Australian Journal of Forensic Sciences

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/tajf20

Evaluating intuitive decision-making in non-metric sex estimation from the cranium: an exploratory study

Nicole A. Mantl, Sherry Nakhaeizadeh, Rebecca Watts, Carolyn Rando & Ruth M. Morgan

To cite this article: Nicole A. Mantl, Sherry Nakhaeizadeh, Rebecca Watts, Carolyn Rando & Ruth M. Morgan (2022): Evaluating intuitive decision-making in non-metric sex estimation from the cranium: an exploratory study, Australian Journal of Forensic Sciences, DOI: 10.1080/00450618.2022.2104371

To link to this article: https://doi.org/10.1080/00450618.2022.2104371

Published online: 01 Aug 2022.

AUSTRALIAN JOURNAL OF FORENSIC SCIENCES
https://doi.org/10.1080/00450618.2022.2104371

Taylor & Francis
Taylor & Francis Group

OPEN ACCESS Check for updates

Evaluating intuitive decision-making in non-metric sex estimation from the cranium: an exploratory study

Nicole A. Mantl[a,b,c], Sherry Nakhaeizadeh[a,b], Rebecca Watts[c], Carolyn Rando[c] and Ruth M. Morgan[a,b]

[a]Department of Security and Crime Science, University College London, London, UK; [b]Centre for Forensic Sciences, University College London, London, UK; [c]Institute of Archaeology, University College London, London, UK

ABSTRACT

In recent years, forensic science has seen a rise in the number of multidisciplinary studies examining the effect of human cognition on the evaluation of forensic evidence. Notably, the House of Lords' report highlighted the need for further investigation of the consequences of cognitive factors on decision-making processes involved in evidence evaluation and interpretation. Utilizing the concepts of intuitive and methodical decision-making, this pilot study investigated this dichotomy within the field of forensic anthropology, applied to sex estimation from the skull. Participants were asked to estimate the biological sex of six crania in two experiments: once 'intuitively' under time-pressure, and once by rationally applying the Acsádi and Nemeskéri method with no time-pressure. The potential influence of experience and its correlation with the participants' confidence levels was also explored. The results demonstrate that intuitive and methodical evaluations can be consistent with each other, yet consistency decreases as ambiguity increases. Confidence was affected more by time availability, and less by level of experience. The insights from this exploratory study address how decision-making processes are involved in the examination of skeletal remains and offers justification for future exploration into the value of applying wider decision-making theories in the field of forensic anthropology.

ARTICLE HISTORY
Received 2 April 2022
Accepted 18 July 2022

KEYWORDS
Forensic anthropology; sex estimation; visual assessment; human cognition; decision-making; subjectivity

Introduction

Human cognitive reasoning and interpretation of stimuli are increasingly being considered within forensic science, as awareness has grown of the effects of subconscious factors that can impact decision-making in investigations and result in wrongful convictions[1]. The field of cognitive forensics is relevant to all stages of the forensic science process that spans judicial and analytical systems, from the development and improvement of training programmes to robust and reproducible interpretation of evidence in court[2]. The published literature addressing expertise, decision-making, and situation awareness within the social, psychological, and behavioural economics domains[3–8] offers an important insight into human judgement in practitioner performance. These studies

CONTACT Nicole A. Mantl nicole.mantl.18@ucl.ac.uk

provide a foundation for emerging cognitive research within healthcare, law enforcement, and forensic science[2,9–17]. This has led to an increased awareness of how decision-making processes can potentially impact expert evidence, and numerous government reports have been produced that highlight the importance of addressing cognitive biases in forensic science practice, evaluation, and interpretation[18–21]. Forensic anthropology is a well-established field of forensic science that offers a clear example of the importance of human expertise, and the application of both qualitative and quantitative methods to inform the evaluation of materials that are relevant to a crime reconstruction. It is a field where visual examination of relevant materials is a core practice and so human decision-making is a key component of forensic anthropology practice and evaluative conclusions. In many instances both intuitive and methodical mechanisms can be utilized to evaluate materials and make inferences and reach conclusions. The aim of this paper is to explore decision-making processes involved in forensic anthropology methods, by comparing estimations of biological sex that have been assessed through 'intuitive' and 'methodical' mechanisms whilst examining human skulls.

Experts can come to intuitive conclusions without a full and conscious awareness of their thought process or the ability to articulate reasons for their judgement[22]. Scholars of heuristics describe the act of performing an action or making a decision as carrying the result of an involuntary judgment and/or a deliberate choice[5]. The two entities exist in a dichotomy of contradicting systems, that agree or disagree depending on the nature of the task[5]. In this dichotomy, the logical answer, based on mathematical likelihood, and the intuitive prediction, based on representativeness (congruity between the outcome and the model) may from time to time disagree[3,4]. However, they are not mutually exclusive, nor is either restricted to any particular area of decision-making[5,6]. Intuitive mental association, opposed to rational and methodological information processing, is widely recognized as the natural response to external stimuli such as time-pressure, despite the acknowledged higher chances of achieving poor or incorrect results[7]. Depending on the specific task, available time, and the individual's knowledge, skill, and experience, the influence of one system on the other may vary[3–6,8]. By manipulating the response, through priming or time-pressure, it is thought possible for one system to prevail on the other. However, clearly distinguishing between the two systems becomes more difficult when the normal nature of a task negates the nature of the response being forced[6,23]. Grounding the existence of this dichotomy in the concepts of conscious and unconscious thinking is in itself a controversial practice[24]. What could be interpreted as intuition performing better than reason, may be instead be explained by an inefficient use of conscious decision-making processes[24].

It is believed that by understanding the mechanisms underlying gross and detailed viewpoints, one can improve complex cognitive tasks such as facial recognition where an individual is required to identify similarity or dissimilarity between faces[23,25,26]. A face can be processed holistically after minimal exposure (intuitively), and it is suggested that, empirically, the same applies to identifications based on individual traits[27]. It has been argued that broadening and deepening knowledge of decision-making in forensic science (beyond contextual biases) can offer a more holistic approach to improving the transparency and reproducibility of forensic interpretations[28]. Therefore, understanding how intuitive mental association, and rational and methodological information processing operate within a field could be beneficial to improve how interpretations are communicated to a lay audience in court[22].

While previous studies have focused on establishing the existence of cognitive biases and their influence in the analysis of skeletal remains[29], there is a lack of published research addressing how mental mechanisms (dichotomies like intuition and reason) contribute to the decision-making process involved, specifically in osteological identification. The skill required to assess skeletal remains presents a methodological and intuitive approach[30]. Experts possess a set of skills that develop through time which often results in the development of routines that are created during repeated exposure to skeletal remains, as well as through practical experience working with different skeletal collections[31]. The accuracy of the biological profile that is established following the analysis of skeletal remains relies not only on the method(s) employed, but also on the ability of the observer[30,32]. Arguably, forensic anthropologists, in a manner consistent with experts in other domains, rely on 'experience' to understand and decode the features that are used to estimate, for example, sex of skeletal remains[33].

The most common visual non-metric assessment for sex estimation on the skull is the method adapted from Acsádi and Nemeskéri in 1970, found in Buikstra and Ubelaker (1994) standards[34,35]. The method idealizes the morphology of the male and female skull based on the degree of expression of superficial sexually dimorphic features; the procedure scores each feature on a scale of 1–5, from least to most expression. Despite having become a standard procedure utilized by many forensic anthropologists, some limitations within the method have been highlighted within the literature. For example, previous studies considering population variation suggest that the use of ordinal scores can lead to the under or overestimation of the number of individuals of either sex within a particular group[35–37]. Trait expression scores are never fully representative of the full variation among and within populations, and the issue of misidentification of individuals has been discussed within the literature, alongside the vulnerability of these methods to cognitive factors[35,38–40].

Despite the wide availability of studies and psychological experiments on the topics of cognitive decision-making[2–4,7] there appears to be a general lack of adoption of this type of broader empirical research in forensic anthropology and forensic science in general. Integrating an understanding of cognition in to decision-making offers opportunities to gain insight into the role of experience[22]. This initial exploratory study was undertaken to investigate if previous knowledge and experience of using a specific method would generate an intuitive decision that would be similar to the interpretative conclusion reached when applying the method formally. To achieve this, an empirical experiment was designed to explore how the length of time of exposure can affect decision-making processes involved in visual non-metric sex assessment on the skull. The study assessed how time limitations can also affect observer confidence to address whether confidence varied based on the level of experience of applying the method and providing evaluative interpretations.

Methods

Materials

Six crania were selected for this study on the basis of four inclusion criteria: adult age (>18 years); no sign of pathological, taphonomic, or traumatic damage; good level of preservation of the cranium, including the facial bones; good level of preservation of sexually dimorphic traits.

All crania originated from post-medieval (17th-19th century) cemeteries in the UK and comprise part of archaeological human remains teaching collections. Due to the archaeological nature of the materials used, the identities of these crania is unknown, and the representativeness of the sexually dimorphic traits of these six specimens within the originating population can therefore not be confirmed. Crania were selected to represent the full range of expression of sexually dimorphic traits as identified by Acsádi and Nemeskéri (1970), in addition to taking into consideration time constrains to complete the study. Therefore, crania with clear male and clear female trait expressions were used, in addition to crania with intermediate trait expressions. This resulted in a total of two female, two male and two indeterminate crania (Figure 1); the suggested sex of the specimens was collected from previous records, while the scores presented in Table 1 are based on personal observations of the primary researcher.

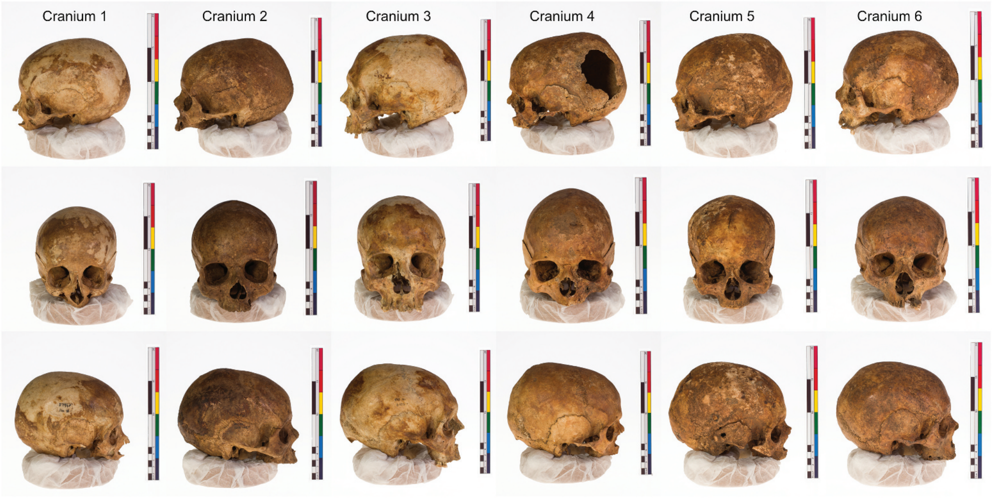

Figure 1. Images of all crania utilized in the study. Each column represents a different cranium. Each row shows one of three sides of each different cranium.

Table 1. Suitable skull specimens and descriptions. These are scores given as a general indication by the author to suggest why each skull was chosen to represent female, male or indeterminate sex.

Identifier	NC Score	MP Score	SOM Score	SOR Score	Suggested Sex
Skull 1	1	1	2	2	Female
Skull 2	2	1	2	2	Female
Skull 3	5	4	5	5	Male
Skull 4	5	4	5	4	Male
Skull 5	3	2	3	3	Indeterminate*
Skull 6	2	3	3	2	Indeterminate

NC nuchal crest; MP mastoid process; SOM supra-orbital margin; SOR supra-orbital ridge; *known male named individual

Procedure

The study was divided into two separate experiments (Phase 1, the intuitive experiment, followed by Phase 2, the methodological experiment), with two distinct time intervals set as the distinguishing factor. The two phases were carried out over a one-week interval. The one-week interval was introduced in order to give time between participants intuitive decision vs. the results of their methodical analysis to study each phase separately. The crania were placed on cork or foam bases for stability, lined up on a bench, and covered with a dark sheet. The crania were positioned in lateral view to offer a comprehensive view of the cranial features including the method used. During both phases of the experiment each cranium was uncovered and examined individually and covered again before moving on to the next one. This was done so that each cranium was examined independently from its neighbours, avoiding direct comparison. After the participant reached a decision on the sex of each cranium, they were asked to rate their self-reported confidence in their decision. This was measured on a scale of 0–100, where 0 was the least confident and 100 was the most confident. The scores were recorded on the participant answer sheet in numerical form.

The structure and purpose of the experiment was communicated to all participants in full transparency, as there was no deception involved. This study has been approved by the University College London, Institute of Archaeology Departmental Research Ethics Committee (Project ID number: 2019.009).

Phase 1: the intuitive experiment

Participants were allowed 5 s to view each of the six crania and provide a sex estimation, choosing between either male, female, or indeterminate sex. Participants were asked to do so by giving their first impression or intuitive opinion, without applying the Acsádi and Nemeskéri (1970) feature scoring method[34]. The time limit was chosen based on peer feedback, from a prior test run, which deemed 5 s to be adequate in that the pressure forced intuitive decisions to be used, and also in alignment with similar experiments within other fields[10,41].

Phase 2: the analytical experiment

In the second experiment, participants were given 30 minutes (in total) to examine all six crania by applying an adapted version of the scoring system designed by Acsádi and Nemeskéri (1970), found in Buikstra and Ubelaker (1994)[34,35]. For this experiment they were asked to score four of the five sexually dimorphic traits outlined in the method: i) the nuchal crest or external occipital protuberance; ii) the mastoid processes of the temporal bone; iii) the supra-orbital ridge or glabella; iv) the supra-orbital margin of the eye orbit. [Note that mandibles were not present for scoring.] Each of these traits was scored on a scale of 1 to 5, where 1 represents minimal expression (female) and 5 is the strongest (male). While these four traits are important for sex estimation from skulls it is important to note that they form pairs that share the same developmental pathway – muscle attachment (nuchal region, mastoid process) and extended maturation (glabella, supraorbital margin). It is therefore necessary to cognizant of how sexual dimorphism covaries within these pairs. This particular methodology was chosen because it is a common technique[37] often taught in academic courses, so all participants had at least an experience and understanding of the procedure. During this phase of the experiment participants were allowed to interact and handle the crania to simulate a real-life examination. Participants were asked to provide a final sex estimation of each cranium based on the modified methodology choosing between either: female, female?, indeterminate, male? or male.

Participants

A total of 17 participants initially volunteered to take part in the study, with 14 completing both experiments. Only data collected from the 14 participants who completed both experiments of the study was used. Using a demographic survey, it was recorded that participants were MSc (n = 10) or PhD (n = 4) students all with training and experience of using osteological non-metric sex assessment methods for the skull. Participant experience with the method was on average 3.17 years (range = 1–7 years), the median being 3.25 years. This allowed for a gross distinction between two groups (<3.25 years) and (>3.25 years). More experienced individuals should not be mistaken for professional practitioners as it is related to the experience within the participant group. Seven participants had <3.25 years of experience and seven had ≥3.25 years of experience. All participants volunteered to participate and gave their informed consent

Analysis

Statistical testing was carried out using SPSS (IBM). Due to the unequal variables in Phase 1 and Phase 2, for statistical tests that involve a direct comparison (Cohen's Kappa – for intra-observer error – and Chi-Square test) of intuitive and methodical choices the probable female? and male? variables were combined with their corresponding extreme, female and male respectively. A series of Cohen's weighted Kappa test were carried out to determine each participant's intra-observer agreement for the direct comparison of their intuitive and methodical choices made for each cranium. Another set of Cohen's weighted Kappa tests were also run to determine interobserver agreement rates for all participants by comparing them against each other in pairs. Two separate sets of results were produced for Phase 1 and Phase 2 to better distinguish the efficiency of either intuitive or methodological procedures. Further, each dataset was averaged arithmetically to calculate Light's Kappa and Fleiss' Kappa to assess the overall agreement of all participants in each Phase of the experiment.

Mixed ANOVA tests (95% confidence interval) for each cranium were computed analysing the mean differences of sex estimations to investigate the influence of the different time conditions (within-participant variable) and experience groups (between-participant variable) on each participant's answers. Ultimately, the statistical relationship resulting from the cross-classification of time and experience was evaluated to determine whether experience was a decerning factor within each experimental phase.

To evaluate the overall consistency between the sex estimations given for each observed cranium in the intuitive experiment and those given in the methodical experiment, Chi-Square tests were run at 95% confidence intervals. To assess the effect of time variation and experience on participants' confidence levels, paired and independent student t-tests were run at 95% confidence intervals.

Results

Intra-observer and inter-observer agreement rate

Each participant's consistency between Phase 1 and Phase 2 was assessed using Cohen's weighted Kappa and the results are shown in Table 2.

To establish interobserver agreement rates among participants, Cohen's weighted Kappa was utilized (Table 3). This test is useful in identify any noise in the data caused by participants that had either perfect agreement or poor agreement with the rest of the group. The test considers all assessments made in either Phase 1 or Phase 2 by each individual participant and compares them to the answers given by every single other participant in the group for that specific experiment. Light's Kappa (average weighted Kappa) indicated that in the intuitive experiment, agreement between all participants was substantial (0.643), and in the methodical experiment, agreement was moderate (0.575). Fleiss' Kappa shows moderate (0.520) and slight (0.322) agreement in the intuitive and methodical estimations, respectively.

Table 3 summarizes the frequency with which agreement rates fell within each of the significance categories determined by the Cohen's Kappa test (Poor, 0.001–0.200; Slight, 0.201–0.400; Moderate, 0.401–0.600; Substantial, 0.601–0.800; Almost perfect, 0.801–0.999; Perfect, 1.000). The results indicate that a large proportion of participants in both experiments presented substantial agreement, suggesting that the majority of pairs agreed on the sex estimation given for at least three crania (Phase 1: n = 34/91; Phase 2: n = 35/91 – See Table 4). Perfect agreement was only achieved between three participants in Phase 1 – the intuitive experiment.

Table 2. Results for Cohen's weighted Kappa assessing each participant's intra-observer error between Phase 1 and Phase 2.

Participant	Kappa	Confidence Interval (95%)
#1	0.800	0.491 | 1.109
#2	0.250	−0.311 | 0.811
#3	0.333	−0.198 | 0.864
#4	0.647	0.265 | 1.029
#5	0.833	0.544 | 1.122
#6	0.471	0.042 | 0.899
#7	0.500	0.117 | 0.883
#8	0.625	0.201 | 1.049
#9	1.000	1.000 | 1.000
#10	0.800	0.449 | 1.151
#11	0.625	0.206 | 1.049
#12	0.625	0.216 | 1.034
#13	0.438	−0.010 | 0.885
#14	0.667	0.346 | 0.987

Table 3. Frequency of significance level for Cohen's weighted Kappa test pairings classified by significance level.

Significance	Value range		# Of pairings in Phase 1	# Of pairings in Phase 2
Poor	0.001	0.200	0 (0.0%)	3 (3.3%)
Fair	0.201	0.400	9 (9.9%)	16 (17.6%)
Moderate	0.401	0.600	23 (25.3%)	30 (33.0%)
Substantial	0.601	0.800	34 (37.4%)	35 (38.5%)
Almost Perfect	0.801	0.999	19 (20.9%)	7 (7.7%)
Perfect	1.000		6 (6.6%)	0 (0.0%)

Table 4. Sex estimation results for intuitive (5 seconds) and methodical (30 minutes) experiments.

	Intuitive			Methodical				
Specimen	F	I	M	F	F?	I	M?	M
Cranium 1	14			12*	2†			
Cranium 2	9	5		2*+2§	3†+1((	2*+3‡	1†	
Cranium 3			14				1†	13*
Cranium 4		2	12			3‡	4†+2((	5*
Cranium 5	4	8	2	1*	1†+2((	2*+2‡	2†+4((	
Cranium 6	6	8		1§	3†+4((	2*+3‡	1((	

*Remained consistent; †Determinate to probable sex; ‡Determinate to indeterminate sex; §indeterminate to determinate sex; ((indeterminate to probable sex.

Sex estimation

All results from the sex estimation of the six crania are summarized in Table 4.

Cranium 1 (*female*). An individual presenting traits consistent with what can be described as an 'obvious' female. The results reflect the clarity of the four sexual traits (see Figure 1 for reference) as the sex estimation remains consistently female (or female?) among all 14 participants independently from the time of exposure. In Phase 1 all 14 participants scored this cranium as female. In Phase 2, 12 participants scored it as female with two participants scoring it as female?. In each experience group, less or more experienced participants, six observers (85.7%; n = 6/7) stayed fully consistent across experiments and only one (14.3%; n = 1/7) who scored it female? instead of female in Phase 2.

Cranium 2 (*female*). An individual presenting traits consistent with a female individual. In Phase 1, nine participants scored this cranium as female and five participants scored it as indeterminate. In Phase 2, four scored the cranium as female, four as female?, five as indeterminate and one as male?. Only four participants (28.6%) remained consistent in their assessment across phases, two who scored the cranium as female in both phases, and two who scored it as indeterminate in both phases. In each experience group, less or more experienced participants, two observers (28.6%; n = 2/7) stayed fully consistent across experiments and five (71.4%; n = 5/7) who scored it female? instead of female in Phase 2.

Cranium 3 (male). An individual consistent with an extreme male trait expression. The results reflect the clarity of the four sexual traits (see Figure 1 for reference). Similarly to Cranium 1, the results suggest that independently from the time of exposure the sex estimations remain consistently male or male? among all participants. In Phase 1 all 14 participants scored this cranium as male. In Phase 2, 13 participants (92.9%) remained consistent with an estimation of male, with only one participant scoring the cranium as male? Across both experiments all the less experienced participants remained consistent in their answers, whereas one (14.3%; n = 1/7) of the more experienced participants scored it male? instead of male in Phase 2.

Cranium 4 (male). Specimen presenting traits consistent with a male individual. In Phase 1, 12 participants scored this cranium as male and two as indeterminate. In Phase 2, five participants scored the cranium as male, six as male? and three scored the cranium as indeterminate. Only the five participants (35.7%; n = 5/14) who scored the cranium as male remained consistent in their assessment across phases. Three (42.9%; n = 3/7) of the less experienced participant remained consistent in their answers across experiments, and two (28.6%; n = 2/7) chose male? instead of male in Phase 2. In comparison, only one (14.3%; n = 1/7) of the more experienced participants remained consistent across experiments along two (28.6%; n = 2/7) others who answered male? instead of male in Phase 2.

Cranium 5 (ambiguous). Based on church records this cranium was a known male individual with indeterminate sex features. In Phase 1, four participants scored the cranium as female, eight as indeterminate and two as male. In Phase 2, only one participant scored the cranium as female, three chose female?, four indeterminate and six scored it as male?. Within the sex estimations given for this cranium only three participants (21.4%; n = 3/14) remained consistent with their initial estimation. Of these, two (66.6%; n = 2/3) remained indeterminate, and one (33.7%; n = 1/3) female. Both participants who scored the cranium as male in Phase 1 later scored it as male? in Phase 2. Two (28.6%; n = 2/7) less and two (28.6%; n = 2/7) more experienced participants remained consistent across experiments; in Phase 2 a more experienced observer answered male? instead of male, and another answered female? instead of female.

Cranium 6 (ambiguous). This cranium had indeterminate morphological features. In Phase 1, six participants scored the cranium as female and eight as indeterminate. In Phase 2, only one participant scored the cranium as female, seven answered female?, five indeterminate and one scored it as male?. Only two (14.3%; n = 2/14) participants (one more and one less experienced) remained consistent with their first estimation: both identified the cranium as indeterminate. Another less experienced participant (14.3%; n = 1/7) remained marginally consistent across experiments answering female? instead of female in Phase 2. Similarly, two (28.6%; n = 2/7) of the more experienced participants answered female? instead of female in Phase 2.

Mixed ANOVA and chi-square test on time and sex estimation

In the assessment of time as a within-participant variable, the mixed ANOVA test (Table 5) shows that the mean intuitive and methodical sex estimations were only statistically different in the case of Cranium 4 (4.714 and 4.143, respectively, Table 6), where p = 0.01. The test reported that a distinction of sex estimations based on the participant experience group as a between-participant variable was not observed. However, when the time conditions of the intuitive and rational experiments are considered as a sub-classification of the experience group (Table 7), a significant distinction can be observed between participants' sex estimations for Cranium 4 (less experienced: 4.429 [5 seconds], 4.571 [30 minutes]; more experienced: 5.000 [5 seconds], 3.714 [30 minutes]), where p = 0.00, and Cranium 6 (less experienced: 1.857 [5 seconds], 2.857 [30 minutes]; more experienced: 2.429 [5 seconds], 2.000 [30 minutes]), where p = 0.04.

Chi-square tests (Table 8) indicated that there was no statistically significant difference (p-value<0.05) between the sex estimations made in 5 seconds (intuitively) and those made in 30 minutes (methodically). The Chi-square value could not be calculated for Cranium 1 and Cranium 3 as all the sex estimations in Phase 1 were identical for all participants.

Table 5. Mixed Anova computation for sex estimation of each cranium considering time and the time*experience relation as within-subject variables and experience as a between-subjects variable. The time variable distinguishes sex estimations given intuitively (5 seconds) or rationally (30 minutes). The experience variable divides the participants into less (<3.25 years) and more (>3.25 years) experienced observers. The time*experience relation investigates if there is any significant difference between experience groups over time.

	Tests of Within-Subjects Effects				Tests of Between-Subjects Effects	
	Time		Time * Experience		Experience	
Specimen	F	Sig.	F	Sig.	F	Sig.
Cranium 1	2.00	0.18	0.00	1.00	0.00	1.00
Cranium 2	1.45	0.25	0.27	0.61	0.04	0.84
Cranium 3	1.00	0.34	1.00	0.34	1.00	0.34
Cranium 4	8.73	0.01	13.64	0.00	0.22	0.65
Cranium 5	1.47	0.25	0.53	0.48	0.14	0.72
Cranium 6	0.87	0.37	5.45	0.04	0.18	0.68

Table 6. Tabulation of estimated marginal means of sex estimations over time. This is an overview of how intuitive (5 seconds) and methodical (30 minutes) sex estimations varied for the whole group of participants.

Specimen	Time	Mean[a]	Std. Error	95% Confidence Interval
Cranium 1	5 Seconds	1.000	0.000	1.000–1.000
	30 Minutes	1.143	0.101	0.923–1.363
Cranium 2	5 Seconds	1.714	0.274	1.118–2.310
	30 Minutes	2.214	0.270	1.625–2.803
Cranium 3	5 Seconds	5.000	0.000	5.000–5.000
	30 Minutes	4.929	0.071	4.773–5.084
Cranium 4	5 Seconds	4.714	0.184	4.312–5.116
	30 Minutes	4.143	0.175	3.762–4.524
Cranium 5	5 Seconds	2.714	0.369	1.911–3.518
	30 Minutes	3.071	0.270	2.482–3.661
Cranium 6	5 Seconds	2.143	0.274	1.547–2.739
	30 Minutes	2.429	0.170	2.058–2.799

[a]Mean calculated as the average score if: female = 1; female? = 2; indeterminate = 3; male? = 4; male = 5.

Paired sample t-test on confidence assessment and experience

Confidence levels were assessed using a paired sample t-test to determine any difference between the phases. Participants were on average, 9.49% more confident in their assessment in Phase 2, a modest difference but one that reached the threshold for statistical significance (p = 0.045) (Table 9). The correlation between confidence levels of the two phases is not consistent nor proportional and does not reach 50%.

When confidence was examined in conjunction with level of experience (Table 10), the more experienced group were more confident in their sex estimations in both phases of the experiment (Phase 1 = 80.00 vs. 66.43; Phase 2 = 83.25 vs. 76.79). In Phase 2, confidence was highest for both groups when trait expression was strongest (Cranium 3 (more experienced = 90.43; less experienced = 90.71) and Cranium 1 (more experienced = 87.43; less experienced = 83.57). Ambiguous crania were challenging, especially for the less experienced observers: participants in this group reported confidence of 66.43 and 65.00 out of 100 for their assessments of Cranium 5 and 6, respectively. The more experienced participants recorded greater confidence in their examination of Cranium 6 with an average confidence of 83.70, while Cranium 5 recorded the lowest average at 77.14.

Table 7. Tabulation of estimated marginal means of sex estimations over time per experience group. This is an overview of how intuitive (5 seconds) and methodical (30 minutes) sex estimations varied for either the less (<3.25 years) or more (>3.25 years) experienced participant group.

Specimen	Experience	Time	Mean[a]	Std. Error	95% C.I.
Cranium 1	Less experienced	5 Seconds	1.000	0.000	1.000–1.000
		30 Minutes	1.143	0.143	0.832–1.454
	More experienced	5 Seconds	1.000	0.000	1.000–1.000
		30 Minutes	1.143	0.143	0.832–1.454
Cranium 2	Less experienced	5 Seconds	1.857	0.387	1.014–2.700
		30 Minutes	2.143	0.382	1.310–2.976
	More experienced	5 Seconds	1.571	0.387	0.729–2.414
		30 Minutes	2.286	0.382	1.452–3.119
Cranium 3	Less experienced	5 Seconds	5.000	0.000	5.000–5.000
		30 Minutes	5.000	0.101	4.780–5.220
	More experienced	5 Seconds	5.000	0.000	5.000–5.000
		30 Minutes	4.857	0.101	4.637–5.077
Cranium 4	Less experienced	5 Seconds	4.429	0.261	3.860–4.997
		30 Minutes	4.571	0.247	4.032–5.111
	More experienced	5 Seconds	5.000	0.261	4.432–5.568
		30 Minutes	3.714	0.247	3.175–4.253
Cranium 5	Less experienced	5 Seconds	2.714	0.522	1.578–3.851
		30 Minutes	3.286	0.382	2.452–4.119
	More experienced	5 Seconds	2.714	0.522	1.578–3.851
		30 Minutes	2.857	0.382	2.024–3.690
Cranium 6	Less experienced	5 Seconds	1.857	0.387	1.014–2.700
		30 Minutes	2.857	0.240	2.333–3.381
	More experienced	5 Seconds	2.429	0.387	1.586–3.271
		30 Minutes	2.000	0.240	1.476–2.524

[a]Mean calculated as the average score if: female = 1; female? = 2; indeterminate = 3; male? = 4; male = 5.

Table 8. Chi-square test results for sex estimations of each skull. The table shows the Pearson's chi-square, likelihood ratio and Fisher's exact values and their respective significance values.

	Pearson's Chi-square Test			Likelihood Ratio			Fisher's Exact Test	
Specimen	Value	df	Asymptotic Sig.[b]	Value	df	Asymptotic Sig.[b]	Value	Exact Sig.
Cranium 1	0.000[a]			0.000[a]			0.000[a]	
Cranium 2	0.607	2	0.738	0.934	2	0.627	0.765	1.000[b]
Cranium 3	0.000[a]			0.000[a]			0.000[a]	
Cranium 4	0.636	1	0.425	1.052	1	0.305	n/a	1.000[b]/0.604[c]
Cranium 5	5.833	4	0.212	8.031	4	0.090	5.162	0.253[b]
Cranium 6	1.444	2	0.486	1.806	2	0.405	1.447	0.767[b]

[a]No statistics are computed because sex estimation results in Phase 1 are constant
[b]2-sided [c] 1-sided

Table 9. Paired sample t-test results for confidence levels reported for all crania in the intuitive (5 seconds) and methodical (30 minutes) experiments.

	Paired Sample Statistic				Paired Differences		
Time	Mean	N	Std. Dev.	Std. Error Mean	t	df	Sig. (2-tailed)
5 Sec	73.21	14	15.764	7.33	−2.223	13	0.045
30 Min	80.15	14	7.741	2.069			

Table 10. Group statistics for confidence levels and independent t-test results for all crania. Participants are divided into their relative groups of experience for the group statistics.

Group statistics					
Time	Experience (years)	N	Mean	Std. Dev.	Std. Error Mean
5 Sec	≥ 3.25	7	80.00	10.41	3.93
	< 3.25	7	66.43	17.96	3.79
30 Min	≥ 3.25	7	83.52	7.42	2.81
	< 3.25	7	76.79	6.95	2.63
	Levene's Test for Equality of Variances			t-test for Equality of Means	
Time	F	Sig.	t	df	Sig. (2-tailed)
5 Sec	1.866	0.197	1.730	12	0.109
30 Min	0.025	0.877	1.753	12	0.105

Table 11. Linear correlation between experience and confidence highlighting the trend and correlation coefficient (R^2) calculated for the two variables either for all crania or per individual.

	Trend	Correlation R^2
All crania (Phase 1)	2.36 (+)	0.0966
All crania (Phase 2)	1.27 (+)	0.1147
Cranium 1 (Phase 2)	0.68 (+)	0.0205
Cranium 2 (Phase 2)	0.87 (+)	0.0359
Cranium 3 (Phase 2)	−0.55 (-)	0.0272
Cranium 4 (Phase 2)	−0.80 (-)	0.0211
Cranium 5 (Phase 2)	2.70 (+)	0.2281
Cranium 6 (Phase 2)	4.69 (+)	0.4561

The relationship between experience and confidence was determined to be positive, yet weak, approaching zero in both Phase 1 (R^2 = 0.097) and in Phase 2 (R^2 = 0.115) (Table 11). The same is observed for each individual cranium in Phase 2. Correlation is very weak, positively for Cranium 1–2 (R^2 < 0.04) and negatively for Cranium 3–4 (R^2 < 0.03); Cranium 5–6 show a positive and slightly stronger (0.2< R^2 < 0.5) correlation in comparison, yet a weak correlation, nonetheless.

Discussion

Krogman and Işcan (1986) state that 'in sexing the skull the initial impression often is the deciding factor [. . .]' (p.144)[42].

The results of this preliminary study show that when sexually dimorphic trait expression is at its extreme (either male or female), sex estimation is straightforward and explicit, leading to higher chances of consistency between intuitive, time-sensitive responses and outcomes being reached, following the application of a systematic methodology. As the strength of trait expression regresses to a mean, consistency drops considerably.

In regard to Cranium 1 and 3 (respectively, the female and male extremes), the background and experience of the participant for this study does not appear to be an influencing factor, and unlikely to have a significant impact on the outcome, as all sex estimations reported for these two specimens were the same in both phase 1 and phase 2. Identifications for Craniums 2 and 4 were more varied, despite the unambiguous morphology of the two crania: only 50.0% (Cranium 2) and 64.3% (Cranium 4) of all participants remained consistent between the two phases. For both the ambiguous cases, less than half of participants (42.9% and 35.7%, for Cranium 5 and 6 respectively) remained consistent in their estimations (female/female?, indeterminate or male/male?), with agreement being lowest when trait expression was most unclear. It is observed that the more experienced individuals only accounted for 57.1% and 44.4% of consistent identifications for Cranium 2 and Cranium 4 respectively, and 66.7% and 60.0% for the ambiguous specimens (Cranium 5 and 6 respectively). One would expect the participants with more experience to be more likely to give the same answer in both phases, as studies suggest that information encoding, and processing abilities are gained and improved with increasing levels of experience[43]. The results for Craniums 5 and 6 suggest that experience is, not always conducive to increased rates of consistency, only becoming moderately relevant when observers are presented with dubious or inconclusive traits.

It is noted that on certain occasions when the method yielded an inconclusive result (ambiguous identification), participants preferred to report a probable sex estimation (either female? or male?). The participants were unable to consciously elaborate on their decision to bypass the method in favour of what appears to be an intuition; this suggests that given sufficient time to form an empirical and rule-based decision, the difficulty of the task overshadows the normal functioning of the intuitive (System 1) and rational (System 2) decision-making dichotomy[5]. The involuntary, and therefore arguably 'unconscious', path that brought participants to their conclusion is consistent with the failure of the structural processing of System 2 in favour of associative thinking. This reversal of roles is often observed in highly complex tasks, such as facial recognition[26]. While the decision appears to be unconscious, the experiment was not carried out under conditions that allowed the awareness of the decision-maker to be determined[24]. The distinction between intuitive and reasoned thinking may be 'consciously mediated', yet as a consequence of the examiners' lack of necessary insight or practical language to explain their thinking process may appear unconscious[23,24].

The efficiency of intuitive judgements

The results from this pilot study appear to disagree to some extent with previously published studies supporting the ability of the brain to adapt to time limits by increasing the speed of information processing in response to a short deadline[44]. The results reported in this study indicate that there is not sufficient evidence to show the existence of 'biasing' consequences of time-pressure (as a proxy to distinguish intuition from rational behaviour) on an observer's ability to estimate the sex of a cranium consistently (Table 9). Chi-Square tests of two of the crania (Cranium 1 and 3) showed almost perfect agreement among participants in both experiments. It should be noted that the necessary combining of probable and extreme estimations throughout statistical testing, due to the unequal number of variables between Phase 1 and Phase 2, may have skewed the test results away from a statistically significant outcome.

Overall, the results reported in this preliminary study by the chi-square test suggest that time has little to no influence on the sex estimations provided by participants. However, these results may also be a reflection of the limitations of this study. First, the time limit of 5 seconds may not have been short enough to induce genuine intuitive answers[31], instead allowing the participants sufficient time to still apply to some extent the modified Acsadi and Nemeskeri (1970) sexing method from memory [8]. Secondly, the small sample size may have been too small for a clear distinction to emerge, and additionally it only reflects the behaviour of students, which may differ from expert practitioners' cognitive processes. Similarly, the small difference in level of experience of the methods between the two groups may also have impacted on the results, due to there not being a large enough gap between the two groups. Finally, having only three options in Phase 1 and five in Phase 2 did not allow for a balanced measurement of consistency due to the uneven choice categories.

The premise of this experiment was the hypothesis that it would be possible to identify a separation of intuitive association from mnemonic and logical reasoning based on generalized vs. feature-by-feature identification. When guiding the participants through the task, participants were observed to be attempting to view every sexually dimorphic trait, as if applying the method, in spite of being instructed not to do so. Due to the hierarchical nature of the method, sex estimation may not necessarily be suited to the gross processing that may be expected from intuitive decision-making[45], especially when the final aim is to learn how to arguably prevent bias in decision-making. Ultimately, identification of the sex of a skull solely through comparison with the typical male or female cranium will reduce performance and increase the threshold of uncertainty[25]. Such a grossly intuitive approach may however be the only course of action when assessing exceptionally ambiguous crania. The ability to relate ambiguous features to sex estimation, even if only probable, appears to imply that this conclusion is reached intuitively and not by weighing each feature 'equally' in its individuality, rather through holistic elaboration, much like faces[25]. Arguably, the visual scope becomes too narrow, and the human mind attempts to come to terms with an inconclusive result by switching to a more inclusive strategy, going against its normal behaviour[25]. Ultimately, when the incoming information is ambiguous, it is difficult to determine a clear distinction between the two processing strategies.

Implications of confidence levels

The results of this study reported that when trait expression is strongest (Cranium 1 and Cranium 3), self-reported confidence is highest, and as expression of the traits regressed to the mean (score of 3) confidence decreased accordingly. When the time-condition is considered a contributing factor, a difference is observed between confidence reported across the two experiments of the study. A substantial increase was observed in the average confidence level reported after the 30-minute examination, compared to the values reported after 5 seconds. The increase by 10.36 recorded by the less experienced group after the completion of both experiments reduced the overall disparity with the more experienced group's confidence by 50.3%. The increase in confidence with time is not a strong or consistently proportional correlation. The results indicate that ten of the 14 (71.4%) participants reported an increase in confidence between experiments.

The second factor considered to affect confidence in participants was their experience[8]. Confidence is implicitly believed to increase with experience[46]. The highest and lowest confidence values reported by a less experienced participant for an individual cranium in Phase 2 were 100 and 40 respectively, although the latter was the only occurrence of a reported score below 60. The highest and lowest confidence ratings reported by a more experienced observer for an individual cranium in Phase 2 were 98 and 60. The substantial overlap between the confidence levels reported by the two groups does not allow for comparison through mean values or the statistical identification of definite cut-off points. Statistical testing revealed that these values are not strong enough to show significant changes before and after time-pressure is removed, albeit a difference of 13.57 points in Phase 1 and 6.73 points in Phase 2, yet this may simply be an artefact of the small participant group size. Time-pressure remains a relevant, although not consequential, factor affecting confidence levels. Linear correlation computations for both phases support the previous results of the statistical testing. Confidence and experience correlations for each cranium indicate that each participant's estimation of the correctness of their conclusions was random. In this case, the difficulty of the task may once again be the only diverging factor in this assessment.

Overall, participants with less than 3 years of experience with the method were much less confident than more experienced examiners in their assessments, especially when assessing more ambiguous exhibits, showing similar results to previously published studies in cognitive science[47].

Decision-making and expertise

The susceptibility of decision-making in forensic science contexts to cognitive biases has been increasingly recognized and studied, and the need for a wider understanding of processes involved in decision-making has been extensively advocated by researchers and governments alike[12–21]. Within the broader context of cognitive research in forensic anthropology, this pilot study addresses a gap in the published literature to date, as studies on cognitive reasoning involved in decision-making and judgement are necessary to explain how observers make decisions and what potential pitfalls may arise during the analysis of skeletal remains[40,48,49]. Moreover, this study suggests that perhaps both the problem and the solution lie in how we are taught to think about the analytical method utilized, both in theory and in practice[18–21]. Some research has shown that cognitive biases affect experts in different fields of forensic science and that experts often underestimate their susceptibility to arguably unconscious bias outcomes[12,50]. In some cases, bias has not been acknowledged as an issue, some experts have purported to see it in others but believe that they are not susceptible, and others are aware of their own susceptibility but believe that they are able to control it[51]. Nevertheless, all decision-makers are susceptible to a range of extrinsic and intrinsic factors that can lead to biased outcomes, from supervision-related stress and pressure to reach expected conclusions to the impact of extraneous information[52]; however, not all sources of bias are related to human nature or the work environment[53].

Raising awareness and implementing new control practices do not address the underlying processes that lead to cognitive bias, and its impacts[12,53]. Notably in the UK, the House of Lords Science and Technology Select Committee (2019) report from their inquiry into forensic science addresses the concerns of many forensic science practitioners and researchers regarding the need for achieving a 'better understanding of the cognitive process of pattern recognition, the psychological nature of 'expertise', as well as sources, causes, and consequences of cognitive bias (p.42)[20]. It is a challenge to academic, law enforcement, and specialized professional organizations to address the limitations within forensic science decision-making to create a more reproducible, transparent, and empirically evidence-based reconstruction process[31,54,55]. Indeed, by articulating the mental mechanisms involved in decision-making, through the application of cognitive reasoning studies, it is possible to appreciate more fully the complexity of subjective interpretation of forensic science materials such as skeletal material[51].

The 'expertise' involved in the osteological decision-making process stems from the development of both knowledge and practical skill[31,54]. This initial study suggests that it is arguably possible for forensic anthropologists to develop their observation skills to form an informed automated mental mechanism, drawing on 'rational intuition'. This study did not support the hypothesis that experience with a particular methodology has more influence on the overall ability of the analyst to consistently identify skeletal remains, however more research is necessary to understand this relationship fully. Expanding this experiment on a greater scale, by encompassing a wider range of experience levels including professional, and inexperienced observers from different backgrounds, will offer additional insight addressing the influence of experience on decision-making and confidence[8,45,46]. The findings from this initial study indicate that reducing the time allowed for intuitive judgements, to 1–2 seconds, may increase the chances that observers will rely solely on intuitive association, to prevent them from trying to speed up their reasoning to adapt to the method they are accustomed to using.

Empirical research addressing the use of cognitive processes will increase the understanding of mental mechanisms used in information processing and decision-making involved in the identification of skeletal remains of forensic significance. This study focuses only on one of the many branches of cognitive psychology but has highlighted the application of decision-making theories beyond cognitive biases to forensic anthropology, in order to further our understandings of the cognitive mechanisms involved in sex estimations of the skull. Like studies in facial recognition[27], looking at different processing mechanisms of information could contribute to a more holistic understanding of the mechanisms involved in the interpretations of skeletal remains.

Conclusion

This initial study aimed to further the understanding of the decision-making process and determine affecting factors that may influence judgement within the field of forensic anthropology. The results suggest that intuitive and methodical judgements are distinct and only carried over from one phase to another when features are clearest. Consistency was largely unaffected by the experience of the observer, whereas confidence increased when time limitations were removed. The results of this experiment also highlight that consistency and self-reported confidence levels varied to a greater degree according to the ambiguity of the cranial features, rather than to observers' experience and time limitations. It is possible to argue that ambiguous specimens were complex and difficult to estimate sex for any observer in this study, whether more or less experienced within the group, and both can utilize intuition if depending on the threshold of risk. These initial findings may be helpful in order to expand on research that will aid in a better understanding of how performance is measured within forensic anthropology specifically, and forensic science more broadly where experts are required to make perceptual judgements of similarity.

Stress, experience and confidence are variables that could benefit from being considered through the lens of decision-making theories in order to better understand their relationship with the accuracy, transparency and reproducibility of evaluative interpretations. Ultimately, this preliminary study found that, in the estimation of sex from the skull, intuition rarely persists in rational judgement beyond what fits extreme representation, and that the difficulty of a task affects judgement more than experience and time-pressure factors. A potentially profitable next step would be to investigate the parallels between the examination of skulls and facial recognition in order to identify more insights into a distinction between intuition and reason.

Acknowledgments

The authors would like to thank all the participants that contributed to this research by donating their personal time. Finally we would like to acknowledge helpful and insightful anonymous reviewer comments and recommendations.

Disclosure statement

No potential conflict of interest was reported by the author(s).

ORCID

Nicole A. Mantl http://orcid.org/0000-0003-3097-591X
Ruth M. Morgan http://orcid.org/0000-0002-4146-654X

References

1. Innocence project: policy reform [online]. New York: Innocence Project; 2019 [cited 2019 July 29]. Available from: https://www.innocenceproject.org/policy/
2. Dror IE. Cognitive neuroscience in forensic science: understanding and utilizing the human element. Philos Trans of the Royal Society B: Biol Sci. 2015;370(1674):20140255. doi:10.1098/rstb.2014.0255.
3. Tversky A, Kahneman D. Judgment under uncertainty: heuristics and biases. Sci. 1974;185 (4157):1124–1131. doi:10.1126/science.185.4157.1124.
4. Tversky A, Kahneman D. Extensional versus intuitive reasoning: the conjunction fallacy in probability judgment. Psychol Rev. 1983;90(4):293–315. doi:10.1037/0033-295X.90.4.293.
5. Frederick S. Automated choice heuristics. In: Gilowich T, Kahneman D, Griffin DW, editors. Heuristics and biases. Cambridge: Cambridge University Press; 2012. p. 548–558.
6. Sloman SA. Two systems of reasoning. In: Gilowich T, Kahneman D, Griffin DW, editors. Heuristics and biases. Cambridge: Cambridge University Press; 2012. p. 379–396.
7. Devine PG, Sherman SJ. Intuitive versus rational judgment and the role of stereotyping in the human condition: kirk or spock? Psychol Inq. 1992;3(2):153–159. doi:10.1207/s15327965pli0302_13.
8. Kahneman D, and Frederick S. Representativeness revisited: attribute substitution in intuitive judgment. In: Gilowich T, Kahneman D, Griffin DW, editors. Heuristics and biases. Cambridge: Cambridge University Press; 2012. p. 49–81.
9. Rabin M, Schrag JL. First impressions matter: a model of confirmatory bias. Q J Econ. 1999;114 (1):37–82. doi:10.1162/003355399555945.

10. Trueblood JS, Holmes WR, Seegmiller AC, Douds J, Compton M, Woodruff M, Huang W, Stratton C, Eichbaum Q. The impact of speed and bias on the cognitive processes of experts and novices in medical image decision-making. Cogn Res: Princ and Implic. 2018;3(28). doi:10.1186/s41235-018-0119-2.
11. ALQahtani DA, Rotgans JI, Ahmed NE, Alalwan IA, Magzoub MEM. The influence of time pressure and case complexity on physicians' diagnostic performance. Health Prof Educ. 2016;2(2):99–105. doi:10.1016/j.hpe.2016.01.006.
12. Dror IE, Cole SA. The vision in "blind" justice: expert perception, judgment, and visual cognition in forensic pattern recognition. Psychon Bull and Rev. 2010;17(2):161–167. doi:10.3758/PBR.17.2.161.
13. Dror IE. Cognitive forensics and experimental research about bias in forensic casework. Sci and Justice. 2012;52(2):128–130. doi:10.1016/j.scijus.2012.03.006.
14. Charlton D, Fraser-Mackenzie PAF, Dror IE. Emotional experiences and motivating factors associated with fingerprint analysis. J Forensic Sci. 2010;55(2):385–393. doi:10.1111/j.1556-4029.2009.01295.x.
15. Hamnett HJ, Dror IE. The effect of contextual information on decision-making in forensic toxicology. Forensic Sci Int: Synergy. 2;2020:339–348.
16. Sunde N, Dror IE. A hierarchy of expert performance (HEP) applied to digital forensics: reliability and biasability in digital forensics decision-making. Forensic Sci Int: Digit Investig. 37;2021:301175.
17. Dror IE, Melinek J, Arden JL, Kukucka J, Hawkins S, Carter J, Atherton DS. Cognitive bias in forensic pathology decisions. J Forensic Sci. 2021;66(5):1751–1757. doi:10.1111/1556-4029.14697.
18. National Academy of Sciences (NAS). Strengthening forensic science in the United States: a path forward. Washington D.C: The National Academies Press; 2006.
19. Report to the president: forensic science in criminal courts: ensuring scientific validity of feature comparison methods. Washington D.C: Executive Office of the President's Council of Advisors on Science and Technology; 2016 [cited 2022 Jan 17]. Available from: https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/PCAST/pcast_forensic_science_report_final.pdf
20. House of Lords Science and Technology Select Committee. Forensic science and the criminal justice system: a blueprint for change. London: The Authority of the House of Lords; 2019 [cited 2022 Jan 17]. Available from: https://publications.parliament.uk/pa/ld201719/ldselect/ldsctech/333/333.pdf
21. Pollanen MS, Bowes MJ, VanLaerhoven SL, Wallace J. Forensic science in Canada: a report of multidisciplinary discussion Toronto:University of Toronto; 2013 [cited 2022 Jan 17]. Available from: https://www.crime-scene-investigator.net/forensic-science-in-canada.pdf
22. Edmond G, Towler A, Growns B, Ribeiro G, Found B, White D, Ballantyne K, Searston RA, Thompson MB, Tangen JM, et al. Thinking forensics: cognitive science for forensic practitioners. Sci and Justice. 2017;57(2):144–154. doi:10.1016/j.scijus.2016.11.005.
23. Dale G, Arnell KM. Lost in the forest, stuck in the trees: dispositional global/local bias is resistant to exposure to high and low spatial frequencies. PLoS ONE. 2014;9(7):e98625. doi:10.1371/journal.pone.0098625.
24. Newell BR, Shanks DR. Unconscious influences on decision-making: a critical review. Behav Brain Sci. 2014;37(1):1–19. doi:10.1017/S0140525X12003214.
25. Förster J. Relations between perceptual and conceptual scope: how global versus local processing fits a focus on similarity versus dissimilarity. J of Exp Psychol: Gen. 2009;138(1):88–111. doi:10.1037/a0014484.
26. Macrae CN, Lewis HL. Do I know you? Processing orientation and face recognition. Psychol Sci. 2002;13(2):194–196. doi:10.1111/1467-9280.00436.
27. Willis J, Todorov A. First impressions: making up your mind after a 100-ms exposure to a face. Psychol Sci. 2006;17(7):592–598. doi:10.1111/j.1467-9280.2006.01750.x.

28. Earwaker H, Nakhaeizadeh S, Smit N, Morgan RM. A cultural change to enable improved decision-making in forensic science: a six phased approach. Sci and Justice. 2020;60(1):9–19. doi:10.1016/j.scijus.2019.08.006.
29. Nakhaeizadeh S Cognitive bias and forensic anthropology: the power of context in the interpretation of skeletal remains [PhD thesis]. London: University College London; 2017.
30. Duríc M, Rakočevíc Z, Doníc D. The reliability of sex estimation of skeletons from forensic context in the balkans. Forensic Sci Int. 2005;47(2–3):159–164. doi:10.1016/j.forsciint.2004.09.111.
31. Morgan RM. Conceptualising forensic science and forensic reconstruction. Part II: the critical interaction between research, policy/law and practice. Sci and Justice. 2017;57(6):460–467. doi:10.1016/j.scijus.2017.06.003.
32. Swami S, Patnaik VV, Kaushal S, Sharma D. Evaluation of non-metric traits of supraorbital region in sexual dimorphism: a study in 60 adult human skulls. Medico-Legal Update. 2010;10 (1):68–72.
33. Nakhaeizadeh S, Morgan RM, Olsson V, Arvidsson M, Thompson T. The value of eye-tracking technology in the analysis and interpretations of skeletal remains: a pilot study. Sci and Justice. 2020;60(1):36–42. doi:10.1016/j.scijus.2019.08.005.
34. Acsádi G, Nemeskéri J. History of human life span and mortality. BUDAPEST: Akademiaikiado; 1970.
35. Buikstra JE. Ubelaker DH standards for data collection from human skeletal remains. Fayetteville: Arkansas Archaeological Survey; 1994. Research Series No. 44.
36. Inskip S, Scheib CL, Wohns AW, Ge X, Kivisild T, Robb J. Evaluating macroscopic sex estimation methods using genetically sexed archaeological material: the medieval skeletal collection from St John's divinity school, Cambridge. Am J Phys Anthr. 2019;168(2):340–351. doi:10.1002/ajpa.23753.
37. Garvin HM, Sholts SB, Mosca LA. Sexual dimorphism in human cranial trait scores: effects of population, age, and body size. Am J Phys Anthr. 2014;154(2):259–269. doi:10.1002/ajpa.22502.
38. Walker PL. Sexing skulls using discriminant function analysis of visually assessed traits. Am J Phys Anthr. 2008;136(1):39–50. doi:10.1002/ajpa.20776.
39. Weiss KM. On the systematic bias in skeletal sexing. Am J Phys Anthr. 1972;37(2):239–249. doi:10.1002/ajpa.1330370208.
40. Nakhaeizadeh S, Morgan RM, Rando C, Dror IE. Cascading bias of initial exposure to information at the crime scene to the subsequent evaluation of skeletal remains. J Forensic Sci. 2018;63(2):403–411. doi:10.1111/1556-4029.13569.
41. Dane E, Rockmann KW, Pratt MG. When should I trust my gut? Linking domain expertise to intuitive decision-making effectiveness. Organ Behav and Hum Decis Process. 2012;119 (2):187–194. doi:10.1016/j.obhdp.2012.07.009.
42. Krogman WK, Işcan MY. The human skeleton in forensic medicine. 2nd ed. Springfield: Charles C. Thomas; 1986.
43. Dror IE. The paradox of human expertise: why experts get it wrong. In: Kapur N, editor. The paradoxical brain. Cambridge: Cambridge University Press; 2011. p. 177–188.
44. Maule AJ, Hockey GRJ, Bdzola L. Effects of time-pressure on decision-making under uncertainty: changes in affective state and information processing strategy. Acta Psychol. 2000;104 (3):283–301. doi:10.1016/S0001-6918(00)00033-0.
45. Marr D. Vision: a computational investigation into the human representation and processing of visual information. San Francisco: W. H. Freeman; 1982.
46. Dunning D, Meyerowitz JA, Holzberg AD. Ambiguity and self-evaluation: the role of idiosyncratic trait definitions in self-serving assessments of ability. J Pers and Soc Psychol. 2005;57 (6):1082–1090. doi:10.1037/0022-3514.57.6.1082.
47. Griffin D, Tversky A. The weighing of evidence and the determinants of confidence. Cogn Psychol. 1992;24(2):4114–4135. doi:10.1016/0010-0285(92)90013-R.
48. Nakhaeizadeh S, Dror IE, Morgan RM. Cognitive bias in forensic anthropology: visual assessment of skeletal remains is susceptible to confirmation bias. Sci and Justice. 2014;54 (3):208–214. doi:10.1016/j.scijus.2013.11.003.

49. Nakhaeizadeh S, Hanson I, Dozzi N. The power of contextual effects in forensic anthropology: a study of biasability in the visual interpretations of trauma analysis on skeletal remains. J Forensic Sci. 2014;59(5):1177–1183. doi:10.1111/1556-4029.12473.
50. Murrie DC, Gardner BO, Kelley S, Dror IE. Perceptions and estimates of error rates in forensic science: a survey of forensic analysts. Forensic Sci Int. 2019;302(109887):109887. doi:10.1016/j.forsciint.2019.109887.
51. Dror IE. Cognitive and human factors in expert decision-making: six fallacies and the eight sources of bias. Anal Chem. 2020;92(12):7998–8004. doi:10.1021/acs.analchem.0c00704.
52. Almazrouei MA, Dror IE, Morgan RM. Organizational and human factors affecting forensic decision-making: workplace stress and feedback. J Forensic Sci. 2020;65(5):1968–1977. doi:10.1111/1556-4029.14542.
53. Dror IE, Pierce ML. ISO standards addressing issues of bias and impartiality in forensic work. J Forensic Sci. 2020;65(3):800–808. doi:10.1111/1556-4029.14265.
54. Morgan R. Conceptualising forensic science and forensic reconstruction. Part I: A conceptual model. Science & Justice. 2017;57(6):455–459. doi:10.1016/j.scijus.2017.06.002.
55. Georgiou N, Morgan RM, French JC. Conceptualising, evaluating and communicating uncertainty in forensic science: identifying commonly used tools through an interdisciplinary configurative review. Sci and Justice. 2020;60(4):313–336. doi:10.1016/j.scijus.2020.04.002.

Appendix 1

Results of Cohen's weighted Kappa for all participant pairings. Results for Phase 1 are listed in the left table and Phase 2 on the right. The table also shows the result obtained for Light's Kappa (arithmetic average of all Cohen's weighted Kappa results) and for Fleiss' Kappa measure of reliability between multiple observers).

Phase 1						Phase 2					
Pairing	Kappa	95% C.I.	Pairing	Kappa	95% C.I.	Pairing	Kappa	95% C.I.	Pairing	Kappa	95% C.I.
Part #1 - Part #2	0.625	0.216 | 1034	Part #5 - Part #6	0.625	0.201 | 1.049	Part #1 - Part #2	0.516	0.116 | 0.916	Part #5 - Part #6	0.625	0.270 | 0.980
Part #1 - Part #3	0.667	0.283 | 1051	Part #5 - Part #7	1.000	1.000 | 1.000	Part #1 - Part #3	0.351	-0.079 | 0.782	Part #5 - Part #7	0.727	0.410 | 1.045
Part #1 - Part #4	0.526	0.010 | 1042	Part #5 - Part #8	0.471	-0.033 | 0.974	Part #1 - Part #4	0.486	0.048 | 0.924	Part #5 - Part #8	0.323	-0.020 | 0.665
Part #1 - Part #5	0.824	0.497 | 1.150	Part #5 - Part #9	1.000	1.000 | 1.000	Part #1 - Part #5	0.559	0.095 | 1.023	Part #5 - Part #9	0.719	0.480 | 0.958
Part #1 - Part #6	0.471	0.042 | 0.899	Part #5 - Part #10	0.824	0.502 | 1.145	Part #1 - Part #6	0.690	0.380 | 0.999	Part #5 - Part #10	0.636	0.293 | 0.980
Part #1 - Part #7	0.824	0.497 | 1.150	Part #5 - Part #11	0.824	0.502 | 1.145	Part #1 - Part #7	0.333	-0.068 | 0.735	Part #5 - Part #11	0.625	0.294 | 0.956
Part #1 - Part #8	0.667	0.283 | 1.051	Part #5 - Part #12	0.647	0.193 | 1.101	Part #1 - Part #8	0.294	-0.059 | 0.647	Part #5 - Part #12	0.300	0.048 | 0.552
Part #1 - Part #9	0.824	0.497 | 1.150	Part #5 - Part #13	0.647	0.265 | 1.029	Part #1 - Part #9	0.600	0.196 | 1.004	Part #5 - Part #13	0.438	-0.032 | 0.907
Part #1 - Part #10	0.667	0.283 | 1.051	Part #5 - Part #14	0.500	0.117 | 0.883	Part #1 - Part #10	0.382	-0.025 | 0.790	Part #5 - Part #14	0.824	0.625 | 1.022
Part #1 - Part #11	0.667	0.283 | 1.051	Part #6 - Part #7	0.625	0.201 | 1.049	Part #1 - Part #11	0.382	-0.050 | 0.814	Part #6 - Part #7	0.531	0.169 | 0.893
Part #1 - Part #12	0.824	0.497 | 1.150	Part #6 - Part #8	0.400	-0.143 | 0.943	Part #1 - Part #12	0.520	0.285 | 0.755	Part #6 - Part #8	0.464	0.086 | 0.842
Part #1 - Part #13	0.526	0.165 | 0.888	Part #6 - Part #9	0.625	0.201-1.049	Part #1 - Part #13	0.679	0.249 | 1.108	Part #6 - Part #9	0.889	0.672 | 1.106
Part #1 - Part #14	0.400	-0.029 | 0.829	Part #6 - Part #10	0.800	0.449 | 1.151	Part #1 - Part #14	0.531	0.031 | 1.032	Part #6 - Part #10	0.613	0.320 | 0.906
Part #2 - Part #3	0.250	-0.311 | 0.811	Part #6 - Part #11	0.800	0.449 | 1.151	Part #2 - Part #3	0.500	0.070 | 0.930	Part #6 - Part #11	0.600	0.202 | 0.998
Part #2 - Part #4	0.500	-0.015 | 1.015	Part #6 - Part #12	0.625	0.201 | 1.049	Part #2 - Part #4	0.710	0.457 | 0.962	Part #6 - Part #12	0.571	0.393 | 0.750
Part #2 - Part #5	0.438	-0.082 | 0.957	Part #6 - Part #13	0.571	0.086 | 1.056	Part #2 - Part #5	0.625	0.356 | 0.894	Part #6 - Part #13	0.769	0.464 | 1.074
Part #2 - Part #6	0.357	-0.325 | 1.040	Part #6 - Part #14	0.438	-0.010 | 0.885	Part #2 - Part #6	0.786	0.501 | 1.071	Part #6 - Part #14	0.625	0.248 | 1.002
Part #2 - Part #7	0.438	-0.082 | 0.957	Part #7 - Part #8	0.471	-0.033 | 0.974	Part #2 - Part #7	0.679	0.354 | 1.003	Part #7 - Part #8	0.500	0.149 | 0.851
Part #2 - Part #8	0.625	0.201 | 1.049	Part #7 - Part #9	1.000	1.000 | 1.000	Part #2 - Part #8	0.625	0.316 | 0.934	Part #7 - Part #9	0.600	0.279 | 0.921
Part #2 - Part #9	0.438	-0.082 | 0.957	Part #7 - Part #10	0.824	0.502 | 1.145	Part #2 - Part #9	0.889	0.673 | 1.105	Part #7 - Part #10	0.889	0.670 | 1.107
Part #2 - Part #10	0.250	-0.311 | 0.811	Part #7 - Part #11	0.824	0.502 | 1.145	Part #2 - Part #10	0.778	0.506 | 1.049	Part #7 - Part #11	0.885	0.668 | 1.101
Part #2 - Part #11	0.250	-0.311 | 0.811	Part #7 - Part #12	0.647	0.193 | 1.101	Part #2 - Part #11	0.769	0.456 | 1.082	Part #7 - Part #12	0.172	-0.043 | 0.387
Part #2 - Part #12	0.438	-0.082 | 0.957	Part #7 - Part #13	0.647	0.265 | 1.029	Part #2 - Part #12	0.348	0.096 | 0.600	Part #7 - Part #13	0.500	0.048 | 0.952
Part #2 - Part #13	0.438	-0.010 | 0.885	Part #7 - Part #14	0.500	0.117 | 0.883	Part #2 - Part #13	0.538	0.071 | 1.006	Part #7 - Part #14	0.559	0.214 | 0.904

(*Continued*)

(Continued).

Phase 1						Phase 2					
Pairing	Kappa	95% C.I.	Pairing	Kappa	95% C.I.	Pairing	Kappa	95% C.I.	Pairing	Kappa	95% C.I.
Part #2 - Part #14	0.333	-0.123 | 0.790	Part #8 - Part #9	0.471	-0.033 | 0.974	Part #2 - Part #14	0.625	0.273 | 0.977	Part #8 - Part #9	0.538	0.139 | 0.938
Part #3 - Part #4	0.471	0.019 | 0.922	Part #8 - Part #10	0.625	0.167 | 1.083	Part #3 - Part #4	0.625	0.145 | 1.105	Part #8 - Part #10	0.625	0.355 | 0.895
Part #3 - Part #5	0.824	0.502 | 1.145	Part #8 - Part #11	0.625	0.167 | 1.083	Part #3 - Part #5	0.727	0.262 | 1.192	Part #8 - Part #11	0.591	0.220 | 0.962
Part #3 - Part #6	0.800	0.449 | 1.151	Part #8 - Part #12	0.824	0.502 | 1.145	Part #3 - Part #6	0.531	0.108 | 0.955	Part #8 - Part #12	0.217	-0.063 | 0.498
Part #3 - Part #7	0.824	0.502 | 1.145	Part #8 - Part #13	0.800	0.449 | 1.151	Part #3 - Part #7	0.793	0.528 | 1.058	Part #8 - Part #13	0.423	0.012 | 0.834
Part #3 - Part #8	0.625	0.167 | 1.083	Part #8 - Part #14	0.625	0.027 | 1.223	Part #3 - Part #8	0.308	-0.134 | 0.749	Part #8 - Part #14	0.344	-0.062 | 0.750
Part #3 - Part #9	0.824	0.502 | 1.145	Part #9 - Part #10	0.824	0.502 | 1.145	Part #3 - Part #9	0.613	0.228 | 0.998	Part #9 - Part #10	0.690	0.398 | 0.982
Part #3 - Part #10	1.000	1.000 | 1.000	Part #9 - Part #11	0.824	0.502 | 1.145	Part #3 - Part #10	0.700	0.393 | 1.007	Part #9 - Part #11	0.679	0.346 | 1.011
Part #3 - Part #11	1.000	1.000 | 1.000	Part #9 - Part #12	0.647	0.193 | 1.101	Part #3 - Part #11	0.679	0.348 | 1.009	Part #9 - Part #12	0.455	0.215 | 0.694
Part #3 - Part #12	0.824	0.502 | 1.145	Part #9 - Part #13	0.647	0.265 | 1.029	Part #3 - Part #12	0.200	-0.053 | 0.453	Part #9 - Part #13	0.654	0.257 | 1.051
Part #3 - Part #13	0.800	0.449 | 1.151	Part #9 - Part #14	0.500	0.117 | 0.883	Part #3 - Part #13	0.531	0.119 | 0.944	Part #9 - Part #14	0.719	0.467 | 0.970
Part #3 - Part #14	0.625	0.201 | 1.049	Part #10 - Part #11	1.000	1.000 | 1.000	Part #3 - Part #14	0.559	0.136 | 0.982	Part #10 - Part #11	0.769	0.464 | 1.074
Part #4 - Part #5	0.667	0.289 | 1.044	Part #10 - Part #12	0.824	0.502 | 1.145	Part #4 - Part #5	0.912	0.749 | 1.075	Part #10 - Part #12	0.222	-0.006 | 0.450
Part #4 - Part #6	0.294	-0.160 | 0.748	Part #10 - Part #13	0.800	0.449 | 1.151	Part #4 - Part #6	0.545	0.202 | 0.889	Part #10 - Part #13	0.571	0.135 | 1.008
Part #4 - Part #7	0.667	0.289 | 1.044	Part #10 - Part #14	0.625	0.201 | 1.049	Part #4 - Part #7	0.806	0.477 | 1.136	Part #10 - Part #14	0.471	0.115 | 0.826
Part #4 - Part #8	0.471	-0.134 | 1.076	Part #11 - Part #12	0.824	0.502 | 1.145	Part #4 - Part #8	0.379	0.058 | 0.701	Part #11 - Part #12	0.192	-0.037 | 0.421
Part #4 - Part #9	0.667	0.289 | 1.044	Part #11 - Part #13	0.800	0.449 | 1.151	Part #4 - Part #9	0.625	0.356 | 0.894	Part #11 - Part #13	0.571	0.135 | 1.008
Part #4 - Part #10	0.471	0.019 | 0.922	Part #11 - Part #14	0.625	0.201 | 1.049	Part #4 - Part #10	0.710	0.349 | 1.070	Part #11 - Part #14	0.625	0.243 | 1.007
Part #4 - Part #11	0.471	0.019 | 0.922	Part #12 - Part #13	0.647	0.265 | 1.029	Part #4 - Part #11	0.700	0.345 | 1.055	Part #12 - Part #13	0.550	0.271 | 0.829
Part #4 - Part #12	0.333	-0.231 | 0.898	Part #12 - Part #14	0.500	-0.015 | 1.015	Part #4 - Part #12	0.226	0.001-0 | 450	Part #12 - Part #14	0.300	0.062 | 0.538
Part #4 - Part #13	0.625	0.201 | 1.049	Part #13 - Part #14	0.786	0.385 | 1.187	Part #4 - Part #13	0.344	-0.133 | 0.821	Part #13 - Part #14	0.438	-0.011 | 0.886
Part #4 - Part #14	0.813	0.500 | 1.125				Part #4 - Part #14	0.735	0.414 | 1.056			
Light's Kappa			0.643			Light's Kappa			0.575		
Fleiss' Kappa			0.520			Fleiss' Kappa			0.322		

Part = participant.