Carl N. Stephan,[1] *B.H.Sc. (Hons.)*

Do Resemblance Ratings Measure the Accuracy of Facial Approximations?*

REFERENCE: Stephan CN. Do resemblance ratings measure the accuracy of facial approximations? J Forensic Sci 2002;47(2): 239–243.

ABSTRACT: Since forensic facial approximations are used to promote recognition of a deceased person, an accurate forensic facial approximation (FFA) should be easily recognized as the person to whom the skull belonged (target individual). However, the accuracy of FFAs has been previously assessed by the direct comparison of an FFA to the corresponding target individual for similarity (i.e., a resemblance rating). Resemblance ratings may not indicate a facial approximation's accuracy since the resemblance of non-target individuals is not accounted for. This experiment tests the validity of using resemblance ratings to assess the accuracy of FFAs. The study indicates that there is no statistically significant difference between: (a) resemblance ratings of FFAs to target individuals and (b) resemblance ratings of FFAs to individuals incorrectly identified as the target individual. It is concluded that it is not possible from resemblance ratings to determine the accuracy and/or quality of a facial approximation since a non-target individual may receive a resemblance rating equal to, or higher than, the target individual.

KEYWORDS: forensic science, facial reconstruction, facial reproduction, skull, face, human identification, face recognition, face pool comparison, direct comparison

Facial approximation is a technique used to build a person's face, from their skull, approximating their facial appearance before death. Facial approximation may be used in forensic scenarios to promote recognition of a deceased person in an attempt to gather information that may aid the process of identifying the skeletal remains. Since forensic facial approximation (FFA) is used to promote recognition, an accurate FFA should be easily recognized as the person to whom the skull belonged (target individual).

The accuracy of FFAs has been assessed experimentally by face pool comparisons (1–3). This method requires showing a facial approximation to a group of assessors who attempt to identify the target individual out of a number of presented faces (i.e., the face pool). The face pool is made up of a number of non-target faces, of the same age, sex, and population of origin as the target individual. Depending on the experimental method, the target face may or may not be present in all face pools. Once the assessors have attempted to identify the target individual, the confidence at which identification rates for each face can be considered to be above, below, or equal to chance, can be calculated using statistical methods (e.g., chi squared test, Fisher's exact test). The higher the identification rate of the target individual above chance, at statistically significant levels, the more accurate the facial approximation.

The accuracy of forensic facial approximations as measured by face pool comparisons appears to be generally low. Stephan and Henneberg (3) found only one of 16 facial approximations to be identified at a statistically significant rate (25%) above chance ($p < 0.05$), with many non-target individuals identified for all facial approximations, some significantly above chance rates ($p < 0.05$). Snow et al. (1) found two of two facial approximations to be identified significantly above chance ($p < 0.05$), one identified 12% above chance rates, the other 54% above chance rates. Again, non-target individuals were identified in both cases. In Case 3, two non-target individuals (photos 4,6) were selected at rates close to that of the target individual (1) and appear to be above chance at statistically significant levels ($p < 0.07$). Van Rensburg (2) found that 15 facial approximations were, on average, identified at a rate 19% above chance rates, with the remaining identifications being of non-target individuals.

Not all authors have, however, assessed a facial approximation's accuracy by testing its ability to be recognized. Krogman (4), Suzuki (5), Helmer et al. (6), and Prag and Neave (7) have attempted to assess the accuracy of forensic facial approximations by directly comparing the appearance of the facial approximation to the corresponding target individual for similarities (resemblance ratings).

The accuracy of facial approximations, as judged by direct comparisons to a target individual, appears quite high in comparison to the accuracy obtained in experiments previously mentioned. Examples of the judged accuracy by direct comparison are: [the facial approximation is] "recognizable as that of the subject chosen" (4); "the resemblance between the two [the target individual and the facial approximation] was quite striking" (5); and "The reconstructed face bore an uncanny resemblance to the photograph" [of the target individual] (7).

Helmer et al. (6) measured the resemblance of facial approximations to their corresponding target individuals on a scale from 1 to 5 (1 being a great resemblance and 5 no resemblance). They concluded that "in general it can be said that at least a slight (rating of 4) and often even a close resemblance (rating of 2) was achieved" (6).

It appears that resemblance ratings are used to indicate the accuracy of a facial approximation because it seems that when two faces are similar they are recognizable. However, the validity of using resemblance ratings to assess a facial approximation's accuracy can be questioned for two reasons:

(1) Similar faces may not be the only recognizable faces. A face that does not appear to be morphologically similar to another

[1] Department of Anatomical Sciences, University of Adelaide, Adelaide 5005, Australia.

* Portions of this work have been presented at the 9th Biennial Scientific Meeting of the International Association for Craniofacial Identification.

Received 9 June 2000; 4 Dec. 2000; accepted 6 Aug. 2001.

may still be recognizable as belonging to the same person if observers are able to perceive recognizable characters in both faces despite their morphological differences. This may be evidenced in forensic casework where poor quality FFAs, despite bearing limited resemblance to target individuals, are still identified correctly. Caricatures, as well as pixilated images of faces, present a similar scenario since these images are not precise representations of an actual face, yet remain recognizable (8,9). Furthermore, caricatures have actually been shown to increase the ease of recognition of familiar faces (8,9).

(2) Resemblance ratings are not a relative measure, that is, they do not take into account non-target faces which may bear equal or higher resemblance to the FFA making them more recognizable than the actual target face.

These limitations may be the etiology of the discrepancy between the accuracy of FFA as measured by face pool comparison (low accuracy) and direct comparison methods (high accuracy).

This study tests the second limitation by determining if facial approximations correctly identified as the target individual receive a higher resemblance rating than those facial approximations that are incorrectly identified as the target individual.

Materials and Methods

Four skulls were approximated with four different techniques of facial approximation: (a) a 3D American sculpting method; (b) a 3D combination sculpting method; (c) a 2D FACE assisted computer method; and (d) a 2D American drawing method (for details of techniques see Ref 3.)

Thirty-seven assessors, with a background in the medical sciences, attempted to identify target individuals from a face pool for each facial approximation. Face pools consisted of ten photographs. Antemortem photographs were used of the target individuals. Non-target faces in the face pools were of the same sex and approximate age as the target individual. Faces in the face pools were standardized for size although this resulted in some photographs differing in resolution.

Since antemortem photographs were used of target individuals, the choice of photographs was limited and resulted in one photograph of a target individual wearing a hat and another sunglasses. In these cases, the corresponding faces in the face pool also had similar attire e.g., hat or sunglasses. All photographs were developed and printed on Ilford®IS3.1M photographic paper (127 mm by 100 mm) in black and white. In order to keep lighting between the faces in the face pools and the facial approximations consistent, facial approximations were photographed in a fluorescent-lit room without a flash. This was done to simulate an average, indoor, amateur "snap shot," which many of the photographs of the target individual faces appeared to be.

Assessors were presented with a facial approximation and a corresponding face pool and asked if they could identify a face from the face pool that was the individual approximated. Assessors had the option of not being able to make an identification i.e., deciding that the facial approximation did not correspond to any face in the face pool. Not all face pools included the target individual. Of 592 identification scenarios, 472 included the target individual and 120 did not. Face pools that did not include the target face had one of the other faces in the face pool repeated in a slightly altered position so that no new individuals were introduced, keeping face pools as consistent as possible. Repeated faces were developed in black and white on a slightly higher contrasting (Ilford® IS4.1M) photographic paper (127 mm by 100 mm).

Facial approximations, with corresponding face pools, were presented to assessors who followed written instructions and completed a questionnaire regarding which face (if any) the subject could identify as being that of the person approximated; the resemblance of the facial approximation to the face identified by the assessor; and the confidence with which the assessor thought they had made a correct identification. As assessors completed one assessment scenario (one facial approximation compared to the corresponding face pool) they were given another (in random order) until all 16 assessments were completed. Since four different methods of facial approximation were used on each of the four skulls, face pools were repeated for each facial approximation of the same individual. Since assessors were not aware of the total number of faces approximated and that each face pool included only one target individual, assessors were forced to treat each identification assessment as an independent case.

The written instructions indicated that when an identification was made, assessors were to rate the resemblance of the facial approximation to the face identified on a scale from 0 to 10. Weighting of the assessors' judgments was guided by the description that: 0 = no approximation; 2 = slight approximation; 4 = some approximation; 6 = close approximation; 8 = great approximation; and 10 = perfect approximation. Every other number between 0 and 10 represented the median weighting of the numbers immediately higher and lower. Assessors were also asked to rate the level of confidence that they had correctly identified the target individual by indicating if they were not confident, slightly confident, fairly confident, or confident.

The distribution and variance of the two samples (true positive identifications and false positive identifications) was analyzed (to determine if the assumptions for an unpaired t-test were fulfilled) before a two tailed, equal variance, unpaired t-test was used to determine if any statistically significant difference existed between the average resemblance ratings ($p < 0.05$, power = 90%). Pearson's product moment correlation coefficient was also used to determine if any relationship existed between the resemblance and confidence levels of both samples. Microsoft® Excel 98 was used for all data analysis.

Results

Of the total identification scenarios (592) there were 151 instances (26%) where no identification was attempted and 441 instances (74%) where identifications were made. Of these identifications, 354 (80%) were made when the target face was present in the face pool and 87 (20%) when the target face was not present. Thirty-eight (9%) of the identifications made were true positive (correct identifications of the target individuals) and 403 (91%) were false positive (identifications of non-target individuals).

A histogram indicated that both the true positive and the false positive samples were normally distributed i.e., 68% of sample scores fell within ±1 standard deviation of the mean and 99% of samples scores fell within ±3 standard deviations of the mean (Table 1). An f-test indicated that the variance of the average resemblance ratings for correct and incorrect identifications of target individuals was not significantly different ($p < 0.05$). A two tailed, equal variance, unpaired t-test ($p < 0.05$) indicated that there was no significant difference between the average resemblance for correct and incorrect identifications of target individuals at a power of 90% (Fig. 1).

TABLE 1—*Percentage frequency of resemblance ratings for true positive and false positive identifications.*

Resemblance	True Positive (%)	False Positive (%)
0	0.0	0.2
1	0.0	1.5
2	10.5	6.9
3	2.6	8.7
4	10.5	13.9
5	31.6	20.8
6	18.4	19.1
7	18.4	17.9
8	7.9	10.2
9	0.0	0.5
10	0.0	0.2
Average ± SD	5.3 ± 1.7	5.3 ± 1.8

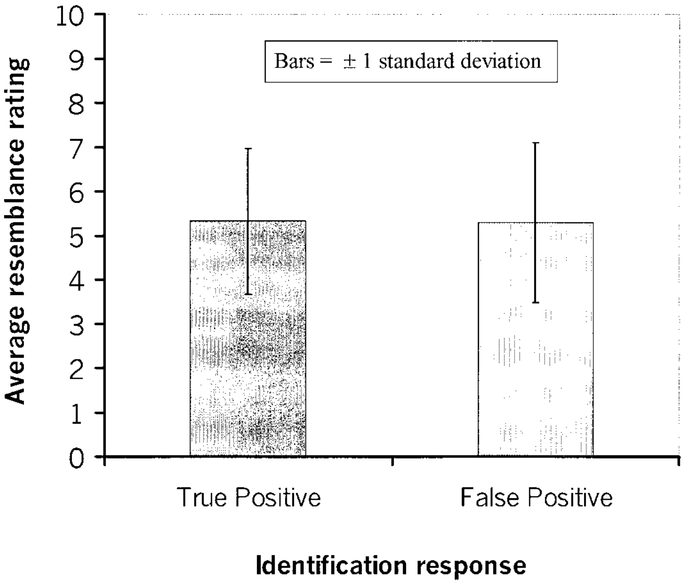

FIG. 1—*Average facial approximation resemblance for true and false positive identifications of faces.*

The resemblance rating of the facial approximation to the face identified tended to increase as the assessor's confidence level increased, independent of whether or not the correct face was identified (Fig. 2). These trends were significantly correlated ($r = 0.95$).

Large standard deviations of the mean for resemblance ratings (as reflected in Figs. 1 and 2) and large ranges in confidence levels were observed, although details have not been included in this paper.

Discussion

The observation that average resemblance ratings for true positive and false positive identifications do not differ at statistically significant levels ($p < 0.05$, $\beta = 0.10$) does not support the hypothesis that resemblance ratings indicate a facial approximation's accuracy, i.e., ability to be recognized as the target individual. It can, therefore, be concluded that resemblance ratings of a facial approximation to a target individual *does not* indicate a facial approximation's accuracy.

The increasing resemblance rating with the assessor's confidence level is to be expected since it is logical to assume that when the resemblance of a chosen individual's face to a facial approximation is perceived as being great, the assessor would believe he/she has made the correct choice and be confident of it. However, the results indicate that an assessor may not have made a correct identification even if an assessor is highly confident they have selected correctly and believes that the resemblance is high. The large range in responses for resemblance and confidence levels is not unexpected since both are subjectively determined.

Since resemblance ratings appear not to be valid in assessing the accuracy of a facial approximation, articles (4–7) that have used such methodology to assess the accuracy and/or quality of facial approximations are probably unreliable. Future studies assessing the accuracy of FFA should use the face pool comparison as opposed to the direct comparison method (resemblance ratings).

This study also demonstrates that printing images of a facial approximation and the corresponding target individual (as has been done in *almost all previous* publications of FFA) to indicate to readers a facial approximation's accuracy is of little use. In such a scenario, only the similarity between the two is indicated, not the "recognizability"/accuracy of the FFA. Authors are advised to print identification rates plus face pools accompanied by the FFAs (if confidentiality is not an issue). Despite the identification rates, images will allow readers to attempt to identify the target individual and compare their selections among themselves, which should help indicate the accuracy of the facial approximation presented.

Since face pool comparisons assess a facial approximation's ability to be recognized, those studies that have used this method appear to give the best indication of an FFA's accuracy and quality. The main disadvantage with this method is that only an unfamiliar identification scenario (the use of assessors who are not familiar with the target individual) has previously been employed. Unfamiliar identification scenarios are not representative of a real forensic environment because a familiar person usually recognizes the facial approximation. It has been shown that recognition of unfamiliar faces is much poorer than is the case with familiar faces (10). Unfamiliar face identification from still video images in a

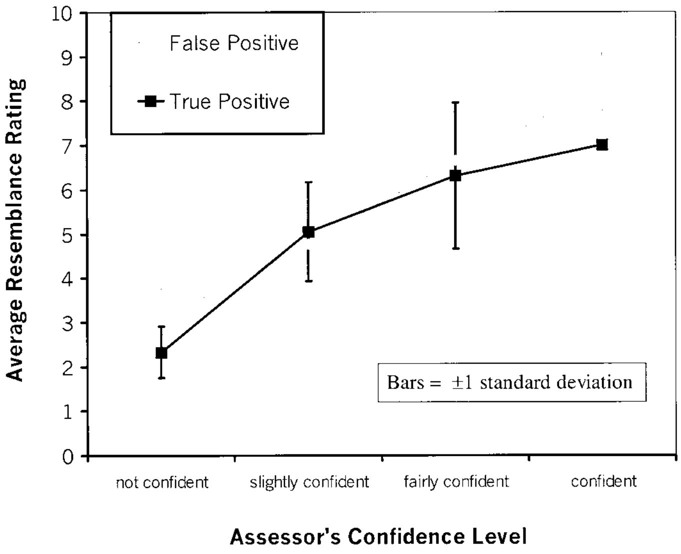

FIG. 2—*Average resemblance ratings by assessors' confidence levels comparing true positive and false positive identifications.*

face pool comparison scenario (of ten faces) shows that approximately 30% of responses can be expected to be incorrect, even if viewpoint of the faces and expression are constant (10). Therefore, assessing a facial approximation's accuracy in a familiar identification scenario may increase accuracy estimates.

Unfamiliar identification scenarios are also not representative of a real identification scenario since in familiar recognition it is the internal features of the face, as opposed to the external features, that play a more significant role in recognition (11). For unfamiliar scenarios the external features of the face appear to contribute as much as, or more than, the internal features for recognition (10,11). Present understanding of face recognition also shows differences in the mechanisms used to recognize a face. In an unfamiliar identification scenario recognition is primarily based on pictorial codes (12). In comparison, recognition of faces in familiar identification scenarios appears to be based on structural codes (12). The use of structural codes in unfamiliar scenarios may be facilitated however, by presenting assessors with several views of each photograph or by presenting all faces in the same exact conditions (e.g., same orientation, lighting, emotion, etc.) (12).

Since achieving a familiar identification scenario for deceased individuals is difficult (3), the face pool comparison method appears to be favorable as it presents assessors with a recognition task (despite the limitations listed above). However, the face pool comparison test is expected to be a more rigorous test of FFA accuracy than would probably occur in real scenarios since the use of structural codes is limited in face pool comparison methods. It is worthy to note if viewpoint or expression changes between the faces in a face pool or the stimulus face misidentifications are expected to increase (10). Also, short video clips in comparison to still images, and color images in comparison to grayscale, appear to offer no advantage, as they seem not to alter identification rates (10).

If the accuracy of facial approximation is to be increased from current low levels (3,8), more scientific guidelines for facial feature approximation need to be developed, with less emphasis on "artistic guidelines" (which have not generally been tested for reliability). Although artistic guidelines can be used to effectively build a face, as has been shown by the great works of many well-known artists, it is improbable that these general guidelines, which are often ideals, are accurate for building the face of an individual of unknown identity. The formulation of scientific guidelines will allow the error of each facial feature approximation to be estimated. Such an estimation of error may make it possible to estimate the confidence with which the final facial approximation can be considered to be accurate, if confidence levels are correlated with identification rates of the target individual. This would enable a quantitative measure of the "recognizability" of a facial approximation to be formulated. Such a measure would be beneficial since it could be generated without the use of timely face pool comparison experiments, which require the identification of the target individual *before* the accuracy of the facial approximation can be determined.

Although it has been reported that the upper face and proportions of the face are the most important for identification (13), it appears that this generalization is unjustified since many exceptions can be found. For example, although it has been reported that the eyes and mouth are important for recognition while the nose and chin are not (14–16), it must be noted that this is for full-face recognition. It may be that the role of the nose and chin may become more significant in profile views (17). Furthermore, not all individuals rely on the upper face and/or the eyes and/or the mouth for identification since it has been found that different subjects may use different features as recognition cues (18,19). Likewise proportions are probably not the most important for recognition as evidenced by studies that have found that both facial configurations and independent parts contribute to recognition (20,21). Again, the contribution of independent features to recognition is evidenced by studies mentioned above (14–16) that show feature salient cues are used in recognition. It is also evidenced by the observation that many people have similar/same facial proportions yet can still be recognized as unique individuals. It is probable that both proportional/holistic and feature specific cues play a substantial role in face recognition (10). It also appears to be of limited import to separate these two functions since they are closely related. That is, it is impossible to alter the proportions of the face without altering feature specific cues (Perrett 2000, personal communication). Likewise, it is impossible to alter local or feature specific cues with out altering proportions (10). Consequently, it is critical that facial approximationists do not place too much or too little emphasis on any feature detail/proportions of the face as all characters appear to contribute considerably to facial recognition. Similarly, studies examining facial relationships need to be numerous to examine all facial characters.

Although science has much to offer the facial approximation method, like reliable facial feature determination with known accuracy, it is not to say that all scientists are qualified to conduct FFAs. Many scientists lack the dexterity needed to produce FFA highly representative of living humans, especially when not using computers. Optimally, people with adequate knowledge of science, facial anatomy, anthropology, medicine, dentistry, face recognition/perception psychology, and high dexterity should be employed to conduct facial approximations. Unfortunately, few of these people exist. Consequently, compromises are often made, with many facial approximations being constructed from collaboration between a scientist (usually a physical anthropologist) and an artist. Although this is preferable in comparison to a scientist with limited dexterity undertaking facial approximation, it is not optimal nor is it essential, as has been claimed (13), if a person with adequate knowledge (in disciplines listed above) and dexterity is available. Collaboration between the artist and scientist increases the probability for errors to be introduced into the FFA due to lack of effective communication and/or cooperation and/or misunderstanding, all of which may happen unintentionally. Collaboration with an anthropologist seems also to have been stressed so that sex, age, and population of origin are accurately determined, as artists are unlikely to have any formal training in this area (13). This may, however, place the facial approximation knowledge and ability of the artist into question. If artists cannot recognize, from the hard tissue, the sex and/or age and/or population of origin of an individual, how can they know how that hard tissue relates to the soft tissue? Artists, by definition, are also creative and imaginative and may unconsciously use this talent in the FFA process, which is not advantageous. FFA methods should not include any creative aspect, but should demand high dexterity and strict adherence to scientifically tested relationships between the hard and soft tissues of the face. However, present knowledge of reliable relationships between the facial hard and soft tissues is not comprehensive and subsequently much subjective interpretation is required in the facial approximation process.

Conclusions

The lack of a statistically significant difference between the average resemblance ratings of a facial approximation identified correctly or incorrectly as the target individual indicates that resemblance ratings are not valid measures of a facial approximation's accuracy. There are three apparent ramifications of this study: (1) printing images of a FFA and the target individual for direct comparison, to indicate a FFA's accuracy to readers, is of little use; (2) results of studies that have used resemblance ratings are unreliable and should be approached cautiously; (3) facial approximation accuracy should be assessed by face pool comparison in the future.

Acknowledgments

Much thanks to Prof. Maciej Henneberg (The University of Adelaide) for his endless support and encouragement. I also wish to thank Dr. Jane Taylor and Dr. Kenneth Brown at the Forensic Odontology Unit (The University of Adelaide) for their invaluable technical assistance. Richard Neave (The University of Manchester), Chris Leigh (The University of Adelaide), Prof. John Clement, David Thomas, and Ronn Taylor (The University of Melbourne) are also thanked for their varied assistance, to allow successful execution of this experiment. Special thanks to Dr. Elanor Pierce (The University of Adelaide) and Dr. Ian Penton-Voak (The University of St-Andrews) for proofreading of this paper. Thanks also to Prof. David Perrett for use of facilities at the Perception Lab, The University of St-Andrews, during drafting of this paper.

References

1. Snow C, Gatliff B, McWilliams K. Reconstruction of facial features from the skull: an evaluation of its usefulness in forensic anthropology. Am J Phys Anthropol 1970;33:221–8.
2. Van Rensburg J. Accuracy of recognition of 3-dimensional plastic reconstruction of faces from skulls (Abstract). Anatomical Society of Southern Africa 1993; 23rd Annual Congress: 20.
3. Stephan C, Henneberg M. Building faces from dry skulls: are they recognized above chance rates? J Forensic Sci 2001;46(3):432–40.
4. Krogman W. The reconstruction of the living head from the skull. FBI Law Enf Bul 1946;17:7–12.
5. Suzuki T. Reconstitution of a skull. Intl Crim Police Rev 1973;264: 76–80.
6. Helmer R, Rohricht S, Petersen D, Mohr F. Assessment of the reliability of facial reconstruction. In: Iscan M, Helmer R, editors. Forensic analysis of the skull: craniofacial analysis, reconstruction, and identification. New York: Wiley-Liss, 1993;229–46.
7. Prag J, Neave R. Making Faces. Using Forensic and Archaeological Evidence. London: British Museum Press, 1997.
8. Rhodes G, Brennan S, Carey S. Identification and ratings of caricatures: implications for mental representations of faces. Cognit Psychol 1987; 19:473–97.
9. Benson P, Perrett D. Perception and recognition of photographic quality facial caricatures: implications for the recognition of natural images. Eur J Cognit Psychol 1991;3:105–35.
10. Hancock P, Bruce V, Burton A. Recognition of unfamiliar faces. TICS 2000;4:330–7.
11. Ellis H, Shepherd J, Davies G. Identification of familiar and unfamiliar faces from internal and external features: some implications for theories of face recognition. Perception 1979;8:431–9.
12. Bruce V, Young A. Understanding face recognition. Br J Psychol 1986; 77:305–27.
13. Taylor K. Forensic art and illustration. Florida: CRC Press, 2001.
14. Fraser I, Parker D. Reaction time measures of feature saliency in a perceptual integration task. In: Ellis H, Jeeres M, Newcombe F, Young A, editors. Aspects of face processing. Dordrecht: Martinus Nijhoft, 1986.
15. Haig N. The effect of feature displacement on face recognition. Perception 1984;13:505–12.
16. Haig N. Exploring recognition with interchanged facial features. Perception 1986;15:235–47.
17. Bruce V. Recognising Faces. UK: Lawrence Eilbaun Associates, 1989.
18. Sergent J. An investigation into component and configural processes underlying face recognition. Br J Psychol 1984;75:221–42.
19. Walker-Smith G, Gale A, Findlay J. Eye movement strategies involved in face perception. Perception 1977;6:313–26.
20. Haig N. How faces differ—a new comparative technique. Perception 1985;14:601–15.
21. Matthews M. Discrimination of Identikit constructions of faces: evidence for a dual processing strategy. Percept Psychophys 1978;23: 153–61.

Additional information and reprint requests:
Carl Stephan
Department of Anatomical Sciences
University of Adelaide
Adelaide 5005
Australia