J Forensic Sci, January 2008, Vol. 53, No. 1
doi: 10.1111/j.1556-4029.2007.00622.x
Available online at: www.blackwell-synergy.com

Carl N. Stephan,[1] *Ph.D.; and Jody Cicolini*[1]

Measuring the Accuracy of Facial Approximations: A Comparative Study of Resemblance Rating and Face Array Methods

ABSTRACT: The success of facial approximation is thought to depend, at least in part, upon the ''accuracy'' of the constructed face. However, methods of accuracy assessment are varied and this range in methods may be responsible for the disparate results reported in the literature. The aim of this study was to determine if the accuracy results of one facial approximation were comparable across two different assessment methods (resemblance ratings and simultaneous face array tests using unfamiliar assessors) and if resemblance ratings co-varied with recognition responses. True-positive recognition performance from the facial approximation was poor (21%) while resemblance scores using the same facial approximation were moderately high (3 out of 5 on a five-point scale). These results are not, therefore, consistent and indicate that either different variables are being evaluated by the methods, or the same variable is being examined but with different weight/calibration. Further resemblance ratings tests of the facial approximation to three foil faces from the face array revealed that resemblance scores were similar irrespective of which face was compared, and did not closely correspond with the degree of recognition performance. This was especially the case for isolated comparisons of single faces to the facial approximation. Collectively, these results indicate that resemblance ratings are: (i) insensitive measures of a facial approximation's accuracy; and (ii) inconsistent with results of unfamiliar simultaneous face-array recognition results. These data suggest that familiar and unfamiliar recognition tests should be given increased weight in contrast to current resemblance rating tests.

KEYWORDS: forensic science, facial reconstruction, facial reproduction, performance, success, recognition

Facial approximation is the method of building from a person's skull, an antemortem facial representation which can be purposefully recognized as the person to whom the skull belonged. Since late 1800's, the potential of this method for helping to establish the identity of skeletonized remains has been recognized (e.g., see 1–3); however, their suggested application to forensic investigations was initially met with some skepticism (4). This skepticism was partly dissipated when facial approximation methods achieved their first casework success *c.* 3 years after concerns had first been raised (5). Despite their validity and current acceptance as an investigation tool (6–8), debate continues as to whether facial approximation casework success is because of the accuracy of the constructed faces or other factors, such as supporting case descriptions and/or the effectiveness of media advertising to draw public attention (9–12).

To determine the role that facial approximation accuracy plays in casework performance it would be advantageous to have a clear definition, and reliable and definitive laboratory tests for facial approximation accuracy. Currently, many different ''accuracy'' assessment methods are employed in professional circles (Table 1), but not all may test the same variable with the same calibration. There are some claims in the literature that resemblance ratings provide little useful information, as incorrectly and correctly identified faces receive similar resemblance rating scores (11), but claims have also been made that disparities exist between the results of resemblance rating and face array tests (13). This study aims to further elucidate the issue by testing: (i) the accuracy of a common facial approximation using resemblance rating and face array tests to determine if comparable results are obtained; and (ii) if resemblance scores of disparately recognized nontarget faces co-vary with their respective identification rates. Thus, this study is split into two parts.

[1]Anatomy and Developmental Biology, School of Biomedical Sciences, The University of Queensland, Brisbane, Australia.

Received 4 Mar. 2007; and in revised form 21 June 2007; accepted 28 July 2007.

Experiment 1: Face Array and Resemblance Rating Tests Using the Same Facial Approximation

Materials and Methods

A facial approximation was constructed from a cast of a male skull under blind conditions by a trainee facial approximation practitioner (the second author) who followed published guidelines. Before facial approximation methods were employed, the edentulous skull (the only skeletal remains of this individual available for study) was subject to age assessment. The skull was determined to belong to an individual who was middle aged (i.e., *c.* 30 to 50 years), but this range could not be narrowed as the sutures were the only available aging characters (see 14–19).

After the mandible had been positioned using dental wax, the face was constructed primarily following Neave's method (20) and soft tissue depth data reported by Helmer for 40-50-year-old males (21). Prosthetic eyeballs were positioned centrally within the orbit (12,22) and protruded anterior to the deepest portion of the lateral orbital wall by 16 mm (23,24). Following Neave's (20) directions, most of the muscles of mastication and facial expression were represented on the plaster cast using clay. These muscles were modeled based on textbook descriptions and relationships evident from prosected wet and plastinated specimens, following the directions of Wilkinson (22; see Fig. 1).

As the canine teeth were not present, their relative positions were estimated and the mouth width calculated using the guideline of Stephan and Henneberg (25). As no bony landmarks were available for mouth width determination, the estimated distances/positions were double checked against the (estimated) locations of the medial

TABLE 1—*Types of facial approximation accuracy assessment methods used in the literature.*

Accuracy Assessment Method	Papers / Sources
a. Qualitative statements of resemblance	Suzuki, 1973 (44); Krogman, 1946 (45); Prag & Neave, 1997 (20)
b. Resemblance ratings	Helmer, 1993 (34); Wilkinson, 2004 (22)
c. Recognition results from face arrays using unfamiliar assessors (simultaneous or sequential presentation)	Snow et al., 1970 (47); Stephan & Henneberg, 2001 (46); Stephan & Henneberg, 2006 (13)
d. Recognition results using familiar assessors	Stephan et al., 2005 (48)
e. Practitioner recognition of target faces	Prag & Neave, 1997 (20)

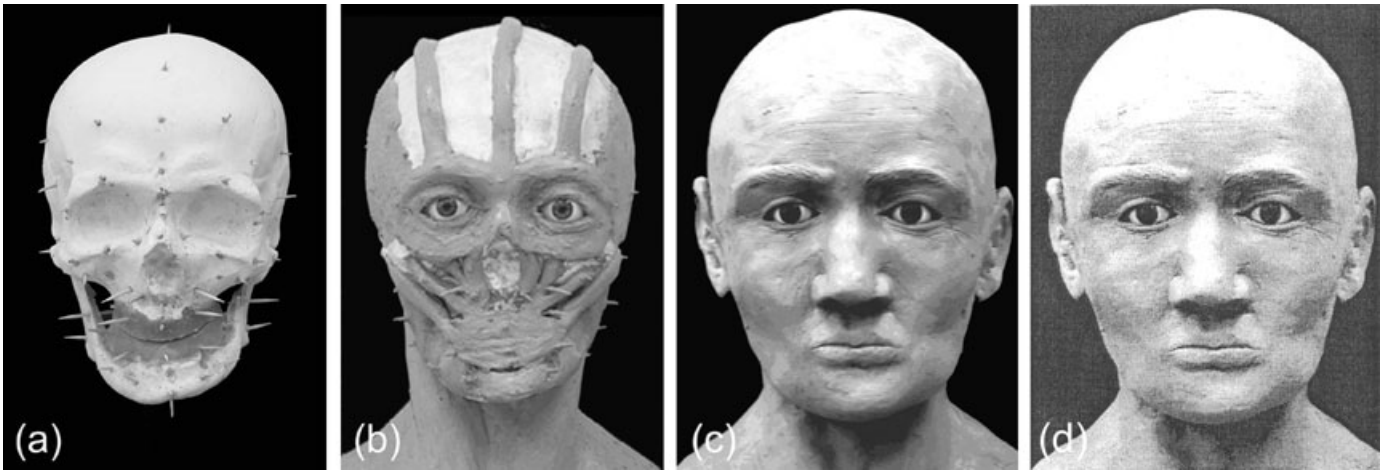

FIG. 1—*Construction of the facial approximation. (a) plaster cast of skull with soft tissue depths in place and mandible secured to cranium using dental wax; (b) partially completed facial approximation; (c) finished facial approximation; (d) the facial approximation as presented to assessors after photocopying.*

iris edges (20,26). The lip closure line and stomion were estimated to lie in the region where the central incisors would have been if they had been present (27). The nose width was determined according to the guideline that the nasal aperture represents three-fifths of the total width of the nose (28) and projection was based on the double tangent guideline (20,29,30).

Sheets of clay (*c.* 5 mm thick) were placed over the muscles to approximate the facial soft tissue contours (20). As the soft tissue depths were only used as guides, small portions (≤0.5 mm) of the indicator pegs remained exposed at several locations after the clay sheets had been set in place; thus, these pegs were trimmed. Eyebrows were represented along the supraorbital margin of the skull (31), but as they are known to show considerable variation (32), they were modeled to be suggestive rather than definite. The auricles were estimated to be slightly larger than the height of the nose (*c.* 10 mm) following data of Farkas et al. (33). After final "touch-ups," the completed facial approximation was photographed in readiness for accuracy tests (Fig. 1).

A face array was constructed which comprised ten faces: the target individual's face and nine other nontarget faces of same sex and approximate age as the target individual. The nontarget faces were strategically selected so that assessors were unlikely to be familiar with them (the nine foil faces came from a nonfamous and non-University cohort). Adobe® Photoshop® 7.0 was used in an attempt to standardize, in terms of lighting and image resolution, the nine "foil" face array photographs against those of the target individual (see Fig. 2). To determine if the face array was balanced in terms of photographic appearances, we informally asked three individuals who were blind as to which face was that of the target individual, if any face in the array stood out as being different from the rest. All three individuals indicated that the face #4 (the target individual) did not seem to fit the "look" of the other photographs. Most of these individuals reported that photograph #4 looked "old-fashioned" in contrast to the other photos (e.g., see differences in hairstyle, dress, pose, and photographic quality; Fig. 2). In addition, the eyes of the target individual were the only ones that deviated from a forward direction. Despite these visual clues as to which face belonged to the target individual, we proceeded with the experiment as there was the chance that these biases would be useful rather than problematic (i.e., in the case that infrequent true-positive recognition responses were obtained).

The facial approximation and the face pool were printed onto a single A4 answer sheet using a laser printer. This answer sheet was photocopied to obtain the desired quantity for the experimental tests. The same procedure was followed for resemblance rating tests except that only the facial approximation (Fig. 1*c*) and the target face (Face #4, Fig. 2) were printed onto the same answer sheet. Figs. 1 and 2, respectively, demonstrate the appearance of the photocopied facial approximation and array faces which assessors examined. The height of the two images shown in the resemblance rating test were identical (85 mm); however, because of space restrictions in the recognition test, the face array images were 46 mm in height while the facial approximation was 72 mm in height.

A total of 80 second-year anatomy students from University of Queensland, who were unfamiliar with the target individual (person to whom the skull belonged) acted as assessors and were thus used to gauge the accuracy of the facial approximation. Forty-eight assessors were used for the recognition tests (14 males and 34 females; mean age = 20 years, s = 4 years, range = 18 to 45 years), while 34 different assessors were used for the resemblance rating tests (17 males and 15 females; mean age = 20 years, s = 4 years, range 17 to 35 years). Prior to taking part in any experiment, all assessors received a brief project information sheet and viewed a short video which demonstrated the construction of the facial approximation used in the current study.

FIG. 2—*The photocopied face array as used for "unfamiliar" recognition tests. The target individual is face #4.*

For the face array test, assessors were asked to examine the facial approximation and use it in an attempt to identify the target face from the array. This evaluation was "forced choice," so all assessors had to identify one of the array faces. The expected chance rate for "guessing" any face in the face pool, including that of the target individual was therefore 10%. Recognition responses for each face array photograph were compared to the chance rate using Fisher's exact tests ($p < 0.05$) in the GraphPad Prism® 4.01 statistical package (GraphPad Software, San Diego, CA). Note here that statistical significance levels were not adjusted for multiple tests.

For the resemblance rating tests, a new group of assessors were asked to score the resemblance between the facial approximation and the target individual using a rating scale from one to five (1 = great resemblance, 2 = close resemblance, 3 = approximate resemblance, 4 = slight resemblance, 5 = no resemblance) as previously described by Helmer (34). This type of scale is commonly employed in facial approximation circles (e.g., 22,34) and yields discrete ordinal data that are nominally coded. The assessors that were used in the resemblance rating tests did not observe any of the nontarget faces included in the face array. In assessing the facial approximation, these assessors were asked: "how accurately do you think the facial approximation resembles the actual individual?," and they wrote their answer on the answer sheet.

Results

Recognition tests showed that face #2 had the highest identification rate (33%) and was the only face selected at rates above chance at statistically significant levels ($p < 0.05$; Fig. 3). Face #6 was the second most identified, but at 26% this rate was not above chance at statistically significant levels ($p > 0.05$). The target face, image #4 was the third most identified face and had an identification rate of 21%, but this rate was also not above chance at statistically significant levels ($p > 0.05$). Four faces in the face array went unidentified throughout the entire test (Face #'s: 1, 5, 8, & 10).

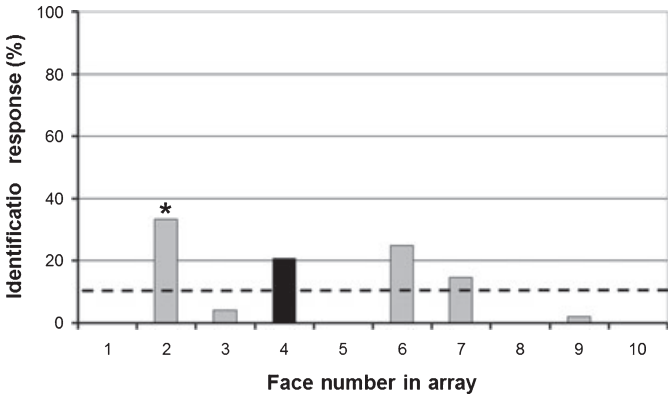

FIG. 3—*Face array recognition results. The black bar indicates recognition responses of the target face and the dashed line represents the chance identification rate. The "*" represents statistically significant difference from the chance at* $p < 0.05$.

Resemblance scores of the facial approximation to the target face ranged from two to five, with a mean, median, and mode of three (equivalent of "approximate resemblance"; Fig. 4). Eighty-eight percent of assessors thought that the facial approximation bore at least a slight resemblance to the target individual with 60% of assessors rating it "approximate resemblance" or better.

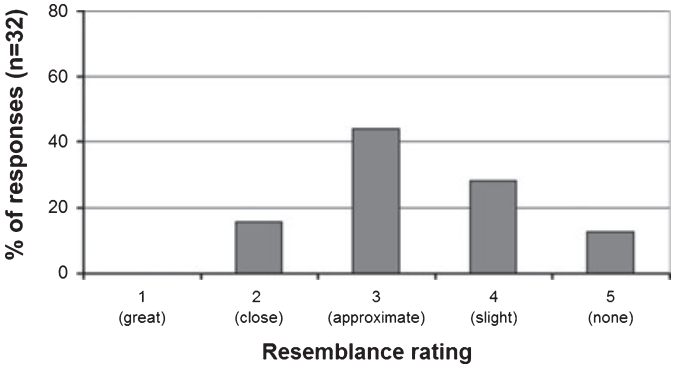

FIG. 4—*Resemblance rating results of the facial approximation to the target individual (face array #4).*

Discussion

Although resemblance ratings and recognition tests have both been employed to measure facial approximation accuracy in the literature (see Table 1), the results of this study indicate that these two assessment methods produce different outcomes, even if the same facial approximation is used. The face array recognition test indicated poor accuracy as the true positive identification rate was low (20%), and was not above chance ($p > 0.05$), and its rank order was high (it was only the third highest identification rate out of six identified faces). In contrast, the resemblance rating responses indicated a moderate degree of accuracy with the mean response rate equal to three on a five-point scale. Although there was some right sided skew in the distribution, the vast majority of responses were positive: *c.* 60% indicated an approximate-to-close resemblance. The mismatch between the results of the two methods indicates that they are, at best, weighing the same variable differently, or at worst, measuring two completely different variables.

The mean resemblance score observed in this study was slightly lower than, but still comparable to resemblance ratings of: (i) facial approximations that have been regarded to be "accurate" (34); (ii) facial approximations that have not been identified above chance rates (35); (iii) facial approximations that have been identified above chance rates (22,35); and (iv) facial approximations that have either been correctly or incorrectly identified (11). The inconsistency of these data do not augur well for resemblance rating tests; however to more precisely determine their nature, an additional experiment was conducted to examine the resemblance rating scores of the three disparately recognized "foil" faces used in the face array.

Experiment 2: Resemblance Rating Scores of Disparately Recognized Foil Faces

Materials and Methods

Three nontarget images from the face array which represented a broad range of false positive identification responses were subject to resemblance rating tests to determine if these scores co-varied with the identification rate of each face. The three "foil" faces comprised: (i) the most frequently identified nontarget face (face array image #2); (ii) the first unidentified nontarget face in the face array sequence (face array image #1); and (iii) a nontarget face that was identified at rates close to (but slightly below) that of the target face (face array image #7).

Resemblance scores for each of the foil faces were obtained by using for each face, new groups of assessors who had not taken part in prior experiments. Furthermore, an additional group of assessors was used to view all three foil faces simultaneously and make a resemblance rating for each. Thus, the three foil faces were subject to tests using both unmatched and matched experimental designs.

The procedure for testing the facial approximation to a single face array image was identical to that used in experiment 1; however, for the group that examined all three faces at once, the question was asked "how accurately do you think the facial approximation resembles each of the three individuals presented below?" The facial approximation and each of the face array images were presented side-by-side as in experiment 1, however, in the case of simultaneous face comparisons, all three images were presented side-by-side and adjacent to the facial approximation image. All images in this second resemblance rating experiment were of identical size to the images used for the original resemblance rating test and all participants received a similar brief on the facial approximation methods before undertaking the experiment.

A total of 93 second-year anatomy students from University of Queensland, who did not take part in experiment 1 and who were unfamiliar with the face array photographs acted as assessors. Face #2 from the face array was evaluated by 16 assessors (nine males and seven females; mean age = 21 years, $s = 9$ years, range = 18 to 56 years). Face #1 was evaluated by 21 different assessors (eight males and 13 females; mean age = 21 years, $s = 6$ years, range = 18 to 40 years). Face #7 was evaluated by 24 different assessors (11 males and 13 females; mean age = 20 years, $s = 4$ years, range = 18 to 36 years). All three faces were evaluated by 32 different assessors (nine males and 22 females; mean age = 22 years, $s = 6$ years, range = 18 to 44 years).

As the data were of the ordinal type and not normally distributed, nonparametric methods were used to test for statistical significance. A Kruskal–Wallis ANOVA (using Dunn's post test) was used to analyze the differences in resemblance scores between each of the single face array images in the unmatched design. A Friedman two-way ANOVA (using Dunn's post test) was used to analyze the matched data (i.e., the differences between resemblance scores for each of the three simultaneously presented face array images). All statistical analyses were conducted within the GraphPad Prism® 4.01 statistical package (GraphPad Software).

Results

Mean and median resemblance scores for the three independently assessed foil faces in the unmatched design were extremely similar (see Fig. 5) and statistical tests revealed that the median score for face #7 was less than the median scores for both face #1 and #2 ($p < 0.05$). There was no statistically significant difference between medians for face #1 and #2. For each comparison the resemblance scores tended to cluster near the middle of the scale with few assessors using either extreme (Fig. 6). This was the case even for face #1 which was *never* identified during the face array tests, and which also received a mean resemblance score identical to that of the *most* identified nontarget face (Fig. 5). A statistical comparison of the median scores for face #4 (the target face) to face #1 and #2 using a Kruskal–Wallis ANOVA (and Dunn's post test) revealed no statistically significant differences between these groups.

For the assessors who viewed all three faces at the same time and rated the facial approximations resemblance to each, median resemblance ratings were found to be most favorable for the poorly-recognized face (face #7, median = 2), followed closely by the most-recognized face (face #2, median = 3) and last by the unrecognized face (face #1, median = 4; see Fig. 5). Statistical tests

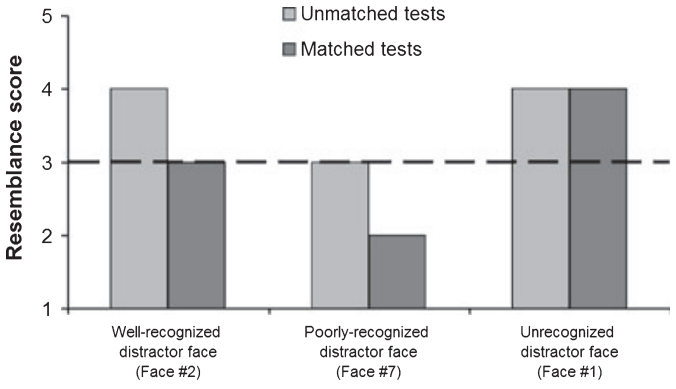

FIG. 5—*Median resemblance rating scores for disparately recognized faces from the face array. The dashed line indicates the median resemblance score previously obtained for the target face (face #4).*

revealed that face #1 received resemblance ratings significantly higher than both face #7 and #2 ($p < 0.01$), but that resemblance scores of face #7 and #2 did not differ ($p > 0.05$).

Discussion

The results of this study clearly indicate that resemblance ratings generated from a five-point scale using a side-by-side comparison of a single face to a facial approximation, as is typically used in current facial approximation procedures, yields little valuable information. Resemblance ratings of different faces are similar (and toward the center of the rating scale) irrespective of the morphology, or the prior recognition performance of the face used for comparison. Furthermore, discrimination between faces using resemblance ratings did not dramatically improve with matched designs where assessors had the opportunity to compare all three faces to one another while making their resemblance rating decisions. This protocol did, however, yield resemblance scores that tended to differentiate unrecognized faces (they received the highest score), but no differentiation was found between poorly or strongly recognized faces. The best recognized face from the face pool (which was recognized above chance rates at statistically significant levels) received a median resemblance score which was higher than that of a poorly recognized face (median of three as compared to a median of two, respectively). Thus, the value of resemblance ratings even in matched comparisons where assessors can simultaneously view a number of faces appears to be small.

General Discussion

This study provides additional support to past claims (11,35) that commonly employed resemblance rating protocols offer little useful information concerning the accuracy of a facial approximation. Firstly, resemblance rating tests do not produce results that are highly consistent with recognition performances and secondly, resemblance rating tests produce similar (and moderately high) resemblance scores irrespective of which face is used for the comparison. This is especially the case when comparison faces are presented singly and in isolation.

It is worth noting here that "show-up" tests in eyewitness identification are also known to be weak because they examine only one person of interest and are thus biased (36–38)—a comparable scenario to the resemblance rating test. Furthermore, without comparison faces or exemplars to calibrate the resemblance scale, assessors may choose to nominate more ambiguous responses toward the middle of the scale as a safety measure (i.e., to ensure they are not entirely incorrect in their decision). This pattern can be observed not only in this study but also in many others (13,22,34) where the extremes of resemblance rating scales are rarely selected. Particularly pronounced was this effect in a recent study by Stephan and Arthur (35), where one of the constructed faces was frequently correctly recognized (true positive identifications = 98%) and another was infrequently correctly recognized (true positive identifications = 12%), yet both facial approximations received similar resemblance rating within the mid-range of the resemblance rating scale.

The findings of this study (and those reported in the literature) do not rule out the possibility that resemblance rating scales may be

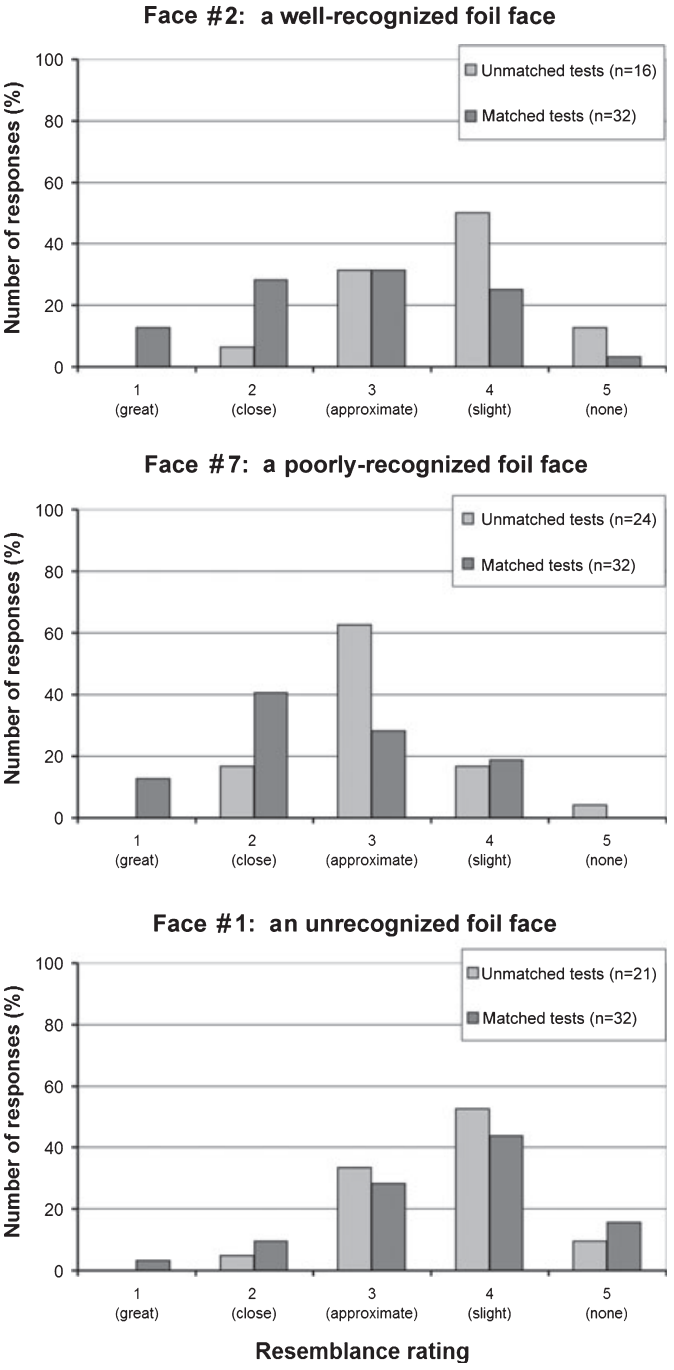

FIG. 6—*The resemblance rating distributions for disparately recognized faces from the face array.*

useful with some alteration, but they strongly suggest that currently employed methods are, at least, insensitive. This lack of sensitivity may be because of the ordinal nature of the scale since the distance to neighboring data points may not be identical within different scale regions. That is, if Helmer's scale is reversed, so 1 becomes no resemblance and 5 becomes great resemblance, then the value of 4 (close resemblance) may not be equal to two times the score of 2 (slight resemblance). Thus, the high density of resemblance rating scores which typically fall between 2 and 3 on Helmer's original rating scale may result from a disproportionately large space between "close" and "approximate" resemblance as viewed by the assessors. Improvements to the scale may, therefore, be possible by: (i) trimming unused extremes and exploding the magnification of the commonly used region; (ii) using a scale that is interval in nature (rather than ordinal); and/or (iii) redefining the discrete values along the ordinal scale. While all/any of these changes may help to improve the sensitivity of the resemblance rating scale, they do not circumvent the problem that when using a single face for comparison, assessors are not given the opportunity to view other nontarget faces. While this problem could be solved by having assessors rate many faces, the procedure becomes cumbersome and time-intensive and offers no additional advantage over recognition trials which better target recognition performance.

It should be noted here that attempts to improve resemblance rating scales by increasing the interval nature have been previously made by Stephan and Arthur (35) but without success. These authors defined the extremes of their 6-point scale as "no" and "high" resemblance and used evenly distributed interval values (i.e., 1–3) between these extremes. The results of their experiments indicate that disparate facial approximations still received similar resemblance rating scores, even though recognition responses were found to be extremely different. These observations suggest that if resemblance rating methods are to be improved, several of the above suggested factors may need to be addressed. Further pursuits in this area remain valuable as a fast and simple test that meaningfully measures the accuracy of facial approximations would be favorable to more time-intensive recognition trials. However, at this time it seems unlikely that a direct comparison between a single face and the facial approximation can accommodate for the limits imposed by the exclusion of other nontarget faces and, therefore, recognition tests using face arrays appear to be most favorable. Here, it is worth noting that multiple formats of recognition tests exist, but that sequential presentation methods are favored because they preclude large numbers of false positive identifications (see for evidence in the facial approximation context [13] or for eyewitness identification contexts [39–43]).

While face array tests hold advantages, it is important to acknowledge that they are not without their limitations. For example, in forensic casework, a member of the public is not presented with an array of faces from which to select, rather the identification is usually made from memory. Furthermore, in such cases, the identification is usually made under familiar conditions (the person making the identification knew the person they are nominating well), which is unrepresentative of the unfamiliar face array test. These limitations could be avoided by using imaging techniques to scan a living subject and capture data to generate a replica skull, upon which a facial approximation could be constructed and tested using relatives of the target individual. However, no such study using a large number of skulls has yet been conducted and thus results of these ultimate tests remain to be determined.

Although this study did not set out to test the accuracy of facial approximation methods themselves, but rather the difference between these assessment methods, this investigation incidentally provides a further example of a poor recognition performance of current facial approximation methods. This result was observed even though the face pool was constructed in such a way as to favor correct responses (i.e., the target individual's antemortem photograph stood out from the other face array images in terms of direction of gaze and photographic style, and up to 40% of the faces included in the face array may have been nonfunctional—that is, as they were never recognized, they may have been too easily excluded as plausible matches).

The lack of above chance recognition of the target face in this study may indicate deficiencies in published facial approximation methods, or alternatively, it may be because of the trainee status of the practitioner as some have previously argued (e.g., see 20). Despite a lack of scientific evidence concerning the role practitioner experience plays in facial approximation method performance, it seems reasonable to conclude that the poor recognition results of facial approximations are, at least in part, due to weaknesses in methods. The relative paucity of empirically tested and published soft tissue prediction guidelines (see e.g., 6,20,22,30) for the estimation of the face (a complex biological feature) appears to be a significant limitation.

Acknowledgments

Thanks are extended to Prof. Maciej Henneberg for permission to use in this study the above described skull and antemortem photograph. This material originally formed a case belonging to Dr. Ram Tulsi and is currently held for teaching/research purposes at the School of Medical Sciences, The University of Adelaide, Australia. Thanks also to Prof. Henneberg for his age assessment on the original skull and to an anonymous reviewer who inspired the second experiment by encouraging resemblance rating studies of array faces using unmatched designs.

References

1. Welcker H. Schiller's Schadel und Todtenmaske, nebst Mittheilungen über Schadel und Todtenmaske Kant's. Braunschweig: Viehweg F and Son, 1883.
2. Welcker H. Zur Kritik des Schillerschadels. Arch Anthropol 1888;17:19–60.
3. His W. Anatomische Forschungen über Johann Sebastian Bach's Gebeine und Antlitz nebst Bemerkungen über dessen Bilder. Abh MathPhysikal KI Kgl Sachs Ges Wiss 1895;22:379–420.
4. Von Eggeling H. Die Leistungsfahigkeit physiognomischer Rekonbstruktionsversuche auf Grundlage des Schadels'. Archiv für Anthropologie 1913;12:44–7.
5. Wilder HH, Wenworth B. Personal identification: methods for the identification of individuals, living or dead. Boston: Gorham Press, 1918.
6. Krogman WM, Işcan MY. The human skeleton in forensic medicine. Illinois: Charles C Thomas, 1986.
7. Reichs KJ, Craig E. Facial approximation: procedures and pitfalls. In: Reichs KJ, editor. Forensic osteology: advances in the identification of human remains. Springfield: Charles C. Thomas, 1998;491–513.
8. Clement JG, Ranson DL. Craniofacial identification in forensic medicine. London: Arnold, 1998.
9. Haglund WD. Forensic "art" in human identification. In: Clement JG, Ranson DL, editors. Craniofacial identification in forensic medicine. London: Arnold, 1998;235–43.
10. Haglund WD, Reay DT. Use of facial approximation techniques in identification of green river serial murder victims. Am J Forensic Med Pathol 1991;12:132–42.
11. Stephan CN. Do resemblance ratings measure the accuracy of facial approximations? J Forensic Sci 2002a;47:239–43.
12. Taylor R, Craig P. The wisdom of bones: facial approximation on the skull. In: Clement JG, Marks MK, editors. Computer-graphic facial reconstruction. Boston: Elsevier Academic Press, 2005;33–55.
13. Stephan CN, Henneberg M. Recognition by facial approximation: case specific examples and empirical tests. Forensic Sci Int 2006;156:182–91.

14. Montagu MFA. Aging of the skull. Am J Phys Anthropol 1938;23:355–75.
15. Singer R. Estimation of age from cranial suture closure. J Forensic Med 1953;1:52–9.
16. Perizonius WRK. Closing and non-closing sutures in 256 crania of known age and sex from Amsterdam (AD 1883–1909). J Hum Evol 1984;13:201–16.
17. Todd TW, Lyon JDW. Cranial suture closure. Part II Ectocranial suture closure in adult males of white stock. Am J Phys Anthropol 1925;8: 23–43.
18. Hershkovitz I, Latimer B, Dutour O, Jellema LM, Wish-Baratz S, Rothschild C, et al. Why do we fail in aging the skull from the sagittal suture? Am J Phys Anthropol 1997;103:393–9.
19. Galera V, Ubelaker DH, Hayek L-AC. Comparison of macroscopic cranial methods of age estimation applied to skeletons from the Terry collection. J Forensic Sci 1998;43:933–9.
20. Prag J, Neave R. Making faces: using forensic and archaeological evidence. London: British Museum Press, 1997.
21. Helmer R. Schadelidentifizierung durch elekronicshe bildmischung: Zugleich ein beitrag zur konstitutionsbiometrie und dickenmessung der gesichtsweichteile. Heidelberg: Krminalistik-Verlag, 1984.
22. Wilkinson C. Forensic facial reconstruction. Cambridge: Cambridge University Press, 2004.
23. Stephan CN. Facial approximation: falsification of globe projection guideline by exophthalmometry literature. J Forensic Sci 2002b;47:1–6.
24. Wilkinson CM, Mautner SA. Measurement of eyeball protrusion and its application in facial reconstruction [technical note]. J Forensic Sci 2003;48:12–6.
25. Stephan CN, Henneberg M. Predicting mouth width from inter-canine width—a 75% rule. J Forensic Sci 2003;48:725–7.
26. Stephan CN. Facial approximation: an evaluation of mouth width determination. Am J Phys Anthropol 2003;121:48–57.
27. George RM. The lateral craniographic method of facial reconstruction. J Forensic Sci 1987;32:1305–30.
28. Hoffman BE, McConathy DA, Coward M, Saddler L. Relationship between the piriform aperture and interalar nasal widths in adult males. J Forensic Sci 1991;36:1152–61.
29. Gerasimov M. Vosstanovlenie lica po cerepu. Moskva: Izdat. Akademii Nauk SSSR, 1955.
30. Gerasimov M. The face finder. London: Hutchinson & Co., 1971.
31. Fedosyutkin BA, Nainys JV. The relationship of skull morphology to facial features. In: Iscan MY, Helmer RP, editors. Forensic analysis of the skull. New York: Wiley-Liss, 1993;199–213.
32. Stephan CN. Position of superciliare in relation to the lateral iris: testing a suggested facial approximation guideline. Forensic Sci Int 2002c;130:29–33.
33. Farkas LG, Forrest CR, Litsas L. Revision of neoclassical facial canons in young adult Afro-Americans. Aesthetic Plast Surg 2000;24:179–84.
34. Helmer RP, Rohricht S, Petersen D, Mohr F. Assessment of the reliability of facial reconstruction. In: Iscan MY, Helmer RP, editors. Forensic analysis of the skull. New York: Wiley-Liss, 1993.
35. Stephan CN, Arthur RS. Assessing facial approximation accuracy: how do resemblance ratings of disparate faces compare to recognition tests? Forensic Sci Int 2006;159S:S159–63.
36. Malpass RS, Devine PG. Measuring the fairness of eyewitness identification lineups. In: Lloyd-Bostock SMA, Clifford BR, editors. Evaluating witness evidence: recent psychological research and new perspectives. New York: John Wiley and Sons, 1983;81–102.
37. Wells GL. What do we know about eyewitness identification? Am Psychol 1993;48:553–71.
38. Gonzalez R, Ellsworth PC, Pembroke M. Response biases in lineups and showups. J Pers Soc Psychol 1993;64:525–37.
39. Lindsay RCL, Lea JA, Nosworthy GJ, Fulford JA, Hector J, LeVan V, et al. Biased lineups: sequential presentation reduces the problem. J Appl Psychol 1991a;76:796–802.
40. Lindsay RCL, Wells GL. Improving eyewitness identifications from lineups: simultaneous versus sequential lineup presentation. J Appl Psychol 1985;70:556–64.
41. Melara RD, DeWitt-Rickards TS, O'Brien TP. Enhancing lineup identification accuracy: two codes are better than one. J Appl Psychol 1989;74:706–13.
42. Lindsay RCL, Lea JA, Fulford JA. Sequential lineup presentation: technique matters. J Appl Psychol 1991b;76:741–5.
43. Cutler BL, Penrod SD. Improving the reliability of eyewitness identification: lineup construction and presentation. J Appl Psychol 1988;73:281–90.
44. Suzuki T. Reconstitution of a skull. Int Crim Police Rev 1973;264: 76–80.
45. Krogman WM. The reconstruction of the living head from the skull. FBI Law Enforce Bull 1946;17:11–7.
46. Stephan CN, Henneberg M. Building faces from dry skulls: are they recognized above chance rates? J Forensic Sci 2001;46:432–40.
47. Snow CC, Gatliff BP, McWilliams KR. Reconstruction of facial features from the skull: an evaluation of its usefulness in forensic anthropology. Am J Phys Anthropol 1970;33:221–8.
48. Stephan CN, Penton-Voak IS, Clement JG, Henneberg M. Ceiling recognition limits of two-dimensional facial approximations constructed using averages. In: Clement JG, Marks M, editors. Computer graphic facial reconstruction. Boston: Academic Press, 2005;199–219.

Additional information and reprint requests:
Carl N Stephan, Ph.D.
Anatomy and Developmental Biology
School of Biomedical Sciences
The University of Queensland
Brisbane, 4072
Australia
E-mail: c.stephan@uq.edu.au