Forensic Science International 201 (2010) 133–137

Contents lists available at ScienceDirect

Forensic Science International

journal homepage: www.elsevier.com/locate/forsciint

The reproducibility of facial approximation accuracy results generated from photo-spread tests[☆]

Carl N. Stephan [a,b,*], Jody Cicolini [a]

[a] Anatomy and Developmental Biology, School of Biomedical Sciences, The University of Queensland, Brisbane 4072, Australia
[b] Joint POW/MIA Accounting Command, Central Identification Laboratory, 310 Worchester Avenue, Building 45, Hickam Air Force Base, HI 96853, USA

ARTICLE INFO

Article history:
Received 14 January 2010
Accepted 23 February 2010
Available online 24 March 2010

Keywords:
Forensic science
Facial reconstruction
Facial reproduction
Face array
Face pool
Face gallery

ABSTRACT

The accuracies of facial approximations have been measured by determining the frequency that examiners recognize correctly matching faces from photo-spreads under blind conditions. However, the reliability of these studies is unknown and warrants investigation since photo-spread results are based on subjective judgements of typically small single groups of examiners (<150 individuals). Moreover, statistical significance tests hold limited value for gauging reliability since these probabilities are only applicable to exactly matched study samples. To redress this issue, this study measured the repeatability of photo-spread results using three previously published facial approximations, the same photo-spread from the original study, and four independent groups of examiners (the original study group (trial 1); and three retest groups: trial 2 = c. 40 individuals; trial 3 = c. 75 individuals; and trial 4 = c. 115 individuals). Across all three facial approximations, differences in recognition rates varied from 0% to 31% between independent samples of examiners. For the nine faces that commonly received high recognition scores, the largest mean difference was 18%. This indicates that when a photo-spread size of 10 faces is used, the results generated from a sample of <115 examiners should be considered approximate, and that any differences in the recognition rates that do not exceed 18% should be considered to be negligible.

Published by Elsevier Ireland Ltd.

The accuracy with which faces can be predicted from bare skulls has been a topic of contentious debate since 1913 (see e.g., [1–8]). However, empirical tests of the method have never been more pertinent since there has been a shift away from expectations that facial approximations replicate broad facial features (i.e., those associated with "race" [1,9,10]), towards claims that facial approximations produce visages that are easily and correctly recognized as the person to whom the skull belonged [11–13].

So that accuracy tests hold maximal value it is essential that they be formulated to address the variable that is deemed to underpin an accurate facial approximation (i.e., the facial approximation's aim). For example, facial approximation accuracy could be defined as: (i) anatomical similarity to the correctly matching antemortem face; (ii) ability to generate forensic casework success; or (iii) as the ability for a facial approximation to generate purposeful and correct facial recognitions in an absence of any other cues as to the person's identity. All of the above definitions are slightly different and demand slightly different methods to properly assess them (i.e., resemblance ratings, casework success and recognition rates respectively; see [7]). Since it is commonly agreed facial approximation targets the production of a correctly recognizable face [12–18], it is recognition (or some other highly correlated factor) that must form the basis for accuracy tests [7].

In forensic casework, it is often observed that the people who report potential matches to law enforcement agencies are close acquaintances of the victim (i.e., people who were once very familiar with the person to whom the skull belongs), therefore, accuracy studies employing "familiar" examiners would be ideal [19–21]. In reality, however, the use of familiar examiners is difficult because the tests must be conducted in the blind (families members of past victims cannot be used) and because ethical issues exist if casework circumstances are simulated (i.e., unsuspecting families are mislead into thinking that one of their loved ones have passed when this has not occurred). For these reasons, facial approximation accuracy has been tested using: (i) an array of antemortem facial images (also known as face galleries, face pools, photo-spreads or lineups); and (ii) examiners who did not know the person to whom the skull belongs (i.e., were "unfamiliar" with them). Such approaches possess limitations since there are differences in the ways the familiar and unfamiliar people process faces [19–22], but without alternatives these approaches form a useful compromise. In facial approximation,

[☆] Portions of this work were presented at the 61st Annual Meeting of the American Academy of Forensic Sciences, Denver, USA, 2009.

* Corresponding author at: Joint POW/MIA Accounting Command, Central Identification Laboratory, 310 Worchester Avenue, Building 45, Hickam Air Force Base, HI 96853, United States.
E-mail address: carl.stephan.AU@jpac.pacom.mil (C.N. Stephan).

0379-0738/$ – see front matter. Published by Elsevier Ireland Ltd.
doi:10.1016/j.forsciint.2010.02.028

such arrays have been constructed using death masks [1,23], scans of faces [8], and antemortem photographs [4,6,24,25]. While each method holds limitations, antemortem photographs hold a number of advantages because they are easily and cheaply obtained, contain texture information, and usually capture the subject in an upright posture.

Face arrays may be administered in one of two fashions. Either all faces can be presented at once (=simultaneous face array or more specifically a photo-spread when photographs are used) or one face can be presented one-after-the-other in sequential order (sequential face array). Despite being more time intensive, evidence favours the second presentation mode since it discourages excessive false positive responses and forces more absolute judgements ([26–28], or for results specific to facial approximation see [24]). In facial approximation, however, photo-spreads have been the most common mode of assessment (see e.g., [4,6,13,24,25]), but their reliability is unknown. To redress this issue, this study measured the repeatability of results across multiple independent samples of examiners when the same exact facial approximation and photo-spread were used.

1. Materials and methods

To avoid copyright issues and so that a broad spectrum of facial approximation "qualities" were covered, three facial approximations previously published by Stephan and colleagues were selected from the literature for testing: facial approximation (FA) 1, see [29]; FA2, see [24]; and FA3, see [25]. These facial approximations included two that were constructed under blind conditions (one by a first time second-year science student [29] the other by a PhD student [24]); and one constructed privy to the target individual's facial appearance by the second aforementioned practitioner [25]. Of course, this last constructed face is not a true facial approximation since it was constructed under conditions that favour close representation of the target individual beyond those normally attainable in forensic casework (i.e., the practitioner was privy to the antemortem facial appearance).

The three facial approximations and their corresponding photo-spreads were printed on A4 pages (one facial approximation with its corresponding photo-spread per page) and presented to three new groups of examiners (undergraduate anatomy students) for evaluation: trial 2 = c. 40 examiners; trial 3 = c. 75 examiners; and trial 4 = c. 115 examiners (for exact sample sizes see Fig. 1). Examiners indicated their age and sex before evaluating the face pools and each examiner was requested to notify the investigators if they recognized (by name) any face/s contained in the assessments. Where any face was thought to be recognized ($n = 1$), data for those assessments were excluded from the analysis.

2. Results

Recognition responses for the three facial approximations using the three new examiner groups and their original published samples are presented in Fig. 1. Across all three facial approximations, differences in recognition rates (for any photo-spread face) varied from 0% to 31% across the trials. Fluctuations in recognition rates > 10% were common, and for the nine faces that commonly received high recognition scores (i.e., recognition rate of 20% for *at least* one trial; see FA1: face array # 2, 4, 6, and 7; FA2: face array # 1, 4, 7, and 10; FA3: face array # 4), the mean value of the largest difference between trials was 18%. Fluctuations of this magnitude were evident even for the larger samples of examiners (see FA2, trials 3 and 4, array face # 4, 7, and 10 in Fig. 1).

Since we had collected much additional data we combined the results of the retest trials with those of the original published study to yield much larger samples ($n > 250$ examiners). These data are presented in Fig. 2. Although chance rates for recognition of FA2 and FA3 are probably closer to 5% than 10% (since examiners had the option of not identifying any face), we use a chance rate of 10% to calculate statistical significance tests to increase the robustness of the conclusions. Two-by-two χ^2 tests indicated that the correctly corresponding face for FA1 was recognized at the chance rate ($\chi^2(1) = 0.00$, $p > 0.05$, power = 89%), the correctly corresponding face for FA2 was recognized above the chance rate at statistically significant levels ($\chi^2(1) = 10.89$, $p < 0.01$), and the correctly corresponding face for FA3 was recognized above the chance rate at highly significant levels ($\chi^2(1) = 199.80$, $p < 0.001$). For FA2 and FA3, no other face in the array received recognition responses that were above chance at statistically significant levels ($p < 0.05$).

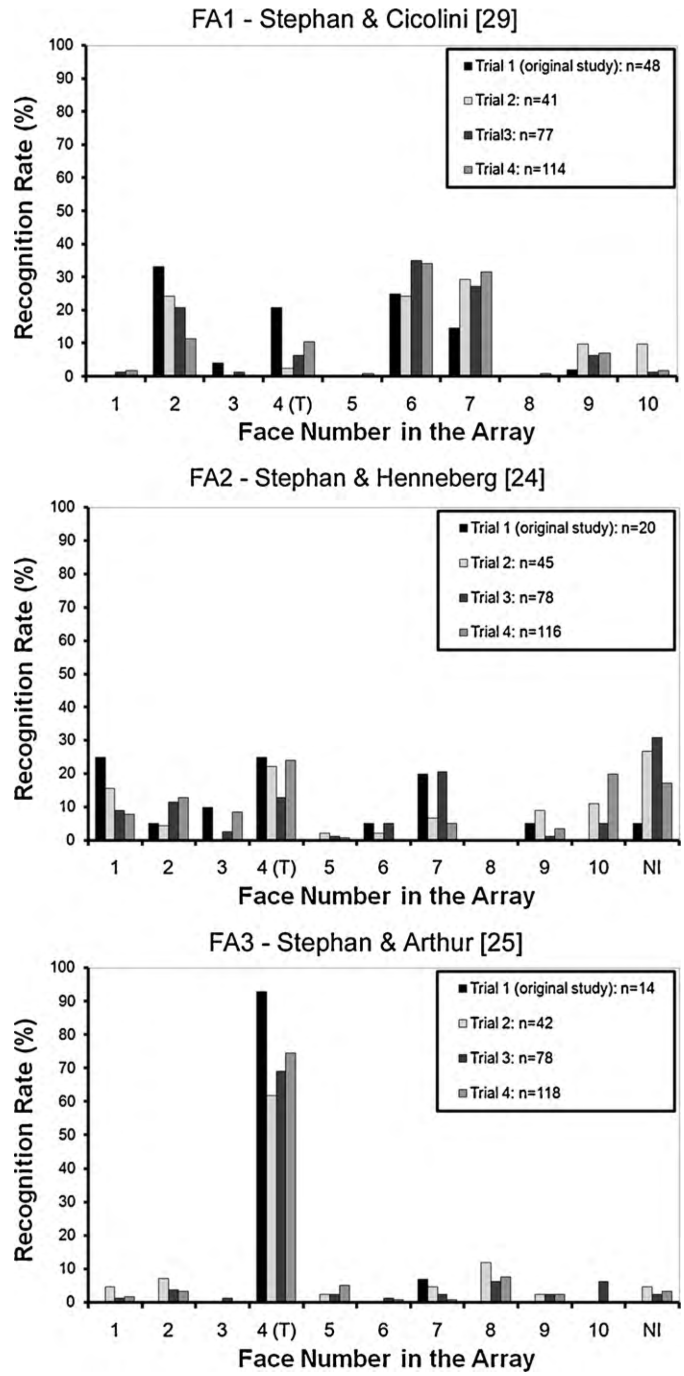

Fig. 1. Recognition rates of three facial approximations produced using their original study samples and three new samples of independent examiners. For each photo-spread the correct match is "4T": "4" = the sequence number in the array; and "T" = target (person to whom the skull belonged). "NI" stands for "no identification possible".

3. Discussion

The results of this study indicate that the recognition rates produced from any single photo-spread study that use relatively