numbers exist, which change over time and can be used selectively to optimise performance in various numerical tasks (Feigenson et al., 2004; Siegler & Opfer, 2003; Dehaene et al. 2008). It has been proposed that there are two types of representational systems. One is an approximate estimation system that is used for representing large or unfamiliar numbers, the same as or similar to that used by nonhuman animals and conforming to scalar variability; thus numbers are represented either as a logarithmic scale with constant generalisation or a linear scale with increasing generalisation between values. The other system is able to discriminate number much more precisely, either through a perceptual subitising-based system for the accurate discrimination of small values, which is purported to exist in nonhuman animals and infants, or a different system which develops with age and experience with number, in which both large and small numbers are represented along a linear scale with constant generalisation between values. This linear

representation is believed to develop after the logarithmic system and the understanding of more complex numerical principles. It represents and discriminates between all numerical values equally allowing for the precise discrimination of both small and large numbers, and the use of arithmetic operations.

The types of representations used by humans in numerical tasks were investigated by Siegler and Opfer (2003). They identified four hypothesised numerical representations: 1) a logarithmic ruler, such as that proposed by Dehaene (1997), where numerical values are represented along a logarithmically spaced scale, with constant generalisation between values; 2) an accumulator model (Meck & Church, 1983) which they explain as representing numbers and other quantities on a linearly spaced scale with increasing generalisation between values; 3) a qualitative-type representation possessed by younger children of 4- and 5-years, e.g. few vs. many, that develops into a linear-rule representation, with linearly spaced values and constant variability by 6-years (Case & Okamoto, 1996); and 4) multiple representations that combine both linear-rule and logarithmic-rule representations, referred to as the overlapping waves theory (Siegler, 1996). These theories postulate that from infancy, children use logarithmic and accumulator-type representations, but develop linear and categorical representations as they gain more experience with the formal number system. Consequently, these different representations coexist and compete, with different representations being utilised in different contexts for optimal performance. Logarithmic representations are ideal for representing unfamiliar ranges of numbers because they are better able to discriminate between lower values than linear representations. Conversely, linear representations provide more accurate discrimination for higher numbers and are useful when more precise discrimination of larger numbers is required for complex mathematical operations, e.g. multiplication, multidigit addition. It was also proposed that linear representations would develop for smaller numbers prior to larger numbers, since greater experience with these values would be gained first.

Siegler and Opfer (2003) tested the use of these representations by testing children of

varying ages and adults on a number to position (NP) and position to number (PN) task using values that ranged from 0-100, and 0-1000. In these tasks, subjects are shown a number and required to locate its position along a number line (NP task) or shown a location on a number line and required to estimate the corresponding number. These should allow the direct testing of three of these hypothesised representations. If either a logarithmic-ruler or accumulator model is used, then responding should show scalar variability; the average response and standard deviations should increase linearly with magnitude. If a logarithmic-ruler is used specifically, then mean estimates should increase logarithmically with numerical magnitude on the NP task, and exponentially with numerical magnitude on the PN task. Responding based on a linear-ruler representation would show linearly increasing mean estimates, but non-scalar variability. Conversely, if responding was based on multiple representations of numerical quantity, the multiple estimation patterns would be observed for different ranges. It was predicted that age and experience would be correlated with an increasing reliance on a linear representation, and that linear estimates would be generated more often on the 0-100 scale than the 0-1000 scale, which primarily would elicit logarithmic patterns of responding.

In the study, there were 32 participants in each of four groups, divided by grade level; 2nd graders, 4th graders, 6th graders and undergraduate university students. Participants were presented with a 25 cm line, with the left end labelled 0 and then right end labelled 100 or 1000. In the NP task, the number to be estimated appeared above the center of the number line, and in the PN task the position to be estimated was shown by a vertical mark intersecting the number line. Two sets of test numbers, with similar distributions, were created for each scale. For the 0-1000 scale, Set A consisted of the values 4, 6, 18, 71, 230, 780, while set B consisted of the values 2, 6, 25, 86, 390 and 810. For the 0-100 scale, Set A included 2, 4, 6, 18, 42, 71, while set B included 2, 3, 6, 25, 67, 86. These numbers were specifically selected to maximise the differentiation of the logarithmic and linear functions in responding, and to minimise the influence of number-specific knowledge, such as 50 is located halfway between 0 and 100. All

participants were tested across 2 sessions, separated by 1 to 2 days. The type of task, set and time limits were counterbalanced across the two sessions. The 0-100 scale problems and 0-1000 scale problems were blocked, so all problems for one range had to be answered before starting problems for the next. The order of items was randomised within these blocks. All participants were given explicit instructions about how to respond correctly in each task.

To compare performance of the different models, the fit of linear, logarithmic and exponential functions to the median estimates for the numbers were calculated. The fit to the estimates changed significantly with ages, particularly for the 0-1000 range, with estimates becoming increasingly linear with age. The estimates for the 2nd graders was better fit by the logarithmic and exponential model significantly better than the linear model for the NP and PN 0-1000 tasks, respectively, providing support for a logarithmic-ruler, but not an accumulator model. The 4th graders' estimates were fit equally well by both the linear and logarithmic functions (variance accounted for equalled 93%, while 6th graders' and adults' estimates were best fit by a linear than logarithmic or exponential functions. It was concluded that this was consistent with an accumulator model, but not a logarithmic-ruler model (Siegler & Opfer, 2003).

A subset of seven numbers was present in both the 0-100 and 0-1000 scales; the analysis of the second graders' responding to these numbers allowed the investigation of whether an individual could represent the same number differently depending on the numerical context. Would the representation of numbers be represented logarithmically in one scale and linear in the other? For this subset, the linear function fit the mean estimates better in the NP 0-100 scale than the NP 0-1000 scale, which was better fit by a logarithmic function. This would suggest that the 2nd graders were able to use a linear number scale when estimating smaller numbers (0-100), but merely had difficulty applying a linear representation to a large numerical scale (0-1000),

Siegler and Opfer (2003) also investigated how linear estimates were generated by the 6th

grade and adult participants, and whether this was based on an accumulator, pure proportionality or landmark based response rule, by analysing response variability. An accumulator model would predict that variability in estimates to increase linearly with number, while a pure proportionality model, where number and spatial position is represented as a pure proportion of the number line, predicts no systematic change in response variability as a function of number. A landmark based proportionality model assumes that participants divide the number line at particular points to use as references to guide responding, and consequently predicts increasing variability in estimates as distance from a reference point increases. Siegler and Opfer (2003) hypothesised that linear estimation patterns on the 0-1000 were a result of participants dividing the line into quarters and using these points to determine the location of numbers.

The data obtained in the NP and PN 0-1000 tasks were used to test the predictions of these three models. The landmark-based proportionality model accounted for a significant amount of variance in responding of the adults and 6th graders on both tasks. Additionally, this model did not fit the estimates generated by the 2nd graders who did not show linear estimates. This suggests that this was directly related to the observed linear estimates in 6th graders and adults, and is not just an artefact of the task. The other two models were poorer predictors of performance, accounting for less than 5% of variance in the variability in the adults and 6th graders.

Thus, results from Siegler and Opfer's (2003) experiment show that children possess and are able to use multiple representations of number, and that with development and experience, numerical discrimination is increasingly dependent on linear, rather than logarithmic representations of number. Responding to the same numerical values can follow either a logarithmic or a linear pattern, depending on the numerical context; with greater experience with number, children are able to learn to use the most appropriate representation for the situation.

Siegler and Booth (2004) further investigated the development of estimation processes in children in a study, in particular parallels in the development of estimation ability; does a parallel

shift from a logarithmic to linear pattern of responding on a 0-1000 number line occur at an earlier age on a 0-100 number line? Additional aims of their study were to: 1)Examine whether there was a relationship between number line estimations and mathematical achievement scores – is linear-based responding related to better mathematical performance? 2) Test the contribution of increasing reliance on linear representations and increasing precision of estimates on the age-related improvement in estimation performance; and 3) Test the malleability of performance on number line estimation; would accuracy improve if participants are given greater exposure to a relevant task?

In a first experiment, kindergarteners, 1st and 2nd graders were presented with 48 number-line estimation items involving numbers between 0 and 100. Siegler and Booth (2004) predicted that there would be a developmental progression from largely logarithmic-based estimates to a mixture of logarithmic and linear patterned estimates to largely linear estimates. It was also predicted that estimation accuracy would be positively correlated with math achievement scores, and improvements in accuracy would be due to both increasing linearity and decreasing variability. Thus, linearity, variability and accuracy should all improve with age and grade.

Eighty-five students were tested; 21 kindergarteners, 33 1st graders and 31 second graders. They were required to locate one of 24 numbers, selected randomly from 3, 4, 5, 8, 12, 17, 21, 23, 25, 29, 33, 39, 43, 48, 52, 57, 61, 64, 72, 79, 81, 84, 90, 96, along a number line with 0 printed at the left end and 100 at the right end. Numbers were specifically selected to allow for discrimination between linear and logarithmic estimation patterns. Participants were instructed to show the experimenter where they thought the numbers would fall on the line by marking the location with a pencil. No feedback was provided about responses, but they were periodically praised throughout the session. Each participant provided two estimates for each of the 24 numbers. Stanford achievement test scores for mathematics were also obtained for each participant, as a measure of mathematical performance.

Results showed that average estimates of kindergarteners were significantly less accurate

than first or second graders. Additionally the median estimates of kindergarteners were better fit by a logarithmic than a linear function. Conversely, estimates of 2nd graders showed the opposite pattern, and estimates of 1st graders were fit equally well by both a logarithmic and linear function. Similar findings were obtained for individual data; the number of estimates where the log function was the best fit decreased with age, whereas the number of estimates where the linear function was the best fit increased with age. This shows that estimates became more linear with age, and interestingly, the slopes of the best fitting function approached 1, suggesting that the equivalence between median estimates and number improved with age also. Analyses with SAT scores also showed that the the fit of a linear function was significantly correlated with math achievement test score at all three grade levels, suggesting the linear estimates was related to higher test scores.

Estimate variability, calculated as the mean difference between the individuals' two estimates for any given number, decreased with age and experience. Variability was unrelated to the magnitude of the number being estimated, contradicting predictions made by the accumulator model (Meck & Church, 1983) but consistent with the findings of Siegler and Opfer (2003). Numerical magnitude accounted for 0%, 2% and 7% of the variance in estimate variability for kindergarteners, 1st graders and 2nd graders, respectively.

Were improvements in linearity and variability responsible for the increase in estimate accuracy? Siegler and Booth (2004) used hierarchical regression analyses to test the individual contributions of linearity and variability in predicting response accuracy.
Linearity of child estimates added at least 20% of variance above and beyond variability in childs estimates, whereas variability of estimates never added more than 1% of variance accounted above beyond linearity of child's estimates. This suggests improvements in linearity primarily responsible for the increase in estimate accuracy observed across individuals.

In second experiment, both accuracy and linearity increased when participants were required to locate multiple numbers along a line, and correct any wrong answers. Participants

were asked to locate multiple evenly spaced numbers along a single number line. An orienting trial was also presented at the beginning of the experiment to ensure participants understood task; participants were asked to estimate the location of number 50 on a number line marked from 0-100 (or 5 for kindergarteners, who used the number scale 1-10). Participants then received feedback on number's correct location and shown the location of their estimate alongside the correct location.

The second experiment involved three testing phases. First, a pretest was conducted, which was identical to the first experiment; all participants were tested with the same 24 numbers on 0-100 number lines, however kindergarteners were also presented with an extra 18 items on 1-10 number lines, 2 of each number from 1-9. This was followed by the experimental manipulation phase. Participants were divided into a control and experimental group. The control group was given exactly the same task as the in the pretest. The 1st and 2nd graders in the experimental group were presented with a single number line with 10 numbers (5, 15, 25, 35, 45, 55, 65, 75, 85, 95) printed in a random order above it. Kindergarteners were presented with numbers 1-10 inclusive printed in a random order above a 1-10 number line. Participants were asked to place a hatch mark to indicate the location of each number on the line and write the number above the mark. They were allowed to erase and replace a mark if they thought it was placed incorrectly. After marking the locations of the 10 numbers, participants were then given an unfilled number line, identical to that presented initially and were asked to place their final estimates for each number on that sheet. The earlier estimates were still available while this was taking place, so children were able to compare their final with their previous estimates. It was thought that this manipulation would improve the performance of kindergarteners, especially, given their relative inexperience with numerical tasks. Following this, a post-test was conducted, in which participants estimated the locations of the same numbers and on the same type of number line as in the pre-test.

Results from the pre-test were much the same as in Experiment 1. Interestingly, Siegler

and Booth (2004) found that the manipulation in Experiment 2 was actually more detrimental than helpful to kindergarteners; the accuracy of those in the experimental group decreased significantly after the experimental manipulation. Consequently, no additional analyses were conducted with their data. The experimental manipulation had a better effect on the performance of the 1st and 2nd graders; the proportion of absolute error in their estimates decreased significantly from the pretest to the experimental task for participants in the experimental, but not the control group. Additionally the fit of a linear function to estimates increased significantly from the pre-test to experimental task for those in the experimental, but not the control group. Similar changes in performance were observed from pre-test to post-test. Thus the estimates of 1st and 2nd graders, but not kindergarteners, for multiple numbers on a single number line were more accurate than their pre-test estimates of a single number on separate number lines, and that this experience improved later performance on the original estimation task. This finding suggests that after a certain experiential threshold, numerical estimation performance is malleable and can be influenced and improved by experience in a relevant task.

Additional evidence for multiple representations, both compressive (logarithmic) and linear numerical scales has been obtained by Lourenco and Longo (2009). Previous research in a numerical bisection task has demonstrated that healthy adults show a slight leftward bias, termed pseudo-neglect, which results in participants underestimating the true midpoint (arithmetic mean) when bisecting physical lines and also intervals between two numbers (Longo & Lourenco, 2007). This bias increases with numerical magnitude, which is consistent with logarithmic scaling; the persistent leftward attentional bias in the bisection task leads to an increasing leftward numerical bias since larger numbers are subjectively closer together.

Lourenco and Longo (2009) conducted a study to investigate whether human adults have access to both logarithmic and linear scales in the same task, and to test what conditions mediate access and use of these scales; is it possible to prime the use of different representations in the same task? Participants were tested in a numerical bisection task under different memory

conditions, requiring the maintenance of small vs. large numbers. Responding in this task has been found to rely largely on logarithmic scaling (Longo & Lourenco, 2007). Discrimination of lower numbers is more accurate than larger numbers with logarithmic scales due to their compression of larger values. Conversely, linear scales represent all values equally. Consequently, the maintenance and resulting greater salience of smaller numbers in memory should increase the likelihood of the use of logarithmic scales, whereas the maintenance of larger numbers in memory should increase the use of linear scaling. Linear scaling should lead to a constant leftward bias in the bisection procedure across all numerical values, since the subjective spacing between numbers remains constant while logarithmic scaling should result in a numerical bias that should increase with increasing numerical magnitude.

Lourenco and Longo (2009) tested 15 university students, who were presented with pairs of "small" and "large" numbers, separated by a horizontal line. These numbers varied from 11 and 99 and were randomly selected. Numbers classified as "small" ranged from 11 to 85, and numbers classified as "large" ranged from 23-99. A wide range of numbers was used to allow the analysis of the effects of differences in numerical magnitude and interval size on bisection; interval ranged from 11 to 87. The location of small and larger numbers on either side of the horizontal line was counterbalanced.

Participants were asked to estimate the number halfway between each pair, but not to compute the answer and to answer as quickly as they could use whichever number seemed "immediately intuitive". Bisection responses were verbal and recorded by the experimenter. Primes were presented before the actual stimulus number pairs; three different numbers were presented sequentially at the top, bottom and center of screen – for 500 ms each with order randomly determined. Participants were required to recall the primes after their bisection response. On half of the trials, primes consisted of small numbers (1-9), while on the other half of trials primes consisted of large numbers (101-109). The value of the primes were randomly selected on each trial and were selected to be outside range of bisection pairs to emphasise the

size of the small and large primes, and also to prevent any direct memory interference between the primes and bisection stimuli.

For each number pair, deviation scores for the bisection responses were computed relative to the arithmetic mean. A significant underestimation of midpoint, that is, a leftward bias was found for both small and large primes. For both conditions, 14 of the 15 participants showed an overall leftward bias in bisection responses.

Lourenco and Longo (2009) examined the change in response bias as a function of magnitude of the number pairs using least squares regression. In the small primes condition, leftward bias increased as numerical magnitude increased, suggesting participants responding was based on a logarithmic numerical scale when primed with small numbers. However, in the large primes condition, no significant relationship between amount of bias and number was found, suggesting subjects were using a linear number scale during number bisection on trials that involved the retention of large number primes. Additionally, no difference in slope for the small and large primes was found on those trials where participants remembered the primes incorrectly, suggesting that the use of logarithmic or linear scaling was directly dependent on the active retention of small vs. large number primes.

It is possible that the size of the interval between the numbers being bisected influences the amount of bias; Siegler and Opfer (2003) found smaller intervals resulted in linear scaling of estimations, whereas a larger interval resulted in logarithmic scaling. In this study, although overall error increased significantly with interval size, there was no significant increase in directional bias with increasing interval size. This suggests bias was largely dependent on the size of the primes and magnitude of the number pairs and not interval size.

The results of Lourenco and Longo (2009) provide evidence that adult humans are able to selectively use logarithmic or linear number scales in a numerical bisection procedure, and that this can be manipulated by changing the context in which these discriminations are made.

The use of numerical scales also appears to be dependent on cultural factors also.

Dehaene, Izard, Spelke and Pica (2008) investigated the structure of numerical scales used in Western and Amazonian cultures, conducting experiments with the Mundurucu. The Mundurucu are an Amazonian indigenous culture that has had little access to education and despite a relatively limited lexicon of number words and little to no access to rulers, measurement devices, etc., they possess sophisticated concepts of number and space, albeit in an approximate and nonverbal manner. The Mundurucu have specific number words for values 1-5, and numbers greater than 5 are labelled with approximate quantifiers, such as some, many. The key research question of Dehaene et al's research was whether number and space would be mapped logarithmically or linearly.

Their experiments tested 33 Mundurucu adults and children in an experiment similar to Siegler and Opfer (2003). Participants were presented with a line segment on a computer screen, with 1 and 10 dots, or 10 and 100 dots on the left and right sides of the line segment, respectively. Then stimulus numbers within the range 1-10 or 1-100, in various forms (sets of dots, sequences of tones, spoken Mundurucu words or Portuguese words) were presented in random order, and participants were required to position these numbers along the line by pointing to the location and the response was recorded by a mouse click. No feedback was given about responses.

Participants only received 2 training trials prior to testing, and these involved sets of dots whose numerosity corresponded to the ends of scales. They were informed that these numerosities belonged at their respective locations, but other stimuli could be placed at any location. Because training did not involve any intermediate numbers, performance during tested would reflect spontaneous mapping of number and space.

Average responses showed the Mundurucu understood the task. Although some participants only responded at the end points of the scale, most used the full response continuum and adopted a reliable strategy of mapping numbers onto their respective locations. A significant positive correlation between stimulus number and response location was found, regardless of

modality of stimulus presentation. Performance was best when number was close to the reference points at each end of the range, but participants were able to map number and space with stimulus values they had not experienced in training.

A linear regression was not the best predictor of the Mundurucu's responses. Response curves were negatively accelerated, and a logarithmic function was a better predictor of performance; numbers were mapped as a log scale where the middle of the interval 1-10 is located at approximately 3 or 4, and not 5 or 6.

The sample of Mundurucu tested was quite heterogenous, varying widely in age and education. However, nonlinearity was observed in responding of the Mundurucu even when analyses only included data from adults, monolingual speakers or uneducated participants. The only trend towards greater linearity occurred as a function of age, however even the oldest Mundurucu adults showed significant nonlinearity in responding in the range 1-10, whereas in Western children, mapping of number becomes linear over a much larger range, 10-100 by the first or second grade (Siegler & Opfer, 2003; Siegler & Booth, 2004).

A separate analysis of performance with Portuguese numerals was used to assess effects of culture on numerical understanding. Although overall performance was consistent with a logarithmic scale, separation by education level found the performance held for participants with 1-2 years of education but not for those with more or no education. Not surprisingly, individuals with no education showed highly variable performance that was only weakly correlated with stimulus number, suggesting they did not know the meaning of the Portuguese number words. However for the most educated group, performance was a strictly linear function. The results of a reanalysis of data, excluding participants with no education, showed greater education significantly changed response patterns to Portuguese number words from logarithmic to linear, while responses to Mundurucu numerals and dot patterns remained logarithmic. This suggests that there is an effect of number notation on responding and culture and education influences the mapping of number onto space, and it is not just a developmental process. Logarithmic-scaled

numerical representations persist into adulthood for Mundurucu, even for very small numbers (less than 10) and regardless of the mode of presentation. Even though the most educated Mundurucu are able to utilise a linear scale which is central to the Portuguese number word system and permits precise measurement and mathematical operations associated with a linear representation, this knowledge is not extended to Mundurucu number words.

Dehaene et al. (2008) also compared the performance of the Mundurucu with that of Western adults in the same task. Unlike the Mundurucu, American adults rated the sets of 1-10 dots linearly. However, with sets of 10-100 dots and sequences of tones, responding became more logarithmic. These results suggest that, for Western adults, numerical judgments are based on a linear scale only when numbers are presented in a manner which allows the precise assessment or counting of number. When larger numerical values are presented, or presented in a manner which is harder to count, responding shows more logarithmic characteristics.

The findings of Dehaene et al. (2008) show an Amazonian indigenous culture is able to map number and space successfully, and largely use a logarithmic scale to do so. More interestingly, their results show that multiple representations of number can be acquired culturally as well as developmentally, with Mundurucu individuals that are well educated in Portuguese adopting a linear number scale when discriminating Portuguese number words, but still continuing to use a logarithmic scale for Mundurucu number words or non-linguistic stimuli.

Cantlon, Cordes, Libertus and Brannon (2008) draw attention to some valid points in a commentary in reply to Dehaene et al's (2008) article. Dehaene et al. contrast linear and logarithmic scales of number, however do not mention the possibility that the representation used by the Mundurucu could also be linear with scalar generalisation, a structure that makes identical predictions of scalar variability as a logarithmic scale with constant generalisation. They also state that these results do not reflect an inherent mapping of space and number; as adult humans tend to be good at mapping between multiple unidimensional properties, such as brightness, loudness, depth. The findings of Dehaene et al. provide evidence that Mundurucu speakers can

map between the unidimensional properties of length and number but this does not provide concrete proof that numerical representations are fundamentally spatial.

Analogue magnitude models of numerical representation are generally proposed to account for human nonverbal data that exhibit scalar variability. An alternative model that predicts more accurate numerical discriminations is the object-file parallel individuation model (Feigenson et al. 2004).

Research by Le Corre and Carey (2007) with 116 3, 4 and 5 year olds with varying levels of numerical understanding, has found support for a object-file parallel individuation system for discriminating small numbers and a approximate system for the discrimination of larger or unfamiliar numbers. Knowledge levels were assessed using various tests, and participants were separated into 5 groups, participants that did not understand the counting principles (referred to as subset knowers) were divided into "one"-knowers, "two"-knowers, "three"-knowers, "four"-knowers, and the rest were counting-principle (CP) knowers

They analysed nonverbal performance on a numerical estimation task; children were presented with sets of 1-10 items, and were asked to provide verbal estimates of the number of items in each set without counting. Children were also presented with pairs of sets of circles, and asked to point to the set containing more circles, without counting the exact number. Of particular interest was whether there would be any differences in performance between individuals that just understood the concepts of "one" or "two", and whether there would be any differences in performance between children that did and did not understand the counting principles.

Prior to actual testing, the experimenter modelled behaviour and correct answers for the participants, to ensure they understood the procedure. Stimuli used for testing were selected from four decks of cards with 1, 2, 3, 4, 6, 8 or 10 circles printed on them. Card presentation was response-dependent; each card was presented for 1s, but if no answer was given, the card was presented for a longer period of time. If still no answer was given, the experimenter told the

participant the correct answer. Trials in which responses were not given straight away were not included in analyses. The total surface area of circles was held constant or negatively correlated with set size, and the presentation of these was alternated. The spatial configuration of sets was arranged so that sets that had an equal number of circles had different configurations, and sets or large numbers of circles could not easily be "chunked" into smaller perceptual groups.

The sets were presented in one of two pseudo-random orders; either the first test cards showed set within the range of parallel individuation (5 circles), or the first test cards showed sets that were within the analog magnitude range (greater than 5 circles). Repetitions of sets containing the same number were separated by at least two trials.

Participants also were exposed to a nonverbal ordinal task, in which they were presented with two cards and asked to identify which card contained more circles. After children made their response, the next trial started. Participants were verbally discouraged from verbally counting, and this proved to be effective in stopping the few children that did attempt to count. Any trials in which children counted were discarded. Note that participants were never given feedback about their responses and were praised after every trial; consequently responding should be based on a spontaneous representation of numerical magnitude.

The pairs tested were 2 vs. 3, 2 vs. 6, 6 vs. 10 and 8 vs. 10. Each pair was presented three times in two pseudorandom orders; each comparison pair never occurred on consecutive trials, and the correct answer was never on the same side for more than two trials in a row. To reduce the likelihood of participants responding to nonnumerical cues, the configuration of circles in each set of the pair made as different as possible. For two exemplars of each number comparison, the more numerous set had a small total surface area, while for the third exemplar the more numerousness set had a larger surface area.

Le Corre and Carey (2007) examined the average estimates for each knower level as a function of set size number. If children were able to represent the larger numerals, the slope of the function in the large set size range should be greater than 0. All subset knowers failed to

discriminate sets containing 6 or more items. "Three"- and "four-knowers were able to provide verbal estimates of the size of sets containing up to 3 or 4 circles, but average estimates were very noisy.

Conversely, average response number functions for the CP knowers could be categorised into showing slopes of 0 and 1 for values greater than 5. Thus these participants could be separated into two groups, based on responding; CP knowers who hadn't mapped numbers beyond four showed response function slopes of 0, whereas CP knowers who had mapped these numbers successfully had response function slopes of 1. While both groups were able to discriminate numerosities 1-4, only the latter group continued to do so for values larger than 4; no referred to CP mappers. The coefficients of variation obtained for CP mappers were constant for the large set sizes (6, 8 and 10); this scalar variability suggests participants were using analogue magnitudes as a basis for responding to these numbers. CVs for the small set sizes (1-4) increased significantly with number for both the CP mappers, CP nonmappers, and"four" knowers, suggesting responding was not based on analogue magnitudes. The other two possible methods of estimating these values are either a counting or parallel individuation process. As "four" knowers did not show they understood counting principles, it is unlikely they were counting values, and a counting process would predict decreasing CVs as number increased, the opposite trend to that found in the current study. Consequently, Le Corre and Carey (2007) concluded that participants were using parallel individuation to discriminate these small values.

The ordinal task allowed Le Corre and Carey (2007) to test whether children who were not able to accurately estimate the numbers of larger sets in the first task were still able to represent large numerals in a less precise manner in a simpler discrimination. If children's responding was determined by nonverbal numerical representations, then the CP mappers' ordinal judgments should be more accurate than the other groups, given their superior performance in that task. Consistent with this, CP mappers were found to be significantly more accurate in the ordinal task than all other groups, except the CP nonmappers. Additionally,

performance of CP knowers was better than subset-knowers. However, with the exception of the "one"-knowers, performance of the subset-knowers was significantly better than chance on all ordinal pairs suggesting these subjects were able to represent these larger magnitudes and successfully use this to select the larger set.

The main results show that all children tested could estimate the numerosity of sets containing up to 4 circles without counting, suggesting the mapping of these smaller values from numbers to magnitudes is part of the process of the development of counting principles. Additionally, this mapping only occurred for the values 1-4 for all subset knowers; they were not able to map numerals for values larger than 4 onto analogue magnitudes, despite being able to recite the consecutive number values at least up to 4. Evidence for multiple representations was found; the variability of CP mappers for sets greater than 6 was scalar, but the variability of estimates of the sets 1-4 was increased with number and was not scalar in any of the groups. Thus CP mappers used analogue magnitudes to represent values larger than 6, but parallel individuation is most likely the process used to estimate the numerosity of the smaller sets.

The fact that both nonhuman animals and infants are able to accurately discriminate small numbers (1-4) has resulted in researchers proposing that this ability reflects a separate subitising/parallel individuation process, rather than a nonverbal analogue magnitude representation. However, an analogue magnitude model also predicts superior discrimination of smaller values; due to its scalar variability, the noise in the representation of small numbers, relative to larger numbers is low and consequently should be discriminated better than larger numbers.

Revkin, Piazza, Izard, Cohen and Dehaene (2008) tested whether there is a shared estimation system for small and larger numerosities in human adults, or whether a separate subitising process is present for small values. They predicted that if there is only one estimation system, then the discrimination of numerosities 1-8 should be as accurate as the discrimination of the decade quantities 10-80, if ratios are kept the same. Similarly, if subjects are trained with

just decade numbers, then the disproportionately higher accuracy seen for discriminations of numbers within the range 1-4 should also be seen in the range 10 to 40; "subitising" type performance should be observed with large numbers, as long as ratios are sufficiently discriminable. Additionally, if subitising results from a nonverbal approximate estimation process, range should be influenced by an individuals' capacity for numerosity discrimination; those with better discriminatory ability should be more precise and have a larger subitising range.

Eighteen participants were tested with three different tasks, a relative numerosity discrimination and two estimation tasks. In all three tasks, stimuli were masked and required a response within a short delay to prevent counting, subgrouping or arithmetic strategies. Due to a natural bias towards underestimating larger quantities, participants received training trials with feedback in order to calibrate participants to the estimation of large quantities. Participants were given extensive training with naming decade stimuli.

The relative numerosity discrimination was a dots comparison task. Revkin et al. (2008) presented participants with two arrays of black dots and they were required to select which array had more dots as accurately and quickly as possible. One of the arrays was kept at a fixed number (16 or 32 for half trials each) while the other array was smaller or larger than the fixed numerosity, with a ratio of 1.06, 1.13, 1.24, or 1.33; harder comparisons involved smaller ratios. These were presented in a random order across blocks. Participants responded by pressing the mouse button which corresponded to the same side as the larger array. To prevent responding to nonnumerical cues, on half of trials dot size of the varying numerosity array was held constant, whereas on the other half of trials the area occupied by the array was held constant. For the fixed numerosity arrays both dots size and area was varied simultaneously. Participants performed 16 training trials with accuracy feedback, before completing 128 experimental trials.

There were two numerosity naming tasks, one involving judgments with all the values from 1-8, another with the decade numerosities 10 through 80. These tasks were completed in

two separate sessions, with orders counterbalanced across sessions and participants. The procedure was identical for the two tasks, and participants were informed which task was going to be performed and which quantities would appear. They were instructed to name the dots as accurately and quickly as possible, with a response time limit of 1s. Any trials with response latencies greater than 1s were discarded. Prior to testing, participants were calibrated with 16 exemplars, consisting of random patterns of dots and their corresponding correct answers. Dot density was kept constant in half of the calibration and test stimuli, while dot size was kept constant on the other half.

Test trials began with a fixation cross presented on the center of the screen, followed by a 150ms presentation of dots. This was followed by a flicker mask and a black screen. Participants' responses were recorded via microphone if they occurred within 1s and feedback was provided. If responses were incorrect, the correct response was displayed on the screen, and if response latency was greater than 1s, a slide was presented encouraging faster responses and showing the correct answer. Participants completed four blocks of each test in each session, each block consisting of 40 experimental trials with each numerosity presented 5 times in random order. Only data from the last two blocks of each session were used for analyses.

Performance on the dots comparison task was used to calculate the internal Weber fraction, a measure of precision of the numerical representation for each participant. Based on data, the participants were divided by a median split into low and high discrimination-precision groups. These two groups did not differ significantly in terms of RT. Responding in the numerosity naming tasks was compared between the low and high discrimination precision groups.

Accuracy on the numerosity naming tasks was significantly higher in the 1-8 number range than in the 10-80 number range, and was higher for participants in the high than low precision group. Errors rates also tended to be lower for the smaller numerosities within each range. There was a significant interaction between range and rank order (number within range),

violating the prediction of similar influence of numerical magnitude across the two ranges. In the 1-8 number range, error rates were close to 0 for values 1-4, and then began to increase for values 5 and greater. However, for the range 10-80, errors were frequent even for the smaller numerosities. Additionally, participants with high precision made fewer errors than those with low precision for most numerosities in then 10-80 discrimination task but only for numerosities 5-7 in the 1-8 discrimination. No difference was found in error rates between high precision and low precision groups for numbers 1-4.

Results of response time analyses mirrored accuracy results. Response times (RTs) were significantly faster in the 1-8 range than 10-80 range, providing evidence for a size effect. High precision participants had longer response times than low precision participants for the 10-80 range only. For the 1-8 range, RTs increased from 1-5 and then stabilised. There was a significant interaction between range and rank order, suggesting differential processing of the small numbers 1-4. Response times for these values suggested they were processed much faster than for values 5-8 or any of the decade numerosities than 10-80. The distinct pattern of processing within the subitising range is contrary to predictions made by Weber's law.

For both number ranges, average response number approached sample number, and response variability increased as numerosity increased. In the 10-80 range, coefficients of variation increased early on and then began to decrease after 30, whereas in the 1-8 range, CVs were constant at about 0 until after 4, when they began to increase. This pattern is not consistent with scalar variability and is not predicted by Weber's law. Additionally, this pattern also differs from that obtained by Le Corre and Carey (2007) who found increasing CVs in their estimation task for values 1-4.

Revkin et al. (2008) concluded that their study provides evidence for a mechanism dedicated to processing small numbers that differs from the mechanism for larger number processing based on two key findings. There was a significant effect of discrimination precision on response variability; high precision participants made fewer errors over most numerosities in

the 10-80 task and outside of the subitising range in the 1-8 task, but there was no difference between high and low precision participants within the subitising range. Additionally CVs were close to 0 for values 1-4, whereas CVs for 10-40 were high, and also the significant effect of discrimination precision on response variability. However, the researchers do not take into account the size effect, or the effect of *absolute* number on representational variability; representational noise will always increase with numerical magnitude, regardless of the ratio of the values being discriminated. Thus, discrimination would be expected to be worse for values within the range 10-40 than 1-4, as numerical magnitude is considerably larger.

Based on these experiments alone, it is difficult to conclude whether there is a true separate, parallel individuation system dedicated to the processing of values less than 5. Given the wealth of research that shows no change in response patterns for values across the subitising range (e.g. Whalen, Gallistel & Gelman, 1999), it seems more reasonable to assume that a single, analogue magnitude system is sufficient to explain performance in numerical discrimination tasks.

7.1.4 Current Experiment

The main purpose of this experiment is to examine human verbal and nonverbal counting processes in bisection and discrimination tasks analogous to those conducted with pigeons reported earlier (Experiments 1 and 2A). It is unclear whether responding would conform to Weber's law; the anomalous findings of binomial variability, and bisection at the arithmetic mean obtained with pigeons in these tasks would suggest they had developed a linear scale with constant variability and were responding similarly to humans verbally counting (Cordes et al., 2001). Thus a major question was whether humans' nonverbal counting would be similar to results obtained with pigeons.

In the current experiment, humans observed a sequence of red and black pictures on a computer screen, and were required to estimate the number of red pictures presented in each

sequence. They responded by either 1) categorising the number of red pictures in the sequence as large or small by pressing one of two keys; 2) reproduce the number in keypresses; or 3) report the number directly by entering it on the keyboard.

The use of human participants allows the testing of whether the observed response variability is due to the nature of the two-key response requirement in the reproduction procedure. Similar to Whalen et al. (1999), a response condition is included where participants only have to report the number of stimuli seen, rather than reproduce that number in keypresses. If performance is similar in both the report and production conditions, it can be assumed that the same processes are underlying responding in both.

The numerical values used in this experiment ranged from 1-20, a larger range than used in the pigeon reproduction experiments. A range of this size allows the investigation of whether responding differs for small or large numbers, and may be able to identify a numerical cross-over point where response variability changes from scalar to nonscalar. Previous research makes different predictions; if subitising processes are active for very small numbers, then a change in responding should occur at approximately 4 or 5 (e.g. Feigenson et al., 2002; Revkin et al., 2008).

Participants were randomly placed into either a verbal or nonverbal counting group. Individuals in the verbal counting group were instructed to count aloud the number of red items seen with every picture presentation. In order to prevent participants in the nonverbal counting group from using covert or overt counting strategies, these participants were explicitly instructed not to count the number of red pictures and to respond as quickly as possible. These participants were also given a verbal distractor task of naming aloud each picture as it was presented. It was believed that this task would require a sufficient cognitive load to prevent any verbal counting.

Results of interest are the location of the bisection points in the bisection condition, particularly whether they will be closer to the arithmetic or geometric mean. For the production and report conditions, the relationship between sample number and the two measures, average

response number and response variability (namely the coefficients of variation) in the production and report conditions will be important. These data will allow us to ask whether average response number increases proportionally to the sample number, and whether the data show scalar or binomial variability.

7.2 Method

7.2.1 Participants

Participants were 30 male and female participants from a wide range of backgrounds, although most were university students. Ages ranged from 20 to 47, with an average of 25.97 years. All had normal colour vision. In return for their participation, each received a $20 petrol voucher.

7.2.2 Apparatus

Experiments were conducted on a Compaq Evo PC, using E-prime software. Responses were recorded on the keyboard. Stimuli were presented on a 15" LCD monitor set back approximately 50cm from where participants were seated.

7.2.3 Design

Participants were randomly assigned to either a verbal or nonverbal counting group. There were 6 participants in the verbal counting group, and 24 participants in the nonverbal counting group. All participants experienced all three response conditions; bisection, reproduction and report, which were presented in blocks in a counterbalanced order. Each block consisted of 20 trials, one trial for each target number from 1-20, inclusive. The order of presentation of each trial type was randomised.

7.2.4 Stimuli

Stimuli were red and black images of 46 common objects, which varied from 8x8cm to 10x10cm in size. Outlines and shading of stimuli were made red for the target stimuli using Microsoft Paint or Microsoft Picture Editor.

7.2.5 Procedure

Before starting the experiment, participants were provided with written instructions outlining the task demands and explaining the different response types required. These instructions were also explained verbally and any questions concerning the procedure were answered. Participants in the nonverbal counting group were asked to name each picture aloud as it was presented, and were asked to keep track of the number of red pictures presented, but not to explicitly count them. Participants in the verbal counting group were instructed to say aloud the number of red pictures they had seen at every picture presentation. Participants were not told the value or range of target numbers that would be presented. The experimenter remained in the room and monitored participants' compliance to the instructions. Any failures to adhere to the instructions were corrected with a verbal reminder of the task requirements.

At the start of the experiment, participants were presented with a screen instructing them to "Press the space bar to begin the first trial". Each trial consisted of two parts, a sample phase and response phase. Sample phases were initiated by a spacebar press, and consisted of a sequence of 40 red and black objects, each with a 500ms onset and offset. The number of red objects was equal to the target number, and these red objects were presented randomly within the sequence. At the end of stimulus presentation, participants were presented with response instructions. There were three different types of responses:

1) Bisection, where participants were required to press the "m" key on the keyboard if it was a large number or "c if it was a small number. Participants were informed at the beginning of the

experiment that 20 would be considered a large number and 1 would be considered a small number.

2) Reproduction, where participants were required to reproduce the target number (number of red objects seen) in keypresses of the "g" key. In order to prevent participants from actively counting the number of keypresses made and to increase similarity with responding of pigeons in a reproduction procedure, participants were asked to press the "g" key as quickly as possible, while saying "gee" out loud.

3) Report, where participants were required to type in the target number using the number keys located at the top of the keyboard.

After participants had made their response, they had to press the space bar to begin the next trial. Participants did not receive any feedback about their response and continued through the experiment at their own pace; they were able to stop for a break if necessary.

7.3 Results

7.3.1 Verbal condition

Due to the small number of participants in the verbal group, and the lack of variability in responding, data were plotted individually where possible.

7.3.1.1 Discrimination

For all subjects, the proportion of "large" choices increased as sample number increased. A plot of the average proportion of "large" responses as a function of number is shown in Figure 7.1. The relationship was not a perfect step function, suggesting that discrimination difficulty increased with proximity to the arithmetic mean

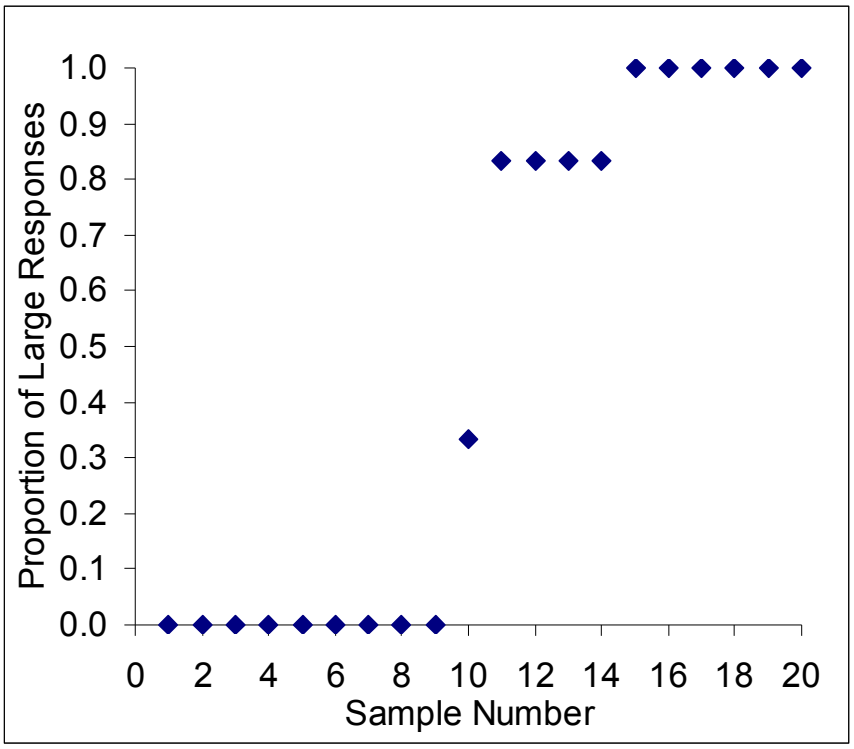

Figure 7.1. Average proportion of large responses plotted as a function of sample number.

Bisection points were calculated by fitting a logistic function to the average response data. Of interest was whether bisection points would be located closer to the geometric or arithmetic mean of the anchor values. Because participants were told that 1 and 20 were considered small and large numbers, these were used as anchor values. All participants but one switched from responding "small" to "large" at 10.5 or 9.5, suggesting that bisection of the numerical scale occurred at approximately 10. To confirm this, a logistic function was fitted to the group mean using Solver in Excel, minimising the squared error. The predicted function was a good fit to the obtained data, accounting for 98.65% of the total variance. Because participants were not explicitly trained with just two anchor values, the geometric mean was calculated using 1 and 20 as anchor values (geometric mean = 4.47) as well as the geometric mean of all the numbers used in the experiment (8.30). The arithmetic mean was 10.50. The predicted bisection point was 10.30, much closer to the arithmetic mean than both geometric mean values.

7.3.1.2 Production

The number of keypresses made in the production condition increased as sample number increased. Individual data are shown in Figure 7.2. A linear regression found a significant relationship between sample number and response number, $\beta = 0.93$, $p < .001$, $R^2 = 85.69\%$. Response number matched sample number perfectly for values 1-6, but became increasingly more variable as sample number increased from 6 to 20. The responding of participant 31 (empty squares) was the most variable, but still followed the same general pattern. A linear regression found a significant relationship between the standard deviation in response number and sample number, $\beta = 0.81$, $p < .001$, suggesting response variability increased linearly with magnitude. As variability increased with sample number, performance also decreased; a linear regression analysis found a significant negative relationship between proportion correct and sample number, $\beta = -0.44$, $p < .001$. These results suggest that participants found it harder to produce the correct number of responses as the numerical requirement increased, resulting in increased variability and poorer performance.

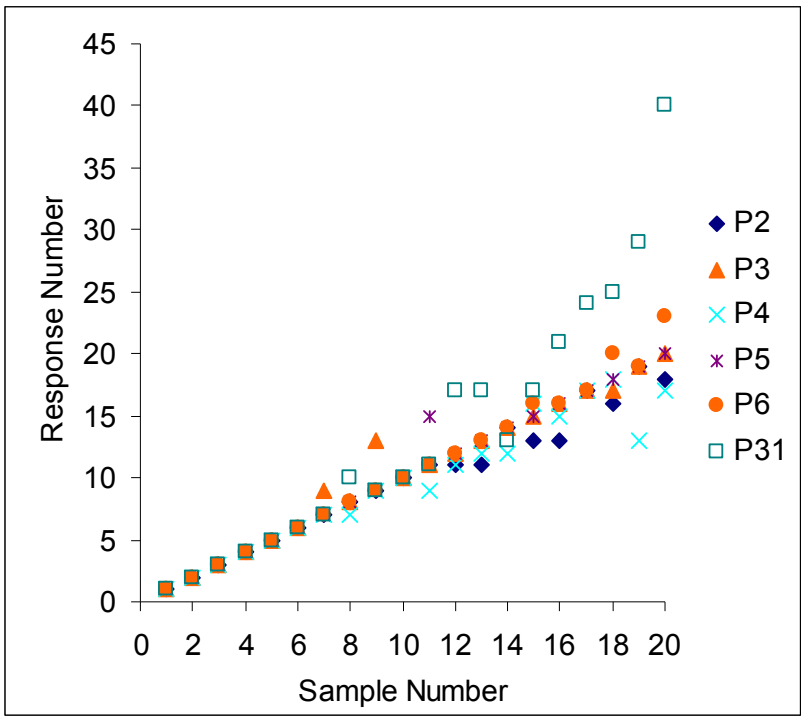

Figure 7.2. Individual response numbers produced by participants as a function of sample number in the verbal production condition.

.3.1.3 *Report*

Responding in the report condition followed a similar pattern to that in the production condition. A plot of individual response numbers is shown in Figure 7.3. It was expected that performance would be more accurate here than in the production condition, as there was a lower possibility of response error since subjects just had to type in their response. Regression analyses showed that the number of responses increased with sample number, matching it more or less perfectly, $\beta = 1.00$, $p < .001$. This was to be expected, given the easiness of the task. Performance was significantly higher in the report condition than the production condition, $t(38) = 3.32$, $p < .001$, but was still not perfect (average proportion correct = 93%, cf. 73% in the production condition). However, sample number did not appear to have any significant effect on accuracy; a linear regression analysis found no significant relationship between these variables, $\beta = -0.14$, $p = .11$. Thus, incorrect responses made in the report condition did not appear to be related to number. Additionally, no significant relationship between response number standard deviation and sample number was found, $\beta = 0.31$, $p = 0.17$, suggesting response variability was not systematically affected by number.

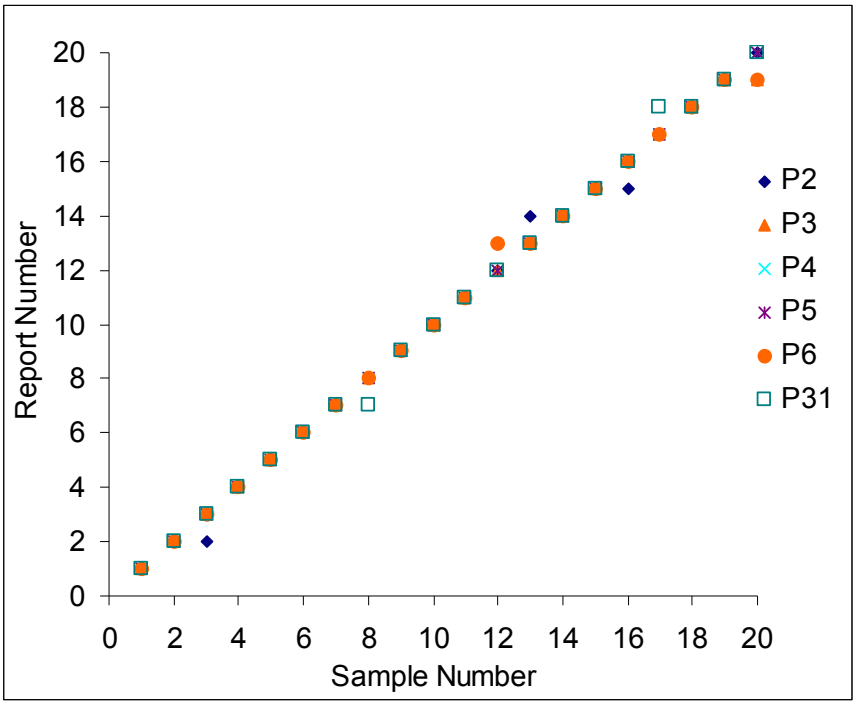

Figure 7.3. Individual report numbers plotted as a function of sample number in the verbal report condition

7.3.2. Nonverbal Group

Data from 3 participants were excluded for analyses in the nonverbal group, due to their failure to comply with instructions. This appeared to be limited to particular response types, and consequently their data was only removed from the affected conditions. Data from one participant were excluded from the discrimination condition, two participants' data were excluded from the production condition, and data from one of those were also excluded from the report condition.

7.3.2.1 Discrimination

Not surprisingly, performance of the nonverbal group was less accurate than the verbal group. The overall pattern of responding was similar, however the proportion of "large" responses increased as sample number increased. A plot of average response data for the discrimination for the groups that experienced the discrimination condition first, second or third and averaged across all groups can be seen in the left and right panels of Figure 7.4, respectively.

To test for bisection point location, logistic functions were fitted to the average data of the groups that experienced the discrimination condition first, second, or third (D1, D2 and D3, respectively), in the same manner as for the verbal condition. Logistic functions provided a good fit of the data, accounting for 83%, 89% and 93% of the data for the D1, D2 and D3 groups respectively. Bisection points for groups D1, D2 and D3 were 12.95, 11.49 and 9.79, respectively; all were located closer to the arithmetic mean (10.5) than the larger geometric mean value (8.3). Results of *t* tests showed that whereas the average bisection points did not differ significantly from 10.5, $t(2) = 0.99$, *n.s*, the difference between the bisection points and the geometric mean approached significance, $t(2) = 3.41$, $p = 0.076$.

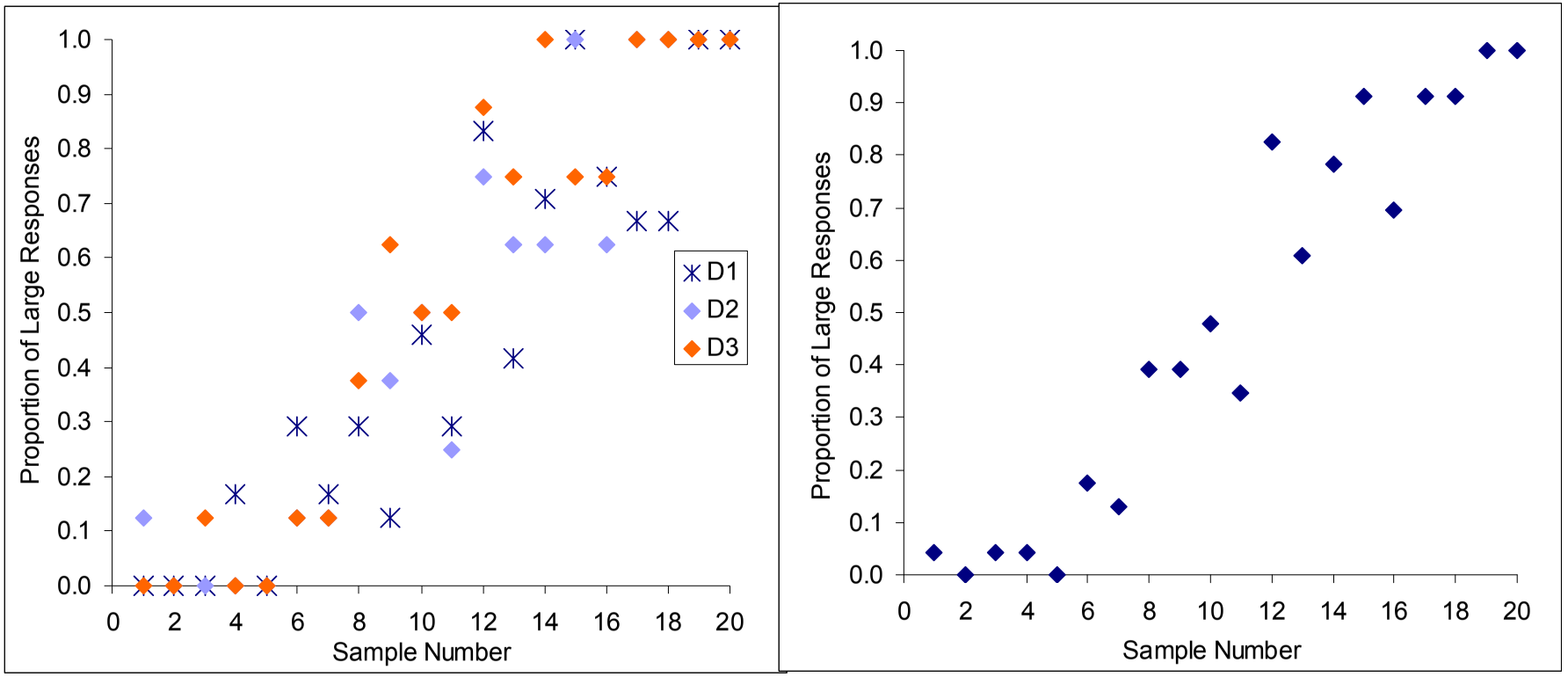

Figure 7.4. Average proportion of large responses plotted as a function of sample number for groups that experienced the discrimination condition first (D1), second (D2) or third (D3) shown in the left panel, and averaged across all groups in the right panel..

7.3.2.2 Production

The numbers of responses in the nonverbal production condition tended to increase as a function of number for all groups, regardless of order. A repeated-measures ANOVA on average response number found a significant effect of number, $F(19,361) = 45.33$, $p < .001$, but no significant effect of order $F(2,19) = 0.77$, *n.s*, or interaction, $F(38,361) = 0.77$, *n.s.* Because no significant effect of order was obtained, data from all groups were collated for consequent analyses of response number. A plot of the average response number for all the participants can be seen in Figure 7.5. Response number tended to increase equivalently with sample number up to 5 or 6, but beyond that point sample number was generally underestimated. A trend analysis showed response number increased as a linear function of sample number, $F(1,19) = 68.81$, $p < .001$. A linear regression analysis revealed a significant positive relationship with slope of $\beta = 0.69$, $p < .001$, confirming that the ratio of response to sample number was less than 1:1.

Proportion of correct responses tended to decrease with sample number, and a significant correlation between these variables was obtained, $r = -.15$, $p < .001$. These are plotted in Figure 7.6. Average performance never exceeded 85%, and for values greater than approximately 7 or 8, accuracy reached a plateau at about 15%. A repeated measures ANOVA found a significant effect of number, $F(19,361) = 7.90$, $p < .001$, but no significant effect of order, $F(2,19) = 0.10$, *n.s.* or nuber/order interaction, $F(38,361) = 1.25$, *n.s.* A trend analysis showed accuracy decreased linearly as sample number increased, $F(1,19) = 64.81$, $p < .001$.

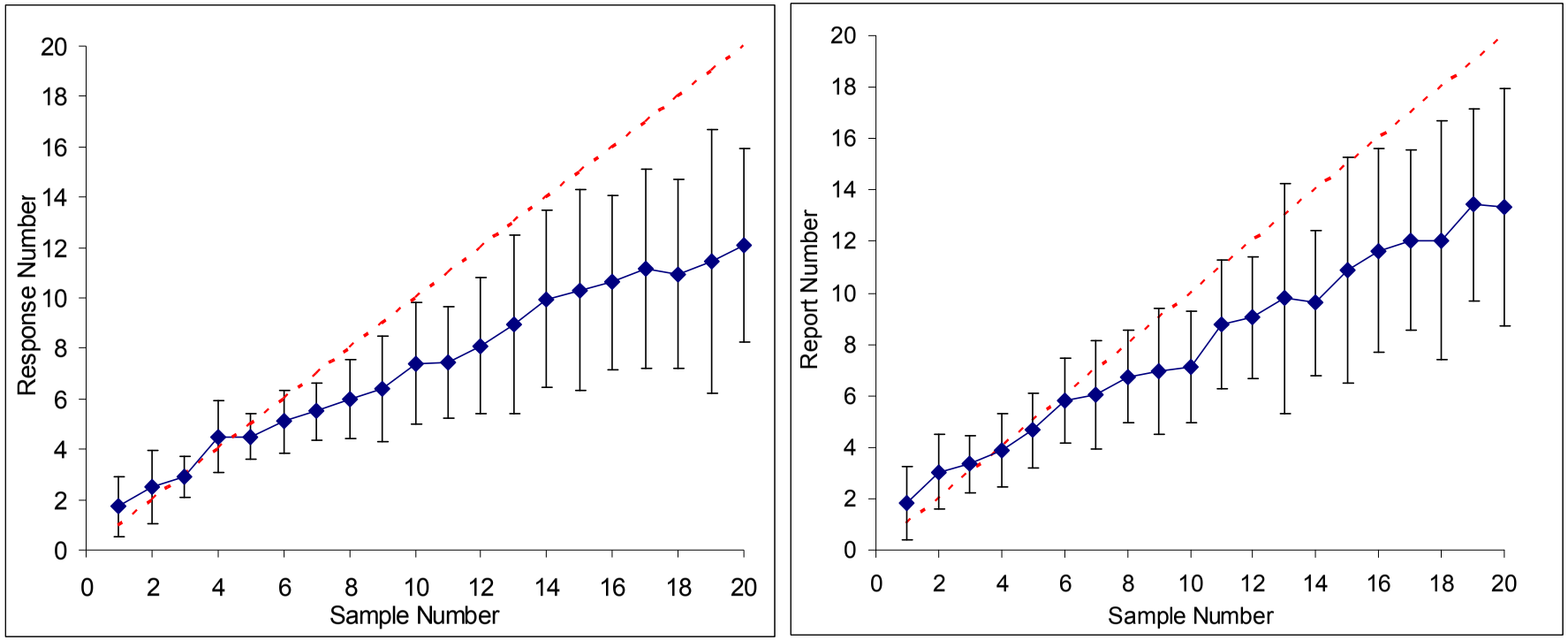

Figure 7.5. Average response number plotted as a function of sample number in the production condition (left panel) and the report condition (right panel). Error bars show + 1 S.D.

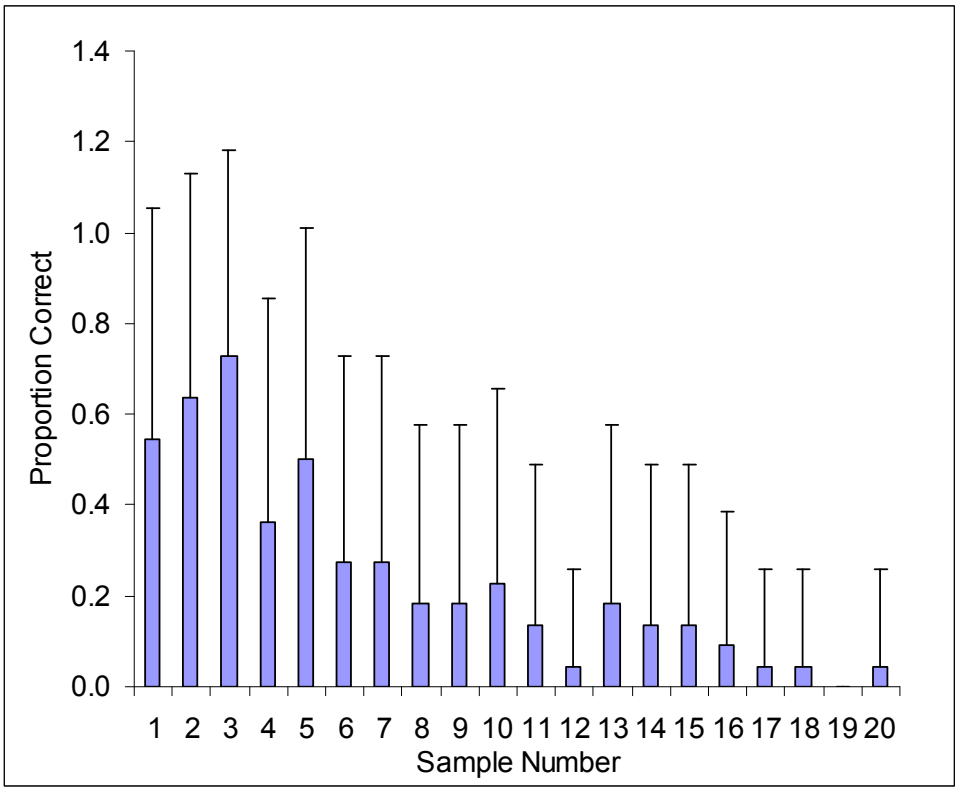

Figure 7.6. Average proportion correct plotted as a function of sample number for the production condition. Error bars show + 1 S.D.

Coefficients of variation were calculated by dividing the standard deviation by the mean response number, and overall group data are plotted against average response number on a log-log scale, in Figure 7.7. As no effects of order on responding had been found in previous analyses, only group mean data were examined. Two distinct patterns can be seen in the data; CVs clearly decrease for the first 7 or 8 values, and then increase slightly for larger values.

Polynomial regression analyses were conducted to calculate the fit of linear and quadratic models to the log CVs and log average response number data. A linear model did not provide a good account of the data, $F(1,18) = 0.92$, $p = 0.35$, whereas a quadratic model did, $F(2,17) = 15.07$, $p < .001$. This confirms that there were two distinct slopes in the data.

To test this further and to identify the inflection point where the slopes changed from negative to positive, a bilinear "broken stick" model was fit to the data and its performance compared with a linear model. Using the log CVs and the log average response number, slopes and intercepts and the variance accounted for (VAC) for both the linear and bilinear models were

calculated using the least squares method, using Solver to minimise the sum of squared deviations of both the models. The slope and intercept obtained for the linear model was -0.11 and -0.38, respectively. This model performed poorly, only accounting for 4.85% of the total variance. Conversely, the bilinear model was able to account for 74.35% of the total variance, with slopes equaling -1.02 and 0.35 and intercepts -0.07, and -0.80 for values less than and greater than the inflection point, respectively. These results were obtained with the inflection point located at 0.75, 0.76 and 0.77 log average number of responses. These inflection points correspond to a flash number of 7 or 8. The fit of the bilinear function, with an inflection point of 0.75 is plotted as a red line in Figure 7.7.

To examine whether the obtained slopes in the bilinear model were significantly different from 0, linear regression analyses were used to calculate the slopes for first 7 log CV values (equivalent to flash numbers 1-7), and remaining log CV values. The slope of the first 7 values was significant and negative, $\beta = -0.90$, $p < .01$, $R^2 = 0.80$. A linear function fitted to log CV values for values larger than 7 revealed a significant positive relationship with log average response number, $\beta = 0.59$, $p < .05$, $R^2 = 0.35$. This suggests that there may be two different processes operating over different parts of the numerical range tested.

7.3.2.3 Report

The number reported by participants in the report condition increased with sample number in a similar manner to the production condition. Plots of average report number can be seen in Figure 7.5. Response number was largely equivalent to sample number for values up to approximately 6, and for values larger than 6, sample number was consistently underestimated. A repeated-measures ANOVA showed a significant effect of number on responding, $F(19,380) = 48.60$, $p < .001$, but no significant effect of order, $F(2,20) = 0.19$, *n.s.*, and no significant interaction $F(38,380) = 0.75$, *n.s.* The slope of the function when response number was plotted against sample number was steeper than in the production condition, $\beta = 0.76$, $p < .001$.

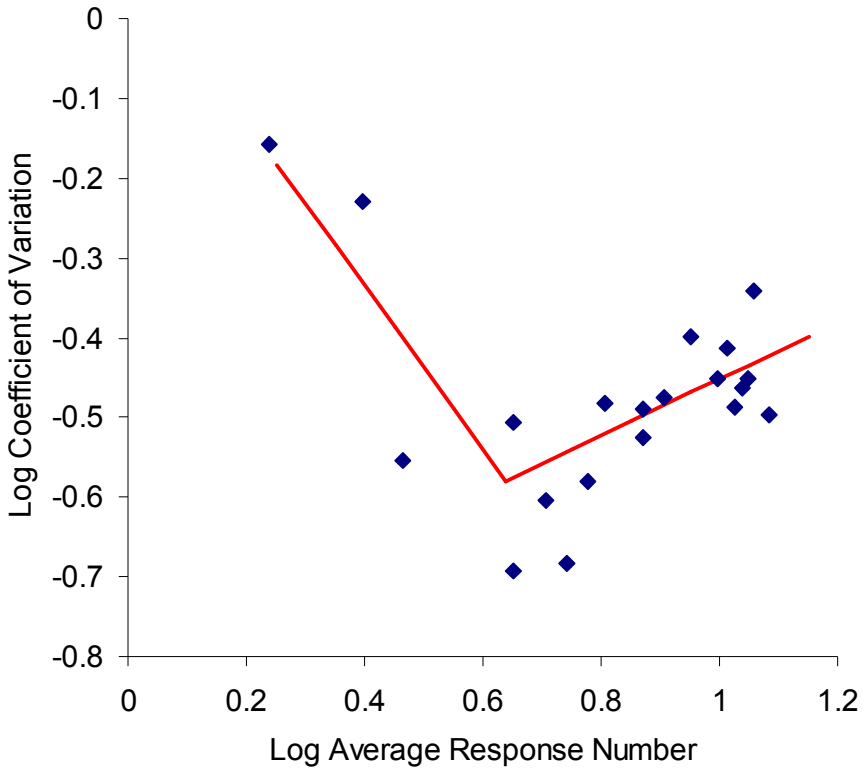

Figure 7.7. Log coefficients of variation (standard deviation of responding/average response number) plotted as a function of log average response number for the production condition. Red line shows fit of bilinear function.

Interestingly, participants tended to report larger numbers in this condition than in the production condition; a repeated measures ANOVA showed that this difference in response number was significant, $F(1,21) = 13.24$, $p < .005$. No significant interaction between number and response type was found, $F(19,399) = 1.27$, *n.s.* This would account for the more accurate responding seen in this condition relative to the production condition.

A plot of average proportion correct as a function of sample number can be seen in Figure 7.8. Proportion of correct trials decreased as flash number increased. A repeated measures ANOVA revealed a significant effect of number $(19, 380) = 3.78$, $p < .001$, and a significant linear trend, $F(1,20) = 41.97$, $p < .01$. There was no significant effect of order on proportion correct, $F(2,20) = 0.81$, *n.s.*, and no interaction, $F(38,380) = 0.89$, *n.s.* There was no significant difference in overall proportion correct in the production and report conditions, $t(19) = 0.36$, *n.s.*

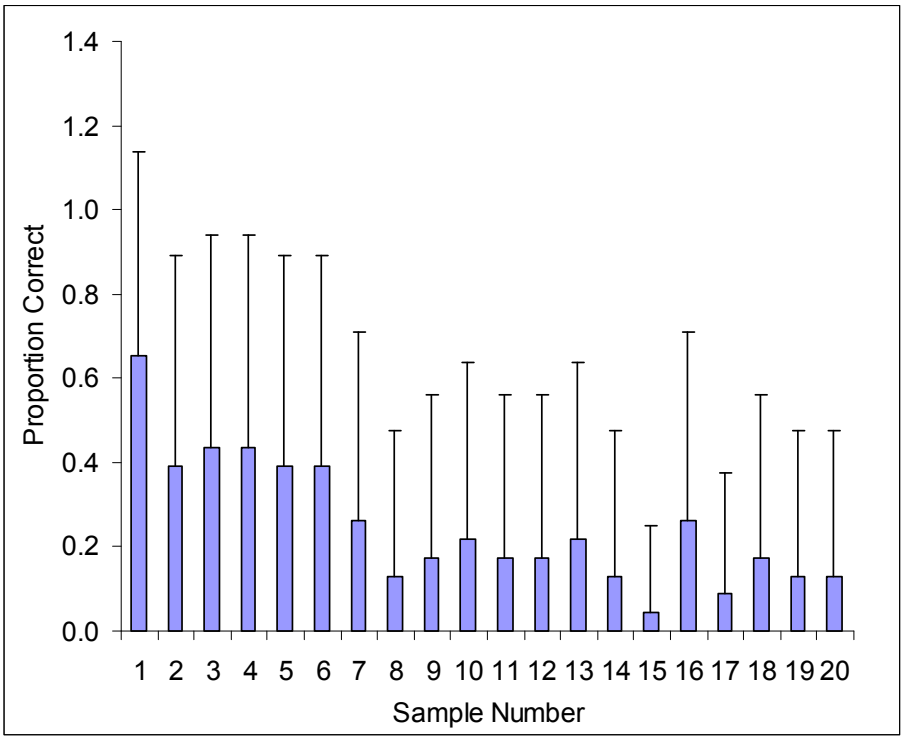

Figure 7.8. Average proportion correct for all participants in the report condition. Error bars show + 1 S.D.

Coefficients of variations for the report condition are plotted in Figure 7.9. CVs decreased for the first 7 or 8 values, but did not vary systematically for larger values. Data were analysed in the same manner as production data. Results of polynomial regression analyses showed a significant linear relationship between log CVs and log average response number, $F(1,18) = 7.65$, $p < .05$, as well as a significant quadratic relationship, $F(2,17) = 14.81$, $p < .001$. When a broken stick model was fitted to the data, the bilinear function outperformed the linear function, accounting for 66.93% and 29.82% of variance, respectively. The slope and intercept of the best fitting linear function was -0.24 and -0.25 respectively. For the bilinear function, the best fitting slopes were -0.96 and 0.01, and intercepts of 0.13 and -0.50 for values less than and greater than an inflection point of 0.5, respectively. An inflection point of 0.5 corresponded to a flash number of approximately 3. Note that assuming a greater inflection point within a range of 0.6-0.65, or a flash number of approximately 4-5, obtained similar results; VAC of 66.17%, slopes of -1.09 and 0.10, and intercepts of 0.17 and -0.59 for values less and greater than the inflection point, respectively.

Regression analyses of the log CV values and log average response number revealed a negative slope of β = -0.96 for the first 3 values, and a slope of β= 0.06, for the larger values. However neither was significant. When an inflection located at a flash number of 5 was assumed, a significant negative slope of β = -0.95, $p < .05$ was obtained for flash number values 1-5 and no significant slope for values greater than 5, β = 0.21, *n.s.* Thus, CV patterns in the report and production condition were similar in that they were best accounted for by a bilinear, rather than linear function with decreasing CVs for values less than the inflection point. However, the CVs also differed in several ways; a lower inflection point was obtained in the report condition, and the slopes of CVs for average response numbers greater than the inflection point were constant in the report condition, rather than increasing.

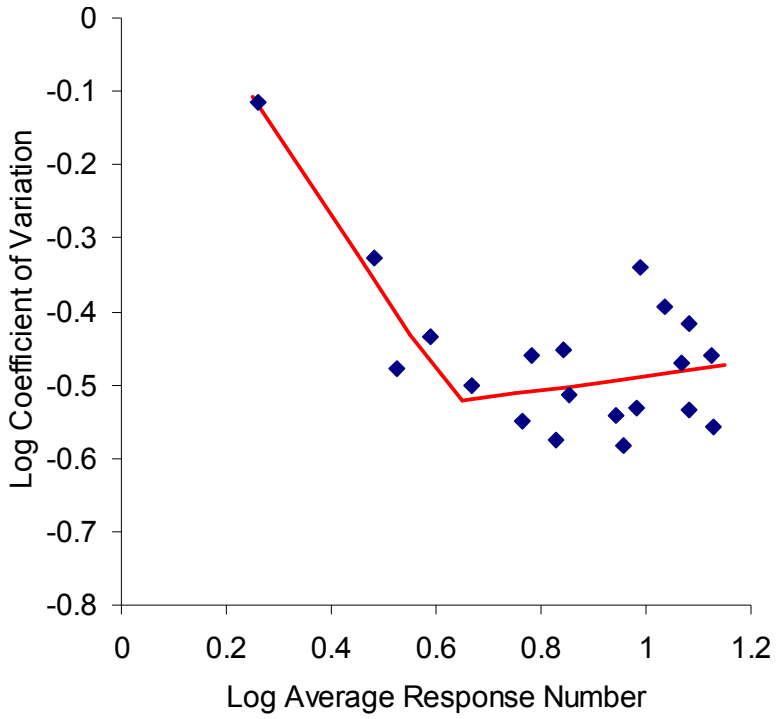

Figure 7.9. Log coefficients of variation (standard deviation of responding/average response number) plotted as a function of log average response number for the report condition. Red line shows fit of bilinear function.

Pigeon performance and people performance

To compare performance of participants in the nonverbal condition of this experiment to that of pigeons trained in previous experiments (Experiment 2A and 3), response data for the

numerical values common to the human and pigeon experiments were collated and compared qualitatively, or quantitatively as appropriate.

Below in Figure 7.10 are the psychometric plots for the pigeons in Experiment 1, plotted on a relative scale, and psychometric plots for the participants in the nonverbal discrimination condition. Although different numerical ranges were used in these two experiments, the general forms of the psychometric plots are similar. In both experiments, bisection points were located at the arithmetic, not geometric mean.

Average response number is plotted in the left panel Figure 7.11. For both the pigeons and people in the nonverbal production and report conditions, average response number increased with sample number. Human participants seemed to be more sensitive to number; relatively, pigeons tended to overestimate smaller numbers and underestimate larger numbers. Consistent with this, a repeated measures ANOVA found no significant effect of species, $F(2,6) = 2.09$, *n.s.*, but obtain a significant interaction, $F(12,36) = 6.43$, $p < .001$.

Coefficients of variation calculated for the pigeons were similar to those obtained with the human data. Plots of log CV as a function of log average response number for both pigeons and human are shown in the right panel of Figure 7.12. To compare CVs across species, regression analyses were used to calculate slopes and intercepts. The slope for the average log CVs for pigeons was -0.94, $p < .01$, with an intercept of -0.11, $p < .05$. The slopes for the first seven log CV values were -0.895, $p < .01$, and -0.89, $p < .01$, for the production and report data, respectively. Both intercepts did not differ significantly from 0. The log CVs obtained for pigeons were regressed onto both the human report and production log CVs to quantify the strength of their relationship, and found significant positive relationships between both pigeon and human report CVs , $\beta = 0.78$, $p<.05$, $R^2 = .61$ and pigeon and human production CVs, $\beta = 0.89$, $p<.01$, $R^2 = .78$, respectively. These findings suggest that although pigeons showed greater compression in the response number scale than human participants, relative variability patterns for both species showed similar slopes and were strongly correlated.

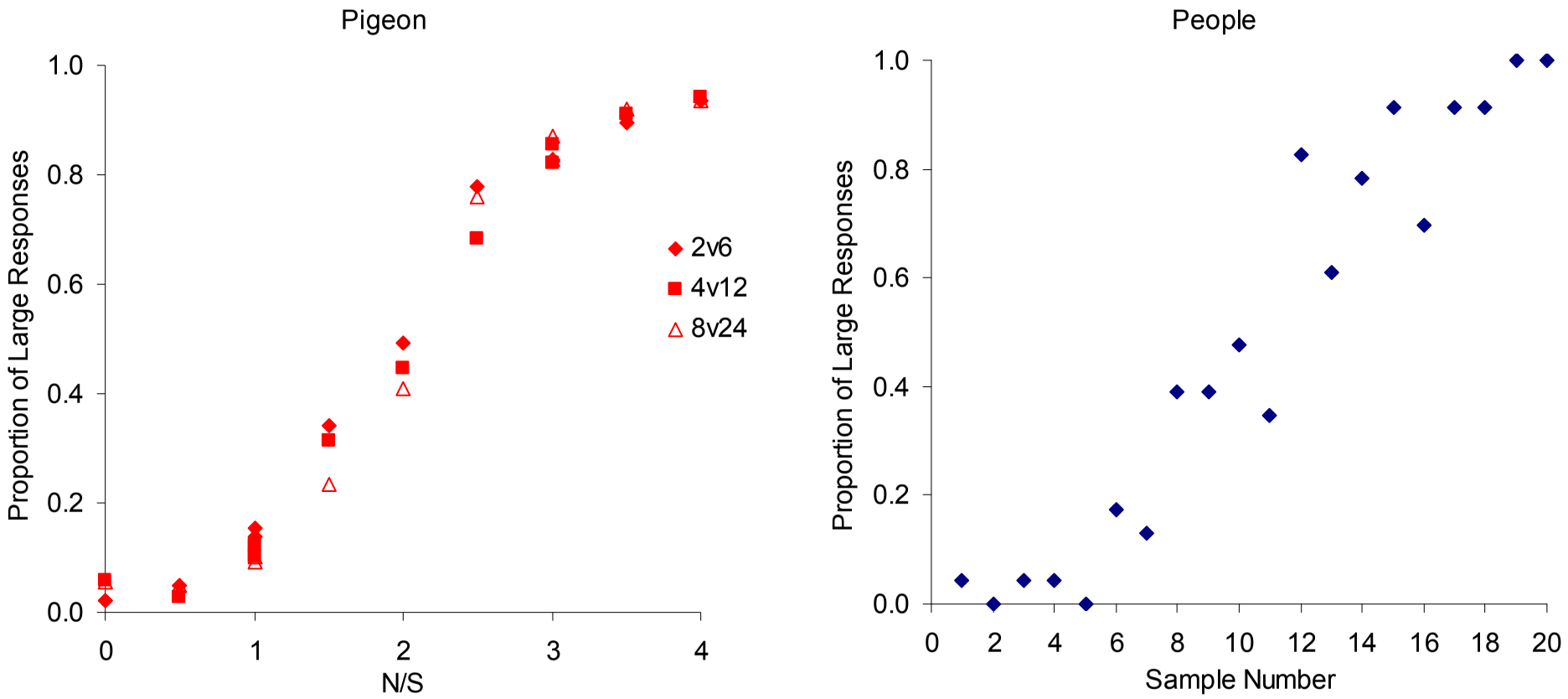

Figure 7.10. Proportion of large responses plotted as a function of sample number, on a relative scale for pigeons in Experiment 1 (left panel), and obtained for humans in the nonverbal condition in the current experiment (right panel).

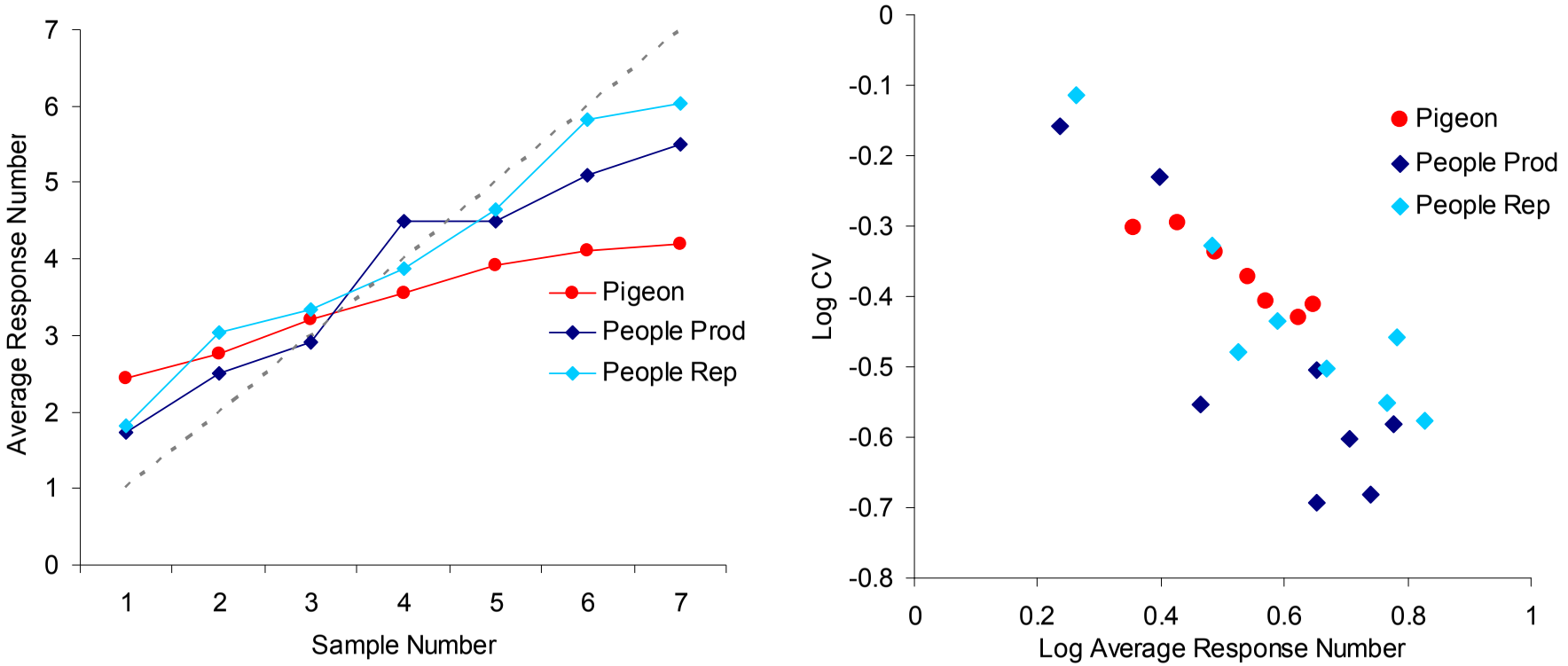

Figure 7.11. Left panel shows average response number obtained for pigeons (red series), humans in the nonverbal production (dark blue series) and nonverbal report (light blue series) conditions. Right panel shows log CVs plotted as a function of log response number for the same groups.

7.4 Discussion

7.4.1 Human performance

Results from this experiment showed that humans are able to respond on the basis of number in numerical bisection, production and report tasks using both verbal and nonverbal counting. Additionally, characteristics of responding in all three tasks suggest discrimination of the values 1-7, at least, were based on a linear scale of number, with constant generalisation between values.

Our nonverbal manipulation was effective in preventing covert and overt counting strategies. Anecdotal evidence suggests the task of naming the pictures was sufficiently demanding of mental resources; often participants would have trouble merely keeping up with the rate of stimulus presentation, and the difference in performance in the verbal and nonverbal conditions is evident in the obtained data.

As expected, responding in the verbal counting condition was precise, with consistently high accuracy and as a consequence, low variability. Responding in all three response conditions covaried systematically with number, with larger response numbers increasing with sample number. Variability also covaried systematically with number in two of the response conditions. In the discrimination condition, accuracy decreased with increasing proximity to the subjective midpoint, suggesting the discrimination of the "correct" categorisation response became more difficult with increasing distance from the anchor values. In the production condition, although performance was perfect for values 1-6, response variability tended to increase as numerical magnitude increased. This pattern was not evident in the report condition; errors and variability did not vary systematically with number, suggesting that variability was due to errors in the response number discrimination, rather than the discrimination of the number of red objects.

There were similarities and differences among the different response types in verbal and nonverbal conditions. Bisection points in both verbal and nonverbal discrimination tasks were located closer to the arithmetic than geometric mean, consistent with Droit-Volet et al. (2003),

and Experiment 1 of this dissertation. This suggests that participants were not using a numerical scale with scalar characteristics that conforms to Weber's law, but rather a linear number scale with constant variability. It is possible that the sequential presentation of the stimuli may be responsible for this finding, as acknowledged by Jordan and Brannon (2006), and further research using stimuli presented simultaneously is necessary to investigate this further.

Responding in the verbal and nonverbal production and report conditions shared some key features: 1) average response number increased with sample number; and 2) with the exception of the verbal report condition, response variability also increased with sample number and consequently proportion correct decreased as sample number increased. The strong correlation between performance in the production and report conditions is consistent with Whalen, Gallistel and Gelman (1999). These results show that participants generally were able to discriminate the number of red stimuli both verbally and nonverbally and could successfully reproduce that number in keypresses or report it.

However, there were some critical differences in performance between the verbal and nonverbal conditions. In both the nonverbal numerical production and report tasks, for values smaller than 6 average response number was equivalent to sample number, whereas for values larger than 6, average response number increasingly underestimated sample number. This pattern was much more marked in the production than report task. It is unclear what is responsible for this effect, although it appears to be specific to nonverbal numerical processing. Although there was a similar increase in response variability in the verbal production condition, there did not appear to be any systematic tendency to produce fewer numbers of responses. The underestimation may be a result of the compression of the nonverbal numerical scale, or memorial decay, which is more likely when nonverbal numerical processes are used. Memory effects can be tested directly by the manipulation of the retention interval separating the sample and response phases, as memorical decay which should be correlated with the delay between stimulus and response. Additionally, using simultaneously presented stimuli would avoid the

issue of decay that is encountered with sequentially presented stimuli; if similar patterns are found regardless of the length of the retention interval, or the mode of presentation, then the role of memory effects in underestimation can be ruled out.

Another important difference in performance between the verbal and nonverbal conditions was the differences in relative variability. In both the nonverbal report and production conditions, two trends emerged. For relatively small numbers (i.e., values less than 5 and 7), decreasing coefficients of variation were obtained, consistent with previous experiments (2, 3, 4) and suggested binomial variability: As number increased, relative variability decreased. However, for values larger than 5 or 7, in the report and production conditions respectively, coefficients of variation were constant or increasing, suggesting relative variability was at least proportional to or increasing with flash number.

The difference in CV patterns in the report and production conditions is likely due to differences in response requirements in the two conditions. Two sources of variability are present in the production condition; as well as the discrimination of the sample number, variability is also introduced in the discrimination of the number of responses generated during the response phase. Relatedly, participants showed a greater tendency to underestimate sample number when responding in the production condition than in the report condition, and this is most likely responsible for the increasing pattern of CVs observed.

The finding of multiple variability patterns in responding has considerable implications for the structure of the underlying numerical representation developed in this procedure. The finding of binomial variability suggests that for values 1-7, at least, responding is based on a linear scale with constant generalisation between values, rather than a logarithmic scale with constant generalisation or a linear scale with scalar generalisation, both of which would predict scalar variability. The latter two scales are supported by response data for values larger than 8, which show scalar variability.

The decreasing, then scalar response variability is partially consistent with previous

research using Mechner-type procedure with nonhuman animals (Machado & Rodrigues, 2008), although decreasing CVs were reported for a larger range, up to approximately 10. Consequently, it would be reasonable to hypothesise that the decreasing CVs may be a result of the structure of the response phase, however the fact that this result was found in both the production and report conditions provides evidence that the changes in relative response variability are not just an artefact of the response production task, but rather reflect a difference in numerical processing across different ranges.

The finding of these different variability patterns presents a contrast with human research. The change in response variability from binomial to scalar did not occur for values for 1-4 and greater than 4, which would be predicted if subjects were subitising the smaller values (see Feigenson et al, 2004; Le Corre & Carey, 2007). Consequently these data provide evidence against the object-file, perceptual-memory based hypotheses, although these hypotheses may be limited to simultaneously-presented visual stimuli only.

These results partially support previous human research, which has found evidence for multiple representations that change and develop with age and experience (Siegler & Opfer, 2003; Siegler & Booth, 2004). Siegler and Booth (2004) predict the use of a linear scale (namely, a scale conforming to Weber's law) primarily for the discrimination of familiar, small numbers and a logarithmic scale with larger numbers, which are only represented inexactly. The response variability observed in this experiment is in line with this.

The present results are also consistent with Feigenson et al.'s (2004) hypothesis that two systems are used for numerical representation; two systems, one for the precise discrimination of small numbers, and another for the approximate discrimination of larger numbers. It is possible that previous research has failed to provide evidence for both within the same task due to the use of numerical values that are located in the range of only one of these systems, and not both (e.g. Whalen, et al., 1999, only used values greater than 7). Additionally, numerical values are often spaced quite far apart, instead of being consecutive numbers, which may reduce the resolution of

variability analyses. Consequently, future research needs to examine wider ranges of numbers with less distance between values in order to investigate changes in variability and numerical processes as a function of numerical magnitude properly.

From visual inspection of data from the verbal production condition, it would appear that similar variability patterns are also present when participants were required to nonverbally reproduce numbers that were verbally counted. As can be seen in the plot of average standard deviations for the verbal and nonverbal production and report conditions in Figure 7.12, no variability in responding is seen for values 1-6, but for values larger than 6, response variability appears to increase proportionally with number. Unfortunately, due to the lack of variability in this condition, it is not possible to test this statistically. However, the present results do suggest that this pattern is not solely limited to nonverbal numerical discriminations.

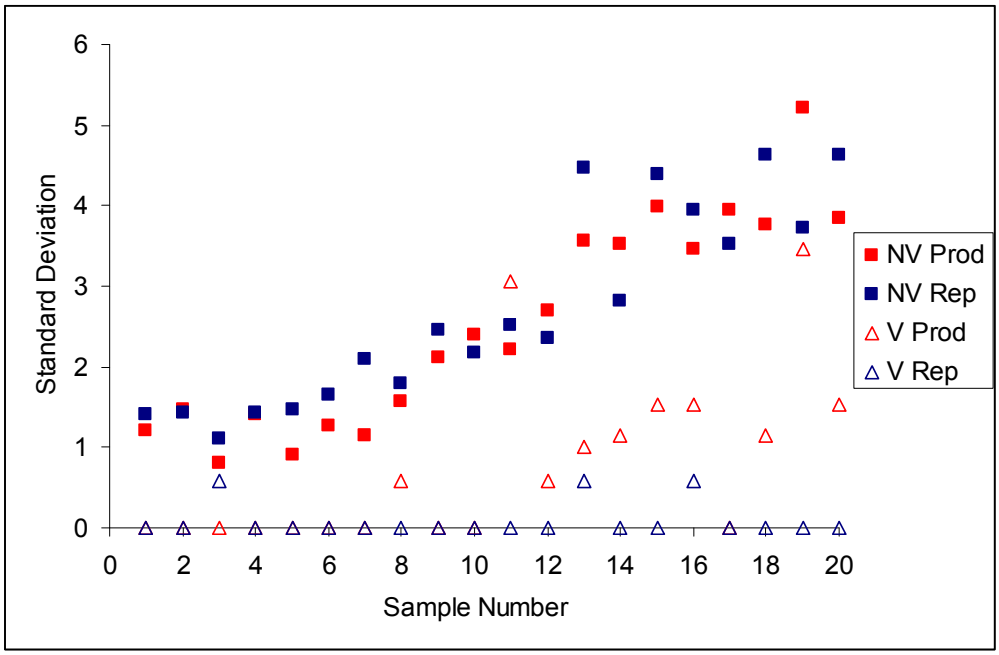

Figure 7.12. Average standard deviations for the verbal and nonverbal conditions.

It would be possible to differentiate which response characteristics are due to variability in the nonverbal discrimination of stimulus number and which are due to variability in the nonverbal discrimination of response number by comparing responding in the verbal production and nonverbal report conditions. The plot of average response number for these two conditions

is plotted in Figure 7.13 below. The underestimation of sample number appears to be specific to the nonverbal discrimination of stimulus number, suggesting it affects processing at the stimulus presentation stage. As can be seen in Figure 7.12 above, the proportional increase in response variability appears to be present with any nonverbal discrimination of number, with standard deviations in the verbal production condition showing the same pattern as the nonverbal production and report conditions, although shifted downward vertically. Thus, it appears that effects of variability in stimulus and response number discrimination are additive.

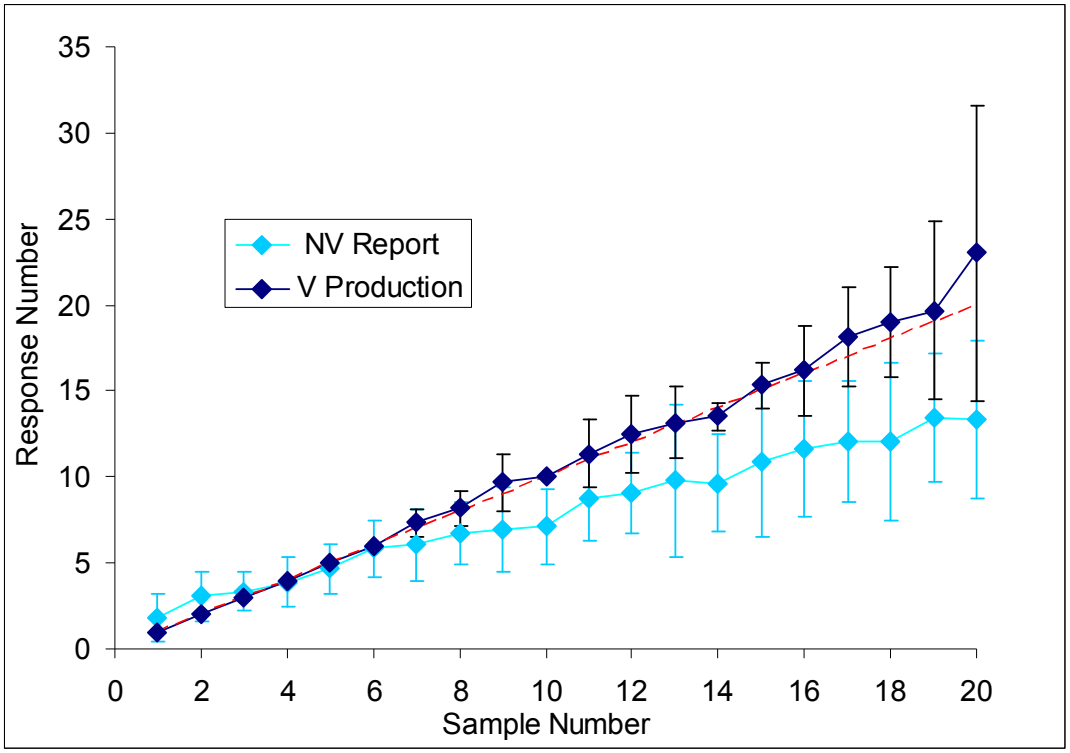

Figure 7.13. Average response numbers in the nonverbal report and verbal production condition. Errors bars show ± 1 S.E.

7.4.2 Pigeon vs. people performance

There were strong parallels in performance of pigeons and human participants in these tasks, hinting at possible similar processes underlying their performance.

Logistic functions provided good fits of the data of both people and pigeons in the discrimination task, and notably, bisection points were all located closer to the arithmetic than geometric mean in all experimental conditions, regardless of species and whether processing was

verbal or nonverbal. This would suggest responding was based on a representation of number that was linearly spaced, with constant generalization between adjacent values.

Human responding in both the production and report conditions resembled that obtained with pigeons in the same numerical range. Average response number increased linearly with sample number in both species, although pigeons appeared to be less sensitive to number shower greater overestimation of smaller values and underestimation of larger values. This may be due to the greater experience of human participants in number-related tasks, and also the explicit instructions they received to respond on the basis of number, which has been shown to increase numerical sensitivity (Droit-Volet et al,. 2003; Roitman et al., 2007). Thus, it is possible that if experimental conditions were changed such that participants learnt task requirements through trial and error, in a manner more similar to pigeons in the reproduction procedure, responding would show similar sensitivity to number.

One clear consistent finding across species is that of bisection at the arithmetic mean and strongly decreasing relative variability for values 1-7 in production and report tasks, which suggest a linear scale with constant generalisation between values. This is especially noteworthy due to its inconsistency with the majority of previous research on nonverbal numerical discriminations with both nonhumans (e.g. Meck & Church, 1983; Fetterman, 1993; Emmerton & Renner, 2006) and humans (e.g. Whalen, et al., 1999, Cordes et al., 2001), which have primarily found scalar variability; bisection at the geometric mean and constant CV. across different numbers. It is unclear what is responsible for this anomalous finding. The decreasing response variability does not appear to be an artefact of the (re)production task, since similar patterns are also obtained when participants are only required to report, rather than reproduce the sample number. It does seem to be related to nonverbal discrimination of data with the pattern shown in responding in the verbal production, nonverbal production and report conditions, but not the verbal report condition.

One common feature of previous experiments that have previously found decreasing

relative variability is the use of sequential stimuli. Droit-Volet et al. (2003) used sequential stimuli in their bisection task, and similarly, the discrimination of the number of responses constituting a response run in the Mechner FCN procedure (Mechner, 1958, Machado & Rodrigues, 2007) can also be considered the discrimination of sequential stimuli. Jordan & Brannon (2006) suggested that sequential stimuli in Droit-Volet et al., (2003) may promote the development of a more serial, linear-based representation of number and consequently results in responding becomes relatively more accurate as number increases, instead of responding with constant relative variability that conforms to Weber's law. Consistent with this notion, is the finding of scalar variability in similar production and response tasks used by Whalen et al. (1999), who used Arabic numerals as indicators of response requirement, and other studies which have found scalar variability using simultaneously presented stimuli (e.g. Feigenson et al., 2004; Beran & Rumbaugh, 2001; Emmerton & Renner, 2006). However, responding conforming to Weber's law has also been found in several studies with sequential stimuli (e.g. Meck & Church, 1983; Fetterman, 1993; Roberts, 2005; Boisvert, Abroms, Roberts, 2005), so sequential stimulus presentation cannot be the sole explanation for these response variability findings.

More research needs to be conducted to further investigate influences on response variability. The numerical reproduction experiments conducted with pigeons reported previously only used values up to 7, and so it is unclear whether the same CV patterns found for numbers greater than 7 with humans would also be obtained with nonhumans. Additional experiments manipulating the method of stimulus presentation and the influence of correction trials would be worthwhile in teasing out their effects on response variability.

In summary, the findings of this experiment provide strong evidence for nonverbal discrimination of both absolute and relative numerosity in human participants that parallels performance of pigeons in analogous tasks. Analyses of response variability suggest there may be two separate processes or representations operating across different ranges, producing binomial variability for values less than 8 and scalar variability for values greater than 8. This is

consistent with previous research (e.g. Sielger & Booth, 2004) which has found that humans are able to selectively use both logarithmic and linear number scales within the same task, depending on the context. Further research would be valuable in further elucidating the processes behind both human and nonhuman performance in tasks requiring the discrimination of absolute and relative numerosity.

8 Chapter 8: General Discussion

The primary aims of the current research were to investigate nonverbal numerical discrimination in both humans and nonhuman animals, to explore factors which influence performance, and possibly to elucidate the underlying representation and processes that allow the discrimination of relative and absolute number. This chapter will summarise and integrate the key findings of the six experiments described in this thesis with respect to the major theories in the field.

8.1 Did they discriminate number?

A considerable amount of research has investigated numerical abilities in nonhumans - whether, and if so, to what extent, they are able to discriminate relative and absolute numerosity. However, one recurring issue is that of experimental confounds; the covariation of other stimulus characteristics with number has often made it difficult to isolate, identify and quantify the degree of numerical control over responding relative to other cues. This has been sufficiently problematic that it has been proposed that numerical cues are only used as a "last-resort"; animals do not attend to number unless there are no other reliable cues on which to base responding (Davis & Memmott, 1982; Breukelaar & Dalrymple-Alford, 1998; Seron & Pesenti, 2001). One persistent confound, which applies specifically to sequentially presented stimuli, is that between number and temporal variables; unless otherwise controlled for or randomised, sample duration and flash rate will covary with number. This is particularly problematic because temporal cues have been shown to be at least as salient as, if not more than, numerical cues (Roberts & Mitchell, 1994; Breukelaar & Dalrymple-Alford, 1998; Roberts, 2005), and it has also been proposed that numerical and temporal information is processed by the same mechanism (Meck & Church, 1983). If time and number are highly correlated, then it would be difficult to distinguish between responding based on one or both of these variables.

Researchers have generally dealt with these problems in a variety of ways. One approach involves training subjects with two sets of either number-relevant or time-relevant stimuli (e.g. Meck and Church, 1983), and then testing them with stimuli where time or number is held constant while the other varies. Numerical and temporal control can be assessed by the examination of responding to transfer test stimuli. Alternatively, researchers have varied confounding stimuli systematically with number (e.g. Meck & Church, 1983; Breukelaar & Dalrymple-Alford, 1998; Cantlon & Brannon, 2007), or randomised confounding stimuli so that there is only a weak relationship with number (e.g. Emmerton & Renner, 2006; Tan et al., 2007). However, after these manipulations, relatively few researchers have quantitatively tested the effectiveness of their manipulations in decreasing control by confounding/extraneous cues over responding. A notable exception is Fetterman (1993), who used regression analyses to investigate the relative contributions of numerical and temporal variables in accounting for variability in responding.

Experiment 2 and 2A were conducted specifically to examine the control by time and number over responding in a numerical reproduction task when flash rate or sample phase duration was perfectly correlated with number (Exp 2) and when temporal variables were unreliable cues for responding (Exp 2A). Pigeons were trained to make 2, 4, or 6 responses on a center key and an additional completion response on a right key, following the presentation of 2, 4, or 6 response-dependent flashes, respectively. After baseline training, subjects experienced transfer tests with novel numbers of flashes (1, 3, 5 & 7).

These experiments extend research previously conducted with FCN procedures (Mechner, 1958; Platt & Johnson, 1971, Machado & Rodrigues, 2007), which have demonstrated that nonhuman animals are able to discriminate three different numbers of responses generated to a manipulandum. The results also extend the research of Xia et al. (2000), where pigeons were trained to generate a certain number of responses following the presentation of their respective numerical symbol. In the experiments described here, stimuli that varied in number were used

instead of abstract symbols and response limits were not restricted as in Xia et al., such that response variability could be examined.

Results of Experiment 2 showed that subjects would rely primarily on temporal cues if they were correlated with number and reinforcement, such that performance deteriorated when subjects were placed in transfer tests where conditions were different from that experienced in training. If subjects that had been trained in rate-controlled conditions were placed in duration- or time-controlled transfer tests, or subjects that had been trained in duration-controlled conditions were placed in rate-controlled transfer tests, subjects were unable to transfer responding to novel values. These results are consistent with the last-resort hypothesis (Davis & Memmott, 1982) and suggest that temporal variables are preferentially attended to over number. However, in Experiment 2A, subjects were able to perform the same numerical discrimination with similar accuracy when flash rate and sample phase duration were randomised and only weakly correlated with number, suggesting that although there may be a bias towards temporal cues when they are available, it is possible to train pigeons to respond primarily to numerical cues.

Additionally, hierarchical regression analyses in both experiments showed that number alone accounted for a significant amount of unique variance when entered into a model predicting response number with sample phase duration and flash rate as predictors. That is, in both experiments, significant control by number over responding was obtained above and beyond the temporal variables of flash rate and sample phase duration, even if temporal variables were correlated with number. Not surprisingly, when subjects were trained with randomised temporal variables, stronger control by number was obtained. These results were also replicated in Experiments 3 and 4.

Similar findings were obtained in a numerical bisection experiment (Experiment 1); pigeons learned to bisect three different numerical intervals, 2 and 6, 4 and 12 and 8 and 32, in a symbolic matching-to-sample procedure with randomised temporal variables. Hierarchical

logistic regression analyses were consistent with Experiment 2B; flash number accounted for a significant amount of unique variance in the probability of a "large" response, above and beyond the variance accounted for by the temporal variables (flash rate and sample phase duration).

Despite the randomization procedure used to generate the inter-flash intervals, the response-dependent nature of the sample phase and the variation in response latencies resulted in some covariation between the temporal variables and number. Thus if number was only used as a last resort, subjects could have responded purely on the basis of flash rate or sample phase duration and still have been reinforced occasionally. This does not seem to have been the case, although it is possible that subjects integrated both temporal and numerical cues to determine the appropriate response.

Overall, these results would suggest that pigeons are able to respond primarily on the basis of number, if the reliability of competing temporal cues is reduced.

8.2 What did they learn?

Given that control by number was shown in the current experiments, an important question is the extent of the numerical understanding subjects developed in these procedures. It has been said that nonhuman animals will only respond as accurately as the procedure requires (Cantlon & Brannon, 2007); did subjects only learn what was necessary to obtain reinforcement, or did they develop response rules or knowledge that could be applied to novel testing conditions?

Subjects in Experiments 1, 2, 3 and 4 were given transfer tests following baseline training. In these tests, randomly reinforced probe trials were presented amongst regular trials and involved novel numerical values both within and outside of the baseline training range. Subjects in Experiment 5 were exposed to retention interval tests to examine the effects of delay on responding in the numerical reproduction procedure.

In Experiment 1, subjects were tested for transfer with values inside and also two values

outside both upper and lower anchor values; in the 2 vs. 6 discrimination, subjects were tested with 0, 1, 3, 4, 5, 7 and 8. These were multiplied by 2 and 4 for the 4 vs. 12 and 8 vs. 24 discriminations, respectively. Results replicated and extended the results of Emmerton and Renner (2006), who found that pigeons were able to extrapolate to numerosities one value higher than the upper and lower anchor values. For all three discriminations, subjects were able to categorise novel stimuli successfully, regardless of whether they were between, lower or higher than the baseline training values. This suggests subjects had developed some representation of number during baseline training and were able to classify novel stimuli on the basis of that representation. Of particular interest was how subjects extrapolated to values outside the training range. The proportion of "large" responses to the two highest extreme values was significantly greater than the proportion of "large" responses to the higher anchor value in the 2 vs. 6, 4 vs. 12 and 8 vs. 24 conditions, suggesting that subjects extrapolated their numerosity judgments to values as high as 32. A similar pattern was observed with the two lowest extreme values, with subjects producing a lower proportion of "large" responses to the two values lower than the lower anchor value. No significant differences between responding on these trial types were found, although this may have been due to a floor effect.

To our knowledge, Experiment 1 is the first to test responding to 0 in a bisection task. There is relatively little research investigating zero concepts in nonhuman animals (Olthof et al., 1997; Boysen & Berntson, 1989, Merrit, et al., 2009; Biro & Matsuzawa, 2001; Pepperberg & Gordon, 2005; Pepperberg 2006). Although it is unlikely that subjects had spontaneously developed a true understanding of zero, including its cardinal and ordinal properties, responding on the 0-flash trials may provide some information about how subjects characterised a sample phase with no flashes. The proportion of "large" responses on the 0-flash trials tended to be lower than the proportion of "large" responses to the lower anchor value, but also tended to be greater than or equal to the proportion of "large" responses to the second lowest transfer test value. This suggests subjects knew that the number of flashes presented on 0-flash trials were

less than those presented on the trials with the lower anchor value, but the differentiation between the 0-flash trials and the trials containing the next highest number of flashes was either non-existent or in the wrong direction. However, as mentioned before this may have been due to a floor effect, as the proportion of large responses on the 1, 2, and 4 flash trials on the 2 vs. 6 trials (and for the corresponding trials in the 4 vs. 12 and 8 vs. 24 discriminations) was generally close to or equal to 0. Thus, it is unclear whether this lack of discrimination between 0 and 1 (or 0 and 2, or 0 and 4) is due to a failure to represent 0 correctly, or an artefact of the response limits in the bisection procedure. More research is necessary to investigate this further.

Results of Experiments 2-4 showed that subjects had developed an understanding of number that allowed the spontaneous transfer of responding to novel values located both within and outside of the training values 2, 4 and 6. Average response number on transfer test trials increased as flash number increased, and similarly modes of response distributions for each of the trial types generally covaried with flash number. Subjects were able to generate different numbers of responses that corresponded approximately to the flash number on novel trial types, despite never receiving any explicit training with these numbers.

The acquisition of performance and the development of numerical control in the reproduction procedure were specifically investigated in Experiment 3. This experiment showed that control by number developed early in training. Correlations between response number and flash number were greater than correlations between response number and flash rate or sample phase duration from the first block of training, and differentiation between response numbers on the 2, 4, and 6-flash trials emerged after 40-50 sessions of training.

The distribution of the different trial types in baseline training appeared to affect responding; subjects that had relatively greater exposure to 2- flash trials showed a greater tendency to make fewer numbers of responses, which was partially corrected by exposing subjects to just 2- and 6-flash trials for several sessions. Additionally subjects also exhibited a global tendency to make more or fewer numbers of responses across all trial types, which varied

between session blocks.

Experiment 4 examined the effects on responding in the numerical reproduction procedure of varying the retention interval (RI) between the end of stimulus presentation and beginning of the response phase. Previous research on the effect of delays in temporal and numerical bisection tasks has found that increasing the RI results in a choose "small/short" bias, whereas decreasing the RI has the opposite effect; a choose "large/long" bias (choose"short" effect: Gaitan & Wixted, 2000; Lieving et al., 2006; Spetch & Rusak, 1989; Spetch & Wilkie, 1983; Wearden et al., 2002; choose long effect: Roberts et al., 1995, Santi & Hope, 2001; Spetch & Rusak, 1989, 1992; Zentall et al., 2004; choose "small" and choose "large": Fetterman & MacEwen, 1989, Roberts, et al., 1995, Santi & Hope, 2001, Santi, Lellwitz & Gagne, 2006). It was unclear whether similar response patterns would be found in a response reproduction procedure. If responding was based on an analogue magnitude or activation-based representation that was susceptible to memorial decay, then subjects should produce a smaller number of responses when RI delays were increased relative to baseline training delays, analogous to the "subjective shortening" effect (Spetch & Rusak, 1992). However, results showed that when RIs were increased from 2s to 8s, average response numbers increased significantly across all trial types, and was largest for the 2-flash trials. A very slight increase was also observed for responding on the 2-flash trials when retention intervals were decreased to 0.5s, but this difference was not significant. Thus, increasing RIs had the opposite effect on responding to that predicted by most current theories that explain effects of RI on responding in temporal and numerical procedures, whereas decreasing RIs did not change responding significantly.

This finding is also interesting because of the greater effort (and delay to reinforcement) involved in producing larger numbers of responses. A possible explanation for this anomalous "produce-large" effect is a disruption in stimulus control that may have resulted in an increase in response variability and consequently a decrease in response differentiation, consistent with the

"direct remembering" theory of responding in DMTS procedures (White & Wixted, 1999). Additionally, this disruption may have increased reliance on temporal cues. If so, then responding may have been influenced by the duration of the sample phase and retention interval in addition to number, resulting in increased average response numbers when RIs were increased.

Although it is not possible to draw any definitive conclusions about what is responsible for the effect of RI delays on responses in the reproduction procedure, Experiment 5 provides a good starting point for future research examining the role of memory and delays in numerical tasks other than traditional bisection procedures.

8.3 Numerical processes and representations

It has been determined that subjects in the bisection and reproduction procedures were able successfully to discriminate number and extrapolate responding to novel values. What sort of processes and representations were used to determine responding in these tasks? This was another key question the research in this thesis hoped to answer.

There are two main scale structures that are thought to underlie nonverbal responding in numerical tasks, both of which predict scalar variability in responding (i.e., response/representational variability increases proportionally to number). Some researchers believe responding is based on a logarithmically-spaced scale with constant generalisation between numbers, in which the spacing between values decreases with increases in number but the spread of distributions for each value do not change (Dehaene, 1997). Others believe responding is based on a linearly-spaced scale with increasing generalisation between values (Brannon et al., 2000, Gallistel & Gelman, 2001). Both these scales make similar predictions and are generally difficult to differentiate empirically. Thus, responding of both nonhuman animals and humans in nonverbal numerical discrimination procedures studied previously is usually consistent with both of these scales. Points of subjective equality in bisection tasks are

normally located at the geometric mean (Fetterman, 1993; Emmerton & Renner, 2006; Roberts, 2005) and bisection functions superimpose when plotted on a relative scale; in absolute numerical discriminations, standard deviations increase proportionally to number and relative variability, as measured by the coefficients of variation, remains constant as numbers increase (Brannon, 2006).

In this regard, the majority of results in the present experiments were in opposition to previous findings; generally, responding did not conform to Weber's law and did not exhibit scalar variability. The only exception was Experiment 1; points of subjective equality were obtained at the arithmetic, not geometric mean, but functions for the three different discriminations (2 vs. 6, 4 vs. 12 and 8 vs. 24) still superimposed. Relative response variability in the reproduction procedures generally *decreased* with increases in number, instead of remaining constant. The slopes of the coefficients of variation, when plotted against average response number on a log-log scale was approximated or exceeded -0.5, suggesting binomial, rather than scalar variability. This was replicated across Experiments 2, 3 and 4.

The location of the bisection points, and relative response variability shown by pigeons, were more consistent with results obtained with human verbal counting (Cordes et al., 2001). If adult humans were asked to calculate the halfway point between two numerical values, it is likely that they would report the arithmetic and not geometric mean. In verbal response production procedures, adult human responding typically shows binomial variability; relative accuracy increases as number increases. Note that decreasing CVs are also reported in fixed consecutive number (FCN) procedures (Machado & Rodrigues, 2007), however the slopes were only halfway between binomial and scalar variability.

To investigate this further, responding of pigeons in the numerical bisection and reproduction procedures was compared with responding of humans in analogous tasks (Experiment 5). Generally human performance showed similar response patterns; points of subjective equality were located at the arithmetic mean in both verbal and nonverbal bisection

tasks, and response variability in the nonverbal production and report, and verbal production tasks showed similar decreasing CVs for the numerical values 1-7, although CVs did show scalar variability for values greater than 8. The results of the human experiment suggest that participants were using two forms of representations, one for the values less than 8, and another for values greater than 8.

The most parsimonious explanation for the similarity across species is that humans and pigeons share a similar nonverbal numerical discrimination process. The characteristics of responding for values 1-7 are not consistent with either of the scales that predict scalar variability, thus it is unlikely that the numerical representation developed in these procedures is either logarithmic with constant generalisation or linear with scalar generalisation. As a result, previously suggested models of numerical discrimination, such as the pacemaker-accumulator model (Meck & Church, 1983), or neural network models (Dehaene & Changeux, 2001) are unable to account for these findings. Rather, the location of the bisection points at the arithmetic mean, and the decreasing relative variability patterns akin to human verbal counting suggest that subjects' responding was based on a linear scale with constant variability.

This sort of numerical scale is believed to develop in humans after considerable experience with numbers and training with a linear number scale (Siegler & Opfer, 2003; Siegler & Booth, 2004; Dehaene et al. 2008). It is unlikely that the performance of the pigeons was based on true counting behaviour, as seen in humans, due to the relatively low accuracy in the task, and the failure to find positive transfer in the inconsistent transfer tests in Experiment 1. However, if subjects were treating numerical stimuli and their associated responses as categories, rather than as continuous stimuli, then the obtained results would also be expected. A categorical discrimination of number would predict decreasing relative response variability, since the probability of the assignment of any given sample number to a representational numerical category should not covary with number. Additionally, if categorisation were determined by the similarity of the samples' and representation's stimulus characteristics, then transfer to novel

numbers and stimuli would be expected to the extent that the sample and representations are similar.

8.3.1 The Prototype Response Class Model

A prototype category-learning model (prototype response class [PRC] model) was developed that successfully accounted for performance in the numerical reproduction tasks. In this model, subjects learned to associate response classes, which varied in terms of response number with the different trial-type sample phases; generating a category scale of "small", "medium" and "large" response runs associated with different numbers of presented flashes. These response classes and their associated sample phase are represented as prototypes and are used to determine responding during the production phase on any given trial based on the similarity of the most recent sample phase to the prototypes.

These prototypes develop over baseline training, through reinforcement of correct responses and also through correction trials; the probability of making either 2, 4, or 6 responses following the presentation of 2, 4, or 6 flashes respectively results in the higher-order response units that are ordered in terms of increasing numerosity. The sample-and-comparison process assumed by the model generates response distributions associated with each trial type. These distributions have some degree of overlap, and the modes of the distributions increase as the number of flashes increases.

The PRC model consists of two stages. In the input or prototype-activation stage, the recently-completed sample phase is compared to the prototypes stored in memory. Prototypes consist of a stimulus and response component and are developed for each trial type presented in baseline training. The stimulus component represents the sample phase during each trial type, and is defined in terms of three dimensions – sample duration, average inter-flash interval, and stimulus number. The response component is the higher-order response class associated with each trial type – "small", "medium" and "large" response bursts. Thus, the PRC model assumes

that subjects have, in effect, a categorical representation of number.

Stimulus dimensions acquire strength or associative value via reinforcement learning according to an incremental rule (e.g., Bush & Mosteller, 1955; Rescorla & Wagner, 1972). It was assumed that the stimulus dimensions are scaled logarithmically, however tests showed the fits of the models were not influenced by whether linear or logarithmic scales were assumed. The values of the sample phase duration and average inter-flash intervals associated with each prototype were calculated as the averages of the obtained values during the last 10 sessions of baseline training.

A different response class is associated with each prototype. It is assumed that subjects respond at a constant rate until they terminate the trial by pecking the right key. The probability of pecking the right key increases as the number of responses made during the production phase approaches the number associated with the prototype. According to Equation 8.1 below, the conditional probability of pecking the right key increases exponentially as the difference between the log number of responses associated with the prototype and the log number of responses made during the production phase decreases:

$$p(\text{"stop"}|\, n) = \exp\left(-\lambda \cdot (\ln N - \ln n) + \delta\right) \qquad \text{for } 0 < n < N$$

$$p(\text{"stop"}|\, n) = \exp(\delta) \qquad \text{for } n >= N$$

Equation 8.1

where p("stop"|n) is the probability of terminating the production phase after n responses, N is the number of responses associated with the activated prototype, and λ and δ are parameters (λ ,$\delta \geq 0$). This conditional probability reaches a maximum value of exp(δ) when $n = N$ and remains constant until the pigeon pecks the right key. Note that subjects are always assumed to make at least one response during the production phase (i.e., p("stop"|0) = 0). The resulting hazard functions for each prototype are shown in Figure 8.1, for $\lambda = 2.50$, $\delta = 0.25$.

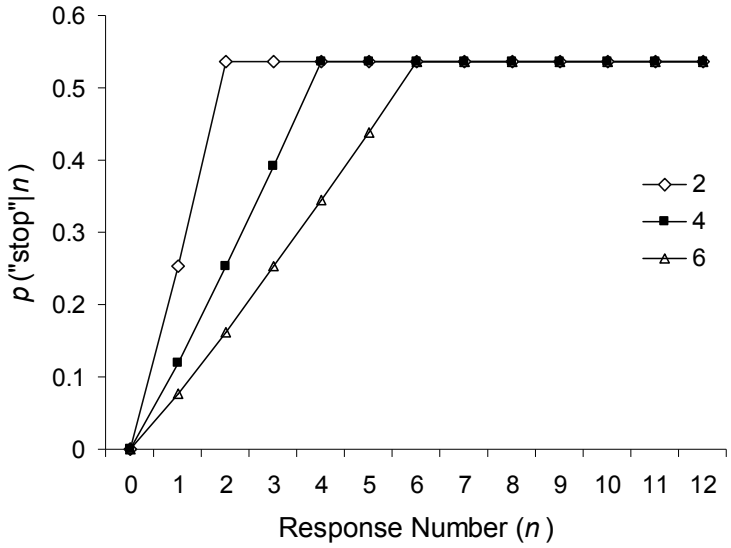

Figure 8.1. Hazard functions generated by Equation 8.1 for $\lambda = 2.50$, $\delta = 0.25$. Shown are the conditional probabilities of stopping a run during the production phase as a function of the number of responses already completed during the run, for each prototype.

Because Equation 8.1 is a discrete hazard function, it can be used to generate a response number distribution:

$$p(n) = p("stop"|\, n) \cdot \left(1 - \sum_{i=1}^{n-1} p("stop"|\, i)\right),$$

Equation 8.2

where $p(n)$ is the probability of making n responses during the production phase. Given specific values for λ and δ, Equations 8.1 and 8.2 predict response number distributions associated with each prototype. Sample prototype response number distributions are shown in Figure 8.2. These distributions represent the response classes learned in baseline training, and are uniquely associated with a single prototype.

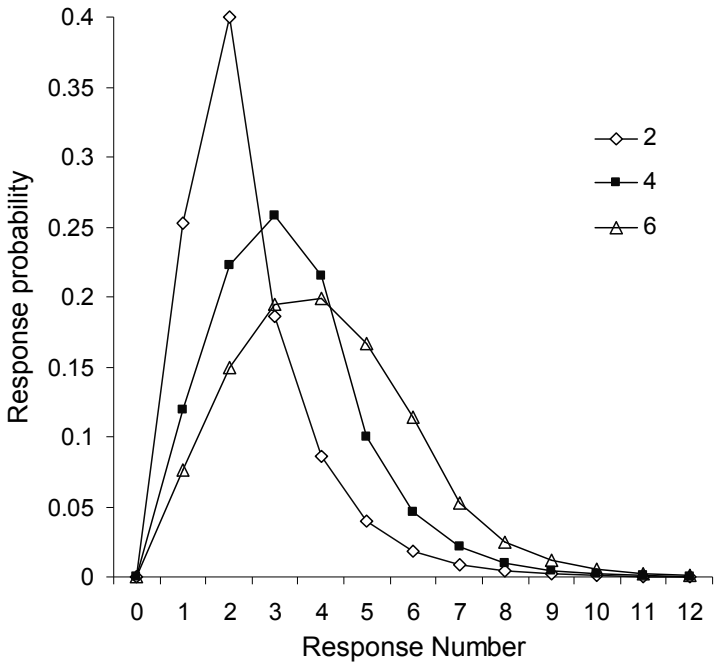

Figure 8.2. Distributions of number of responses during the production phase associated with the hazard functions in Figure 14, for each prototype. Data were generated by Equation 8.1, assuming $\lambda = 2.50$, $\delta = 0.25$.

Prototypes are activated based on the similarity of the just-completed sample phase with the prototype sample phases in terms of number, duration and average inter-flash interval. According to the model, the similarity score of the sample to prototype p, sim_p, is computed for each prototype prior to the production phase as follows:

$$\text{sim}_p = \exp\left(-\left(d_n \cdot \left|\ln n - \ln N_p\right| + d_d \cdot \left|\ln d - \ln D_p\right| + d_i \cdot \left|\ln i - \ln I_p\right|\right)\right) .$$

Equation 8.3

In Equation 8.3, n, d, and i are the sample number, duration, and average interval between flashes in the sample phase, and N_p, D_p and I_p are the corresponding values for prototype p. There are three parameters – d_n, d_d, and d_i, which allow for differential weighting of the three dimensions. These parameters are assumed to be greater than or equal to zero, and to vary depending on training. For example, sufficient exposure to a condition in which number and duration, but not interval, were correlated with reinforcement would result in a reduced value of d_i. Thus, Equation 8.3 calculates the similarity between a sample phase and prototype p as a

decreasing exponential function of the weighted absolute distances on the dimensions of number, duration, and interval.

The likelihood of prototype activation is also dependent on the relative frequency of the trial types during baseline. Similarity scores can also be weighted multiplicatively by the relative frequency of the respective trial type during baseline training. The model multiplies the similarity score by the relative frequency of the prototype during baseline (Equation 8.4):

$$\text{percent}_p = \#\,\text{trials}_p \Big/ \sum_{i=1}^{P} \#\,\text{trials}_i \ ,$$

Equation 8.4

where P is the number of different prototypes. The probability of activation for prototype p is then computed as follows in Equation 8.5:

$$p(\text{Act}_p) = \frac{\text{sim}_p \cdot \text{percent}_p}{\sum_{i=1}^{P} \text{sim}_i \cdot \text{percent}_p} \ .$$

Equation 8.5

The final predicted response number distribution is calculated, using Equation 8.6, as a mixture of the response distributions associated with each prototype, weighted by their respective activation probabilities.

$$p(n) = \sum_{i=1}^{P} p(\text{Act}_i) p_i(n) \ ,$$

Equation 8.6

where $p_i(n)$ is the probability of making n responses associated with prototype P.

The PRC model was fitted to the data from Experiments 2 and 2A using maximum likelihood estimation and successfully predicted the main features of responding, accounting for 80.6% and 80.8% of the variance in the rate- and time-controlled conditions of Experiment 2 in this dissertation, respectively, and 85% of the variance in Experiment 2A. The model

successfully predicted the patterns in average response number, as well as the obtained response distributions and decreasing CVs. It also predicted greater transfer performance in the consistent than inconsistent tests as found in Experiment 2. The PRC model also includes a parameter that is able to explain an interesting finding obtained in the acquisition experiment (Experiment 3); the covariation in the average number of responses generated on the 2-, 4- and 6-flash trials across session blocks. The δ parameter, used to calculate the response hazard functions, provides a means of adjusting the final probability of stopping for any given trial type; the patterns in average response number could be explained by varying δ across sessions, while keeping λ, a parameter that calculates the differentiation between trial types, at a roughly constant value.

The PRC model provides a good starting point for an explanation for the results obtained in Experiments 2-3. Unlike previously proposed numerical discrimination models, it is able to account for the main aspects of responding in the reproduction procedure, and takes into account aspects of the acquisition process that may influence responding. The model includes parameters that account for the relative distribution of trial types during baseline training and the correlation of the numerical and temporal variables, sample phase duration and flash rate (inter-flash interval) with reinforcement. A categorical numerical discrimination process is also consistent with the results of Experiment 4; if the changes in retention interval disrupted numerical control in behaviour, then the effect on responding would be determined by the generalisation of performance from the training delay to novel delays, with an increase in response variability and decreased differentiation between trial types. The size of this effect should be expected to increase with the difference in the novel RI from baseline. Consistent with this prediction, the "produce-large" effect was much greater when the RI was increased to 8s, than when it was decreased to 0.5s.

The PRC model may also be adapted to explain responding in a numerical bisection procedure, as in Experiment 1, by altering the response component from response classes to dichotomous "small" or "large" responses. Some further investigation would be required to

determine whether activation would be based on the difference or ratio of the prototype values and the sample phase values of the current trial's numerical and temporal variables, and the form of the response distributions. If the modified PRC model were able to predict obtained performance in the bisection procedure, in particular the novel results of extrapolation to novel values outside of the training range and bisection at the arithmetic mean, then this would provide additional evidence supporting a category-based account of numerical discrimination.

Given that the human participants showed similar response patterns to pigeon subjects, it is likely the PRC model also would provide a good description of human performance in the production tasks. However, it is unclear whether the PRC model would account for the scalar variability found in responding to higher values, and whether it could be applied to performance in the report tasks.

Nevertheless, the PRC model shows some promise in providing an alternative explanation for performance in numerical discrimination procedures; it is able to account for the main characteristics of performance in the numerical reproduction procedure and includes parameters that account for the influences of the proportion of baseline trial types in baseline training. It is yet unclear whether it could also be adapted to explain performance in the numerical bisection procedure, and the performance of humans in bisection, production and report tasks. However, these are empirical questions that can be addressed in future research.

A category-based representation of number appears to be able to explain the binomial variability obtained in the numerical reproduction procedure. However, as mentioned above, response variability in humans changed from binomial to scalar once values reached approximately 7 or 8. Assuming a common response process in this task, it would be expected that pigeons would show the same pattern. This would suggest that the human participants, at least, were using multiple representations of number when responding in the reproduction tasks; one representation that generates binomial variability for relatively small values, and one that generates scalar variability for larger values. This is similar to the findings obtained by Machado

and Rodrigues (2007), who found different response variability patterns for values less than and greater than approximately 10 in an FCN procedure. It is possible that more exact representations were not possible for values greater than 7, and consequently participants relied on an analogue magnitude representation that resulted in the scalar variability. The fact that participants showed the same variability in the report task suggests that the scalar variability is not a result of increased error made in the generation of larger numbers of keypresses required on those trial types, but rather is attributable to discrimination and representational processes. The finding of multiple representations would be consistent with previous research (Siegler & Opfer, 2003, Siegler and Booth, 2004), which has found that different types of numerical representations are developed and can be used selectively depending on the context of the task.

8.4 Future research

The current experiments have revealed original results that are interesting in their own right. They have shown that both pigeons and humans are able nonverbally to discriminate and bisect numerical smaller values with relative variability that decreases with number, consistent with a linearly spaced numerical scale with constant variability, and human participants discriminate larger values with performance that exhibits scalar variability.

These results also bring to light aspects of the discrimination and representation of number that need greater investigation. Experiment 1 showed that bisection of 3 different pairs of anchor values with a ratio of 1:3 occurred at the arithmetic, not geometric, mean and that psychometric functions superimposed; however it is unclear whether these findings would be replicated with anchor values comprising different numerical ratios. Additionally, it would be worthwhile to test whether RI manipulations would result in similar choose-large effects as those obtained in the reproduction experiment.

One of two aspects of the numerical reproduction procedure that may be important in generating the response patterns observed is the use of correction trials, which required the

repeitition of a trial and the generation of the correct number of responses following an previously incorrect trial. This may have facilitated the development of more precise numerical representations and their associated response runs.

The second characteristic is the method of stimulus presentation. The sequential and response-dependent nature of the flash sequences presented in the sample phase may have had some influence in eliciting nonscalar responding. Droit-Volet, Clement and Fayol (2008) compared bisection with stimuli presented both sequentially or simultaneously, finding that bisection points were located closer to the geometric mean in adults and 8-year olds when stimuli were presented non-sequentially; but with sequential stimuli, bisection points were located closer to the arithmetic than geometric mean. This provides some evidence supporting Jordan and Brannon's (2006) claim that the sequential presentation of number may have elicited a more linear form of representation. Thus, investigating whether responding in the numerical reproduction task would still exhibit the same characteristics when the stimuli presented in the sample phase are simultaneously presented, or even symbolic (Xia et al., 2000), would provide a further test of this notion. Additionally, if transfer between different types of stimulus presentation could be obtained, this would provide strong evidence that a concept of number had been developed in this procedure.

Response-dependent stimulus presentation may also play a role in increasing the accuracy of responding in the reproduction task. It is believed that applying behavioural tags to each item being counted may assist numerical discrimination and the acquisition of counting principles by facilitating the application of the counting principles, for example, cardinality, abstraction and one-to-one correspondence (Alilbali & DiRusso, 1999; Boysen et al., 1995). Thus, requiring a key-peck for every flash presentation may have helped subjects keep track of the number of flashes and consequently improved discriminative ability. If behavioural tags play a critical role in the accurate discrimination of flash number in the reproduction procedure, then performance should drop or become more variable when stimulus presentation is made response-

independent, and the number of responses made to flashes in the sample phase is not correlated with the number of responses required in the production phase.

The numerical understanding developed in the reproduction procedure could also be tested in additional experiments. Experiment 2 showed that transfer performance was weak when testing conditions, i.e. the temporal organisation of the sample phase, were different from baseline training conditions. However, if subjects trained with randomised temporal variables had developed a greater reliance on numerical than temporal cues, then it is possible that strong positive transfer should be obtained if they were placed in transfer tests in which duration of the sample phase or flash rate was controlled, despite the different testing conditions relative to baseline training. If significant control by number was still obtained in these transfer tests, this would provide stronger evidence that subjects were responding on the basis of number alone.

It is unclear how subjects would respond to 0 flashes in the numerical reproduction procedure; in the numerical bisection experiment, responding to 0 flashes did not differ significantly from responding to trials with the next two lowest numbers. If subjects had learned to reproduce the number of flashes presented in the sample phase in keypecks, then hypothetically, they should be able to respond accurately to 0 flashes. If positive transfer can be demonstrated, this would show that nonhuman animals are able to develop some understanding of zero without any explicit training.

Another aspect of the numerical reproduction procedure worthy of further investigation is the values that subjects experience during baseline training and testing. It is possible that the spacing and range of numbers used in training affects the nature of responding. In all the experiments, subjects were trained and tested with values spaced arithmetically; this may have biased subjects towards an arithmetic/linearly spaced scale. Would performance still be the same if subjects were trained with values that were spaced logarithmically, or would responding exhibit scalar characteristics? Additionally, it is unclear whether performance improve or worsen if subjects were trained with values that encompassed a larger range. Values separated by a

larger difference should be more discriminable; however, it is unclear whether transfer to novel numbers would also improve.

The structure of the numerical representation developed in these procedures can be examined in greater detail using a response variable that allows more analysis at a finer resolution. A number of recent studies with humans have used a number-space mapping procedure, where participants are required to locate a number along a horizontal number line (Siegler & Opfer, 2003; Siegler & Booth, 2004; Dehaene et al, 2008). Much more variability can be seen in this type of procedure, as the responding now falls along a properly continuous, rather than discrete variable. Consequently, this task would be useful in elucidating the structure of the number scale developed with both nonhumans and humans, because it would allow the direct examination of the location and structure of response distributions as a function of number.

8.5 Final thoughts

The current experiments have demonstrated that nonhuman animals are able to discriminate both relative and absolute number in bisection and reproduction procedures. Response patterns resembled those of humans in analogous bisection and reproduction tasks, exhibiting the same variability patterns that were more consistent with human verbal counting and a linear scale with constant generalisation between values. Additionally, responding of humans in the numerical reproduction and report procedures provided some evidence for the use of multiple representations of number within one task, one resulting in binomial response variability for values less than 8, and another resulting in scalar response variability for values greater than 8. A category-learning model provided a good account of responding in the reproduction task and shows promise in describing human and nonhuman performance in bisection, reproduction and report tasks. These experiments provide a strong foundation for future research into the representation of number and numerical discrimination processes, particularly the investigation of a nonverbal linear numerical scale.

References Cited

Allan, L.G. (2002). The location and interpretation of the bisection point. *Quarterly Journal of Experimental Psychology, 55*(B), 43-60.

Allan, L.G., & Gibbon, J. (1991). Human bisection at the geometric mean. *Learning and Motivation, 22*, 39-58.

Alibali, M.W., & DiRusso, A.A. (1999). The function of gesture in learning to count: More than keeping track. *Cognitive Development, 14*, 37-56.

Alsop, B., & Honig, W. K. (1991). Sequential stimuli and relative numerosity discriminations in pigeons. *Journal of Experimental Psychology: Animal Behavior Processes, 17*(4), 386-395.

Beran, M.J. (2001). Summation and numerousness judgments of sequentially preseted sets of items by chimpanzees (*Pan troglodytes*). *Journal of Comparative Psychology, 115*, 181-191.

Beran, M.J. (2004). Chimpanzees (*Pan troglodytes*) respond to nonvisible sets after one-by-one addition and removal of items. *Journal of Comparative Psychology, 118*, 25-36.

Beran, M.J. (2007). Rhesus monkeys (*Macaca mulatta*) enumerate large and small sequentially presented sets of items using analog numerical representations. *Journal of Experimental Psychology: Animal Behavior Processes*, *33,* 42-54.

Beran, M.J., Johnson-Pynn,. J.S., & Ready, C. (2008). Quantity representations in children and rhesus monkeys: Linear versus logarithmic scales. *Journal of Experimental Child Psychology*, *100*, 225-233.

Beran, M.J., & Rumbaugh, D.M. (2001). "Constructive" enumeration by chimpanzees (*Pan troglodytes*) on a computerized task. *Animal Cognition*, *4*, 81-89.

Beran, M.J.,Rumbaugh, D.M., & Savage-Rumbaugh, S. (1998). Chimpanzee (*Pan troglodytes*) counting in a computerized testing paradigm.*The Psychological Record, 48*, 3-19.

Beran, M.J., Taglialatela, L.A., Flemming, T.M., James, F.M. & Washburn, D.A. (2006). Nonverbal estimation during numerosity judgments by adult humans. *The Quarterly Journal of Experimental Psychology, 59*, 2065-2082.

Biro, D., & Matsuzawa, T. (2001). Use of numerical symbols by the chimpanzee (*Pan troglodytes*): Cardinals, ordinals and the introduction of zero. *Animal Cognition*, *4*, 193-199.

Boisvert, M.W., Abroms, B.D., & Roberts, W.A. (2003). Human nonverbal counting estimated by response production and verbal report. *Psychonomic Bulletin and Review, 10*, 683-690.

Boysen, S.T., & Berntson, G.G. (1989). Numerical competence in a chimpanzee (*Pan troglodytes)*. *Journal of Comparative Psychology*, *103*(1), 23-31.

Boysen, S.T. & Bernston, G.G. (1995). Responses to quantity: Perceptual versus cognitive mechanisms in chimpanzees (*Pan trogolodytes)*. *Journal of Experimental Psychology: Animal Behavior Processes, 21*(1), 82-86.

Boysen, S.T., Berntson, G.G., Shreyer, T.A. & Hannan, M.B. (1995). Indicating acts during counting by a chimpanzee (*Pan troglodytes*). *Journal of Comparative Psychology, 109(1)*, 47-51.

Brannon, E.M. (2005). What animals know about numbers in Campbell, J.D. (Ed.) *Handbook of Mathematical Cognition,* pp85-107. New York, US. Psychology Press.

Brannon, E.M. (2006). The representation of numerical magnitude. *Current Opinion in Neurobiology*, *16*, 222-229.

Brannon, E.M., & Terrace, H.S., (1998). Ordering of the numerosities 1 to 9 by monkeys. *Science*, *282*, 746-749.

Brannon, E.M. & Roitman, J.D. (2003). Nonverbal representations of time and number in animals and human infants in Meck, W.H (Ed.) *Functional and neural mechanisms of*

interval timing. Florida, US: CRC Press.
Brannon, E.M., & Terrace, H.S. (2000). Representation of the numerosities 1-9 by rhesus macaques (*Macaca mulatta*). *Journal of Experimental Psychology: Animal Behavior Processes*, *26*, 31-49.
Brannon, E. M., Wusthoff, C. J., Gallistel, C. R., & Gibbon, J. (2001). Numerical subtraction in the pigeon: Evidence for a linear subjective number scale. *Psychological Science, 12*(3), 238-243.
Breukelaar, J. W. C., & Dalrymple-Alford, J. C. (1998). Timing ability and numerical competence in rats. *Journal of Experimental Psychology: Animal Behavior Processes, 24*(1), 84-97.
Brissiaud, R., & Greenbaum, C. (1992). A tool for number construction: Finger symbol sets. In Bideaud, J, Meljac, C., & Fischer, J-P (Eds.) *Pathways to number: Children's developing numerical abilities.* Hillsdale, NJ, England: Lawrence Erlbaum Associates, Inc. vii.
Broadbent, H.A., Church, R.M., Meck, W.H., & Ratikin, B.C. (1993). Quantitative relationships between timing and counting in Boysen, S.T. & Capaldi, J.E. (Eds). *The development of numerical competence: Animal and human models.* Hillsdale, NJ, England: Lawrence Erlbaum Associates, Inc.
Brown, G.D.A., McCormack, T., Smith, M. & Stewart, N. (2005). Identification and bisection of temporal durations and tone frequencies: Common models for temporal and nontemporal stimuli. *Journal of Experimental Psychology: Human Perception and Performance*, *31*, 919-938.
Bush, R.R. & Mosteller, F. (1955). *Stochastic models for learning.* New York, NY. Wiley
Cantlon, J.C. & Brannon, E.M. (2006). Shared system for ordering small and large numbers in monkeys and humans. *Psychological Science*, *17,* 401-406.
Cantlon, J.C., & Brannon, E.M. (2007). How much does number matter to a monkey (*Macaca mulatta*)? *Journal of Experimental Psychology: Animal Behavior Processes, 33*, 32-41.
Cantlon, J.F., Cordes, S.,. Libertus, M.E., & Brannon, E.M. (2009). Comment on "Log or linear? Distinct intuitions of the number scale in Western and Amazonian indigene cultures". *Science, 323*, 38b-
Case, R., & Okamoto, Y. (1996). The role of conceptual structures in the development of chidlren's thought. *Monographs of the Society for Research in Child Development, 61*, Nos 1-2.
Clearfield, M.W., & Mix, K.S. (2001). Amount versus number: Infants' use of area and contour length to discriminate small sets. *Journal of Cognition and Development, 2,* 243-260.
Cohen, L.B. & Cashon, C.H. (2003). Infant perception and cognition in Lerner, R.M., Easterbrooks, M.A., & Mistry, J. (Eds). *Handbook of Psychology: Developmental Psychology*. New York: Wiley.
Cordes, S., Gelman, R., Gallistel C.R., & Whalen, J. (2001). Variability signatures distinguish verbal from nonverbal counting for both large and small numbers. *Psychonomic Bulletin & Review*, *8*, 698-707.
Davis, H. & Bradford, S. (1986). Counting behavior by rats in a simulated natural environment. *Ethology*, *73*, 365-280.
Davis, H., & Memmott, J. (1982). Counting behavior in animals: A critical evaluation. *Psychological Bulletin, 92*, 547-571.
Davis, H., & Perusse, R. (1988). Numerical competence in animals: definitional issues, current evidence and a new research agenda. *Behavioural and Brain Sciences*, *11*, 561-579.
Dehaene, S. (2001). Subtracting pigeons: Logarithmic or linear? *Psychological Science, 12*, 244-247.
Dehaene, S. & Changeaux, J.P. (1993). Development of elementary numerical abilities: A nueronal model *Behavioural Processes, 3,*, 216-228.
Dehaene, S., Izard, V., Spelke, E., & Pica, P. (2008). Log or linear? Distinct intuitions of the number scale in Western and Amazonian indigene cultures. *Science,320*, 1217-1220.

Droit-Volet, S., Clement, A., & Fayol, M. (2003). Time and number discrimination in a bisection task with a sequence of stimuli: a developmental approach. *Journal of Experimental Child Psychology, 84*, 63-76.

Emmerton, J. (1998). Numerosity differences and effects of stimulus density on pigeons' discrimination performance.*Learning and Behavior, 26,* 3, 243-256.

Emmerton, J. & Renner, J.C. (2006). Scalar effects in the discrimination of relative numerosity in pigeons. *Learning and Behavior,* 34, 176-192.

Fantz, R.L. (1964). Visual experience in infants: decreased attention to familiar patterns relative to novel ones. *Science*, *146,* 668-670.

Feigenson, L. (2008). Parallel non-verbal enumeration is constrained by a set-based limit. *Cognition, 107*, 1-18.

Feigenson, L., & Carey, S. (2005). On the limits of infants' quantification of small object arrays. *Cognition*, *97*m 295-313.

Feigenson, L., Carey, S., & Hauser, M. (2002). The representations underlying infants' choice of more: Object files versus analogue magnitudes. *Psychological Science, 13*, 150-156.

Feigenson, L., Carey, S., & Spelke, E., (2002). Infants' discrimination of number vs. continuous extent. *Cognitive Psychology, 44*, 33-66.

Feigenson, L., Dehaene, S., & Spelke, E. (2004). Core systems of number. *Trends in Cognitive Sciences, 8*, 307-314.

Fernandes, D. M., & Church, R. M. (1982). Discrimination of the number of sequential events by rats. *Animal Learning & Behavior, 10*(2), 171-176.

Fetterman, J.G. (1993). Numerosity discrimination: Both time and number matter. *Journal of Experimental Psychology: Animal Behavior Processes, 19*, 2, 149-61.

Fetterman, J.G., & Killeen, P.R. (1992). Time discrimination in *Columba livia* and *Homo sapiens*. *Journal of Experimental Psychology:Animal Behavior Processes, 18,* 80-94.

Fetterman, J.G. & MacEwen, (1989). Short-term memory for responses: The "choose-small" effect. *Journal of the Experimental Analysis of Behavior, 52*, 311-324.

Fetterman. J.G., Dreyfus, L.R., & Stubbs, A.D. (1985). Scaling of response-based events. *Journal of Experimental Psychology: Animal Behavior Processes*, *11*, 388-404.

Fetterman, J.G., Stubbs, A.D., & Dreyfus, L.R. (1986). Scaling of events spaced in time. *Behavioural Processes, 13*, 53-68.

Fleshler, M., & Hoffman, H.S. (1962). A progression for generating variable-interval schedules. *Journal of the Experimental Analysis of Behavior, 5*, 529-530.

Flombaum, J.I., Junge, J.A. & Hauser, M.D. (2005). Rhesus monkeys (*Macaca mulatta*) spontaneously compute addition operations over large numbers. *Cognition*, *97*, 315-325.

Gaitan, S.C., & Wixted, J.T. (2000). The role of 'nothing' in memory for event duration in pigeons. *Animal Learning and Behavior, 28*, 147-161.

Gallistel, C.R. & Gelman, R. (1992). Preverbal and verbal counting and computation. *Cognition*, *44*, 43-74.

Gallistel, C.R. & Gelman, R. (2000). Non-verbal numerical cognition: From reals to integers. *Trends in Cognitive Sciences, 4*(2), 59-65.

Gelman, R., & Gallistel, C.R. (1978). *The child's understanding of number*. Cambridge, MA: Harvard University Press.

Gibbon, J. (1977). Scalar expectancy theory and Weber's law in animal timing. *Psychological Review*, *84,* 279-325.

Gibbon, J. (1981). On the form and location of the psychometric bisection function for time. *Journal of Mathematical Psychology, 24*, 58-87.

Gibbon, J. & Church, R.M. (1981). Time left: Linear versus logarithmic subjective time. *Journal of Experimental Psychology: Animal Behavior Processes*, 7*,* 87-108.

Hanus, D., & Call, J. (2007). Discrete quantity judgments in the Great Apes (*Pan paniscus, Pantroglodytes, Gorilla gorilla, Pongo pygmaeus*): The effect of presenting whole sets versus item-by-item. *Journal of Comparative Psychology, 121*, 241-249.

Hauser, M.D. & Carey, S. (2003). Spontaneous representations of small numbers of objects by rhesus macaques: Examinations of content and format. *Cognitive Psychology*, *47*, 367-401.

Hauser, M.D., Carey, S., & Hauser, L.B. (2000). Spontaneous number representation in semi-free-ranging rhesus monkeys. *Proceedings of the Royal Society of London B, 267*, 829-833.

Hauser, M.D., MacNeilage, P., & Ware, M. (1996). Numerical representations in primates. *Proceedings of the National Academy of Sciences, 93*, 1514-1517.

Hauser, M.D., Tsao, F., Garcia, P., & Spelke, E. (2003). Evolutionary foundations of number: spontaneous representation of numerical magnitudes by cotton-top tamarins. *Proceedings of the Royal Society of London B, 270,* 1441-1446.

Hobson, S.L., & Newman, F. (1981). Fixed-ratio counting schedules: Response and time measures considered. In M.L. Commons, & J.A. Nevin (Eds.), *Quantitative analysis of behaviour: Vol. 1. DIscriminative properties of reinforcement schedules* (pp. 193-224). Cambridge, MA: Ballinger.

Honig, W.K., & Stewart, K.E. (1989). Discrimination of relative numerosity by pigeons. *Animal Learning and Behavior*, *17*, 134-146

Huntley-Fenner, G. (2001). Children's understanding of number is similar to adults' and rats': Numerical estimation by 5-7 year olds. *Cognition, 78*, B27-B40.

Ifrah, G. (1985). *From one to zero: A universal history of numbers*. New York: Viking Press

Jaakkola, K., Fellner, W., Erb, L., Rodriguez, M., & Guarino, E.(2005). Understanding of the concept of numerically "less" by bottlenose dolphins (*Tursiops truncatus). Journal of Comparative Psychology*, *119*, 296-303.

Jordan, K.E., & Brannon, E.M. (2006). Weber's law influences the numerical representations in rhesus macaques (*Macaca mulatta). Animal Cognition, 9*, 159-172.

Jordan, K.E., & Brannon, E.M. (2006b). A common representational system governed by Weber's law: Nonverbal numerical similarity judgments in 6-year-olds and rhesus macacques. *Journal of Experimental Child Psychology*, *95*, 215-229.

Jordan, K.E., Brannon, E.M., Logothetis, N.K., & Ghazanfar, A.A. (2005). Monkeys match the number of voices they hear to the number of faces they see. *Current Biology*, *15*, 1034-1038.

Kahneman, D., Treisman, A., & Gibbs, B.J. (1992). The reviewing of object files: Object-specific integration of information. *Cognitive Psychology*, *24*, 175-219.

Kaufman, E. L., Lord, M. W., Reese, T. W., & Volkman, J. (1949). The discrimination of visual number. American Journal of Psychology, 62, 498-525.

Keen, R., & Machado, A. (1999). How pigeons discriminate the relative frequency of events. *Journal of the Experimental Analysis of Behavior, 72*(2), 151-175.

Klahr, D., & Wallace, J.G. (1973). The role of quantitification operators in the development of conservation of quantity. *Cognitive Psychology, 4*, 301-327.

Koehler, O. (1950). The ability of birds to “count”. *Bulletin of Animal Behaviour*, *9*, 41-45.

Laties, V.G. (1972). The modification of drug effects on behavior by external discriminative stimuli. *Journal of Pharmacology and Experimental Therapeutics, 183*, 1-13.

Le Corre, M. & carey, S., (2007). One, two, three, four, nothing more: An investigation of the conceptual sources of the verbal counting principles. *Cognition, 105*, 395-438.

Lieving, L.M., Lane, S.D., Cherek, D.R. & Tcheremissine, O.V. (2006). Effects of delays on human performance on a temporal discrimination procedure: Evidence of a choose-short effect. *Behavioural Processes, 71*, 135-143.

Longo, M.R., & Lourenco, S.F. (2007). Spatial attention and the mental number line: Evidence for characteristic biases and compression. *Neuropsychologia, 45*, 1400-1407.

Lourenco, S.F. & Longo, M.R. (2009). Multiple spatial representations of number: evidence for co-existing compressive and linear scales. *Experimental Brain Research*, *193*, 151-156.

Machado, A. & Rodrigues, P. (2007). The differentiation of response numerosities in the pigeon.

Journal of the Experimental Analysis of Behavior, 88, 153-178.
Mechner, F. (1958). Probability relations within response sequences under ratio reinforcement. *Journal of the Experiment Analysis of Behavior, 1*(2), 109-121.
Meck, W.H (1997). Application of a mode-control model of temporal integration to counting and timing behaviour. In C.M. Bradhsaw & E.Z. Szabadi (Eds.), *Time and Behaviour: Psychological and neurobehavioural analyses* (pp133-184). Amsterdam: Elsevier.
Meck, W. H., & Church, R. M. (1983). A mode control model of counting and timing processes. *Journal of Experimental Psychology: Animal Behavior Processes, 9*(3), 320-334.
Meck, W.H., Church, R.M., & Gibbon, J.(1985). Temporal integration in time and number discrimination. *Journal of Experimental Psychology: Animal BehaviorProcessses, 11*, 591-597.
Menzel, E.W. Jr., (1960). Selection of food by size in the chimpanzee and comparison with human judgments. *Science, 131*, 1527-1528.
Menzel, E.W. Jr., (1961). Perception of food size in the chimpanzee. *Journal of Comparative and Physiologica Psychology, 54*, 588-591.
Merrit, D.J., Rugani, R., & Brannon, E.M. (2009). Empty sets as part of the numerical continuum: Conceptual precursors to the zero concept in rhesus monkeys. *Journal of Experimental Psychology: General, 138*, 258-269.
Mix, K.S., Huttenlocher, J., & Levine, S.C. (2002). Multiple cues for quantification in infancy: Is number one of them? *Psychological Bulletin, 128*, 278-294.
Moyer, R.S., & Landauer, T.K. (1967). Time required to judgments of numerical inequality. *Nature, 215*, 1519-1520.
Nieder, A., & Miller, E.K. (2003). Coding of cognitive magnitude: Compressed scaling of numerical information in the primate prefrontal cortex. *Neuron*, *37*, 149-157.w
Olthof A., & Roberts, W.A. (2000). Summation of symbols by pigeons (*Columbia livia)*: The importance of number and mass of reward items. *Journal of Comparative Psychology, 114*, 158-166.
Olthof, A. Iden, C.M., & Roberts, W.A. (1997). Judgments of ordinality and summation of number symbols by squirrel monkeys (*Saimiri sciureus). Journal of Experimental Psychology: Animal Behavior Processes, 23*, 325-339.
Pepperberg, I.M. (2006). Grey parrot (*Psittacu erithacus*) numerical abilities: Addition and further experiments on a zero-like concept. *Journal of Comparative Psychology, 120*, 1-11.
Pepperberg, I.M. & Gordon, J.D. (2005). Number comprehension by a grey parrot (*Psittacus erithacus*) including a zero-like concept. *Journal of Comparative Psychology, 119*, 197-209.
Pfungst, O. (1911) *Clever Hans* (C.L. Rahn, trans.) New York: Holt.
Platt, J.R., & Johnson, D.M. (1971). Localization of position within a homogenous behavior chain: Effect of error contingincies. *Learning and Motivation, 2*, 386-414.
Platt, J.R., & Senkowski, P.C. (1970). Effect of discrete-trials reinforcement frequency and changes in reinforcement frequence on preceding and subsequent fixed-ratio performance. *Journal of Experimental Psychology, 85*, 95-104.
Raslear, T.G. (1985). A test of the Pfanagl bisection model in rats. *Journal of Experimental Psychology: Animal Behavior Processes, 9*, 49-62.
Rescorla, R.A., & Wagner, A.R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A.H. Black & W.F Prokasky (Eds.) *Classical Conditioning II*, New York: Appleton-Century-Crofts.
Revkin, S.K., Piazza,M., Izard,V., Cohen, L., & Dehaene, S. (2008). Does subitizing reflect numerical estimation? *Psychological Science*, *19*, 607-614.
Roberts, W. A. (2005). How do pigeons represent numbers? Studies of number scale bisection. *Behavioural Processes. Special Issue: Stimulus Control in Animals: A Tribute to the Contributions of Donald S. Blough, 69*(1), 33-43.

Roberts, W.A. (2006). Evidence that pigeons represent both time and number on a logarithmic scale. *Behavioural Processes, 72,*207-214.

Roberts, W.A., & Boisvert, M.J. (1998). Using the peak procedure to measure timing and counting processes in pigeons. *Journal of Experimental Psychology: Animal Behavior Processes*, *24*, 416-430.

Roberts, W.A., Macuda, T. & Brodbeck, D.R. (1995). Memory for number of light flashes in the pigeon. *Animal Learning and Behavior, 23*, 182-188

Roberts, W.A. & Mitchell, S. (1994). Can a pigeon simultaneously process temporal and numerical infromation? *Journal of Experimental Psychology: Animal Behavior Processes, 20*, 66-78.

Roitman, J.D., Brannon, E.M., Andrews, J.R., & Platt, M.L. (2007). Nonverbal representation of time and number in adults. *Acta Psychologica, 124*, 296-318.

Rumbaugh,D.M., Savage-Rumbaugh, S. & Hegel, M.T, (1987). Summation in the chimpanzee (*Pan troglodytes). Journal of Experimental Psychology: Animal Behavior Processes, 13(2)*, 107-115.

Santi, A. & Hope, C. (2001). Errors in pigeons' memory for number of events. Animal *Learning and Behavior. 29*, 208-220.

Santi, A., Lellwitz, J. & Gagne, S. (2006). Pigeons' memory for sequences of light flashes: Reliance on temporal properties and evidence for delay interval/gap confusion. *Behavioural Processes, 72*, 128-138.

Sargisson, R.J. & White, K.G. (2001). Generalization of delayed matching to sample following training at different delays. Jo*urnal of the Experimental Analysis of Behavior, 75*, 1-14.

Saxe, G. (1982). Developing forms of arithmetical thought among the Oksapmin of Papua New Guinea. *Developmental Psychology, 18*, 583-594.

Siegel, S.F. (1986) A test of the similarity rule model of temporal bisection. *Learning and Motivation, 17*, 59-75.

Siegler, R.S. (1996). *Emerging minds: The process of change in children's thinking*. New York: Oxford University Press.

Siegler, R.S., & Booth, J.L (2004). Development of numerical estimation in young children. *Child Development, 75*, 428-444.

Siegler, R.S. & Opfer, J.E. (2003). The development of numerical estimation: Evidence for multiple representations of numerical quantity. *Psychological Science, 14*, 237-243.

Simion, T.J. (1997). Reconceptualizing the origins of number knowledge: A 'non-numerical' account. *Cognitive Development, 12*, 349-372.

Spetch, M.L. & Rusak, B. (1989). Pigeons memory for event duration: Intertrial interval and delay effects. *Animal Learning and Behavior, 17*, 147-156.

Spetch, M.L. & Rusak, B. (1992). Temporal context effects in pigeons' memory for event duration. *Learning and Motivation*, *23*, 117-144.

Spetch, M.L., & Wilkie, D.M. (1983). Subjective shortening: A model of pigeons' memory for event duration. *Journal of Experimental Psychology: Animal Behavior Processes, 9*, 14-30.

Starkey, P., Spelke, E.S., & Gelman, G. (1990). Numerical abstraction by human infants. *Cognition, 36*, 97-127.

Suzuki, K. & Kobayashi, T. (2000). Numerical competence in rats (*Rattus* norvegicus): Davis & Bradford (1986) extended. *Journal of Comparative Psychology, 115*, 83-91.

Tan, L., (2008). Effects of retention interval on performance in a numerical reproduction task. *Behavioural Processes, 78*, 279-284.

Tan, L., Grace, R., Holland, S. & McLean, A.P. (2007). Numerical reproduction in pigeons. *Journal of Experimental Psychology, Animal Behaviour Processes, 33,* 409-427.

Trick, L.M. & Pylyshyn Z.W. (1994). Why are small and large numbers enumerated differently? A limited-capacity preattentive stage in vsion. *Psychological review*, *101*, 80-102.

Uller, C., Carey, S., Huntley-Fenner, G., Klatt, L. (1999). What representations might underlie

infant numerical knowledge? *Cognitive Development*, *14*, 1-36.
van Marle, K., Aw, J., McCrink, K. & Santos, L. (2006) How capuchin monkeys (*Cebus apella*) quantify objects and substances. *Journal of Comparative Psychology, 120*, 416–426,
von Glaserfield, E. (1982). Subitizing: The role of figural patterns in the development of numerical concepts. *Archives de Psychologie*, 50, 191-218.
Washburn, D.A. & Rumbaugh, D.M (1991). Ordinal judgments of numerical symbols by macaques (*Macaca mulatta). Psychological Science, 2(3)*,190-193.
Wearden. J.H., & Ferrara, A. (1995). Stimulus spacing effects in temporal bisection by humans. *Quarterly Journal of Experimental Psychology, 48B*, 289-310.
Wearden, J.H., & Ferrara, A. (1996). Stimulus range effects in temporal bisection by humans. *The Quarterly Journal of Experimental Psychology, 49B*, 24-44.
Wellman, H.M. & Miller, K.F. (1986). Thinking about nothing: Developmental concepts of zero. *British Journal of Developmental Psychology, 4*, 31-42.
Wearden, J.H., Parry, A. & Stamp, L. (2002). Is subjective shortening in human memory unique to time representations? *The Quarterly Journal of Experimental Psychology 55B*, 1-25.
Whalen, J., Gallistel, C.R.. & Gelman, R. (1999). Nonverbal counting in humans: the psychophysics of number representation. *Psychological Science, 10*, 130-137.
White, K.G., & Wixted, J.T. (1999). Psychophysics of remembering. *Journal of the Experimental Analysis of Behavior,* 71, 91-113
Wilkie, D.M., Webster, J.B., & Leader, L.G. (1979). Unconfounding time and number discrimination in a Mechner counting schedule. *Bulletin of the Psychonomic Society, 13*(6), 390-392.
Wynn, K. (1992). Addition and subtraction by human infants. *Nature*, *358*, 749-750.
Wynn, K. (1992a). Children's acquisition of the number words and the counting system. *Cognitive Psychology, 24*, 220-251.
Xia, L., Emmerton, J., Siemann, M. & Delius, J.D. (2001). Pigeons (*Columbia livia*) learn to link numerosities with symbols. *Journal of Comparative Psychology, 115*(1), 83-91.
Xia, L., Siemann, M. & Delius, J.D. (2000). Matching of numerical symbols with number of responses by pigeons. *Animal Cognition*, *3*, 35-43.
Xu, F. (2003). Numerosity discrimination in infants: Evidence for two systems of representations. *Cognition, 81*, B15-B25.
Zeiler, M.D., & Hoyert, M.S. (1989). Temporal reproduction. *Journal of the Experimental Analysis of Behavior*, *52*, 81-95.
Zentall, T.R. (1999). Support for a theory of memory for event duration must distinguish between test-trial ambiguity and actual memory loss. *Journal of the Experimental Analysis of Behavior, 72*, 467-472.
Zentall, T.R. (2007). Temporal discrimination learning by pigeons. *Behavioural Processes, 74*, 286-292.
Zentall, T.R., Klein, E.D. & Singer, R.A. (2004). Evidence for detection of one duration sample and default responding to other duration samples by pigeons may result from an artefact of retention-test ambiguity. *Journal of Experimental Psychology: Animal Behavior Processes, 30*, 129-134.