Current Biology, Vol. 15, 1034–1038, June 7, 2005, ©2005 Elsevier Ltd All rights reserved. DOI 10.1016/j.cub.2005.04.056

Monkeys Match the Number of Voices They Hear to the Number of Faces They See

Kerry E. Jordan,[1,2] Elizabeth M. Brannon,[1,2,*] Nikos K. Logothetis,[3] and Asif A. Ghazanfar[3,4,*]
[1]Center for Cognitive Neuroscience
[2]Department of Psychological and Brain Sciences
Duke University
Box 90999
Durham, North Carolina 27708
[3]Max Planck Institute for Biological Cybernetics
Spemannstrasse 38
72076 Tuebingen
Germany

Summary

Convergent evidence demonstrates that adult humans possess numerical representations that are independent of language [1–6]. Human infants and nonhuman animals can also make purely numerical discriminations, implicating both developmental and evolutionary bases for adult humans' language-independent representations of number [7, 8]. Recent evidence suggests that the nonverbal representations of number held by human adults are not constrained by the sensory modality in which they were perceived [9]. Previous studies, however, have yielded conflicting results concerning whether the number representations held by nonhuman animals and human infants are tied to the modality in which they were established [10–15]. Here, we report that untrained monkeys preferentially looked at a dynamic video display depicting the number of conspecifics that matched the number of vocalizations they heard. These findings suggest that number representations held by monkeys, like those held by adult humans, are unfettered by stimulus modality.

Results and Discussion

If nonverbal number representations are independent of stimulus modality, prelinguistic human infants and nonlinguistic animals should be able to detect the numerical correspondence between sets of entities presented in different sensory modalities. In other words, even without linguistic coding, infants and monkeys should appreciate the correspondence between, for example, three dog barks and three wags of a tail. Data from studies concerning this prediction, however, are controversial. Starkey and colleagues tested whether infants detect numerical correspondences between visual and auditory stimuli by presenting side-by-side slides of two or three household objects while a hidden experimenter hit a drum two or three times [10, 11]. Infants preferentially looked toward the visual display that numerically matched the number of drumbeats. Unfortunately, other researchers had difficulty replicating these results, finding in some cases no preference for the matching visual array and in others a reverse preference for the nonmatching array [12, 13].

The few laboratory studies of crossmodal number representation in animals have also yielded conflicting results. Church and Meck found that rats trained to discriminate two from four sounds or light flashes later responded to compound cues of two lights and two sounds as if four events had occurred, suggesting that rats can transfer numerical representations across modalities [14]. Yet, when Davis and Albert trained rats to discriminate three sequentially presented sounds from two or four sounds and then exposed rats to sequences of two, three, and four lights, they found no evidence that the rats transferred their auditory numerical discrimination to the visual modality [15]. The Davis and Albert [15] results raise the possibility that the rats in the Church and Meck study [14] made a dichotomous discrimination that was purely intensity based (i.e., they equated the less-intense sound with the less-intense light).

Field playback studies have yielded suggestive evidence that animals predict the number of intruders they expect to see on the basis of the number of vocalizing intruders they hear. In these studies, the probability that an animal from a focal group will approach a speaker emitting vocalizations from foreign conspecifics depends on the relationship between the number of vocalizing foreign animals and the number of animals present in the focal group [16, 17]. For example, McComb and colleagues found that lions were more likely to approach a speaker emitting the roar of a single unfamiliar lion than a chorus of three unfamiliar lions, suggesting that lions decide whether to defend their territory on the basis of the perceived number of intruders [16]. However, such studies did not control for all possible nonnumerical auditory cues that covary with number (e.g., some aspects of duration), leaving open the question of whether the calculations made by the animals were in fact based on number. Thus, the existence of evolutionarily primitive, modality-independent, nonverbal numerical representations remains an open question.

Here, we explicitly test whether rhesus monkeys (*Macaca mulatta*) match the number of animals they see with the number of vocalizations they hear. To test this, we employed a preferential-looking paradigm that has been used extensively with human infants. Both infants and monkeys prefer to look at the visual stimulus that matches the auditory stimulus they hear [18–22]. For example, when either human infants [19] or rhesus monkeys [21] hear a conspecific vocalization, they look preferentially at a face that articulates the vocalization they hear in comparison with a face that articulates a different vocalization. Given the established social expertise of nonhuman primates [23–26] and the suggestion that wild animals use the number of vocalizations they hear to make defensive decisions [16, 17], we ex-

*Correspondence: brannon@duke.edu (E.M.B.); asifg@princeton.edu (A.A.G.)
[4]Present address: Department of Psychology, Green Hall, Princeton University, Princeton, New Jersey 08544.

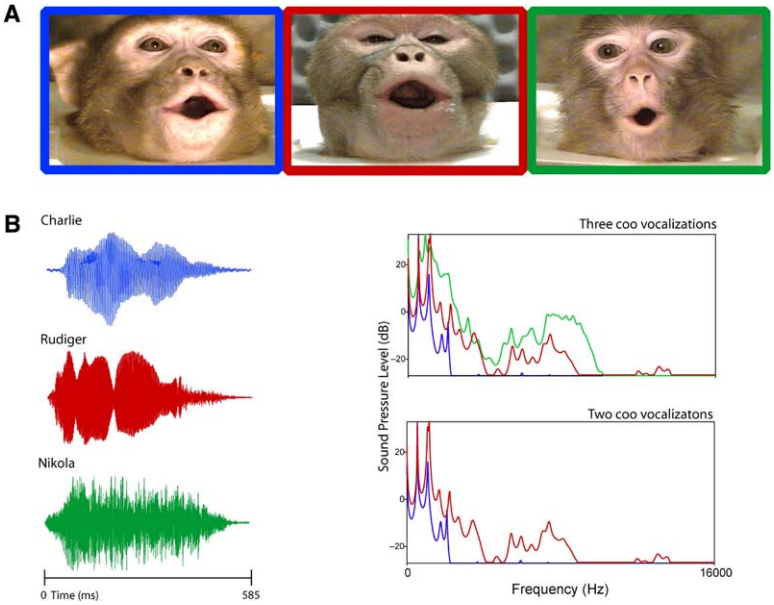

Figure 1. Faces and Voices during Concurrent Coo Vocalizations

(A) Still frames extracted from a stimulus set used in Experiment 1.

(B) Coo vocalizations are tonal, harmonically rich calls produced in affiliative contexts. The first panel shows the time-amplitude waveforms of the coo vocalizations from the individuals depicted in (A). The second panel shows the distinct but overlapping power spectra of the concurrent coos. The spectra were smoothed with a linear predicitive coding (LPC) filter in Praat 4.2 (www.praat.org) for display purposes only.

pected that framing a numerical problem within a social context would increase the probability of successful matching across stimulus modalities. We specifically tested whether monkeys would preferentially attend to dynamic visual displays featuring the number of unfamiliar conspecifics they simultaneously heard vocalizing (Figures 1A and 1B).

We chose to test discrimination between the quantities two versus three because these were the quantities used in all previous studies of this sort with human infants [10–13]. Each of 20 subjects was seated in front of two liquid crystal display (LCD) monitors and a hidden speaker located between the monitors. One monitor displayed a dynamic 1 s video of two simultaneously vocalizing monkey faces, and the other monitor displayed a dynamic 1 s video of three simultaneously vocalizing monkey faces. Each video played in a continuous loop for 60 s. The two videos contained two common animals such that the two-animal display was a subset of the three-animal display (Figure 1A). Videos were edited so that the onset and offset of all individuals' mouth movements were synchronous. Synchronously with the videos, subjects heard either two or three of these monkeys simultaneously producing coo calls. Three different stimulus sets were used (each composed of one two-animal display and one three-animal display). Sets 1 and 2 contained female rhesus monkeys (at Duke University), whereas set 3 contained male monkeys (from the Max Planck Institute for Biological Cybernetics) that had long been deceased and were thus unknown to the subjects. Individual coos were equated for duration, and composite auditory stimuli were equated for amplitude (Figure 1B). Because all visual and auditory components were identical in duration and synchronized, the subjects could not use amodal cues (such as rate) to make a match. All trials were recorded on digital video tape and later acquired and scored blind by independent observers. Thus, our paradigm addressed whether monkeys would spontaneously preferentially attend to a visual stimulus that was numerically equivalent to the number of coo calls they heard. Because the two or three coo calls were heard concurrently, this pattern of looking would also demonstrate auditory-stream segregation of the voices.

Monkeys spent a greater proportion of time looking at the display that numerically matched the number of vocalizers they heard than at the numerically nonmatching display. Monkeys looked at the matching display for 60% of the total time that they spent looking at either screen; this proportion differed significantly from chance [$t(19) = 3.00$, $p < 0.01$] (Figure 2A). On average, monkeys looked at the matching display for 14.2 ± 2.0 s and the nonmatching display for 9.2 ± 1.2 s (Figure 2B). A 2 (match versus nonmatch) × 2 (two sounds versus three sounds) × 3 (stimulus set 1, 2, or 3) analysis of variance (ANOVA) revealed that the monkeys looked longer at the numerically matching display than at the nonmatching display [$F(1,15) = 7.5$, $p < 0.02$] and that there were no other main effects or interactions. Thus, the effect held both for the monkeys who heard two calls and for the monkeys who heard three calls. Finally, 15 out of the 20 monkeys tested looked longer at the matching display than at the nonmatching display [$p < 0.022$, sign test] (Figure 2C).

These results suggest that rhesus monkeys segregated two or three simultaneously presented vocalizations and detected the numerical correspondences between the calls they heard and the vocalizing faces they saw. Importantly, because we used a between-subjects design (see Experimental Procedures), each monkey experienced only one trial and heard only the two- *or* only the three-composite call stimulus—not both. Thus, the subjects could not have learned to map the more intense or the more complex auditory stimulus to the more intense/complex visual stimulus. Previous studies with human infants have fallen prey to this argument because they used a within-subject design; infants consequently heard two- and three-sound stimuli and

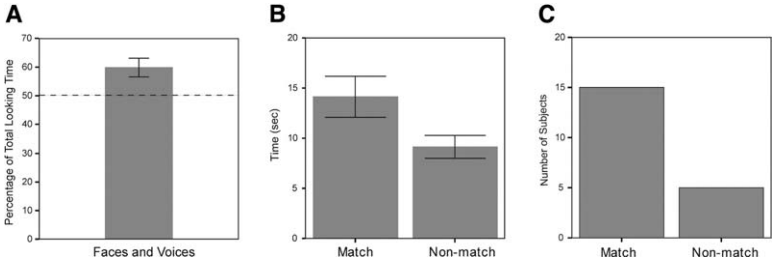

Figure 2. Monkeys Match Number of Faces with Number of Voices

(A) The mean percentage (± standard error of the mean [SEM]) of total looking time spent looking at the matching video display; the dotted line indicates chance expectation.

(B) The mean duration (± SEM) of time spent looking at the match versus the nonmatch displays.

(C) A significant proportion of subjects looked longer at the match than the nonmatch.

saw two- and three-element arrays in a single session and could have learned to match the more intense/complex stimuli in each modality. In contrast, our design provided no basis for learning to match more intense/complex stimuli within the experimental setting. Controlling for auditory cues that often covary with number also made it unlikely that monkeys could use a priori expectations to spontaneously map two (or three) sounds to a continuous property of the visual stimulus (e.g., calls of this amplitude are usually paired with a certain surface area of monkey face). Furthermore, because the visual and auditory components were identical in duration and synchronized, the monkeys could not have used rate, duration, or synchrony cues as a basis for matching. It was also not possible to match auditory to visual stimuli on the basis of the presence or absence of a particular individual monkey because all stimulus animals were unknown to subjects. From these data, we posit that, without any explicit training, rhesus monkeys (A) can represent the equivalence between the number of voices they hear and the number of faces they see and (B) are capable of concurrent-stream segregation of voices with overlapping spectra at a level comparable to that of humans [27, 28].

Our experimental design was motivated by an intuition that a monkey would be more likely to numerically match across modalities if the problem were made socioecologically relevant. The nonarbitrary connection between our visual and auditory stimuli is in direct contrast to all previous studies with human infants [10–13]. These earlier studies used either slides of randomly selected household objects [10–12] or black dots [13] and paired them with sequential drumbeats. Our results suggest the possibility that this experiment succeeded because it tapped a socioecologically relevant scenario [23, 24, 29]. In their everyday lives, gregarious and territorial animals would be aided by the ability to segregate overlapping vocal signals and predict the number of individuals they will likely encounter on the basis of the number of individuals heard [16, 17]. However, it is impossible to conclusively state that the social nature of the stimuli in our experiment was critical for numerical matching; it is possible that *any* nonarbitrary auditory-visual pairing of stimuli familiar to the subjects may also be effective (e.g., an impact sound synchronized with dynamic videos of variable numbers of fruits hitting the ground).

Conclusions

The results of our experiment suggest three important conclusions. First, rhesus monkeys recognize the correspondence between three (or two) vocalizations and three (or two) faces; this spontaneous, multisensory number representation in nonhuman animals is an important, clear parallel to adult-human nonverbal number representations [9]. Second, these results suggest that rhesus monkeys can segregate simultaneously presented conspecific coo vocalizations, even though the power spectra of the calls are highly overlapping. This capability is on par with the perceptual separation of voices by humans via pitch differences (i.e., fundamental frequency) and harmonicity [27, 30]. This is notable because a previous study found that highly trained monkeys could discriminate concurrent sequences of *artificial* sounds only when their frequency ranges did not overlap [31]. Last, our results also suggest that the use of auditory and visual stimuli that are ecologically relevant and/or have nonarbitrary associations may be important for subjects to detect the numerical correspondence between modalities. Future studies should attempt to determine whether it is the social or more generally the nonarbitrary nature of the stimuli that allows for crossmodal numerical matching. Future studies should also extend this paradigm to other numerical values, which may help determine the numerical representational system underlying this ability [8]. Regardless, these data strongly support the contention that monkeys share with adult humans language-independent number representations that are unfettered by stimulus modality.

Experimental Procedures

Subjects

We tested 20 male rhesus macaques (age range: 4–13 years) from a large colony housed at the Max Planck Institute for Biological Cybernetics. All monkeys tested looked at both screens during the course of the 60 s trial and were therefore all included in the final analyses. Animals are socially housed and provided with enrichment objects (toys, hammocks, ropes, etc.). All experimental procedures were performed in accordance with the local authorities (Regierungspraesidium) and the European Community (EUVD 86/609/EEC) for the care and use of laboratory animals.

Stimuli

The stimuli were digital video recordings of seated rhesus monkeys spontaneously producing coo vocalizations. Two stimulus sets were based on videos of female monkeys from Duke University, and one stimulus set was based on 3-year-old digital videos of now-deceased male monkeys from the Max Planck Institute for Biological Cybernetics. These videos were then acquired onto a computer and manipulated as needed in Adobe Premiere 6.0 (www.adobe.com). We extracted the audio track from the digital video samples. Calls were sampled at 32 kHz and normalized to the peak amplitude, and then two of the three calls were temporally ex-

panded to match the call of the longest duration; the two corresponding videos were also expanded accordingly. The fundamental frequencies of the calls within a stimulus set did not overlap. We constructed two- or three-call tracks by mixing them down in Adobe Audition 1.0 (www.adobe.com) and then equating their average root mean square (RMS) power.

Stimulus Presentation and Testing Procedure

The "two" versus "three" visual stimuli were played simultaneously on side-by-side 15 inch LCD monitors (Acer FP559, www.global.acer.com). Audio tracks were synchronized with both videos and played through a hidden speaker (a self-powered Advent AV750 speaker) placed directly between and slightly behind the monitors. The RadLight 3.03 Special Edition software video player (www.radlight.net) was used to play the videos in synchrony. Sounds were presented at an intensity of 72–75 dB (A-weighted) SPL as measured with a Brüel & Kjær 2238 Mediator sound-level meter (www.bksv.com) at a distance of 72 cm. For testing, a subject was brought to the testing room and placed in front of the two monitors at a distance of 72 cm. The monitors were 65 cm apart (center-to-center distance) and at eye-level with the subject. All trials were videotaped with a digital video camera placed above and between the monitors. All equipment, except for the monitor screens and the lens of the camera, was concealed by a thick black curtain. The experimenter monitored subject activity from outside of the room. During this time, the subject's attention was directed to the center by the flashing of a 1.2 W light placed centrally between the two monitors. A test session began when the subject looked centrally. A trial consisted of the two videos played in a continuous loop for 60 s with one of the two sounds also played in a loop through the speaker. The left-right position of the two dynamic visual stimuli was counterbalanced. Each subject was only tested once, and all trials were recorded on digital video. We used a between-groups design because, as in all studies that examine the spontaneous behavior of animals and prelinguistic human infants, the subjects quickly habituate to the testing environment [21, 22]. No reward or training was provided.

Video Scoring

We collected high-quality, close-up digital videos of the subjects' behavior with a JVC GR-DVL805 digital camera (www.jvc.com). Videos were acquired at 30 frames/s (frame size: 720 × 480 pixels) onto a PC via an IEEE 1394a input and Adobe Premiere 6.0 software (www.adobe.com). The audio tracks were acquired at a 32 kHz sampling rate and 16-bit resolution. Clips for analysis were edited down to 60 s, starting with the onset of the auditory track.

The total duration of a subject's looking toward each video (left or right) was recorded and expressed as the proportion of total time spent looking at either screen. Scoring which of the screens the monkey subject was looking toward was absolutely unambiguous. The screens were far apart in the horizontal dimension, fairly close to the monkey's face and at eye level. Thus, the monkey had to make large eye and head movements to look to one screen or the other, and it was similarly clear when he was not looking at either screen. To validate this, 50% of videos were scored by a second observer blind to the experimental condition in order to determine interobserver reliability, which was 0.952 ($p < 0.0001$) as measured by a Pearson r test.

Acknowledgments

We thank Evan MacLean for coding videos and Susan Carey, Marc Hauser, David Lewkowicz, and Joost Maier for comments on an earlier draft of the manuscript. This work was supported by a National Science Foundation Graduate Fellowship to K.E.J., a John Merck Fund fellowship and NICHD ROI (HD49912) to E.M.B., and the Max Planck Society (A.A.G. and N.K.L.).

Received: February 18, 2005
Revised: March 31, 2005
Accepted: April 22, 2005
Published: June 7, 2005

References

1. Varley, R.A., Nicolai, J.C., Klessinger, C., Romanowski, A.J., and Siegal, M. (2005). Agrammatic but numerate. Proc. Natl. Acad. Sci. USA *102*, 3519–3524.
2. Gelman, R., and Butterworth, B. (2005). Number and language: How are they related? Trends Cogn. Sci. *9*, 6–10.
3. Pica, P., Lemer, C., Izard, V., and Dehaene, S. (2004). Exact and approximate arithmetic in an Amazonian indigene group. Science *306*, 499–503.
4. Gordon, P. (2004). Numerical cognition without words: Evidence from Amazonia. Science *306*, 496–499.
5. Gelman, R., and Gallistel, C.R. (2004). Language and the origin of numerical concepts. Science *306*, 441–443.
6. Moyer, R., and Landauer, T. (1967). Time required for judgments of numerical inequality. Nature *215*, 1519–1520.
7. Brannon, E., and Roitman, J. (2003). Nonverbal representations of time and number in non-human animals and human infants. In Functional and Neural Mechanisms of Interval Timing, W. Meck, ed. (New York, NY: CRC Press), pp. 143–182.
8. Feigenson, L., Dehaene, S., and Spelke, E. (2004). Core systems of number. Trends Cogn. Sci. *8*, 307–314.
9. Barth, H., Kanwisher, N., and Spelke, E. (2003). The construction of large number representations in adults. Cognition *86*, 201–221.
10. Starkey, P., Spelke, E., and Gelman, R. (1983). Detection of intermodal numerical correspondences by human infants. Science *222*, 179–181.
11. Starkey, P., Spelke, E., and Gelman, R. (1990). Numerical abstraction by human infants. Cognition *36*, 97–127.
12. Moore, D., Benenson, J., Reznick, J., Peterson, M., and Kagan, J. (1987). Effect of auditory numerical information on infants' looking behavior: Contradictory evidence. Dev. Psychol. *23*, 665–670.
13. Mix, K., Levine, S., and Huttenlocher, J. (1997). Numerical abstraction in infants: Another look. Dev. Psychol. *33*, 423–428.
14. Church, R., and Meck, W. (1984). The numerical attribute of stimuli. In Animal Cognition, H.L. Roitblat, T.G. Bever, and H.S. Terrace, eds. (Hillsdale, NJ: Erlbaum), pp. 445–464.
15. Davis, H., and Albert, M. (1987). Failure to transfer or train a numerical discrimination using sequential visual stimuli in rats. Bull. Psychon. Soc. *25*, 472–474.
16. McComb, K., Packer, C., and Pusey, A. (1994). Roaring and numerical assessment in contests between groups of female lions, *Panthera leo*. Anim. Behav. *47*, 379–387.
17. Kitchen, D.M. (2004). Alpha male black howler monkey responses to loud calls: Effect of numeric odds, male companion behaviour and reproductive investment. Anim. Behav. *67*, 125–139.
18. Spelke, E. (1976). Infants' intermodal perception of events. Cognit. Psychol. *8*, 533–560.
19. Kuhl, P., and Meltzoff, A. (1982). The bimodal perception of speech in infancy. Science *218*, 1138–1141.
20. Patterson, M.L., and Werker, J.F. (2002). Infants' ability to match dynamic phonetic and gender information in the face and voice. J. Exp. Child Psychol. *81*, 93–115.
21. Ghazanfar, A.A., and Logothetis, N.K. (2003). Facial expressions linked to monkey calls. Nature *423*, 937–938.
22. Maier, J.X., Neuhoff, J.G., Logothetis, N.K., and Ghazanfar, A.A. (2004). Multisensory integration of looming signals by rhesus monkeys. Neuron *43*, 177–181.
23. Cheney, D., Seyfarth, R., and Smuts, B. (1986). Social relationships and social cognition in nonhuman primates. Science *234*, 1361–1366.
24. Hare, B. (2001). Can competitive paradigms increase the validity of experiments on primate social cognition? Anim. Cogn. *4*, 269–280.
25. Hare, B., and Tomasello, M. (2004). Chimpanzees are more skilful in competitive than in cooperative cognitive tasks. Anim. Behav. *68*, 571–581.
26. Flombaum, J.I., and Santos, L.R. (2005). Rhesus monkeys attribute perceptions to others. Curr. Biol. *15*, 447–452.
27. Brokx, J.P., and Nooteboom, S.G. (1982). Intonation and the perceptual separation of simultaneous voices. J. Phonetics *10*, 23–36.

28. Parsons, T.W. (1976). Separation of speech from interfering speech by means of harmonic selection. J. Acoust. Soc. Am. *60*, 911–918.
29. Ghazanfar, A.A., and Santos, L.R. (2004). Primate brains in the wild: The sensory bases for social interactions. Nat. Rev. Neurosci. *5*, 603–616.
30. Summerfield, Q., Culling, J.F., and Fourcin, A.J. (1992). Auditory segregation of competing voices: Absence of effects of Fm or Am coherence. Philos. Trans. R. Soc. Lond. B Biol. Sci. *336*, 357–366.
31. Izumi, A. (2002). Auditory stream segregation in Japanese monkeys. Cognition *82*, B113–B122.