A Horse's Eye View:
Size and Shape Discrimination Compared with Other Mammals

Masaki Tomonaga
Primate Research Institute, Kyoto University, Japan

Kiyonori Kumazaki
Horseman Kakamigahara, Japan

Florine Camus
Ecole Nationale Supérieure d'Agronomie et des Industries Alimentaires, France

Sophie Nicod
Iinstitut du Cheval et de l'Équitation Portugaise, France

Carlos Pereira
Université Paris III Sorbonne Nouvelle,
Institute National de la Recherche Agronomique, and
Iinstitut du Cheval et de l'Équitation Portugaise, France

and

Tetsuro Matsuzawa
Primate Research Institute, Kyoto University, Japan

Supplementary Material:
Detailed Descriptions of the Methods and Results

1. Methods for the Horse Experiments

(a) Participants

Three horses (*Equus caballus*) participated in the present experiments: Ponyo (female, 9 years old, 222 kg, see Fig. 2*a* of the main text), Nemo (female, 1 year old, 120 kg, see Fig. 1*a* of the main text), and Thomas (male, 4 years old, 430 kg). Nemo was an offspring of Ponyo and Thomas. They were ponies. The distances between the eye and mouth, which approximately corresponded to the minimum looking distance to the monitor, were 30 cm, 30 cm, and 40 cm for Ponyo, Nemo, and Thomas, respectively. The horses lived in the Horseman Kakamigahara, facility for horse riding which is located in Kakamigahara City, Gifu, Japan. Ponyo and Thomas routinely participated in horse-riding activity but all three horses were naïve in terms of perceptual-cognitive experiments. The horses were fed three times a day ad libitum during the current study.

(b) Ethics Statement

The experimental procedure for the horses was approved by the Animal Welfare and Animal Care Committee of KUPRI and by the Animal Research Committee of Kyoto University (#2015-134). All procedures adhered to the Ethical Guidelines for the Conduct of Research on Animals by Zoos and Aquariums issued by the World Association of Zoos and Aquariums (WAZA) [1], the Code of Ethics issued by the Japanese Association of Zoos and Aquariums (JAZA), and the Japanese Act on the Welfare and Management of Animals.

(c) Apparatus

The computer-controlled touchscreen experiments used a 42-inch LCD touchscreen monitor controlled by the Surface Acoustic Wave system (ET4201L-8UWA-0-GY-G, Elo Touch Solutions, Yokohma, Japan]; 995 mm [W] × 588 mm [H] × 79 mm [D], 1920 × 1080 pixels, pixel size: 0.4845mm × 0.4845mm). The monitor was set on a portable stand (H-669; Hayami Industry, Shiga, Japan; 760 mm [W] × 745 mm [D] × 1331-1681 mm [H]) and the luminance level was 350.0 cd/m^2 for the white background and 4.483 cd/m^2 for the black stimuli. For every experiment, the monitor was set in front of a stall (see Figs. 1*a* and 2*a* of the main text) and a universal feeder (BUF-310-P25, Bio-Medica, Osaka, Japan) delivered a small piece of carrot as a reward via a food tray set below

the monitor. All equipment and experimental procedures were controlled by a laptop personal computer (PC).

(d) Procedure

(i) Initial shaping of nose-touch responses

The horses were initially trained to perform the nose-touch responses using a successive approximation procedure [2]. In the first step, all three horses were presented a circular disk (130 mm in diameter) on a stick (200 mm long) and were required to touch the disk with their nose or mouth. This training continued for 4-5 days and Ponyo, Nemo, and Thomas performed 121, 108, and 127 trials, respectively.

(ii) Shaping of touchscreen responses

After establishing the nose-touch response, a touchscreen monitor was introduced and set in front of the stall for each horse on each experimental day. During the training period, a filled black circle was presented at the center of the monitor and, using the successive approximation procedure, the nose-touch response to the circle was shaped. When the horses touched the circle, they were given a piece of carrot along with a chime sound. The size of the circle was initially set at a diameter of 195 mm (300 pixels) and it was then flexibly changed based on the individual horse's behaviour; all horses successfully touched the circle. Each experimental session consisted of 18 trials (range: 5-25); Ponyo performed 100 trials over 2 days, Nemo performed 85 trials over 1 day, and Thomas performed 222 trials over 7 days (see Table S1).

(iii) Acquisition training of size discrimination using an errorless learning procedure

During this phase, the "fade-in" technique for the training of size discrimination was introduced [3]. Angle bars were set in front of the monitor to guide the horse's response to a stimulus (Figures 1*a* and 2*a* of the main text). Each trial was initiated by the experimenter, who delivered a piece of carrot at the onset of each session. As the horse explored the food tray and ate the carrot, the experimenter presented the stimulus pair in conjunction with a beep sound by pressing the space bar of the PC. When the horse touched either of the circles, both stimuli disappeared. If the horse chose the larger one, he/she was rewarded with food and a chime sound while only the

buzzer sound was given when he/she made an error. Each session consisted of either 12 or 24 trials but the horses sometimes stopped the experiment before finishing the session. For Ponyo and Nemo, the initial sizes of the circles were 195 mm vs. 13 mm and for Thomas they were 260 mm vs. 3.3 mm.

(iv) Measurements of the discrimination threshold for circle size

After completing the errorless learning training, all the horses were shifted to the psychophysical measurement of the discrimination threshold (difference limen [DL]) using a modified version of the up-down method [4,5]. In this modified procedure, the stimulus value was changed according to session-based accuracy rates rather than trial-based correct/incorrect choices. Each session consisted of 12 trials and each horse received four to seven sessions per day. Within a single session, the stimulus value did not change irrespective of the horse's performance. However, when the accuracy rate within a particular session was better than or equal to 83.3% (10/12), the diameter size of the smaller circle was increased by one step and when the accuracy rate was worse than or equal to 58.3% (7/12) the size was decreased by one step. When the accuracy was 66.7% (8/12) or 75.0% (9/12), the circle size was not changed from the previous session. In this modified up-down procedure, the participant's accuracy rate was maintained at around 70.8% and, thus, it was possible to measure the DL without disrupting the behaviour of the horses due to low accuracy, such as chains of error responses or a sudden stop of the experiment.

The psychophysical measurements were performed twice using different sizes for the larger circle. In the first set of experiments, the initial circle sizes were set at a diameter of 130 mm (200 pixels) vs. 13 mm (20 pixels) and the step size was set to 3.25 mm (5 pixels). In the second set, the circle sizes were set to 65 mm (100 pixels) vs. 32.5 mm (50 pixels) and the step size was set to 1.625 mm (2.5 pixels, on average). The Weber fractions (DL divided by the standard size) were calculated using the area and length differences as follows:

Weber fraction (area) = [(diameter of the larger circle)^2 – (diameter of the smaller circle)^2] / [(diameter of the larger circle)^2].

Weber fraction (length) = [(diameter of the larger circle) – (diameter of the smaller circle)] / [(diameter of the larger circle)].

The relationship between these values was as follows:

$$\text{Weber fraction (area)} = 1-[1-(\text{Weber fraction (length)})]^2.$$

To calculate the DL values, the accuracy rates of 12 successive sessions were compared. These 12 sessions were grouped into three four-session blocks and the mean accuracy was calculated for each block. If the accuracies of these three blocks neither exhibited an increasing nor decreasing trend (e.g., 75%-80%-72% or 75%-69%-72%), the horse's behaviour was judged as stable. The DLs were calculated based on the mean values of the area and length differences across these sessions.

(v) Shape discrimination

Concurrent with the second set of the size discrimination task, the horses were given a new task in which they were required to discriminate various shapes. Eight stimuli were prepared for this task, one of which was an open circle with a black line. As in the size discrimination training, the initial discrimination of the circle (O) and the cross (X) was introduced using the "fade-in" procedure and the horses were required to touch the circle. Each session consisted of 12 trials. At first, the diameter of the O was set to 130 mm (200 pixels) and the size of the X was set to 52 mm (80 pixels), 65 mm (100 pixels), and then finally 130mm.

After finishing these initial training sessions, all the horses performed the shape discrimination tests. During this phase, several sets of training/testing pairs were prepared. Initially, the horses were trained to discriminate O versus X and, thus, the other six stimuli in the testing sessions were paired as negative stimuli with the O as a positive stimulus; in each session, only one type of negative stimulus appeared. Each horse was given four to seven sessions each day and the baseline training with the O and X was always given in the first session of the day. The horse received four sessions for each stimulus pair and, after finishing the O-X sets, the horses were successively shown a D-shape versus X, S versus X, triangle versus X, square versus X, H versus X, and Z versus X, as during the baseline training (see Table S7). The X was always set as a negative stimulus. For each baseline set, the horses were initially trained with the baseline pair for at least eight sessions and then the testing pairs (positive

stimulus and the other stimulus) were given. Reversed pairs were not given; that is, if O versus X was shown, then X versus O was not shown, except for some stimulus pairs with the triangle because the performances tended to be worse when the triangle was set as the positive stimulus (see Table S7 and Figure S2). Overall, the horses were given 31 pairs in which three pairs were reversed pairs that included the triangle.

On the basis of these accuracy data, the perceptual similarities among these eight shapes were analyzed using multidimensional scaling (MDS) analyses with an INDSCAL procedure [6]. This method yielded spatial representations for the stimuli as well as weights for each dimension of the representation for each horse. For each pair, the mean accuracy was calculated across four sessions and for the seven baseline training pairs the mean accuracy was based on the data from the first four sessions. For the three symmetrically trained pairs with the triangle, the better accuracy of the two was used. A two-dimensional solution for the present analyses was adopted using SPSS 19.0J. To evaluate the goodness of fit, the stress values and coefficients of determination (RSQ) were determined.

To evaluate the relative contribution of each feature, such as the vertical/horizontal lines, curvatures, or closures, to the perceptual grouping of shapes, the mean percentages of errors for the pairs in which both stimuli had the same features were calculated. These values were standardized using means and standard deviations, referred to as standardized similarities, and then compared with the data from the previous dolphin, chimpanzee, and human experiments [6].

2. Methods for Chimpanzee Experiments

(a) Participants

Three adult female chimpanzees (*Pan troglodytes*) participated in the present size discrimination experiment; Chloe, Cleo, and Pendesa. Chloe was 34, Cleo was 15, and Pendesa was 38 years of age at the onset of the experiment. They lived in a social group of 13 individuals (including themselves) indoors and in an environmentally enriched outdoor compound (770 m^2) at the Primate Research Institute at Kyoto University (KUPRI) in Japan [7]. The chimpanzees were fed various kinds of foods three times per day ad libitum. The

chimpanzees had previously experienced various computer-controlled perceptual and cognitive tasks, including psychophysical measurements and shape discrimination [5,6,8-12]. The data from Tomonaga et al. [6] were reanalyzed for comparisons with the horses.

(b) Ethics Statement

The care and use of the chimpanzees adhered to the 2010 version of the Guide for the Care and Use of Laboratory Primates issued by KUPRI, which are compatible with the guidelines issued by the National Institutes of Health (Bethesda, MD, USA). The research design was approved by the Animal Welfare and Animal Care Committee of KUPRI and by the Animal Research Committee of Kyoto University (#2015-044). All procedures adhered to the Japanese Act on the Welfare and Management of Animals.

(c) Apparatus

All experimental sessions were conducted in a booth (1.8 × 2.15 × 1.75 m) in an experimental room adjacent to the chimpanzee facility. Each chimpanzee came to the booth via an overhead walkway connecting the facility and the booth. A 17-inch LCD monitor with a resistive membrane system touchscreen (LCD-AD172F2-T, I-O Data, Tokyo, Japan; 384mm [W] ×51 mm [D] × 349 mm [H], 1280 × 1024 pixels, pixel size: 0.264 mm × 0.264 mm) was installed on the wall of the booth with a viewing distance of approximately 40 cm. The luminance level was 124.4 cd/m^2 for the white background and 3.362 cd/m^2 for the black stimuli and the food reward was delivered via a universal feeder (BUF-310-P100, Bio-Medica, Osaka, Japan). All equipment and experimental events were controlled by a PC.

(d) Procedure

The chimpanzees also performed the circle-size discrimination task. The initial size of the larger circle (filled black) was 65 mm (260 pixels) in diameter, the smaller one was 32.5 mm (130 pixels), and the viewing distance was approximately 40 cm. Thus, the visual angle for each stimulus was very similar to the stimuli in the second set of the horse experiments.

Each trial began with the presentation of a blue square (26 mm × 26 mm) at the bottom center of the monitor. When the chimpanzee touched the square twice, two circles appeared horizontally. If the chimpanzee touched the larger circle, all stimuli disappeared and the sound of a chime and presentation of a food reward (a small piece of apple or raisin) followed. If the chimpanzee touched the smaller circle, a buzzer sound was presented as error feedback. The inter-trial interval was 2 seconds and each session consisted of 12 trials. The chimpanzees were given an average of six sessions per day. As with the horse experiments, the modified up-down method was used to measure the DL. The DL was calculated in the same manner as in the horse experiments and the step size was set to 1.625 mm (6.5 pixels on average).

3. Methods for Human Experiments

(a) Participants and Apparatus

Six adult humans (three females and three males) with normal or corrected-to-normal vision voluntarily participated in the circle-size discrimination experiment. Informed consent was obtained from all participants and all experimental protocols were consistent with the Guide for Experimentation with Humans and were approved by the Human Research Ethics Committee of KUPRI (#2015-05). Informed consent was obtained from all participants prior to experiments. The experimental apparatus was the same as that used for the chimpanzee experiments.

(b) Procedure

Unlike the horse and chimpanzee experiments, a trial-based modified up-down method was used to measure the DL of the humans. Each session consisted of 100 trials and the procedure of the trials was the same as that for the chimpanzees, except that no feedback (food and chime or buzzer) was presented and the inter-trial interval was 0.5 seconds. The initial size of the larger circle was 65 mm (260 pixels) in diameter, the diameter of the smaller circle was 60 mm (240 mm), and the step size was 4.5 mm (two pixels). When the participant made two successive correct choices, there was a one-step increase in the size of the smaller circle whereas a single error resulted in a one-step decrease in the size of the smaller circle. If the size difference of the circles was 4.5 mm, then the size did not change even if the participant made two successive correct choices. The calculation of the DL was

the same as for the horses and chimpanzees except that trial-based accuracy was used to judge performance stability rather than session-based accuracy.

4. Reanalysis of the Perceptual Similarities of the Dolphin, Chimpanzee, and Human Data

To compare the shape perception of the horses with that of other species, previous data from dolphins, chimpanzees, and humans [6] were reanalyzed. These authors had used nine shapes and six of these shapes were the same as those used in the horse experiment (O, D, square, triangle, H, and X). The dolphins (n = 3) and chimpanzees (n = 7) were tested in a delayed matching-to-sample task and the humans (n = 20) were tested using visual analog scaling. Using the data from the 15 pairs resulting from various combinations of the six common stimuli, the perceptual similarities among the species for these shapes were reanalyzed using an INDSCAL procedure. Additionally, to evaluate the similarities among the four species, intraclass correlation coefficients (ICC) were calculated based on the averaged data from each species.

5. Statistical analysis

Due to the small number of samples in the current study, instead of using parametric statistical significance tests such as analysis of variance, we applied statistical significance tests on the basis of bootstrap resampling procedure to our data [13-15]. All input data were average values for each individual.

In our data (for example, Weber fraction value obtained from each individual), there were more than two conditions (for example, we tested three species). Thus, in this example, we set the null and alternative hypotheses as follows

H0: $\mu_1=\mu_2=\mu_3$

H1: $\mu_1\neq\mu_2$ or $\mu_1\neq\mu_3$ or $\mu_2\neq\mu_3$

Significance tests using bootstrap resampling proceeds as follows.

1) Calculate the test statistics $T_{1,2}$, $T_{1,3}$, $T_{2,3}$ from the original data $\{x_{11}, \ldots, x_{1l}\}$, $\{x_{21}, \ldots, x_{2m}\}$, and $\{x_{31}, \ldots, x_{3n}\}$.

   In the current study, we used an absolute value of the difference of means (e.g., $T_{1,2}=|\overline{x}_1-\overline{x}_2|$).

2) Create bootstrap samples $\{x_{11}^*, \ldots, x_{1l}^*\}$, $\{x_{21}^*, \ldots, x_{2m}^*\}$, and $\{x_{31}^*, \ldots, x_{3n}^*\}$, via resampling with replacement from the *combined* sample $\{x_{11}, \ldots, x_{1l}, x_{21}, \ldots, x_{2m}, x_{31}, \ldots, x_{3n}\}$.

3) Calculate bootstrap test statistics $T_{1,2}^*$, $T_{1,3}^*$, $T_{2,3}^*$ from the bootstrap samples $\{x_{11}^*, \ldots, x_{1l}^*\}$, $\{x_{21}^*, \ldots, x_{2m}^*\}$, and $\{x_{31}^*, \ldots, x_{3n}^*\}$.

4) Repeat 2) and 3) for B times (in our case, B=10,000).

5) Count the number ($C_{1,2}$, $C_{1,3}$, and $C_{2,3}$) of the bootstrap test statistics which are equal to or greater than the original test statistics. Obtained p value for each comparison is defined as $p_{1,2}=C_{1,2}/B$, $p_{1,3}=C_{1,3}/B$, and $p_{2,3}=C_{2,3}/B$, respectively.

In this case, since these are multiple comparisons, we corrected each p value using the false discovery rate (FDR) set at 0.05 [16].

5. Results

(a) Size discrimination using the errorless learning procedure in horses

The size discrimination training began with a large circle and a very small circle that had an initial difference in area that was larger than 0.9. During the first 212–216 trials, all the horses performed very well: Ponyo scored 89.8% (216 trials, mean area difference = 0.931), Nemo scored 84.8% (212 trials, mean area difference = 0.931), and Thomas scored 73.0% (212 trials, mean area difference = 0.974). Ultimately, Ponyo performed 2164 trials, Nemo performed 2066 trials, and Thomas performed 1802 trials during the errorless learning training; the individual results are shown in Table S2.

(b) Discrimination threshold for circle size in horses, chimpanzees, and humans

(i) Horses

Figure S1*a* shows the results for the first set of experiments. The left vertical axis shows the Weber fraction based on area difference, the right vertical axis shows the Weber fraction based on length difference, and the

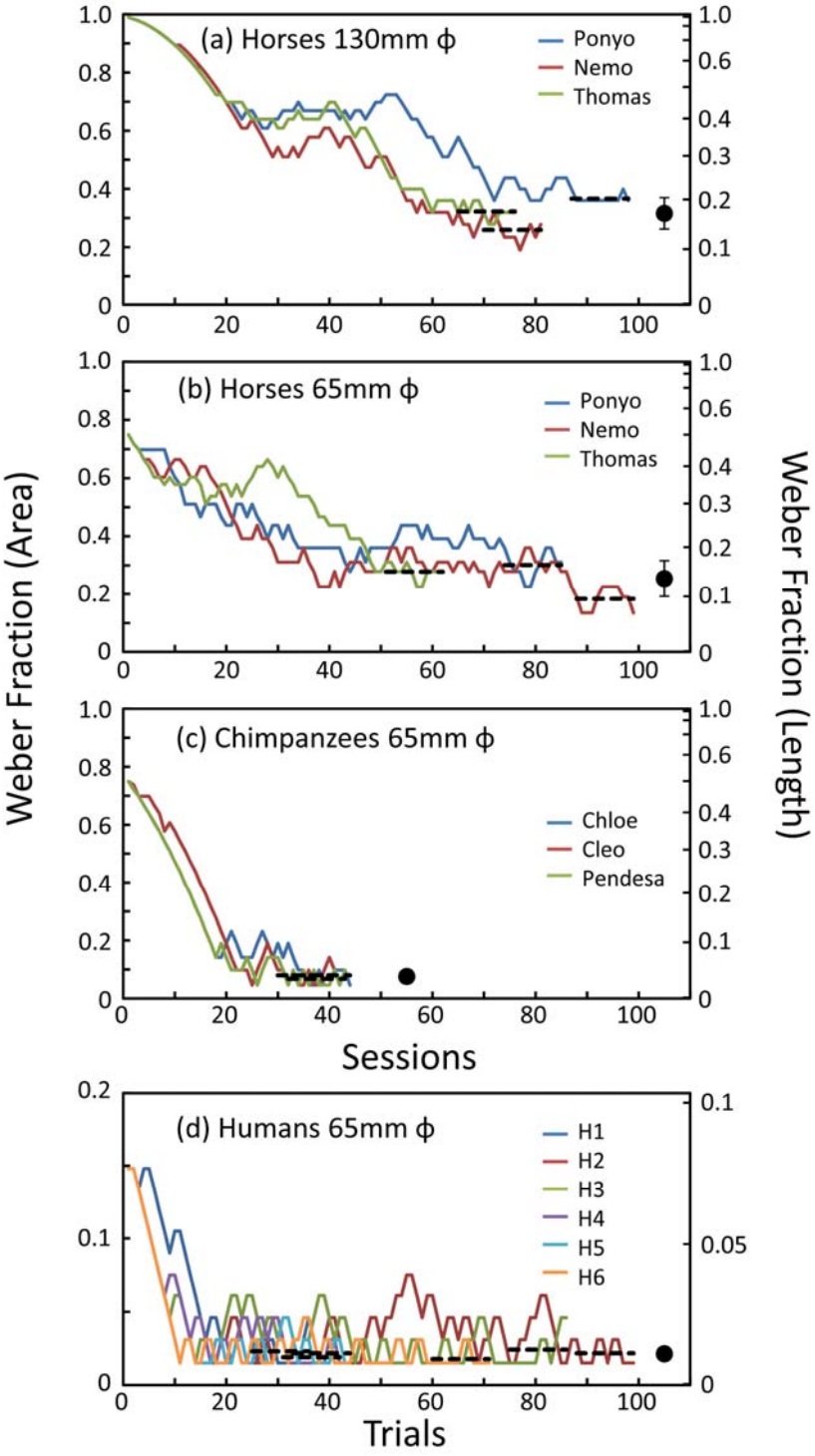

Figure S1. Results of the size discrimination experiments in horses (*a*,*b*), chimpanzees (*c*), and humans (*d*). The left vertical axis shows the Weber fraction based on area difference, the right vertical axis shows the Weber fraction based on length difference, and the horizontal axis shows the sessions for horses and chimpanzees and trials for humans. Dotted lines depict the criterial sessions for each horse and the black circle located at the right of the panel depicts the mean Weber fractions (with standard deviations).

horizontal axis shows the sessions. The dotted lines depict the criterial sessions for each horse and the black circle located at the right of the panel represents the mean Weber fractions (with standard deviations); the individual results are shown in Table S3. For the first set of experiments, the horses were tested with a larger circle that was 130 mm in diameter and a smaller circle that was initially 13 mm in diameter. All horses showed a pseudo-floor effect from approximately the 30^{th}–50^{th} sessions (Figure S1*a*). The mean area differences for Ponyo, Nemo, and Thomas were 0.657, 0.564, and 0.642, respectively, and the mean accuracy rates for Ponyo, Nemo, and Thomas were 68.3%, 72.1%, and 70.0%, respectively. The performances of all the horses improved after the 50^{th} session and resulted in mean accuracy rates of 74.2%, 77.9%, and 75.8% for Ponyo, Nemo, and Thomas, respectively, between the 51^{st} and 70^{th} sessions.

Ultimately, Ponyo reached the stability criterion during the 103^{rd} session, Nemo during the 87^{th} session, and Thomas during the 81^{st} session and the accuracy rates during the 12 criterial sessions were 69.5%, 72.9%, and 70.1% for Ponyo, Nemo, and Thomas, respectively. The obtained Weber fractions based on area were 0.367, 0.259, and 0.322 for Ponyo, Nemo, and Thomas, respectively. The mean Weber fraction averaged across the horses was 0.316 and the mean Weber fraction based on length was 0.174.

Figure S1*b* shows the results for the second set of experiments; the individual results are shown in Table S4. There was no clear pseudo-floor effect observed in the second set of experiments. Ponyo reached the stability criterion during the 85^{th} session, Nemo during the 99^{th} session, and Thomas during the 62^{nd} session. The mean accuracy rates for the criterial sessions were 72.2%, 72.2%, and 71.5% for Ponyo, Nemo, and Thomas, respectively, and the obtained DLs based on area were 0.301, 0.184, and 0.278 for Ponyo, Nemo, and Thomas, respectively. The averaged Weber fraction based on area was 0.254, the mean Weber fraction for length was 0.137, and all the horses showed better DLs for the second set of the experiments than for the first set.

Timney and Keil [17] measured the DL of the relative line length at a 70% correct performance level and the obtained Weber fraction averaged across two horses was 0.258. The horses in the present study showed comparable but better Weber fractions based on length than Timney and Keil [17] and statistical significance test based on 10,000 bootstrap samples revealed a significant difference between Timney & Keil's data [17] and the present data from the second set of experiments ($p = 0.045$, correction based on the false discovery rate set at 0.05).

(ii) Chimpanzees

Figure S1*c* shows the results from the three chimpanzees in the present study; the individual results are shown in Table S5. The number of sessions to reach the criterion were 44, 41, and 43 for Chloe, Cleo, and Pendesa, respectively, and the mean accuracy rates during the criterial sessions were 69.5%, 72.2%, and 67.4% for Chloe, Cleo, and Pendesa, respectively. The obtained Weber fractions were 0.08, 0.08, and 0.068 for Chloe, Cleo, and Pendesa, respectively, and the mean value was 0.076. The mean Weber fraction based on length was 0.039.

(iv) Humans

Figure S1*d* and Table S6 show the results for the human participants in the present study. Note that the horizontal axis depicts the trials rather than the sessions. All the human participants received a single 100-trial session and the achievement of criterion levels of performance was determined once each participant finished the session. The participants reached the criterion level after 63.0 trials on average (range: 36 to 99 trials) and the mean accuracy rate of the 12 criterial trials across the participants was 82.0%. The obtained Weber fraction based on area was 0.021 on average (SD = 0.002) and the Weber fraction based on length was 0.011 (SD = 0.0011). The difference in area in the present study was explicitly higher than that observed in previous studies [18] but when the present results were compared with the previous data on the basis of diameter length, the dissociation was smaller. It is possible that the current human participants discriminated size not based on area difference but on the difference in diameter length [18].

(v) Comparison among species.

As shown in Figure 1*b* of the main text, the Weber fractions differed among the species and, thus, a randomisation test was conducted to assess these data; note that only the second set of data (standard circle with a diameter of 65 mm) from the horses was used for this analysis. As a result, horses showed significantly worse Weber fractions than the other species (statistical significance test based on 10,000 bootstrap samples; horse vs. chimpanzee, $p = 0.046$; horse vs, human, $p = 0.005$; correction based on the false discovery rate set at 0.05).

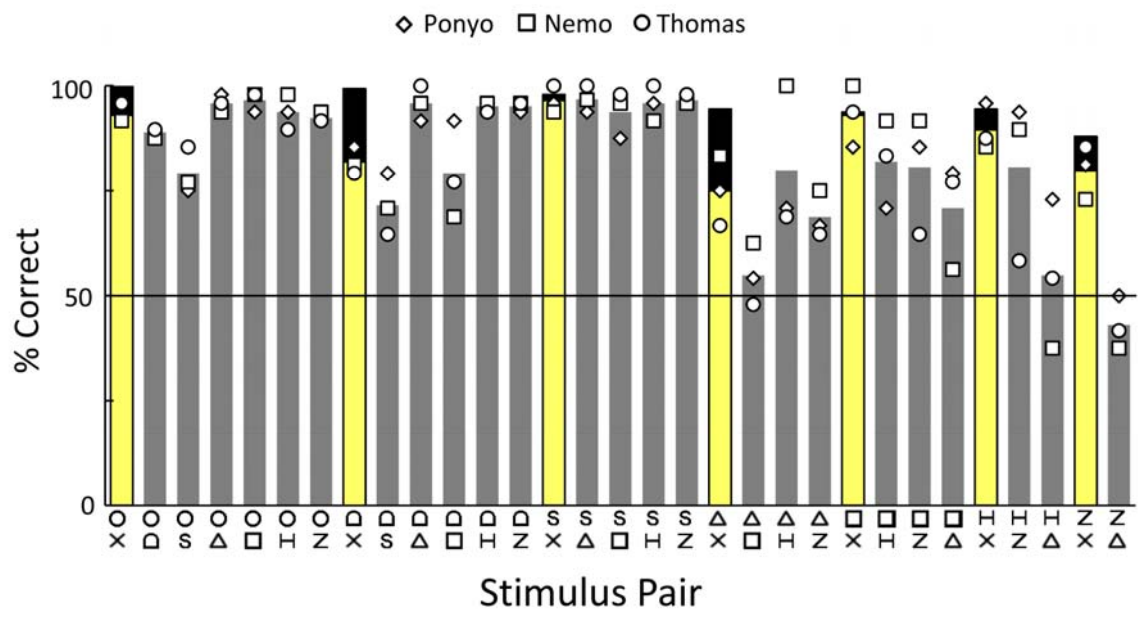

Figure S2. Results of all the training and tested pairs. The yellow and black bars depict the accuracy rates of the baseline-training pairs. The yellow bars depict the accuracy rates of the first four pairs and the black bars depict the rates of the last four sessions of baseline pairs. The gray bars depict the results of the tested pairs on the basis of four sessions of testing. The data for each horse are also illustrated in this figure.

(c) Shape discrimination

(i) Horses

The initial errorless learning training was given over 20 sessions to Ponyo, 20 sessions to Nemo, and 19 sessions to Thomas. The size of the negative stimulus (X) began at 52 mm, increased to 60 mm, and ultimately equalized with the positive stimulus at 130 mm. The mean accuracy rates averaged across horses were 92.1% when the size of the negative stimulus was 52 mm, 90.3% when it was 65 mm, and 96.9% when it was 130 mm.

Table S7 shows the individual results of the horse experiments and Figure S2 illustrates the results of all the training and tested pairs. The yellow and black bars depict the accuracy rates of the baseline-training pairs. The yellow bars depict the accuracy rates of the first four pairs and the black bars depict the rates of the last four sessions of baseline pairs. The gray bars depict the results of the tested pairs on the basis of four sessions of testing. The data

for each horse are also illustrated in this figure. On the basis of these data, the ICC values among the horses were calculated and the value was significantly above 0 ($ICC_{2,3} = 0.783$, $p < 0.001$).

An MDS analysis was also conducted using the INDSCAL method to visualize the perceptual similarities among the shapes; Figure 2*b* of the main text shows the two-dimensional solution. The RSQ and stress values, which are measures of goodness of fit, were 0.510 and 0.256, respectively. Three perceptual categories were identified: shapes with curvature (O, D, and S), shapes with only vertical and horizontal lines (square and H), and shapes with diagonal lines (X, Z, and triangle; Figure 2*b*).

(ii) Comparisons of the dolphin, chimpanzee, and human data

To make comparisons with the horse data, the data obtained from three dolphins, seven chimpanzees, and 20 humans in previous experiments by Tomonaga et al. [6] were re-analyzed. The results related to the stimuli common to the present experiments (O, D, square, H, triangle, and X) and the previous experiments were chosen and the ICCs were initially calculated using the averaged data for each species. The ICC was 0.565, which was significantly above 0 ($p = 0.019$). Next, MDS analyses using INDSCAL were performed for each species using the data from the common stimuli; the resulting two-dimensional solutions are shown in Figure 2*c* of the main text. The RSQ and stress values for each solution were 0.505 and 0.227 for dolphins, 0.596 and 0.218 for chimpanzees, and 0.607 and 0.237 for humans, respectively.

The relative contributions of the various features to perceptual similarities [6] were also further analyzed. Each shape contained various elemental features and, of these, six features were selected: vertical/horizontal line, curvature, diagonal line, closure, right angle, and open end. Using all of the data from the present horse experiments and the previous experiments from dolphins, chimpanzees and humans, a standardized similarity was calculated for each feature and then compared among the species (Figure S3). Statistical significance test based on 10,000 bootstrap samples conducted for each feature category revealed that shapes with open ends (e.g., X and H) were more closely categorized by chimpanzees than by humans ($p < 0.001$) and horses ($p = 0.015$), shapes with diagonal lines (e.g., X, Z, and triangle) were perceived as more similar by horses than by chimpanzees ($p = 0.008$) and humans ($p = 0.026$), shapes with right angles were less closely categorized by horses than by chimpanzees ($p =$

0.031) and humans (p =0.027), and closed shapes (e.g., O, D, square, and triangle) were less closely categorized by horses than by humans (p = 0.002, all p values were corrected based on the false recovery rate set at 0.05).

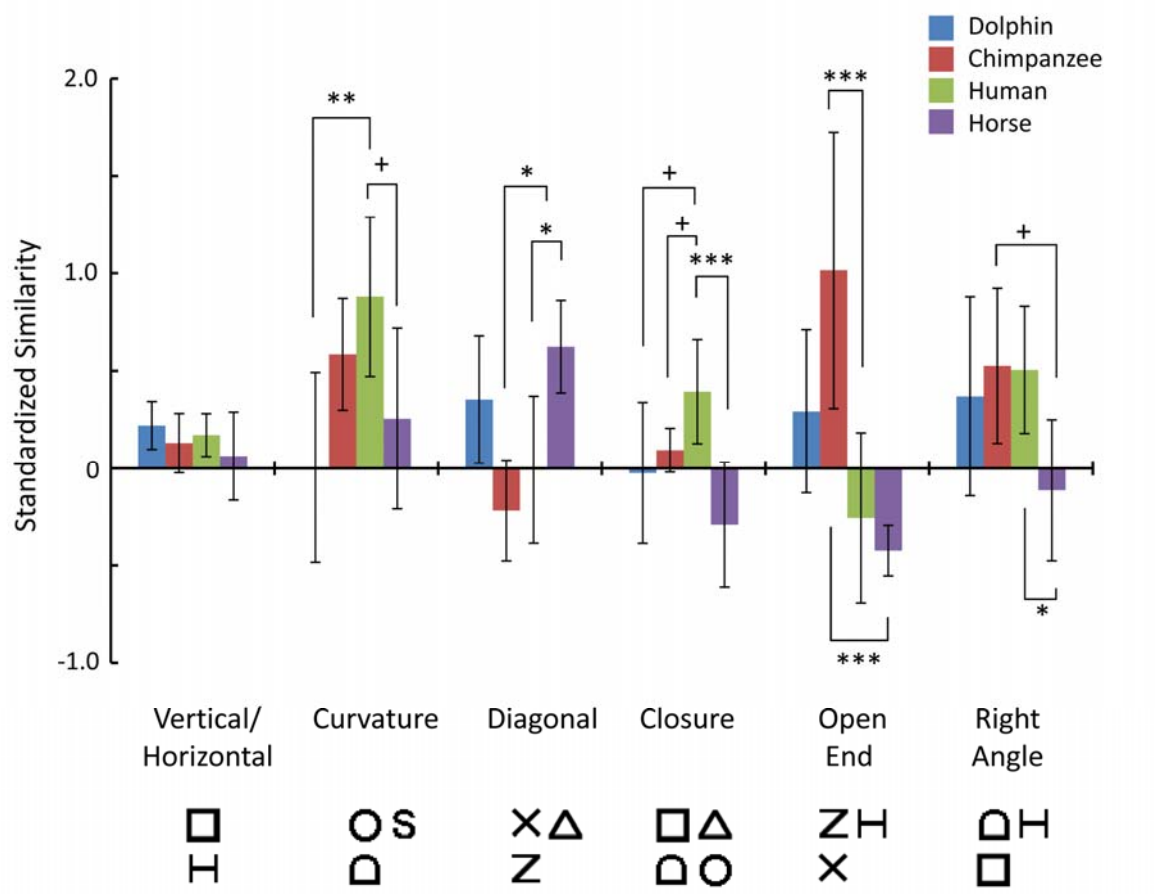

Figure S3.

Standardized similarities of the shapes sharing each feature for each species. +p < 0.1, *p < 0.05, **p < 0.01, ***p < 0.001. Error bar: standard error

References for Supplementary Material

1. Sato A, Tomonaga M. 2010 WAZA (World Association of Zoos and Aquariums) ethical guidelines for the conduct of research on animals by zoos and aquariums. *Jpn. J. Anim. Psychol.*, 60, 139-146.
2. Sidman M. (1962). Operant techniques. In *Experimental foundations of clinical psychology* (Ed. Bachrach AJ), pp. 170-210. New York: Basic Books.
3. Terrace HS. 1963 Discrimination learning with and without "errors" 1. *J. Exp. Anal. Behav.*, 6, 1-27.
4. Levitt HCCH. 1971 Transformed up–down methods in psychoacoustics. *J. Acoust. Soc. Am.*, 49, 467-477.
5. Matsuno T, Tomonaga M. 2006 Measurement of contrast thresholds of chimpanzees using a parameter estimation by sequential testing (PEST) procedure. *Jpn. J. Psychon. Sci.*, 25, 115-116.
6. Tomonaga M, Uwano Y, Saito T. 2014. How dolphins see the world: A comparison with chimpanzees and humans. *Sci. Rep.* 4, 3717.
7. Matsuzawa T. 2006 Sociocognitive development in chimpanzees: A synthesis of laboratory work and field work. In *Cognitive development in chimpanzees* (eds Matsuzawa T, Tomonaga M, Tanaka M), pp. 3-33. Tokyo, Japan: Springer-Verlag.
8. Matsuzawa T. 1990 Form perception and visual acuity in a chimpanzee. *Folia Primatol.*, 55, 24-32.
9. Tomonaga M, Matsuzawa T. 1992. Perception of complex geometric figures in chimpanzees (*Pan troglodytes*) and humans (*Homo sapiens*): Analyses of visual similarity on the basis of choice reaction time. *J. Comp. Psychol.*, 106, 43-52.
10. Matsuno T, Tomonaga M. 2007 An advantage for concavities in shape perception by chimpanzees (*Pan troglodytes*). *Behav. Process.*, 75, 253-258.
11. Fagot J, Tomonaga M. 1999. Global and local processing in humans (*Homo sapiens*) and chimpanzees (*Pan troglodytes*): Use of a visual search task with compound stimuli. *J. Comp. Psychol.*, 113, 3-12.
12. Goto K, Imura T, Tomonaga M. 2012 Perception of emergent configurations in humans (*Homo sapiens*) and chimpanzees (*Pan troglodytes*). *J. Exp. Psychol. Anim. Behav. Process.*, 38, 125-138.
13. Efron B, Tibshirani RJ 1993. *An Introduction to the Bootstrap: Monographs on Statistics and Applied Probability, Vol. 57*. New York and London: Chapman and Hall/CRC.

14. Roff DA 2006. *Introduction to computer-intensive methods of data analysis in biology*. Cambridge, UK: Cambridge University Press.

15. Wang, J, Taguri M 1996. Bootstrap method: An introduction from a two sample problem. *Proc. Inst. Statist. Math.*, 44, 3-18 (Japanese text with English abstract).

16. Benjamini Y, Hochberg Y 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Statist. Soc. B*, 57, 298-300.

17. Timney B, Keil K. 1996 Horses are sensitive to pictorial depth cues. *Perception*, 25, 1121-1128.

18. Nachmias J. 2011 Shape and size discrimination compared. *Vision Res.*, 51, 400-407.