Journal of Experimental Psychology:
Animal Behavior Processes
2006, Vol. 32, No. 2, 120–134

0097-7403/06/$12.00 DOI: 10.1037/0097-7403.32.2.120

The Role of Reference Points in Ordinal Numerical Comparisons by Rhesus Macaques (*Macaca mulatta*)

Elizabeth M. Brannon and Jessica F. Cantlon
Duke University

Herbert S. Terrace
Columbia University and New York State Psychiatric Institute

Two experiments examined ordinal numerical knowledge in rhesus macaques (*Macaca mulatta*). Experiment 1 replicated the finding (E. M. Brannon & H. S. Terrace, 2000) that monkeys trained to respond in descending numerical order (*4→3→2→1*) did not generalize the descending rule to the novel values *5–9* in contrast to monkeys trained to respond in ascending order. Experiment 2 examined whether the failure to generalize a descending rule was due to the direction of the training sequence or to the specific values used in the training sequence. Results implicated 3 factors that characterize a monkey's numerical comparison process: Weber's law, knowledge of ordinal direction, and a comparison of each value in a test pair with the reference point established by the first value of the training sequence.

Keywords: Weber's law, ordinal numerical knowledge, visual arrays, numerical representations, animal cognition

How animals represent number is an important question that has recently stimulated much research by investigators of animal cognition, cognitive development, cognitive psychology, and cognitive neuroscience. Since the discredited claims that a horse named *Clever Hans* could count and perform long division, many experiments have carefully documented the actual numerical abilities of many different nonhuman species (for reviews see Brannon & Roitman, 2003; Brannon, 2005). Examples include studies that show that animals can learn the relation between arbitrary symbols and numerosities (e.g., Boysen & Berntson, 1989; Matsuzawa, 1985; Pepperberg, 1987; Washburn & Rumbaugh, 1991; Xia, Siemann, & Delius, 2000), that animals can compare the relative numerosity of visual stimuli (e.g., Brannon & Terrace, 1998, 2000; Smith, Piel, & Candland, 2003; Judge, Evans, & Vyas, 2005), that animals can perform operations analogous to addition (e.g., Boysen & Berntson, 1989; Olthof, Iden, & Roberts, 1997), that animals may represent number abstractly without regard to the modality in which stimuli are presented (Church & Meck, 1984; Jordan, Brannon, Logothetis, & Ghazanfar, 2005), and that monkeys spontaneously track small numbers of objects by using object file representations (Hauser, Carey, & Hauser, 2000; Hauser & Carey, 2003; Hauser, MacNeilage, & Ware, 1996).

Here we focus on the following question: How do rhesus monkeys compare the numerical values of two or more visual arrays? In previous work, we demonstrated that rhesus monkeys have the capacity to represent ordinal relations (Brannon & Terrace, 1998, 2000). Two rhesus monkeys, Rosencrantz and Macduff, were first trained to respond in ascending order to exemplars of the numerosities *1–4*. They were then tested with all possible pairs of the numerosities *1–9*. We hypothesized that if the monkeys had learned an ordinal ascending rule for one set of numerosities (*1–2–3–4*), they should be able to extrapolate that rule to novel numerical values outside the range in which they were originally trained (i.e., *5–6–7–8–9*).

We took three precautions to ensure that our subjects used a numerical rule to order the stimuli on which they were trained and to order the novel numerosities on which they were tested. First, because all of our stimuli were trial unique they couldn't be memorized. Second, the smaller numerosity had a larger cumulative surface area on 50% of trials. That made it unlikely that the monkeys could have used a nonnumerical perceptual cue. Third, no positive or negative reinforcement was used on trials on which a novel numerosity was presented. That eliminated the possibility that our subjects could have learned the order in which to respond to novel numerosities by trial and error.

Both monkeys chose the smaller of the two novel numerical values more frequently than would be expected by chance (e.g., responding first to *5* when shown the pair *5–9*). Brannon and Terrace (2000) interpreted these findings as evidence that, their subjects having learned to order one set of numerosities, the monkeys abstracted a rule that enabled them to order novel numerosities. Although Brannon and Terrace's monkeys ordered novel numerical comparisons with above chance accuracy, performance was far superior on pairs that involved two familiar values or one familiar and one novel value than on pairs that involved two novel values. This pattern of results might be explained as a generalization decrement whereby performance decreased with decreasing familiarity. An alternative possibility is that performance was worse on pairs containing two novel values because, on average, they had the largest Weber fractions. Weber's law states

Elizabeth M. Brannon and Jessica F. Cantlon, Center for Cognitive Neuroscience and Department of Psychological and Brain Sciences, Duke University; Herbert S. Terrace, Department of Psychology, Columbia University, and New York State Psychiatric Institute, New York, New York.

The research was supported by RO1 MH040462 to Herbert S. Terrace. We thank Ilona Kovary for help in collecting the data and Ed Wasserman for discussions that led to the design of Experiment 2.

Correspondence concerning this article should be addressed to Elizabeth M. Brannon, Center for Cognitive Neuroscience, Duke University, Box 90999, Durham, NC 27708-0999. E-mail: brannon@duke.edu

that the just noticeable difference (jnd) with respect to a particular numerosity (*n*) is proportional to *n*. It follows that the degree to which two numerosities can be discriminated is determined by their ratio. The average Weber fraction for pairs containing two familiar values is .50 (range = .25−.75), whereas the average Weber fraction for pairs containing two novel values is .75 (range = .56−.89). An important question then is the relative contribution to accuracy of numerical novelty and discriminability (as defined by the Weber fraction; e.g., Beran, 2001; Nieder & Miller, 2004; Lewis, Jaffe, & Brannon, 2005).

A second question raised by Brannon and Terrace's (2000) previous research is whether the ordinal direction of training influences the acquisition of an abstract numerical rule. Benedict, a 3rd monkey in Brannon and Terrace's study, was trained to respond to the numerosities *1–4* in descending order. Although Benedict learned the *4→3→2→1* rule as easily as Rosencrantz and Macduff learned the *1→2→3→4* rule, his accuracy on ordering pairs composed of the novel numerosities *5–9* did not exceed the level predicted by chance. Because Benedict was the only subject who learned a descending sequence, it was unclear whether his inability to order novel numerosities was idiosyncratic or whether it was a consequence of learning a descending rule. Even if individual differences can be ruled out, it is not clear why knowledge of an ascending rule would enable a monkey to extrapolate that rule to novel numerosities, whereas knowledge of a descending rule would not.

Two experiments were conducted to address these issues. In Experiment 1, we sought to establish the reliability of our observation that a monkey trained to respond in a descending order to exemplars of the numerosities *1–4* would not generalize that rule to novel numerosities whose values ranged from *5–9*. In Experiment 2, we used a new experimental design to test two hypotheses about the difference in the performance of monkeys trained on ascending and descending rules: the direction of the ordinal rule per se and the absolute values of the numerosities on which the monkeys were trained. In our second experiment, monkeys were trained to order the numerosities *4*, *5*, and *6* in an ascending or a descending order. They were then tested on all of the pairs that could be derived from the numerosities *1–9*. This design allowed us to test each monkey with novel values both larger (e.g., *8* vs. *9*) and smaller (e.g., *1* vs. *2*) than the values that comprised the ascending and the descending training sequences. This design also allowed us to assess whether accuracy in a pairwise test is controlled by the novelty of the numerical values or by the Weber fraction of the pair. If performance to pairs of novel numerosities was controlled by novelty, there should be no difference between novel pairs composed of numerosities that were larger or smaller than the training values. However, if Weber's law were a factor, accuracy on pairs composed of two novel numerical values with small Weber fractions (e.g., *1* vs. *3* or *1* vs. *9*) should exceed accuracy on other pairs with larger Weber fractions, even when these include two familiar values (e.g., *5* vs. *6*).

Experiment 1

Brannon and Terrace (2000) trained a single monkey to order the numerosities *1–4* in a descending order. When tested with the novel values *5–9*, this monkey's performance differed markedly from that of two other monkeys that were trained to respond to the same numerosities in an ascending order. The purpose of Experiment 1 was to assess the reliability of those differences by testing a second monkey with the same experimental design.

Method

With one exception, the task, methods, and procedure used in Experiment 1 were identical to those used by Brannon and Terrace (2000). Like the subjects of the Brannon and Terrace study, the subject of this experiment was trained in his home cage on the descending sequence *4→3→2→1*. Unlike the subjects of the Brannon and Terrace study, the subject of this experiment was pairwise tested in a test chamber in a room adjacent to the colony room.

Subjects

Prospero, the sole subject of this study, was a 4-year-old rhesus monkey (*Macaca mulatta*) that was housed in the same colony room as Rosencrantz, Macduff, and Benedict, the subjects of Brannon and Terrace's (2000) experiment. Indeed, Prospero was Benedict's cage mate for the majority of the study. He was fed daily between 1300 and 1400 hr (Purina Monkey Chow, fruit), and water was available ad libitum.

Apparatus

A Macintosh computer, with PsyScope software (Cohen, MacWhinney, Flatt, & Provost, 1993), controlled experimental events and data collection. Reinforcers were 190-mg Noyes pellets (banana, orange, or grape flavored).

In cage testing. A mobile cart that housed a Microtouch touch-sensitive 15″ video monitor (3M Touch Systems, Methuen, MA) and a Gerbrands (Georgia, VT) pellet dispenser was positioned in front of the subject's cage before each session. The guillotine door in front of the cage was raised after the cart was secured to provide the subject with unimpeded access to the monitor. Prospero was tested in his home cage until the end of *4→3→2→1* training.

Test chamber testing. The subject was tested in a sound-attenuated booth (dimensions: 183 × 64 × 91 cm) that each contained a 72 × 58 × 69-cm chamber made of stainless steel and tempered glass. Each test chamber was fitted with a touch screen video monitor and a pellet dispenser. The subject was unrestrained during both in-cage and out-of-cage training and testing. All sessions, each of which lasted approximately 40 min, were monitored by a video camera positioned above the experimental chamber. Prospero was tested in the testing chamber throughout the pairwise test.

Previous Training

Like the monkeys in the Brannon and Terrace (1998, 2000) experiments, Prospero was first trained to produce 3- and 4-item simultaneous chains (Terrace, 1984) in which the list items were photographs and their order was arbitrarily determined by an experimenter. Details of this training can be found in Terrace, Son, and Brannon (2003). Prospero was also trained on a sequential match-to-sample task in which the stimuli were photographs.

Stimuli

The stimuli used in this experiment were identical to those used by Brannon and Terrace (2000) except for the fact that due to a programming error only five of the seven stimulus classes were used (clip art mixed and random size shape and color were excluded). Each stimulus set contained one exemplar of each of the numerosities *1–4*. Examples are shown in Figure 1 and described in detail in Brannon and Terrace (1998, 2000). Each

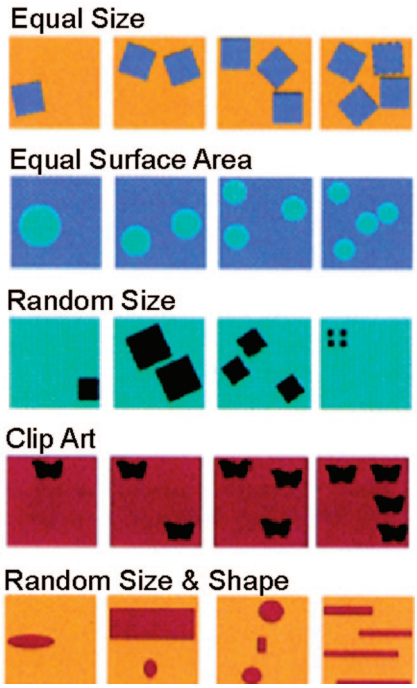

Figure 1. Exemplars of the seven different types of stimulus sets. All types were used with equal frequency in both 4-item training and 4-item testing. Equal size = elements were of same size and shape; equal area = cumulative area of elements was equal; random size = element size varied randomly across stimuli; clip art = identical nongeometric elements selected from clip art software; clip art mixed = clip art elements of variable shape.

stimulus was 3.5 × 3.5 cm and contained from one to nine elements. The first three classes ensured that subjects were not using cumulative surface area as a mechanism for ordering lists. All of the stimuli in a given stimulus set were drawn from the same category.

Pairwise tests contained one to nine elements. Only circles, ellipses, squares, and rectangles were used as elements during these tests. The elements of each stimulus were homogeneous with respect to size, shape, and color. To prevent subjects from using surface area as a cue, the smaller numerosity possessed elements that summed to a larger cumulative surface area compared with the larger numerosity on 50% of the trials. For the other half of trials, the larger numerosity contained a larger cumulative element surface area. Novel exemplars of each of the 36 numerosity pairs were used on each trial during each of 20 sessions (20 trial unique sessions, 720 stimulus pairs).

Task and Procedure

The task we used was identical to that described for Benedict in Brannon and Terrace (2000). Prospero was first trained to respond to each of four stimuli displayed on the video monitor in a descending numerical order (*4→3→2→1*). He was subsequently tested on all possible pairs of the numerosities *1–9*. The experiment was carried out in three phases: training the *4→3→2→1* sequence, transfer to novel exemplars of *4*, *3*, *2*, and *1* sequence, and testing on pairs of numerosities whose values ranged from *1–9*.

Training the 4→3→2→1 response sequence. Four numerical stimuli, each representing a different numerical value between *1* and *4*, were presented simultaneously and continuously on each trial on a touch-sensitive video monitor. The configuration of the stimuli was varied randomly from trial to trial to ensure that the subjects did not learn the sequence as a series of fixed motor responses. Prospero's task was to touch the numerical stimuli in a descending numerical order. Each correct response was followed by brief visual and auditory feedback (a 100-ms green border surrounding the stimulus and a 1200 Hz tone) to indicate that it had been detected. Any error terminated the trial immediately and resulted in a 15-s time out (TO), during which the screen of the video monitor was darkened. A food pellet was dispensed only after the subject responded to all four stimuli in the correct order. The intertrial interval (ITI) varied from 5 to 13 s (M = 8 s).

If the likelihood of the monkey selecting any of the available stimuli at each step of the sequence were equal, the probability of responding correctly to each stimulus from a new stimulus set by chance is .25 × .33 × .33 × .33 = .009. The first response could be A, B, C, or D. If A were selected, the trial would continue and the monkey could then respond correctly to B, or respond incorrectly to C or D.[1] If A and B were selected correctly, the trial would continue and the monkey could then make a correct response to C, a backward error to A, or a forward error to D. Finally, if the monkey had already responded correctly to A, B, and C, it could either complete the sequence correctly by responding to D, or make a backward error to A or B. If the monkey remembered the choices he made within each trial and, as a consequence, did not make any backward errors (such as A→B→A or A→B→C→B), the probability of completing a trial correctly by chance is .25 × .33 × .50 × 1 = .04. Because Prospero was an experienced list learner, who made few backward errors, we used the latter more conservative estimate of chance accuracy.

Each session comprised 60 trials in which the same stimulus set was presented on each trial (albeit in randomly varying configurations). Prospero was trained on 35 different stimulus sets, each selected from five of the seven stimulus control categories shown in Figure 1. As in Brannon and Terrace's (1998, 2000) previous procedures, training on each set was terminated once Prospero completed 20% of the trials correctly during a single session (12/60 trials) or after three sessions of training on a particular set. This criterion was used to encourage learning of the numerical rule and to avoid overtraining particular stimulus features.

Transfer to novel exemplars with the 4→3→2→1 sequence. To prepare for the transfer sessions, Prospero was given multilist training in which 10, 15, or all 35 of the previously trained sequences were tested in blocked or randomized trials. The transfer test consisted of 150 novel stimulus sets that were presented during the course of five successive transfer sessions, 30 novel stimulus sets per session. Each novel stimulus set contained one new exemplar of each of the numerosities *1*, *2*, *3*, and *4*. The 150 novel stimulus sets were composed of elements from one of the seven stimulus categories shown in Figure 1 and presented during each session with approximately equal frequency (21 or 22 sets from each of the seven stimulus control conditions). During the first half of each session, 30 novel stimulus sets were presented in a random order. The same 30 stimulus sets were presented in a different random order during the second half of the session, but we only used first trial data to assess transfer performance. The consequences of correct and incorrect responses during the transfer test were the same as those that were in effect during the training sessions. Correct responses produced brief auditory and visual feedback, errors terminated a trial, and correctly completed sequences produced food reward.

Pairwise test with the numerosities 1–9. During the third phase of the experiment, Prospero was tested on his ability to order all 36 pairs of numerosities that can be derived from the values *1–9*. All aspects of this task were identical to that reported by Brannon and Terrace (2000) for Benedict. Six pairs of stimuli were composed of the familiar numerosities *1–4* (FF pairs). Twenty pairs were made up from one familiar and one

[1] Repeat responses (e.g., A→A or B→B), at any position of the sequence, were quite rare and were not recorded.

novel numerosity (FN or FN pairs). The remaining 10 pairs were composed of two novel numerosities (NN pairs). Only circles, ellipses, squares, and rectangles were used as elements. The elements of each stimulus were homogeneous with respect to size, shape, and color. To minimize nonnumerical differences, the shape of the elements, their color, and the background color were identical for each pair of stimuli. Elements were positioned randomly within each stimulus. To prevent the use of surface area as a cue, the total area of the elements was smaller for the larger numerosity for half of the stimuli and larger for the other half.

Reinforcement was provided only on FF trials. For those six pairs (*1−2*, *1−3*, *1−4*, *2−3*, *2−4*, *3−4*), a correct sequence produced a 190-mg pellet. Incorrect responses were followed by an 8-s blackout period. On FF trials, a *correct sequence* was defined as a response to the larger numerosity followed by a response to the smaller value. By contrast, neither food pellets nor blackouts were provided on trials on which any of the novel numerosities were presented. Instead, brief visual and auditory feedback (a 100-ms green border surrounding the stimulus and a 1200 Hz tone) followed a response to either numerosity, regardless of the order in which the monkey responded. As previously, the sole function of feedback was to inform the subject that its response to the touch screen had been detected.

Each session consisted of 90 trials. To minimize the possibility of a response decrement that could result from the absence of reinforcement on trials during which a novel stimulus was presented, the relative frequency of trials on which reinforcement could occur was set at .66. This was done by presenting each of the six FF pairs (*1−2*, *1−3*, *1−4*, *2−3*, *2−4*, *3−4*) on 60 of the 90 trials. Each pair was presented 10 times per session. The remaining 30 numerical combinations (20 FN and 10 NN pairs) were presented only once per session (30 trials). A total of 1,680 stimuli were used in this phase of the experiment.

Results and Discussion

Prospero's performance was very similar to that of Benedict, the monkey trained on the same *4→3→2→1* sequence by Brannon and Terrace (2000). During the first phase of training overall accuracy for the 35 acquisition lists was above that expected by chance (one sample *t* test that compared accuracy on the first session of each of the 35 lists to 4%, the level predicted by chance; $t[34] = 2.34$, $p < .05$). Prospero's performance, as compared with that of the 3 monkeys trained by Brannon and Terrace (2000), is shown in Figure 2A. Like those monkeys, Prospero continued to respond at the same level of accuracy after the abrupt shift from training sessions (in which all of the stimulus sets were familiar) to transfer sessions (in which all of the stimulus sets were novel). A paired *t* test that compared Prospero's accuracy of responding on the last five blocks of acquisition with the first five sessions of test revealed no significant difference, $t(4) = 0.46$, $p > .05$. The absence of a decrement in accuracy during the novel stimulus set test demonstrates that Prospero, like Benedict, Rosencrantz, and Macduff in the Brannon and Terrace (2000) experiment, discriminated the numerosities *1–4* and learned a descending numerical sequence. As in previous studies, Prospero's accuracy varied as a function of stimulus class. Accuracy on the size constant, area constant, clip art, random shape, and random size conditions was 44%, 40%, 36%, 20%, and 16%, respectively, and, in each instance, exceeded the level expected by chance (4%; see Cantlon &

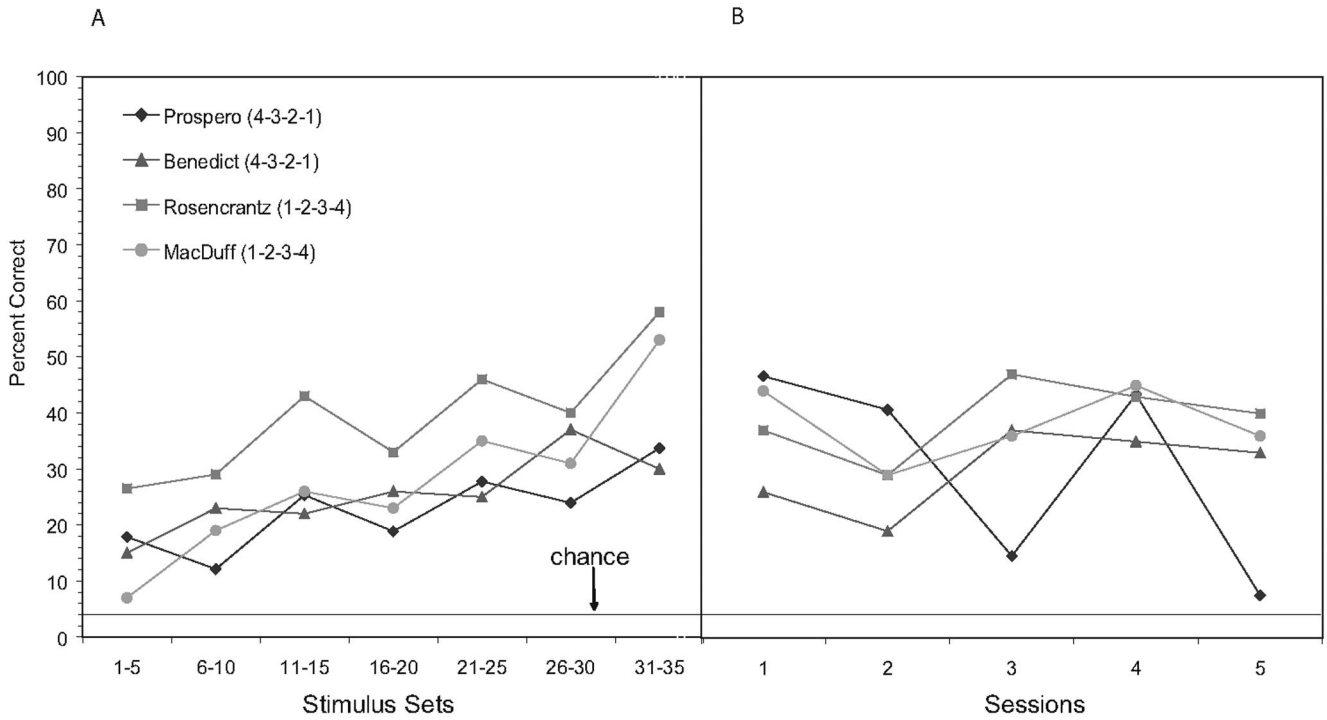

Figure 2. (A) Percentage of correctly completed trials during the first session for each of 35 training stimulus sets in blocks of five sessions for Prospero and 3 other monkeys previously published in Brannon and Terrace (2000). (B) Percentage of correctly completed trials on the 150 test sets over five sessions. Values in parentheses indicate the rule on which each of the monkeys was trained.

Brannon, in press, for further discussion of the effect of stimulus class).

Prospero's performance on the pairwise test of the numerosities *1–9* resembled Benedict's. Figure 3 compares Prospero's performance on pairs of numerosities composed of familiar and novel numerical values (FF, FN, and NN) with that of the 3 monkeys tested by Brannon and Terrace (2000). Chance accuracy was 50%. Although Prospero performed significantly above chance on pairs of familiar values and pairs of one novel and one familiar value, his accuracy on the NN pairs did not differ from the level predicted by chance (one sample t test that compared accuracy for 10 sessions to 50% chance expectation; $t[9] = 0.97$, $p = .36$). Prospero's performance was also uninfluenced by surface area. In addition, Prospero ordered 75% of the pairs correctly when the smaller numerosity had a smaller surface area and 71% when it had a larger surface area.

Additional analyses compared Prospero's performance with that of the subjects from Brannon and Terrace's (2000) experiment. A 2×3 repeated measures analysis of variance (ANOVA) with a between-subjects variable of ordinal direction (ascending vs. descending) and a within-subject variable of stimulus category (FF, FN, and NN) revealed a main effect of ordinal direction, $F(1, 2) = 203.60$, $p < .0001$; a main effect of stimulus category, $F(2, 4) = 112.52$, $p < .001$; and an interaction between ordinal direction of training and stimulus category, $F(2, 4) = 6.60$, $p = .05$. The main effect of ordinal direction of training was due to the higher overall accuracy of the monkeys trained to respond in an ascending order as compared with the monkeys trained to respond in a descending order (87% and 68%, respectively). Inspection of the means shown in Figure 3 suggests that the main effect of stimulus category was due to superior performance on the FF and FN pairs relative to the NN pairs. Figure 3 also suggests that the interaction reflects different patterns in accuracy of responding to each type of test pair as a function of the ascending and the descending rules.

In summary, Prospero's performance mirrored Benedict's, who was trained on the same task. Prospero's performance during the 35 training sets was indistinguishable from that of the other monkeys trained on the ascending and descending sequences in that he showed no decrement in accuracy when presented with novel exemplars of the values *1–4*. Like Benedict, Prospero was unable to extrapolate the *4→3→2→1* rule learned in training to order novel pairs of larger values.

Experiment 2

Experiment 1 showed that a second monkey trained to respond in a descending order to the numerosities *1*, *2*, *3* and *4* failed to generalize that rule to numerosities whose values ranged from *5* to *9*, thereby confirming the findings of Brannon and Terrace (2000). However the reason for this failure remains unclear. Subjects could have failed because they were originally trained to respond in a descending direction. Alternatively, the specific values used to train the descending rule *4→3→2→1* did not allow subjects to continue the sequence in a descending direction. Instead it required subjects to order values that preceded the starting point of the sequence (i.e., values >4).

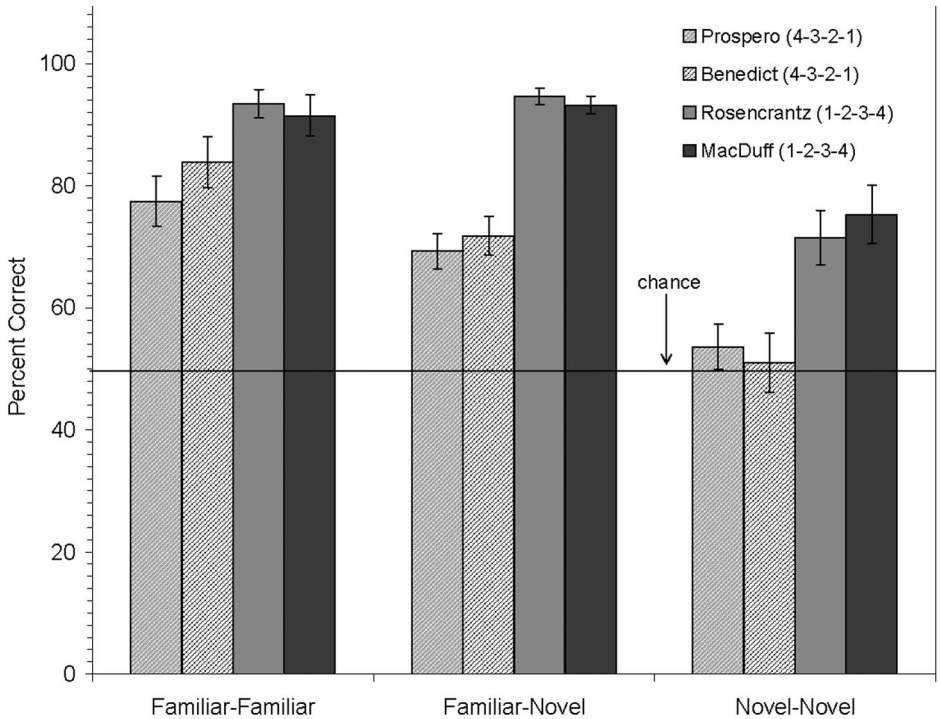

Figure 3. Performance on familiar−familiar, familiar−novel, and novel−novel numerosity pairs for Prospero and 3 other monkeys previously published in Brannon and Terrace (2000). Only familiar−familiar pairs were reinforced. Error bars reflect variance across test sessions. Values in parentheses indicate the rule on which each of the monkeys was trained.

The Brannon and Terrace (1998, 2000) experiments raised another related question regarding performance on NN subsets following ascending training on the *1→2→3→4* sequence. Although accuracy on NN trials exceeded the level predicted by chance, it was lower than that observed on FF and FN trials. The design of that experiment did not allow us to determine whether the reduced performance on NN pairs was due to a generalization decrement, whereby unfamiliar values were difficult to discriminate, or to Weber's law that predicts that discriminability should be controlled by the ratio of two values. Weber's law would predict that, on average, NN pairs that were composed of the values *5–9* would be more difficult to discriminate than FF pairs composed of the values *1–4*. The lowest ratio of an NN pair is 5/9 (.56), which is considerably larger than 1/4 (.25), the lowest ratio of an FF pair. There is a great deal of empirical support for the operation of Weber's law over numerical judgments in many nonhuman species (e.g., see Gallistel & Gelman, 2000). Particularly relevant to the present study is the distance effect that was obtained in Brannon and Terrace's (2000) study. Large numerical distances between two numerical stimuli (given equivalent magnitude) yield low Weber fractions, high levels of accuracy, and rapid reaction times.

To address these questions, we conducted a second experiment in which monkeys were trained to order the values *4–6* in an ascending or a descending order and were then tested on their ability to extrapolate the ordinal response rule to novel numerical values that were both smaller (e.g., *1* vs. *3*) and larger (e.g., *7* vs. *9*) than the training values.

Method

One monkey (Ebbinghaus) was trained to respond in an ascending order (*4→5→6*) and 2 monkeys (Lashley and Horatio) were trained to respond in a descending order (*6→5→4*) on 40 different stimulus sets that contained exemplars of the numerosities *4*, *5*, and *6*.[2] Subjects were then tested with 100 trial-unique stimulus sets to determine whether they had abstracted the ascending and the descending rules on which they were trained. During the final phase of Experiment 2, each monkey was tested with all possible pairs that could be derived from the numerosities *1–9*.

Subjects

The subjects were 3 male rhesus macaques (*Macaca mulatta*). Ebbinghaus and Lashley were 4 years old at the beginning of the experiment; Horatio was 7 years old. They were housed and fed in the same colony room used in Experiment 1.

Apparatus

The apparatus was identical to that used in Experiment 1. Ebbinghaus and Lashley were trained and tested in their home cages. Horatio was trained and tested in the out-of-cage test chamber described for Experiment 1.

Previous Training

All 3 monkeys had previously been trained with the simultaneous chaining paradigm on a list production task (Terrace, Son, & Brannon, 2003). Like Prospero, Horatio also had extensive matching-to-sample training prior to this experiment. None of the monkeys had any experience with numerical stimuli.

Stimuli

Each stimulus contained four, five, or six abstract elements. As in Experiment 1, each stimulus, whose dimensions were 3.5 × 3.5 cm., was programmed to appear in one of 16 positions on the video monitor, each equidistant from one another. On each trial, the stimuli were displayed in a novel configuration that was selected at random from 43,680 possible configurations.

Elements were circles, ellipses, squares, rectangles, or clip art shapes, which were positioned randomly within each stimulus. With the exception of clip art shapes, all elements were black. The background color was blue, green, cyan, lavender, pink, or yellow. A stimulus set contained one exemplar of each of the numerosities *4*, *5*, and *6*. Four stimulus types were used to evaluate the degree of stimulus control that was exerted by nonnumerical dimensions. The four types were equal size, equal surface area, random size, and mixed clip art and examples of each type are shown in Figure 4.

Each stimulus used in the pairwise tests contained between one and nine elements. For one third of the stimuli, the exemplar of the smaller numerosity had the larger cumulative surface area. For another third, it had the smaller cumulative surface area. For the remaining stimuli, the cumulative surface area of the two numerical stimuli was equated. Examples of these stimuli are shown in Figure 4B. Elements were circles, ellipses, squares, or rectangles. Clip art shapes were not used. The elements used to construct each numerical stimulus were homogeneous with respect to size, shape, and color.

Task and Procedure

The task and procedure was similar to that described in Experiment 1 with the exception that the subjects were trained to respond to the numerosities *4*, *5*, and *6* rather than *1*, *2*, *3*, and *4*. Lashley and Horatio were trained on a descending sequence (*6→5→4*); Ebbinghaus on an ascending sequence (*4→5→6*). The experiment was conducted in three phases: acquisition of the *4→5→6* (or the *6→5→4*) sequence, transfer to *4→5→6* (or *6→5→4*) sequence composed of novel stimuli, and testing of pairs composed of the numerical values *1–9*.

Acquisition of the 4→5→6 or 6→5→4 sequence. Subjects were trained on 40 different stimulus sets. Each set was trained until subjects ordered the stimuli on 20% of the trials correctly in a single session (12/60 trials) or until they completed three sessions of training on that stimulus set. Given the assumption that subjects never make the backward error A→B→A, the probability of selecting A, B, and C, in that order, at the start of training on a new list is $0.33 \times 0.50 \times 0.50 = 0.08$ or $0.33 \times 0.50 \times 1 = 0.165$. If the subject selects the first item randomly, its options are A, B, or C (hence, chance = 0.33). After the monkey responds correctly to A, the trial continues and the monkey could then respond to B (correctly) or to C incorrectly (hence, chance = 0.50). If the monkey responded correctly to A and B, the trial would continue and the monkey could then respond correctly to C or incorrectly to A. Backward errors to A were extremely rare. Accordingly, our estimate of chance accuracy was based on the conservative assumption, that having responded correctly to A and B, the subject would always respond to C ($0.33 \times 0.50 \times 1 = 0.165$).

The 40 stimulus sets were trained in four blocks of 10 sets. After each block, subjects were given multilist training in which each of the 10 previously learned sets were presented in random order during the same session. The purpose of multilist training was to prepare subjects for the

[2] Our goal was to have 2 subjects in each condition. However, the 2nd subject assigned to the *4→5→6* condition was also a pilot subject in a study on the effects of electroconvulsive shock on memory. Although the other study did not affect his accuracy level on number tasks, it seems to have reduced his motivation to work. We therefore stopped using this monkey midway through the experiment.

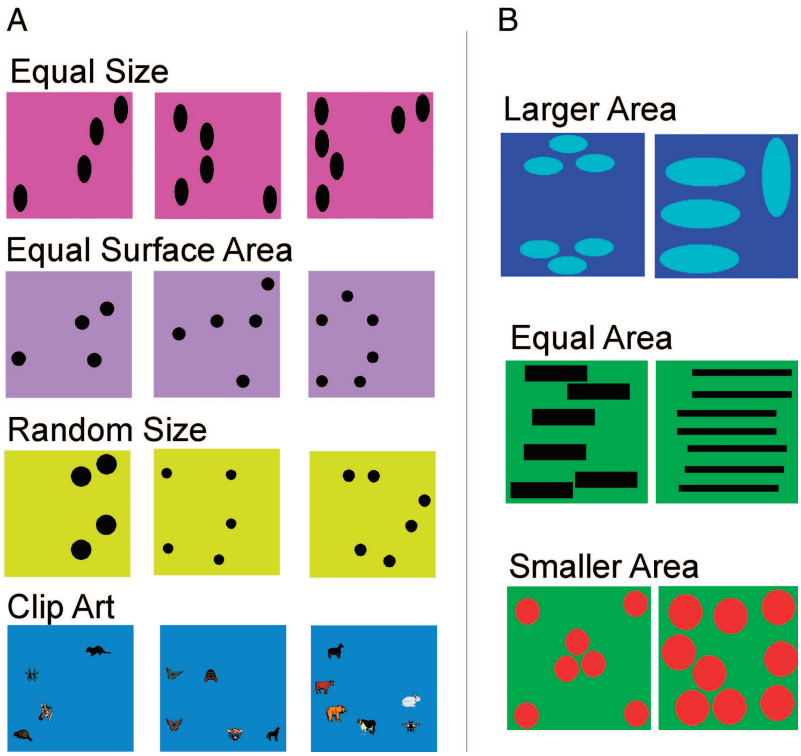

Figure 4. (A) Exemplars of the four different types of stimulus sets used in 3-item training and 3-item testing in Experiment 2. Equal size sets = all of the elements from which numerical stimuli were constructed were of uniform size, irrespective of their numerical value; equal surface area sets = the sum of the area of the elements within each stimulus was equal for the three numerosities within a set; random size = the numerosity *5* had the smallest or largest element size. This meant that neither cumulative surface nor element size could be related to number; mixed clip art stimulus sets = elements were small icons of various colors. Each element in a stimulus set was unique (15 elements per set). (B) Three stimulus types used in pairwise testing in Experiment 2. The smaller numerosity had a larger cumulative surface area on one third of trials, a larger cumulative surface area on one third of trials, and equal surface area on the final one third of trials.

transfer phase of training during which a new stimulus set was presented on each trial. Multilist training continued until the subject responded correctly on 60% of the trials during a single session or until the subject completed five multilist sessions. After the final block of 10 stimulus sets, all 40 stimulus sets were presented in random order during two successive sessions.

Transfer to novel stimuli with the 4→5→6 or 6→5→4 sequence. All 3 subjects were tested with 100 novel stimulus sets that were presented over five successive sessions (20 new lists/session). As was the case for the training sets, the elements used to construct novel stimulus sets were black circles, ellipses, squares, and rectangles. These were presented on blue, pink, lavender, or yellow backgrounds. After the first 20 trials of each session, the same 20 novel stimulus sets were repeated two times in a random order, for a total of 60 trials. Only first trial data were considered in our analyses of each subject's performance. The reinforcement contingencies were the same as those used during the training phase. Correct responses to the first two items produced brief auditory and visual feedback. Food reward was provided only after the sequence was completed correctly, that is following the third correct response. Any error terminated a trial immediately.

Pairwise testing of the numerosities 1–9. During this phase of training, each subject was presented with all possible pairs of the numerosities *1–9* through the use of trial unique stimuli. Two stimuli, each of a different numerosity, were presented on each trial. On an equal proportion of trials the two numerosities had equal surface area, the smaller numerosity had a larger surface area, or the smaller numerosity had a smaller surface area. Elements within each stimulus were homogeneous in color and size. Three numerosity pairs were composed of two familiar values (FF), 18 were composed of one familiar and one novel value (FN), and the remaining 15 pairs were composed of two novel values (NN). The chance probability of ordering any of these 2-item sequences correctly was .50.

Reinforcement, which was only available on FF trials (*4−5*, *4−6* and *5−6*), was a 190-mg food pellet. Following training on the *4→5→6* rule, we considered a pairwise trial correct if the subject responded to the smaller value before responding to the larger value. The reverse contingency was in effect for the monkeys trained on the *6–5–4* sequence. Errors produced an 8-s timeout period. Neither food pellets nor timeouts occurred on FN or NN trials, regardless of the order in which subjects responded. Instead, brief visual and auditory feedback (a 100-ms green border surrounding the stimulus and a 1200 Hz tone) followed each response. The sole function of the feedback was to inform the subject that its response was detected.

The pairwise test was conducted during the course of 23 sessions, each of which was made up of 90 trials. Because there was no reinforcement on any of the FN and NN trials, subject performance on those trials could not be attributed to learning. The relative frequency of trials on which each type of stimulus pair was presented was adjusted to make reinforcement available on 63.3% of the trials during the subset test. Specifically, each of

the three FF pairs was presented 19 times per session (57 trials). The other 33 numerosity combinations were presented only once per session (33 trials). All trials with novel numerical values consisted of trial-unique novel stimuli (33 pairs per session, two stimuli per pair, 23 sessions = 1,518 stimuli). Five unique pairs were used for each of the three FF pairs during each session (690 additional unique stimuli).

Results and Discussion

All 3 subjects readily acquired the *4→5→6* or the *6→5→4* rules and applied those rules appropriately to novel exemplars of the numerosities *4–6*. Overall, subjects also ordered pairs of the numerosities *1–9* accurately, but an analysis of subject performance on those pairs revealed three factors that influenced accuracy on specific pairs. Two of those factors were described previously (Brannon & Terrace, 1998, 2000): Weber's law and the ordinal direction of the training sequence. The third factor, the subject's reference point (RP), emerged from an analysis of results of the present experiment.

Phase 1: Learning the Ascending and Descending Rules (4→5→6 and 6→5→4)

Overall accuracy for the 40 acquisition lists exceeded the level expected by chance for each monkey (Ebbinghaus, $t[39] = 7.21$, $p < .0001$; Lashley, $t[39] = 13.23$, $p < .0001$; Horatio, $t[39] = 6.81$, $p < .0001$). The relevant data are shown in Figure 5A.

Transfer to Novel Exemplars of the Numerosities 4, 5, and 6

Performance on trial-unique sets of the numerosities *4*, *5*, and *6* is shown in Figure 5B. Performance exceeded the level predicted by chance (16.5%) for all 3 monkeys on the 100 novel sets (one sample *t* tests that compared five blocks of test trials with .165 for each monkey: Ebbinghaus, $t[4] = 4.24$; $p < .05$; Lashley, $t[4] = 3.66$; $p < .05$; Horatio, $t[4] = 4.04$; $p < .05$). In addition, there was no difference in accuracy between the last five acquisition blocks and the five test blocks for any of the 3 monkeys (paired dependent *t* test that compared the last five acquisition blocks with five test blocks for each monkey: Ebbinghaus, $t[4] = 1.59$; $p = .19$; Lashley, $t[4] = 2.01$; $p = .11$; Horatio, $t[4] = 2.06$; $p = .11$). However, a comparison of Figures 5A and 5B shows that there was a drop in accuracy from the final block of acquisition to test in Experiment 2 that was not observed in Experiment 1. Table 1 shows that there was variability in performance across the four stimulus control conditions. Although overall accuracy exceeded the level expected by chance during the five test blocks for all 3 monkeys, none of the 3 monkeys performed above chance in the clip art condition (>7 out of 25, $p < .05$, binomial test), and neither Horatio's performance on the area-constant condition nor Ebbinghaus' performance on the random-size condition exceeded the level expected by chance. All told, generalization of the ascending and the descending rules was not as consistent in Experiment 2 as it was in Experiment 1. One possible explanation for this pattern of results is that a lower performance criterion was used in training the 3-item *4→5→6* and *6→5→4* sequences than the 4-item *1→2→3→4* and *4→3→2→1* sequences. Although the criterion for 3- and 4-item sequences was the same in both cases (20%), subjects could satisfy that criterion with essentially chance performance in the case of 3-item sequences. We suspect that subjects may have exhibited a higher level of transfer in response to novel stimuli if the *4→5→6* and the *6→5→4* sequences were trained to a higher criterion.

Pairwise test. Performance on FF, FN, and NN numerical pairs differed from that reported by Brannon & Terrace (1998,

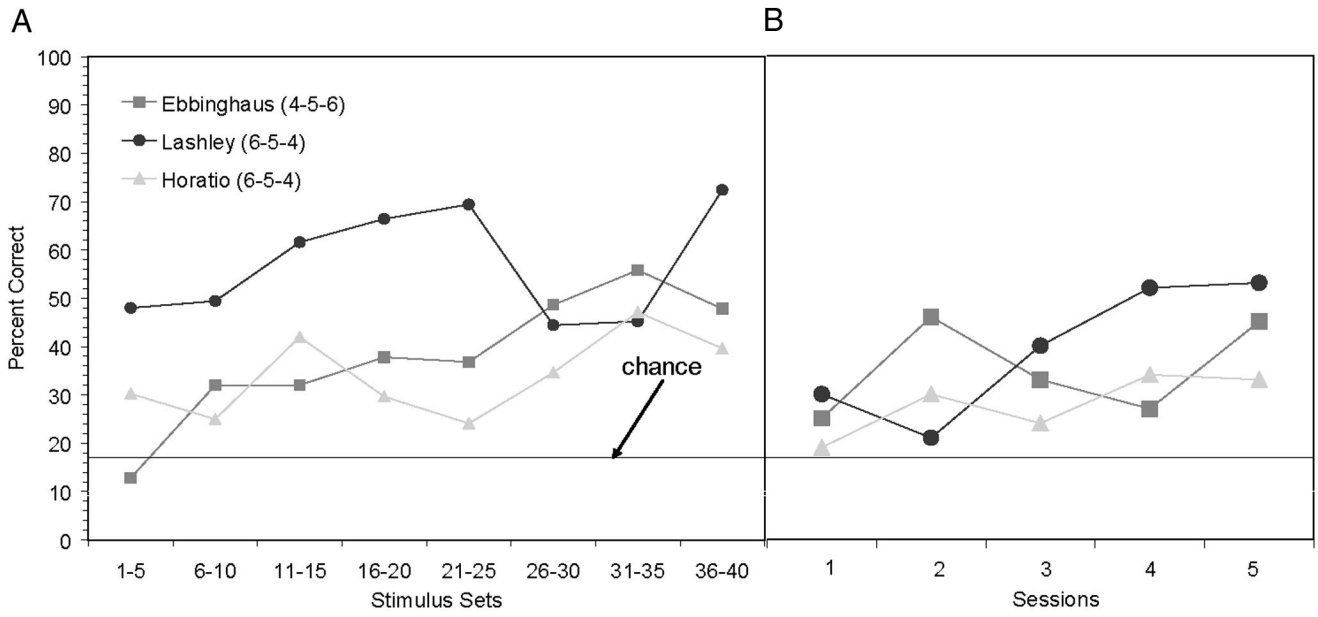

Figure 5. (A) Percentage of correctly completed trials during the first session for each of 40 training stimulus sets in blocks of five sessions in Experiment 2. (B) Percentage of correctly completed trials on the 100 test sets over five sessions. Values in parentheses indicate the rule on which each of the monkeys was trained.

Table 1
Accuracy (in Percentages) as a Function of Stimulus Type by Subject in Experiment 2

Stimulus type	Lashley	Horatio	Ebbinghaus	Average
Clip art	28	24	24	25
Area constant	56*	24	32*	37
Size constant	60*	40*	32*	44
Random size	40*	36*	20	32

* $p < .05$.

2000). As shown in Figure 6, the average accuracy of responding to stimuli from the NN category was at least as good as the accuracy of responding to FF and FN pairs. A 3 × 3 repeated measures ANOVA on data, grouped into five 5-session blocks, with a between-subjects variable of subject and a within-subject variable of pair type (FF, FN, NN), revealed a main effect of pair type, $F(2, 24) = 10.3$, $p < .001$, and a significant Subject × Pair Type interaction, $F(4, 24) = 3.92$, $p < .05$. Fisher's least significant difference (LSD) post hoc tests revealed that the main effect of pair type was due to superior performance on both NN ($p < .001$) and FN ($p < .01$) pairs relative to FF pairs. The interaction was due to different patterns of accuracy for subjects trained on the ascending and descending numerical rules. For Ebbinghaus, accuracy was highest on FN pairs and lowest on NN pairs (FN > FF > NN). For Lashley and Horatio, accuracy was highest on NN pairs and lowest on FF pairs (NN > FN > FF).

A more detailed analysis of accuracy of responding to NN pairs revealed that average accuracy was a misleading measure of performance because accuracy of responding varied systematically across three mutually exclusive types of NN pairs. NN pairs can be divided into small pairs that include two novel numerosities both smaller than the training values (*1–2*, *1–3*, or *2–3*), large pairs that include two novel values larger than the training values (*7–8*, *7–9*, *8–9*), and span pairs that include one value larger than the training values and one smaller than the training values (*1–7*, *1–8*, *1–9*, *2–7*, *2–8*, *2–9*, *3–7*, *3–8*, and *3–9*).

The accuracy of responding to each type of test pair is shown in Figure 7. All 3 monkeys performed at high levels of accuracy on NN pairs, which included one small (*1*, *2*, or *3*) and one large (*7*, *8*, or *9*) novel value (center panel). A more complicated pattern emerged when pairs were composed exclusively of small or large values. The accuracy of Ebbinghaus, the single monkey trained on the *4→5→6* sequence, was well above the level predicted by chance for large, but not for small, novel values (one sample *t* tests that compared accuracy over five test sessions with chance expectation of .50: large = 70%, $t[8] = 4.26$, $p < .001$; small = 41%, $t[8] = 1.95$, $p > .05$). The reverse was true for the two monkeys trained on the *6→5→4* sequence. Their accuracy exceeded the level expected by chance when they were tested with small novel values (91%, $t[8] = 28.97$, $p < .001$) but not with large novel values (57%, $t[8] = 1.09$, $p > .05$).

A 3 × 3 repeated measures ANOVA on the NN data grouped into blocks of five sessions, with a between-subjects variable of subject and a within-subject variable of NN pair type (small, large,

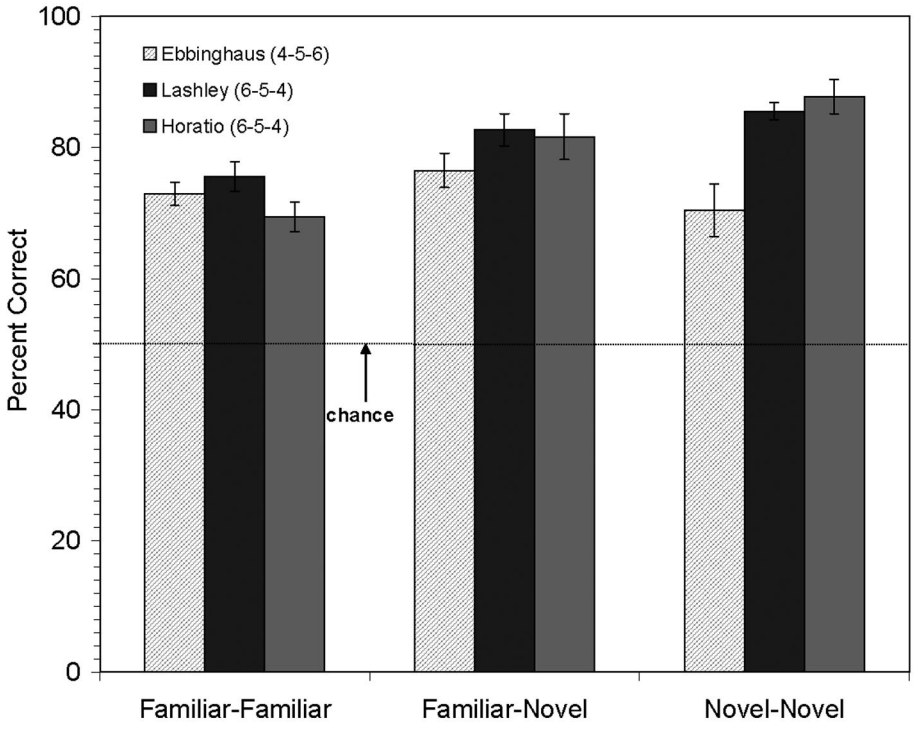

Figure 6. Performance on familiar−familiar, familiar−novel, and novel−novel numerosity pairs for Experiment 2. Only familiar−familiar pairs were reinforced. Error bars reflect variance across test sessions. Values in parentheses indicate the rule on which each of the monkeys was trained.

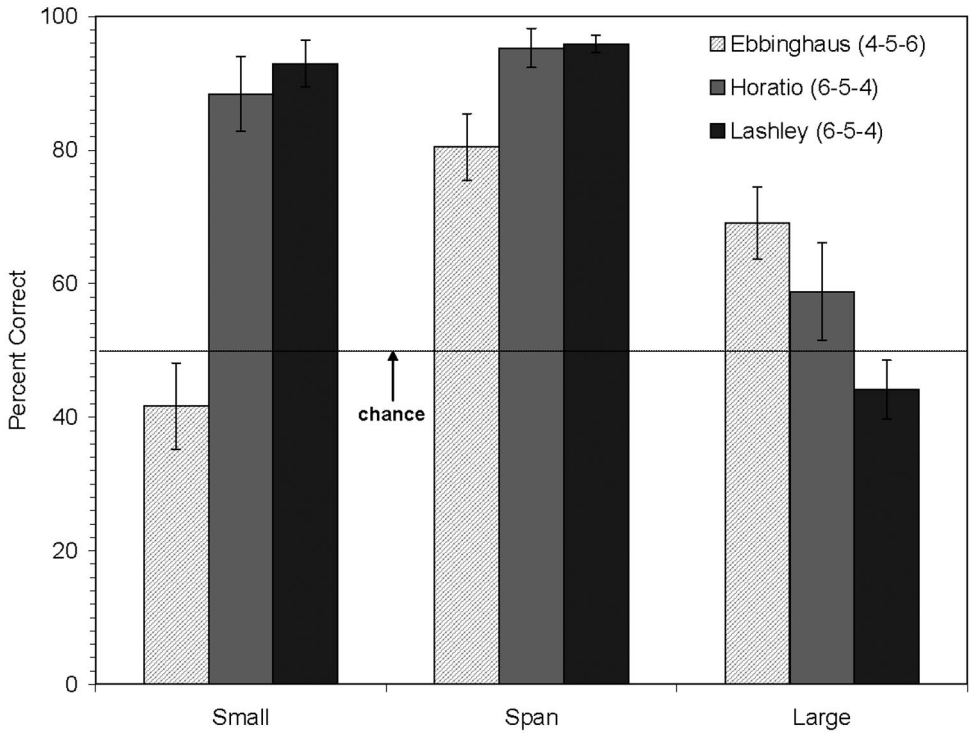

Figure 7. Performance on novel−novel numerosity pairs segregated as a function of whether the numerical values were both small (*1–2*, *2–3*, or *1–3*), both large (*7–8*, *8–9*, or *7–9*), or spanned the training values and involved one small (*1*, *2*, or *3*) and one large (*7*, *8*, or *9*) numerosity. Error bars reflect variance across test sessions. Values in parentheses indicate the rule on which each of the monkeys was trained.

span), revealed a main effect of subject, $F(2, 12) = 12.84$, $p < .01$; and pair type, $F(2, 24) = 23.49$, $p < .001$; and a Subject × Pair Type interaction, $F(4, 24) = 13.15$, $p < .001$. Fisher's LSD post hoc tests revealed that the main effect of subject was due to the overall superior performance of Lashley and Horatio, as compared with Ebbinghaus ($p < .01$). The main effect of pair type was due to superior performance on span pairs, as compared with small ($p < .05$) and large ($p < .0001$) pairs. The interaction between pair type and subject appears to be due to the opposite trends in accuracy for monkeys trained on the ascending and descending rules for pairs composed exclusively of large or small values.

It is also important to note that accuracy on small−small NN pairs for both monkeys trained on the *6→5→4* rule was extremely high and was in fact higher than accuracy on the FF pair. This pattern of results supports the conclusion that familiarity or novelty per se were poor indicators of accuracy. Indeed, accuracy on NN pairs surpassed that observed on FF pairs when the values that were compared allowed a continuation of the training sequence and when they involved comparisons with favorable Weber fractions. These issues are revisited in the next section, which provides a meta-analysis of a larger sample of data.

Experiments 1 and 2. Brannon and Terrace (1998, 2000) suggested that monkeys trained to respond to a set of numerosities in an ascending order learned a simple abstract ordinal rule, such as respond to the smaller value first, for responding on a pairwise test. Thus, a monkey faced with two novel numerical numerosities (X and Y) would simply compare them and respond to X if X < Y. The curious pattern of results obtained in the pairwise test of Experiment 2 suggests that the monkeys used a more complicated comparison rule that required comparison of each value in a pair to a third value.

We compared the performance of all 7 rhesus monkeys trained on ascending or descending numerical sequences (Experiments 1 and 2 of this study and the 3 monkeys described in Brannon & Terrace, 1998, 2000). Our analysis suggested that, in addition to the use of a general rule such as choose X if X < Y, subjects also compared X and Y with a RP, the first value of the training sequence. The basic idea is that the extensive training that each monkey received gave the first value of the training sequence a special status and that each value in a test pair was compared with the RP. The monkey chose whichever value was numerically closest to the RP. Thus the RP computation can be considered a similarity-based decision, however the similarity judgment is based on numerical similarity not a perceptual stimulus-based similarity (e.g., surface area).

The four different training rules used with macaques involved three different reference points (*1*, *4*, and *6*). During the pairwise test, the subject appears to choose the value that has the smallest absolute difference from the first value of the training sequence (RP). The monkey chooses the smaller value (S) if the absolute difference between the RP and S is less than the absolute difference between the RP and the larger value (L): Choose S if $|RP - S| < |RP - L|$ and choose L if $|RP - L| < |RP - S|$.

Consider, for example, the subset *2–3* following training on the sequence *1→2→3→4*. Where RP = *1*, |RP − L| = |*1–3*| = *2* and |RP − S| = |*1–2*| = *1*. Because |RP − S| < |RP − L|, the subject should respond to S. Training on the sequence *4→3→2→1* predicts a different outcome for the same subset. Where RP = *4*, RP − L| = |*4–3*| = *1* and |RP − *2*| = |*4–2*| = *2*. Accordingly, the subject chooses L, because |RP − L| < |RP − S|. In both cases the RP analysis predicts behavior that is in accord with the ascending or descending ordinal rule.

The RP computation predicts a more complicated pattern of accuracy following *4→5→6* and *6→5→4* training. Consider the pair *7–9*. Following training on the sequence *4→5→6*, RP = *4*, the monkey should choose the smaller numerosity because |RP − S| < |RP − L|: |RP − L| = |*4–9*| = *5*, |RP − S| = |*4–7*| = *3*. However, after *6→5→4* training, RP = *6*. Accordingly, the monkey should also choose the smaller numerosity because |RP − S| < |RP − L|: (|RP − L| = |*6–9*| = *3*; |RP − S| = |*6–7*| = *1*). Thus the RP analysis predicts that, for pairs composed of large novel values, a monkey trained on the *4→5→6* rule should exhibit choices that conform to the ordinal rule but that a monkey trained the *6→5→4* rule would not. In contrast, the RP analysis predicts that monkeys trained on the *6→5→4* rule should exhibit high levels of accuracy for subsets composed of small values (*1*, *2*, or *3*) because the RP computation is in accord with the ordinal direction of training. However, monkeys trained on the *4→5→6* rule should have difficulty with pairs composed of small values because the reference point computation predicts ordering that is opposite to that predicted by the ordinal rule.

As predicted by the RP analysis, subjects trained on ascending rules performed accurately whenever |RP − S| < |RP − L| and poorly when the reverse was true. Similarly, monkeys trained on descending sequences performed well when |RP − L| < |RP − S| and poorly when the reverse was true. The relevant data are shown in Figure 8. A 2 × 2 repeated measures ANOVA, with a between-subjects variable of training direction and a within-subject true−false variable indicating whether |RP − S| < |RP − L|, revealed no main effects but did reveal a significant interaction between those variables $F(1, 8) = 66.98$, $p < .001$. Examining the means revealed that the interaction was a consequence of the fact that accuracy was much higher for monkeys trained to respond in ascending order when |RP − S| < |RP − L|. Conversely, accuracy was higher for monkeys trained to respond in a descending order when |RP − L| < |RP − S|. It should be noted however that for 2 of the 3 monkeys trained on ascending rules, |RP − L| was never less than |RP − S|. Thus, they did not contribute to both levels of the true−false variable.

Evidence of the influence of Weber's law can be seen in Figures 9 and 10. Figure 9 shows that the ratio of the two values of each pair was a good predictor of accuracy for 6 of the 7 monkeys. Ebbinghaus was the only monkey whose accuracy was not influenced by the ratio of the values of each numerical pair. As can be seen in Figure 10, the lack of a ratio effect for Ebbinghaus reflects the fact that the ratio of numerical values and the values derived from the RP computation make opposite predictions about performance on pairs of small numerosities. Pairs with small ratios should be easier to discriminate than pairs with large ratios; however, the RP rule predicted poor performance for pairs of small numerosities and good performance for larger values. As can be observed in Figure 10, Ebbinghaus, who was trained on the descending *6→5→4* rule,

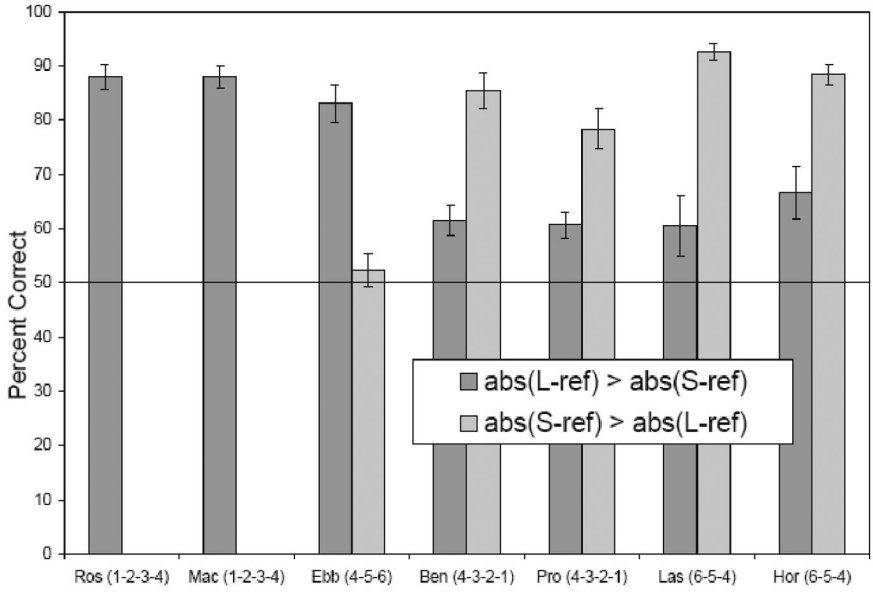

Figure 8. Percentage correct as a function of whether abs(RP − L) < abs(RP − S), shown in dark gray, or vice versa, shown in light gray. Data are from the 3 monkeys tested in Experiment 2, the single monkey tested in Experiment 1, and the 3 monkeys tested in Brannon and Terrace (2000). Values in parentheses indicate the rule on which each of the monkeys was trained. Abs = absolute value; ref = reference; Ros = Rosencrantz; Mac = Macduff; Ebb = Ebbinghaus; Ben = Benedict; Pro = Prospero; Las = Lashley; Hor = Horatio.

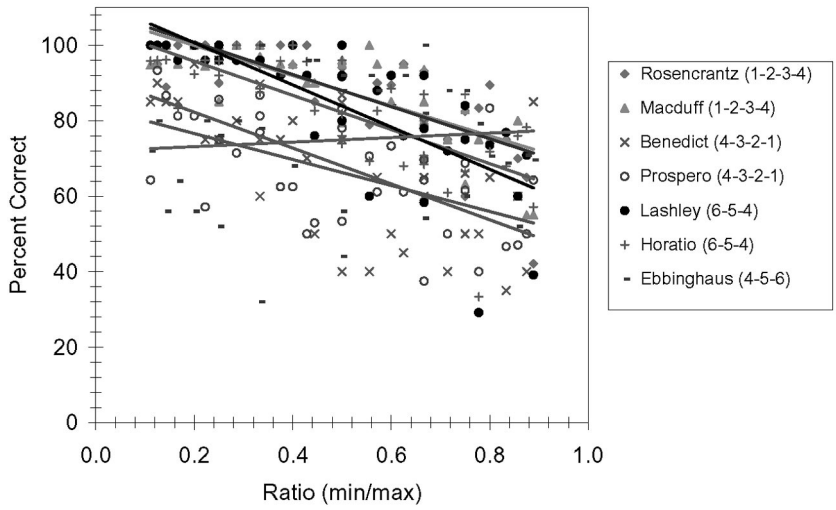

Figure 9. Accuracy as a function of ratio of the two values in a numerical test pair. Values in parentheses indicate the rule on which each of the monkeys was trained.

showed the opposite pattern from all other monkeys. We return to this point later.

To provide a quantitative analysis of the effect of the reference point on numerical judgments, we used linear regression to determine the independent contributions of (a) comparisons of each value with the reference point (|RP − L|) and (|RP − S|) and (b) the ratio of the two numerical values in each pair. With the exception of Ebbinghaus, each data point in the model was the

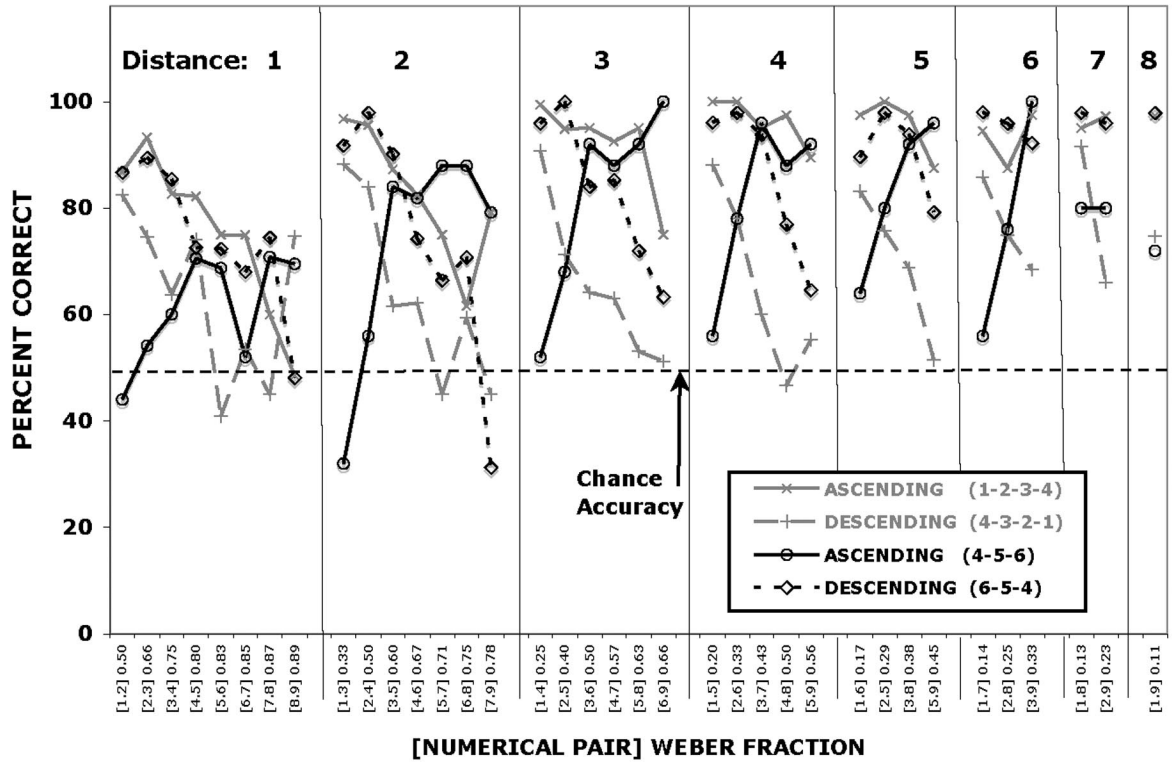

Figure 10. Accuracy as a function of a numerical pair segregated by distance clusters. Data is averaged for 2 monkeys trained on each rule except for the descending *6→5→4* rule for which only Ebbinghaus was trained. All monkeys showed increasing accuracy with distance (across clusters) and decreasing accuracy with magnitude when distance was held constant (within clusters), with the exception of Ebbinghaus, who showed an inverse magnitude effect.

average accuracy of the 2 subjects trained on a particular sequence.[3] Table 2 shows the results of our regression analysis. A model incorporating ratio and the RP similarity index accounted for between 61% and 73% of the variance. The beta values were significant for the Weber ratio for all monkeys except Ebbinghaus, the subject that was trained on the *4→5→6* rule. The beta values for the RP comparison were significant for Ebbinghaus, Lashley, Horatio, Prospero, and Benedict. In other words, including RP in the model accounted for additional variance for all monkeys who were trained on a rule for which the reference point had a value other than *1*. Indeed, it is not surprising that the RP comparison did not account for additional variance when the RP was *1* because, under these circumstances, the RP comparison was exactly equivalent to numerical distance, which is linearly related to ratio. This analysis highlights that both the discriminability of the two values being compared (their ratio) and the relative similarity of each value to RP contributed to the accuracy of rhesus monkeys, whereas making pairwise comparisons after extended training on a numerical rule.

We also obtained suggestive evidence that ordinal direction of the training sequence impacted accuracy during the pairwise test. Consider, for example, accuracy on pairs of the values *1–3*, all of which have very favorable Weber fractions. Horatio and Lashley, both trained on *6→5→4* sequences, were highly accurate when responding to such pairs because both the ordinal rule and the RP comparison predicted that the monkeys would respond in descending order. Horatio and Lashley ordered these pairs accurately on 93% and 85% of trials, respectively. By contrast, Ebbinghaus (the monkey that was trained on *4→5→6* sequences) ordered those pairs correctly on 43% of trials. This suggests that the ordinal rule and the RP computation may have worked additively in the case of the *6→5→4* rule but that they competed with each other in the case of the *4→5→6* rule. A similar pattern was observed for pairs that have relatively high Weber fractions. Both the ordinal rule and the RP analysis predicted that Ebbinghaus would respond accurately to pairs that were composed of the numerosities *7*, *8*, and *9*. He did so on 73% of trials. By contrast, the ordinal rule predicted that Horatio and Lashley (the two *6→5→4* monkeys) would respond to pairs that were composed of the numerosities *7*, *8*, and *9* in a descending order, whereas the RP comparison predicted response to an ascending order. Horatio and Lashley responded to those pairs correctly on only 56% and 46% of the trials, respectively.

Accuracy was considerably above chance when both the RP computation and the ordinal rule predicted the same outcome and was at chance when the two rules had opposite predictions. If the RP computation were the only factor driving monkeys' choice behavior when presented with novel values, then accuracy should be significantly below chance when the RP computation predicts ordering that conflicts with the ordinal rule, just as it is significantly above chance when it predicts ordering that is in accord with the ordinal rule. In contrast, performance was at chance when the two rules made contrary predictions, suggesting that two factors were at play. Another source of evidence that monkeys used an ordinal rule as well as the RP computation comes from their accuracy on pairs that were composed of two values equidistant from the RP. Ebbinghaus, trained on the *4→5→6* rule, ordered two versus eight pairs with 76% accuracy and two versus six pairs with 78% accuracy. These pairs are important because they are equidistant in ratio or linear distance from the reference point. If the RP were the only factor driving choices, then performance should be at chance on at least one of these two pairs depending on whether the monkey relies on a ratio or linear distance calculation. Similarly, Lashley and Horatio, trained on the *6→5→4* rule, ordered three versus nine pairs with 96% and 88.5% accuracy and four versus nine pairs with 76% and 83% accuracy.

Table 2
Regression Values for the Effect of Ratio and RP Computation on Accuracy

Training rule	R^2	β (ratio)	β ($\lvert RP - L \rvert - \lvert RP - S \rvert$)
1→2→3→4	0.64*	−1.0*	−0.30
4→3→2→1	0.74*	−0.58	−0.56*
4→5→6	0.69*	−0.04	0.84*
6→5→4	0.69*	−0.46*	0.45*

Note. RP = reference point; L = large numeric value; S = small numeric value.
* $p < .05$.

Additional evidence that monkeys used both an ordinal rule and a RP computation and that these two variables interacted with each other comes from performance on familiar−novel pairs. If monkeys are choosing the value closest to the RP, then monkeys should predominantly choose the familiar value before the novel value on these pairs. Thus the monkey trained on the *4→5→6* rule should perform better when a familiar value (*4*, *5*, or *6*) is paired with a large novel value because this allows a correct first response to the familiar value. In accordance with that prediction, a dependent sample *t* test comparing the nine FN pairs involving a small novel value with the nine FN pairs involving a large novel value revealed that Ebbinghaus performed significantly better on FN pairs with a large novel value compared with FN pairs with a small novel value, $t(8) = -3.74$; $p < .01$. In contrast, the monkeys trained on the *6→5→4* rule should perform better when the familiar value (*4*, *5*, or *6*) is paired with a small novel value because this allows a first response to the familiar value according to the descending rule. In accordance with that prediction, a dependent sample *t* test comparing the nine FN pairs involving a small novel value with the nine FN pairs involving a large novel value revealed that Horatio and Lashley performed significantly better on FN pairs with a small novel value compared with FN pairs with a large novel value (Horatio, $t[8] = 4.55$, $p < .01$; Lashley, $t[8] = 9.25$; $p < .001$).

In summary, the results of Experiments 2 suggest that the accuracy of the monkeys was affected by three factors: Weber's law, which limited their ability to discriminate the values in a test pair; a computation of relative similarity of each numerical value to a numerical RP; and the ordinal direction of the training sequence.

General Discussion

The results of Experiments 1 and 2 add to our understanding of how rhesus monkeys compare numerical magnitudes. Experiment

[3] The sole motivation for averaging the data obtained from individual subjects was to simplify the presentation of their performance. In no instance did the averaged data distort the performance of individual subjects, and this can be verified by consulting Figure 9 of Brannon and Terrace (2000), which shows individual accuracy functions for Macduff and Rosencrantz.

1 supports the conclusion that monkeys trained to order the values *1*, *2*, *3*, and *4*, in a descending rather than an ascending order, are unable to apply the descending rule to pairs of novel numerical values. In Experiment 2, however, we showed that, under some circumstances, monkeys trained on a descending numerical rule (*6→5→4*) could respond accurately to novel numerical stimuli. In fact, both monkeys trained on the *6→5→4* rule performed better on pairs of the small novel values *1*, *2*, and *3* than they did on pairs of familiar values, demonstrating that neither familiarity nor novelty were as good a predictor of accuracy as was Weber's law.

Our previous reports (Brannon & Terrace, 1998, 2000) have suggested that a monkey's ability to order responses to novel numerical values implies that the monkey had extracted a simple ordinal rule from such training; that is, a greater than rule following training on the sequence *1→2→3→4*. The results of Experiment 2 help to further define the type of computation that rhesus monkeys use when making ordinal numerical comparison. Monkeys trained to respond to the values *4–6* in descending order reliably ordered pairs of the small novel values *1*, *2*, and *3* accurately but performed at chance levels of accuracy with pairs of the large values *7*, *8*, and *9*. By contrast, the single monkey trained to respond in ascending order to the values *4–6* displayed the reverse pattern of results by accurately ordering large but not small novel values.

This seemingly curious pattern of results can be understood by positing that monkeys establish the first value in a training sequence as an RP and that, during a pairwise test, they compare each numerical value with the RP. During the pairwise test, the value of the RP is compared with the values of each test stimulus. Referring to S as the smaller value of the test pair, and L as the larger value, the subject computes the absolute values of RP − L and RP − S. The subject responds to S if $|RP - S| < |RP - L|$ and to L if the reverse is true. As shown in Figure 8, this rule predicted the relative accuracy of responses on the pairwise tests used in this and in previous experiments on numerical comparisons by monkeys (Brannon & Terrace, 1998, 2000).

The concept of an RP has also been used to explain the semantic congruity effect observed in adult humans on a wide variety of magnitude comparison tasks. The general finding is that subjects are faster to make a response when the framing of the instructions is congruent with their internal response codes. For example, subjects are slower to respond to the question *Which is smaller, an elephant or a gorilla?* than to the question *Which is larger, an elephant or a gorilla?* The question, *Which is smaller?* prompts a search for exemplars of small animals, but elephants and gorillas do not normally qualify as small animals (Clark, 1969). A congruence principle has also been used to explain the performance of human subjects asked to make numerical comparisons. For example, Banks and colleagues (Banks, Fuji, & Kayra-Stuart, 1976) showed that subjects choose the larger of two relatively large digits more rapidly than they choose the smaller (e.g., *8* or *9*) and vice versa when tested with small values such as *1* versus *2*. In a separate study, we have recently obtained evidence that monkeys show similar congruence effects when making numerical comparisons. That study provides further evidence of an anchor point in the pairwise numerical judgments of monkeys (Cantlon & Brannon, 2005).

One class of explanations for such congruity effects in comparison tasks highlights the importance of reference points (e.g., Holyoak, 1978; Jamieson & Petrusic, 1975). These models assume that the two values in a test pair are never directly compared, instead the relative distance between each value and an RP is compared and the value with the smaller distance is chosen (Dehaene, 1989). Our data are consistent with these models and suggest that the ascending or descending training that a monkey receives in the laboratory functions to establish a RP that is later used in pairwise comparisons. Although the results of Experiment 2 suggest that monkeys compare each of the two values in a test pair with a RP, there was also evidence that subjects encoded the ordinal direction of the training sequence defined by the relationship between the RP and the next value of the sequence (greater than or less than). If monkeys' accuracy in the pairwise test was based solely on relative similarity to the RP and the discriminability of the two values as defined by Weber's law, accuracy should have been significantly below chance for large NN pairs for the monkeys trained on the *6→5→4* descending rule and the small NN pairs for the monkey trained on the *4→5→6* ascending rule. The fact that accuracy was at chance on these reversal pairs suggests that there were two competing factors in play. The two factors were (a) an abstract ordinal numerical rule and (b) an RP relative-similarity computation. Further evidence that an ordinal rule could overcome the RP computation is that all 3 monkeys reliably ordered pairs of one familiar and one novel numerosity, even when the ordinal rule conflicted with the RP computation and required choosing the novel value before the familiar value.

It is important to note that the role of the RP may have been accentuated by our experimental paradigm. In Experiments 1 and 2, and in the procedure used by Brannon & Terrace (1998, 2000), monkeys were trained on a specific numerical rule (e.g., *4→3→2→1*) for approximately 3,000 trials over the course of 4–5 months before being tested with pairs of novel values. Such overtraining on a specific numerical sequence may have emphasized the first value of the sequence and deemphasized abstract ordinal information. In the absence of overtraining on a single sequence, monkeys might be more likely to extract the abstract relations between training values and ignore absolute numerical information. For example, if monkeys were trained on pairs that randomly varied (e.g., *1–2*, *2–3*, *3–4*), they might be less likely to establish a RP and instead rely on a more general ascending numerical rule.

It is also important to note that the results of these experiments, together with previous studies on ordinal numerical knowledge in nonhuman primates, provide no support for the view that primates use two distinct mechanisms to represent small and large numerical values (e.g., Hauser et al., 2000). Studies with human infants have suggested that infants represent large numerosities as approximate magnitudes and that their discriminations are modulated by Weber's law (e.g., Brannon, Abbott, & Lutz, 2004; Lipton & Spelke, 2003; Xu, 2003; Xu & Spelke, 2000), whereas small numerosities are represented by an object file system (e.g., Feigenson & Carey, 2003; Spelke, 2000; Xu, 2003). In addition, other laboratories have found that infants have specific difficulty differentiating two from four elements. This suggests that they may represent *2* and *4* with incompatible representational systems (e.g., Feigenson & Carey, 2003; Xu, 2003). However, Brannon and Terrace (1998, 2000) found that monkeys trained on the small values *1–2–3–4* had no difficulty extrapolating to the large values *5*, *6*, *7*, *8*, and *9*. Similarly, in Experiment 2, monkeys trained to

respond in descending order to the large values *6→5→4* performed exceedingly well with the small numerical values *1*, *2*, and *3*. Collectively, these data suggest that rules learned within the small number range can be extended to large values, and vice versa, and that there is no qualitative difference between formats for representing small and large numerosities.

In conclusion, our findings provide additional evidence that monkeys have sophisticated mechanisms for extracting the abstract numerical value of visual stimuli and that numerical discrimination in rhesus macaques is modulated by Weber's law. Our results also provide new evidence that the numerical comparison algorithm used by rhesus monkeys is anchored to the specific numerical values learned during training.

References

Banks, W. P., Fuji, M., & Kayra-Stuart, F. (1976). Semantic congruity effects in comparative judgments of magnitudes of digits. *Journal of Experimental Psychology: Human Perception and Performance, 2,* 435–447.

Beran, M. J. (2001). Summation and numerousness judgments of sequentially presented sets of items by chimpanzees (*Pan troglodytes*). *Journal of Comparative Psychology. 115,* 181–191.

Boysen, S. T., & Berntson, G. G. (1989). Numerical competence in a chimpanzee (*Pan troglodytes*). *Journal of Comparative Psychology, 103,* 23–31.

Brannon, E. M. (2005). What animals know about number. In J. I. D. Campbell (Ed.), *Handbook of mathematical cognition* (pp. 85–108). New York: Psychology Press.

Brannon, E. M., Abbott, S., & Lutz, D. (2004). Number bias for the discrimination of large visual sets in infancy. *Cognition, 93,* B59–B68.

Brannon, E. M., & Roitman, J. (2003). Nonverbal representations of time and number in non-human animals and human infants. In W. Meck (Ed.), *Functional and neural mechanisms of interval timing* (pp. 143–182). New York: CRC Press.

Brannon, E. M., & Terrace, H. S. (1998, October 23). Ordering of the numerosities 1–9 by monkeys. *Science, 282,* 746–749.

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.

Cantlon, J. F. & Brannon, E. M. (in press) The effect of heterogeneity on numerical ordering in rhesus monkeys. *Infancy.*

Cantlon, J. F. & Brannon, E. M. (2005). Semantic congruity facilitates number judgments in monkeys. *Proceedings of the National Academy of Sciences, 102,* 16507–16511.

Church, R. M., & Meck, W. H. (1984). The numerical attribute of stimuli. In H. L. Roitblat, T. G. Bever, & H. S. Terrace (Eds.), *Animal cognition* (pp. 445–464). Hillsdale, NJ: Erlbaum.

Clark, H. H. (1969). Linguistic processes in deductive reasoning. *Psychological Review, 76,* 387–404.

Cohen, J., MacWhinney, B., Flatt, M, & Provost, J. (1993). PsyScope: An interactive graphic system for designing and controlling experiments in the psychology laboratory using Macintosh computers. *Behavior, Research, Methods, Instruments, and Computers, 25,* 257–271.

Dehaene, S. (1989). The psychophysics of numerical comparison: A reexamination of apparently incompatible data. *Perception and Psychophysics, 45,* 557–566.

Feigenson, L., & Carey, S. (2003). Tracking individuals via object files: Evidence from infants' manual search. *Developmental Science, 6,* 568–584.

Gallistel, C. R., & Gelman, R. (2000). Non-verbal numerical cognition: From reals to integers. *Trends in Cognitive Sciences, 4,* 59–65.

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 semifree-ranging rhesus monkeys. *Proceedings of the Royal Society London, 267,* 829–833.

Hauser, M. D., MacNeilage, P., & Ware, M. (1996). Numerical representations in primates. *Proceedings of the National Academy of Science, 93,* 1514–1517.

Holyoak, K. (1978). Comparative judgments with numerical reference points. *Cognitive Psychology, 10,* 203–243.

Jamieson, D. G., & Petrusic, W. M. (1975). Relational judgments with remembered stimuli. *Perception and Psychophysics, 18,* 373–378.

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.

Judge, P. G., Evans, T. A., & Vyas, D. K. (2004). Ordinal representation of numeric quantities by brown capuchin monkeys (*Cebus apella*). *Journal of Experimental Psychology: Animal Behavior Processes, 31,* 79–94.

Lewis, K., Jaffe, S., & Brannon, E. M. (2005). Analog number representations in mongoose lemurs (*Eulemur mongoz*): Evidence from a search task. *Animal Cognition, 8,* 247–251.

Lipton, J. S., & Spelke, E. S. (2003). Origins of number sense: Large-number discrimination in human infants. *Psychological Science, 14,* 396–401.

Matsuzawa, T. (1985). Use of numbers by a chimpanzee. *Nature, 315,* 57–59.

Nieder, A., & Miller, E. K. (2004). Analog numerical representations in rhesus monkeys: Evidence for parallel processing. *Journal of Cognitive Neuroscience, 16,* 889–901.

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. (1987). Evidence for conceptual quantitative abilities in the African grey parrot: Labeling of cardinal sets. *Ethology, 75,* 37–61.

Smith, B. R., Piel, A. K., & Candland, D. K. (2003). Numerity of a socially housed hamadryas baboon (*Papio hamadryas*) and a socially housed squirrel monkey (*Saimiri sciureus*). *Journal of Comparative Psychology, 117,* 217–225.

Spelke, E. S. (2000). Core knowledge. *American Psychologist, 55,* 1233–1243.

Terrace, H. S. (1984). Simultaneous chaining: The problem it poses for traditional chaining theory. In M. L. Commons, R. J. Herrnstein, & A. R. Wagner (Eds.), *Quantitative analyses of behavior: Discrimination processes* (pp. 115–138). Cambridge, MA: Ballinger Publishing.

Terrace, H. S., Son, L. K., & Brannon, E. M. (2003). Serial expertise of rhesus macaques. *Psychological Science, 14,* 66–73.

Washburn, D. A., & Rumbaugh, D. M. (1991). Ordinal judgments of numerical symbols by macaques (*Macaca mulatta*). *Psychological Science, 2,* 190–193.

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, 89,* B15–B25.

Xu, F., & Spelke, E. S. (2000). Large number discrimination in 6-month-old infants. *Cognition, 74,* B1–B11.

Received April 15, 2005
Revision received October 5, 2005
Accepted December 6, 2005 ■