OMB Control #0693-0043
Expiration Date: 06/30/2025
NIST Generic Clearance for Usability Data Collections

Assessing the effectiveness of facial features and developing calibrated facial proficiency tests

FOUR STANDARD SURVEY QUESTIONS

1. Explain who will be surveyed and why the group is appropriate to survey.

NIST's Image Group conducts studies to evaluate the facial recognition ability of people in the general population and among face expert groups (henceforth: *examiners*). Prior studies have shown that examiners are highly accurate, but little is understood about their processes. Guidelines from a prominent working group, the Facial Identification Scientific Working Group (FISWG), advise examiners to evaluate and describe individual parts of the face when conducting evaluations of all types [1]. It is unknown if some face parts (features) are better than others for identification. It is also unknown to what extent examiners use different parts of the face to make decisions.

Potential participants will be recruited and sampled from populations of face specialists (e.g., facial forensic examiners and facial forensic reviewers) because this is the population of interest. Comparison groups (i.e., laypersons who are not face specialists) may be recruited for comparison purposes. Recruitment for volunteers will be done using the following methods: 1) direct communication to existing contacts/known face specialists, 2) through representatives at organizations who work with, produce guidance and standards documents for, study, or employ face specialists, and 3) verbal and written announcements at talks, conferences, and other venues. When recruiting through representatives, the information sheet and the exact, approved language to use during recruitment will be provided to the representative(s).

The number of respondents is estimated to be about 90 people. This estimate is based on the number of face specialists who participated in a worldwide study on forensic facial examiners and reviewers [2].

To better understand the face processing system of examiners, we will conduct a study funded as part of an interagency agreement with the Federal Bureau of Investigations (FBI). This study will allow us to systematically measure how face examiners use information from the face to make identification decisions.

References

[1] Facial Identification Scientific Working Group, "Facial Image Comparison Feature List for Morphological Analysis." Sep. 11, 2018. [Online]. Available: https://fiswg.org/FISWG_Morph_Analysis_Feature_List_v2.0_20180911.pdf
[2] P. J. Phillips *et al.*, "Face recognition accuracy of forensic examiners, superrecognizers, and

face recognition algorithms," *Proc. Natl. Acad. Sci.*, vol. 115, no. 24, pp. 6171–6176, Jun. 2018, doi: 10.1073/pnas.1721355115.
[3] PCAST Working Group and E. S. Lander, "Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods," *This Rep. Was Downloaded Httpsobamawhitehousearchivesgovadministrationeopostppcastdocsreportsarchives Spring 2021*, Sep. 2016, Accessed: Aug. 30, 2023. [Online]. Available: https://scholarship.rice.edu/handle/1911/113033
[4] National Research Council, *Strengthening Forensic Science in the United States: A Path Forward*. Washington, D.C.: The National Academies Press, 2009.
[5] J. Deng, J. Krause, and L. Fei-Fei, "Fine-Grained Crowdsourcing for Fine-Grained Recognition," in *2013 IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2013, pp. 580–587. doi: 10.1109/CVPR.2013.81.
[6] A. Towler, D. White, and R. I. Kemp, "Evaluating the feature comparison strategy for forensic face identification." *J. Exp. Psychol. Appl.*, vol. 23, no. 1, pp. 47–58, 2017, doi: 10.1037/xap0000108.
[7] G. Jeckeln *et al.*, "Face identification proficiency test designed using item response theory," *Behav. Res. Methods*, Jun. 2023, doi: 10.3758/s13428-023-02092-7.
[8] R. S. Malpass and J. Kravitz, "Recognition for faces of own and other race," *J. Pers. Soc. Psychol.*, vol. 13, no. 4, pp. 330–334, 1969.
[9] A. J. O'Toole, J. Peterson, and K. A. Deffenbacher, "An 'other-Race Effect' for Categorizing Faces by Sex," *Perception*, vol. 25, no. 6, pp. 669–676, Jun. 1996, doi: 10.1068/p250669.

2. Explain how the survey was developed including consultation with interested parties, pre-testing, and responses to suggestions for improvement.

The study was designed to measure facial recognition accuracy under different viewing conditions and to measure which part(s) of the face examiners prioritize when making facial recognition decisions.

The study was developed as part of an interagency agreement with the Federal Bureau of Investigations (FBI). Bureau members identified a need to better understand how examiners conduct facial comparisons with respect to analyzing the face's features. The need to better understand forensic processes was also expressed in a 2016 report by the President's Council of Advisors on Science and Technology in their report to President Barack Obama [3] and in 2009 by the National Research Council [4].

Based on this need, we developed a test to begin answering questions about examiners' forensic comparison processes. The study was created to allow us to measure which parts of the face (facial features) are prioritized by examiners during an evaluation and which, if any, are best to base identification decisions on. The format was inspired by methods in computer science to crowd-source the relevant parts of images (e.g., birds) for identification and categorization [5]. In their study, users hovered a mouse over a blurred image. The corresponding area would be put into focus, and if the user determined what they saw was useful, they clicked on the mouse to indicate this. The procedure was gamified via a points system. This encouraged participants to be parsimonious in their decisions to obtain the most points possible. From this process, the authors

determined which parts of the image were most useful for bird species classification.

The procedure described above inspired this study's methods. In the experiment conducted in this study, participants will view images of faces. They are tasked with determining which images depict the same identity. At the start of each experimental repetition, a face begins completely occluded. Participants are instructed to select from a list of facial features (e.g., eyes, noses, mouths). After selecting, the corresponding facial feature is revealed on the occluded image. From this, we can determine which features are prioritized and their link to accurate decisions. Nine features were used in this study: cheeks, chin, ears, eyebrows, eyes, forehead, mouth, neck, and nose. This final list of features was refined based on three criteria. First, these are all features that appear in FISWG's guidance to examiners. Second, these were features evaluated in a prior study on examiners [6]. Overlap between these two studies will allow us to compare and examine the extent to which feature preference generalizes across studies on examiners. Third, the set of features needed to be ones with clear boundaries. That is, they needed to be consistently detectable based on their predicted locations on the face across images of different faces (e.g., a nose appears near the center of the face; a neck is under a chin, etc.).

All face images were obtained from the Triad Identity Matching (TIM) test dataset [7]. The NIST Research Protections Office approved the procedures to conduct the experiment and to use the face images in this study. The images shown in the experiment are approved to be used for research purposes and be published.

The test was developed in consultation with the examiner community and face identification subject matter experts. During the planning phase, we outlined the proposed procedures to relevant communities and expert groups, such as Program Managers in the FBI, program updates to NIST's Special Programs Office, and working groups specializing in face recognition. After planning and development based on feedback, we created a preliminary version of the experiment. We demoed these to a small selection of face experts in the FBI. Based on their feedback, we adjusted the boundaries around the isolated facial features. We also removed face images from the study that examiners expressed were not high quality enough to base decisions on. We iterated on their feedback. When the NIST researchers and the FBI face experts were satisfied, we deployed the test to federal government employees. To date, 18 examiners who are employed by the federal government have taken part in the study.

3. Explain how the survey will be conducted, how customers will be sampled if fewer than all customers will be surveyed, expected response rate, and actions your agency plans to take to improve the response rate.

The study has two parts: 1) an experiment to evaluate face feature use during identification decisions and 2) a brief demographic survey. The study is always presented in that order.

The experiment's purpose is to investigate which parts of the face ("features") are prioritized during identification and which features are associated with accurate decisions. The experiment has two versions (conditions). Participants are assigned to one condition using counterbalancing.

In one condition, participants will see pairs of faces on the computer screen. These visible pairs

are always of different people. Next to each pair, a black rectangle is displayed. This rectangle covers a third face image. This third face is of the same identity as one of the two visible people. The participant decides which of those two visible people is the same identity as the third image. To make their decision, they can reveal different facial features on the covered face. They will do so by indicating with a mouse click or keyboard press the name of the face feature they want to see from an available list of nine features: cheeks, chin, ears, eyebrows, eyes, forehead, mouth, neck, and nose.

A gamified system was implemented, inspired by the study referenced above as inspiration (see Question 2 above). This gamification was included to encourage participants to be selective and parsimonious in their decisions to reveal facial features and to discourage guessing. Participants began the experiment with 100 points. For every correct answer, they gain an additional 100 points. They may reveal one feature without losing any points. For every feature revealed after the first, 30 points are deducted from the total score. Therefore, they must be selective to gain the most points possible. To discourage guessing, incorrect answers are penalized by deducting 200 points from the total score. There are 24 repetitions in total, so a perfect score adds up to 2500 points. The participants are instructed that their goal is to accumulate as many points as possible by the end of the experiment.

The second experimental condition is identical to the first, with one exception. In the second condition, participants only have the option to choose if they wish to reveal a feature; they do not have the option to select which one. Therefore, they can choose the total number of features to reveal on each trial, but the order in which the features are revealed is predetermined based on a previous benchmarking study. In the benchmarking study, participants rated facial feature similarity for pairs of faces. The faces rated were those that appear visibly at the start of the trial in the current experiment. From that, the order in which features were revealed were either from most similar to most dissimilar; or most dissimilar to most similar. This will allow us to examine whether feature similarity is linked to accuracy, regardless of a participant's subjective preferences of which feature they wish to see first.

After participating in one of the experimental conditions, participants are provided a survey to collect limited demographic information: age, sex, ethnicity, and race. This information will be connected to a participant identification number and never their real name or personally identifiable information. This information is important because face recognition ability is known to vary based on demographic factors [8], [9]. All questions on this demographic survey will be voluntary, with participants having the option to skip any questions they do not want to answer.

The study will be conducted on computers, either in person or remotely. If conducted in person, a NIST researcher will schedule an appointment with the potential participant at an agreed-upon location in a federal building, such as an office or conference room. If conducted remotely, the participant will receive the experimental materials electronically. The electronic methods are currently undergoing development. Options will include a link to a web-based survey tool (e.g., Qualtrics) or sending materials via an electronic packet or web-based application or environment (e.g., Docker). Factors influencing whether the collection will take place in person or remotely for any given participant include: 1) how many other potential participants are at their site, 2) Whether the participant is in a location easily accessible by travel, 3) the overall cost-benefit to

NIST to travel on location to test in person as opposed to testing remotely.

There are 24 repetitions in the experiment. There are 4 questions in the demographic survey. The amount of time needed to complete the study is no more than 30 minutes based on existing participants who have completed the study. In total, we are anticipating 90 respondents. With 90 respondents taking 30 minutes on the experiment each, the total burden is expected to be 45 hours (2700 minutes).

Participation will be voluntary. Our primary method to improve the response rate will be to specify the purpose of the study and the positive impact of participating on the field of facial forensics at the time of recruitment. Reminder emails may be sent. If a participant requests to be removed from an email list/future reminders, we will remove them from the contact list(s) in a timely manner.

No personally identifiable identification (PII) will be collected or published. To de-identify the data, each participant is assigned a random study number at the time of the data collection. All research is performed on de-identified data with study numbers linking the responses. Therefore, none of the analysis links to participants. Participant logs with names, email addresses, or both will be obtained to ensure the same participant doesn't sign up to take the study more than once. That participant log will never be attached to the data and will be available only to researchers listed on study protocols approved by NIST's Research Protections Office.

4. Describe how the results of the survey will be analyzed and used to generalize the results to the entire customer population.

This study will address concerns raised in the National Academy of Sciences (NAS)'s report, *Strengthening Forensic Science in the United States: A Path Forward* [4], and The PCAST Working Group's *Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods [3]*. Both reports called for more scientific rigor and the need to establish evidence-based practices in forensic science.

This study aims to 1) understand the processes of facial specialists, such as facial examiners and reviewers, when determining identity and 2) understand how different parts of the face (e.g., eyes, noses, etc.) influence the ability to make an identification decision.

Using methods from signal detection theory, analysis of variance, correlation theory, and measures of central tendency and variance, the results of these experiments will be used to make predictions about which parts of the face may or may not be better to use for identification and which parts may or may not be preferred by face specialists. This will support facial forensic practitioners' ability to make informed, evidence-based decisions when conducting facial comparisons.