A laboratory study of individual recognition using Bewick's Swan bill patterns

JOHN BROWN and VIVIEN LEWIS

Introduction

The aims of the experiment described below were to investigate the reliability with which individual Bewick's Swans *Cygnus columbianus bewickii* can be identified from bill markings by the human observer and to study the factors affecting this reliability. Bateson (1977) has verified that one very experienced observer is highly reliable at recognizing individual swans with which she is familiar when colour transparencies of the swan are projected. Evans (1977) has studied the reliability with which black and white photographs of the beaks of Bewick's Swans taken in different seasons can be matched. Our experiment used the same photographic material (which was kindly loaned by her) but our aims and procedures were somewhat different.

The reliability of recognition is not the simple concept it sounds. This is partly because willingness to identify a pair as matching despite the differences present can vary from observer to observer and from condition to condition. It is always possible to increase the number of 'hits' (saying 'Yes' to the matching or target pair) at the cost of increasing the number of 'false alarms' (saying 'Yes' to a non-matching or distractor pair). One way of dealing with this problem is to adopt the standard Signal Detection Theory approach (see Green & Swets 1966). A second method is to use a forced-choice test situation, as did Evans, in which the task is to select the matching pair from several pairs. A third method, adopted in the present study, is to use a rating procedure. Reliability can then be assessed in terms of the consistency with which targets are placed in higher rating categories than distractors (see Brown 1974). In addition to overall reliability, rating data also reveal the relationship between confidence and accuracy.

Method

Slides from two slide projectors, programmed for each of the two testing conditions by a Campden Instruments Chipp Unit, were projected level and close to one another on a clean, white wall at the end of the room.

There were eight slides for practice and illustration, and 200 slides for the experiment proper. Each slide was in black and white, and showed either a front or a side view of the head of a Bewick's Swan. (See Evans 1977.) Half the slides were of 50 swans photographed one year (one front view and one side view for each swan), and the second half were of the same 50 swans photographed during a subsequent year (again one front view and one side view). Front views were always paired with front and side with side.

For each hourly session 25 'front' pairs and 25 'side' pairs were of the same bird ('matched'), and the remaining 25 'front' and 25 'side' were of different birds ('unmatched'). On each subsequent day half of the previous day's matched pairs were changed to unmatched, similarly half of the previous day's unmatched pairs were changed to matched. The 100 pairs for each day were presented in a random order with the one constraint that no more than five matched or unmatched pairs should occur together.

Two conditions of viewing were used. Under the simultaneous condition, both sides of a pair were projected, side by side, for 2 seconds. Under the successive condition, the left-hand slide of each pair was projected for 2 seconds and then, after an interval of 4 seconds, the right-hand slide was projected for 2 seconds. The rating categories were defined as follows: 1 = the two pictures are *definitely* of the same swan, 2 = the two pictures are *probably* of the same swan, 3 = the two pictures are *probably* NOT of the same swan, 4 = the pictures are *definitely* NOT of the same swan. Under the simultaneous condition, the subject was told whether or not the pair matched, once he had given his rating. This 'feedback' both helped to maintain his interest and to facilitate learning. 'Feedback' was not given under the successive condition, partly to prevent the testing session becoming unduly long and partly because it was likely to be less useful to the subject under this condition. It was always given second and used the same 100 pairs.

There were eight subjects, research workers in the Department of Psychology of

Bristol University, tested either singly or in pairs. Each subject was tested on five consecutive days. On Day 1, prior to testing proper, the four practice pairs of slides were shown (two 'front' pairs and two 'side'). The experimenter (VL) drew attention to distinctive features of the bill markings and explained the rating procedure.

Immediately after the final testing session, a short test was conducted to assess the extent to which subjects had become familiar with the individual photographs. Duplicates of the 100 slides used in the right-hand projector were shown for 3 seconds each and the subject was asked to rate on a 4-point scale whether he had seen that particular swan previously.

Results

The main analyses were conducted in terms of the A-index measure of recognition (Brown 1965, 1974). In accordance with the notion that discrimination is shown by the consistency with which target pairs are rated higher than distractor pairs, the basic principle is to weight the number of targets (target pairs) placed in a given rating category by the number of distractors placed in a lower rating category *plus* (for a technical reason) one half of the distractors placed in the same category. If this product is called H and the total number of targets and distractors are n and m respectively, then $A = 2\Sigma H/nm - 1$, where ΣH is the sum of the products for the categories. The A-index ranges from O when discrimination of targets is at the chance level to 1 when discrimination is perfect.

The overall mean values of the A-index for each subject was calculated for pairs in which the temporal separation was (a) one season, (b) two to three seasons and (c) more than three seasons: in these calculations, pairs for a given separation were compared with all distractor pairs. The mean values were (a) 0·73, (b) 0·63 and (c) 0·62. Analysis of variance showed that the effect of temporal separation by more than one season is highly significant ($P < 0{\cdot}001$).

Figure 1A shows that discrimination was better on the basis of side views from Day 3 onwards. Figure 1B shows that discrimination was always better under the simultaneous comparison condition: the tendency for the difference to increase during the course of the experiment may be due to the 'feedback' given under the simultaneous condition. In both conditions performance became reasonably stable after the first two days. Accordingly all analyses given below are for Days 3, 4 and 5 combined. Analysis

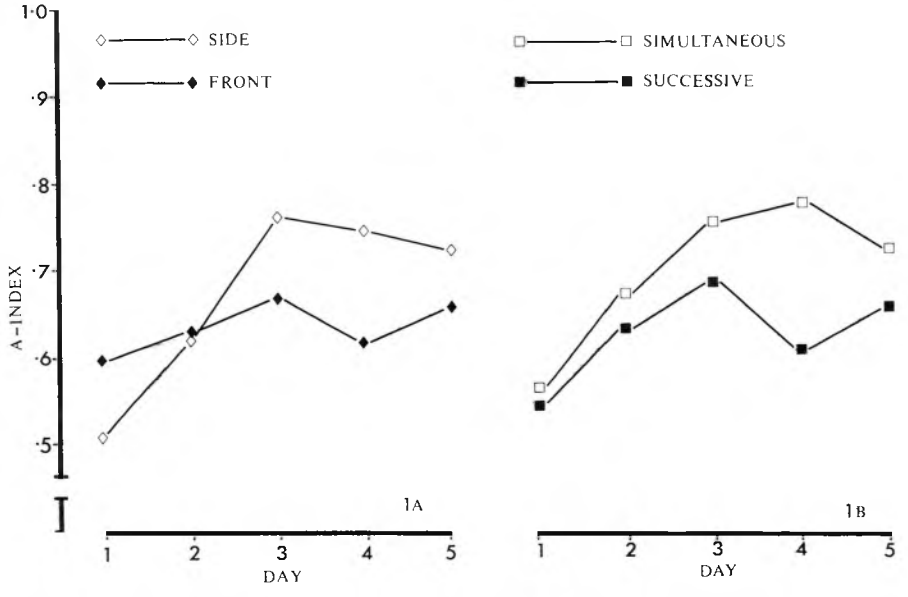

Figure 1A. Matching efficiency assessed by the A-index on successive days for front and side views. 1B. Matching efficiency as a function or whether presentation of the pair was simultaneous or successive.

of variance then showed that side views gave better discrimination than front views ($P < 0{\cdot}001$) and simultaneous presentation than successive presentation ($P < 0{\cdot}001$). There was no sign of an interaction between these variables ($F < 1$). Successive presentation was associated with higher standard errors reflected in the more erratic graph in Figure 1B.

The percentage of hits, defined as matching pairs rated as definitely or probably the same swan (categories 1 or 2), were 81% and 80% for front and side views, in contrast with the A-value result. The explanation for the discrepancy lies in the fact that the percentage of false alarms was 25% for front views but only 16% for side views. A rough alternative to the A-index is to subtract the proportion of false alarms from the proportion of hits. When so adjusted, the proportion of hits for front and side views were 0·56 and 0·64 respectively, which agrees with the A-index result.

The percentages of hits and of false alarms both decline if only pairs placed in category 1 are counted. This helps to emphasize that measures based on hits and misses are liable to be unstable. Moreover, the ways in which the categories were used appeared to vary systematically both by condition and by subject. The proportion placed in category 1 was 58% for the simultaneous-front condition but only 43% for the successive-side condition, despite the fact that the A-value was 0·69 for both conditions. The probability that a pair placed in category 1 was indeed a matching pair was 0·49 and 0·55 for the front and side conditions, and 0·63 and 0·42 for the simultaneous and successive conditions.

The A-value for individual subjects ranged from 0·64 to 0·78 and analysis of variance showed that significant differences were present ($P < 0{\cdot}01$). The mean for five ethologists was 0·71 as compared with 0·67 for three non-ethologists. The major differences between subjects lay in the proportion they placed in the 'definitely-match' category, ranging from 20% to 75% of matching pairs.

In the final test, when subjects were shown single photographs and asked whether they had seen the swans before, two-thirds were put in the top rating category. This implies that a limited familiarity with individual swans and/or photographs had developed.

Discussion

Side views were more effective as a basis for matching than front views. This may have been due to the variable foreshortening of the beak in the latter. An interesting question for further research is whether matching on the basis of *both* front and side views (as used by Evans 1977) is better than matching on the basis of the side view alone.

Simultaneous comparison gave more reliable matching than successive comparison. Strictly, we cannot conclude that simultaneous comparison is easier since this condition was always taken first and 'feedback' was given. However, simultaneous comparison is regularly found to be easier than successive comparison in discrimination tasks. What is of interest is that even successive comparison can show a fair degree of reliability.

We found that when more than one year separated two photographs of the same swan, this had an appreciable effect on matching reliability. This was not shown so strongly by Evans (1977). Since our subjects were able to view photographs for only two seconds, more time may be needed to allow for age-changes when making a comparison.

Subject differences were most apparent in the use of the definitely-match rating category. In contrast, the true variation in ability to discriminate between matching and non-matching pairs was quite small. This suggests that some attempt either to train or calibrate observers in their use of rating categories would be worthwhile.

There has been a number of recent studies of the recognition of human faces (e.g. Ellis 1975). The clinical condition known as prosopagnosia, in which recognition of human faces is alleged to be selectively impaired, has led to the suggestion that we possess a special mechanism to facilitate such recognition (see Yin 1970). The evidence is by no means convincing: one man, adept at recognizing the faces of farm animals, lost this ability simultaneously with losing the ability to recognize human faces (Bornstein, Sroke & Kunitz 1969). There seems to be no good reason why recognition of the faces of Bewick's Swans should not become a highly reliable acquired skill, as indeed the study of Bateson (1977) confirms. In the present study, a moderately good level of performance at the matching task was achieved, despite the limited practice of our subjects and the restricted viewing time.

Summary

Observers were tested over five successive days for their ability to detect whether a pair of black and white slides were of the same or of a different

Bewick's Swan *Cygnus columbianus bewickii*. Matching of side views was more reliable than of front views. Although simultaneous matching was more reliable than successive, quite good matching was achieved under the latter condition. Matching was appreciably worse when the photographs of a matching pair were taken more than a year apart. Substantial variations were found between the observers in their willingness to assert a definite match. Training or calibration of observers in their use of judgmental categories could prove beneficial.

References

Bateson, P. P. G. 1977. Testing an observer's ability to identify individual animals. *Anim. Behav.* 25: 247–8.

Bornstein, B., Sroke, H. & Kunitz, H. 1969. Prosopagnosia with animal face agnosia. *Cortex* 5: 164–9.

Brown, J. 1965. Multiple response evaluation of discrimination. *Br. J. Math. Statist. Psychol.* 18: 125–37.

Brown, J. 1974. Recognition assessed by rating and ranking. *Br. J. Psychol.* 65: 13–22.

Evans, M. E. 1977. Recognizing individual Bewick's Swans by bill pattern. *Wildfowl* 28: 153–158.

Green, D. M. & Swets, J. A. 1966. *Signal Detection Theory and Psychophysics*. John Wiley. London.

Yin, R. K. 1970. Face recognition by brain-injured patients: a dissociable ability? *Neuropsychologia* 8: 395–402.

Prof. John Brown & Miss Vivien Lewis* Department of Psychology, University of Bristol.

* Now at the Applied Psychology Unit, Medical Research Council, 15 Chaucer Road, Cambridge CB2 2EF.

A group of Teal *Anas crecca*, Wigeon *A. penelope*, Mallard *A. platyrhynchos* and Coot *Fulica atra* on the Wildfowl Trust's Welney Refuge last winter. (*Philippa Scott*).