Document Title:	Design and Implementation of Forensic Facial Identification Experts Test
Author(s):	Alice J. O'Toole
Document Number:	254663
Date Received:	April 2020
Award Number:	2015-IJ-CX-K014

Design and Implementation of Forensic Facial Identification Experts Test

2015-IJ-CX-K014: Final Summary Overview - Revised

PI: Alice J. O'Toole

Purpose

The purpose of this project was to develop a multi-phasic "black box" test to measure the identification skills of forensic face examiners. A black box test is a series of small-scale experiments that can be distributed to labs responsible for forensic facial examination in crime prevention and criminal justice applications. We developed and implemented tests that measure skills across a broad range of facial identification tasks with image and video data. Although there have been a small number of laboratory-type tests of facial examiners [1,2], these have often been carried out in highly controlled conditions, under tight time constraints, and without access to environmental resources available to facial examiners when they perform these tasks under normal circumstances. The idea of a black-box test is to allow labs to perform the test *in situ*, using commonly employed processes and tools (e.g., measurement devices, magnification).

The project consisted of three parts. Part 1 was aimed at collecting normative performance data on potential image and video stimuli, using untrained human observers and facial identification software. To achieve this goal, we conducted a series of experiments with untrained human subjects matching facial identity in pairs of images and videos, targeted to specific types of challenge problems. These challenge problems included identification over changes in illumination, pose/viewpoint, and across faces of different ethnic and racial background. The results were analyzed with the goal of finding stimuli appropriate for a streamlined test of examiners that has been able to provide points of comparison between experts, untrained subjects, and computer-based face recognition systems.

The goal of Part 2 was to compile *image* stimuli normed based on the performance of *forensic examiners.* To this end, we performed extensive *item analyses* on the individual stimuli used in a recent test of 27 facial identification experts from the Facial Identification Scientific Working Group (FISWG) Meeting at the FBI Academy (Quantico, VA) in May, 2014 [1]. This study was a collaborative effort by Dr. David White, (University of New South Wales, UNSW), Dr. P. Jonathon Phillips (NIST), and my laboratory. This analysis provided a baseline for individual images judged by highly skilled, practicing facial identification experts in perceptual face identification tasks. This test was done under "laboratory conditions" with time constraints and no access to the tools available in-house. It is, therefore, a perceptual rather than forensic study of face identification. For the present work, our use of these particular items in the black box test provided critical data on the effectiveness of the in-house process of facial identification across forensic laboratories, above and beyond the basic perceptual skills of the experts.

In Part 3 of the proposed work, our goal was to evaluate and refine the black box test through an iterative release to participating labs. In fact, we were actually able to implement this test completely and to publish the results in the *Proceedings of the National Academy of Sciences*. The black box test effort was coordinated through the National Institute of Standards and Technology (NIST) and with the help of Dr. Richard Vorderbruegge at the FBI, who was instrumental in putting us in touch with candidate labs throughout the United States and the world. These labs participated and provided feedback on the tests we developed and administered. Although we did not originally propose a comparison between experts and face recognition algorithms in our original proposal, we were able to include this comparison. This was made possible by a fortuitous combination of projects in the PI's lab. This gave us access to state-of-the-art face recognition software that was being developed under the DOD's Janus IARPA program. This allowed for a full comparison between human experts and computers on an identical task of face identification.

Project Design, Methods, Analysis & Findings

Black Box Test and Comparison

Achieving the upper limits of face identification accuracy in forensic applications can minimize errors that have profound social and personal consequences. Although forensic examiners identify faces in these applications, systematic tests of their accuracy are rare. How can we achieve the most accurate face identification, using people and/or machines, working alone or in collaboration? In a comprehensive comparison of face identification by humans and computers, we found that forensic facial examiners, facial reviewers, and super-recognizers were more accurate than fingerprint examiners and students on a challenging face identification test. Individual performance on the test varied widely. On the same test, four deep convolutional neural networks, developed between 2015 and 2017, identified faces within the range of human accuracy. Accuracy of the algorithms increased steadily over time with the most recent DCNN scoring above the median of the forensic facial examiners. Using crowd-sourcing methods, we fused the judgments of multiple forensic facial examiners by averaging their rating-based identity judgments. Accuracy was substantially better for fused judgments than for individuals working alone. Fusion also served to stabilize performance, boosting the scores of lower performing individuals and decreasing variability. Single forensic facial examiners fused with the best algorithm were more accurate than the combination of two examiners. Therefore, collaboration among humans, and between humans and machines, offers tangible benefits to face identification accuracy in important applications. These results offer an evidence-based roadmap for achieving the most accurate face identification possible.

Work published in ***(ref. #1)****:*

Phillips, P. J., Yates, A. N., Hu, Y., Hahn, C. A., Noyes, E. Jackson, K. Cavazos, J. G., Jeckeln, G. Ranjan, R., Sankaranarayarnan, S., Chen, J.C., Castillo, C., Chellappa, R. White, D., & O'Toole, A. J. (2018). Face Recognition Accuracy of Forensic Examiners, Super-recognizers, and Algorithms. *Proceedings of the National Academy of Sciences*, https://doi.org/10.1073/pnas.1721355115

Item Analysis for Black Box Test

Professional forensic face examiners surpass untrained individuals on challenging face-identity matching tasks. We investigated qualitative/strategic differences in how forensic face examiners versus untrained people perform identity-matching tasks by analyzing *item* responses (ratings of the likelihood that two images show the same person). We developed a novel analysis for quantifying item difficulty for participant groups and establishing group "winners" for items in conditions of interest. "Wisdom-of-the-crowds" effects were explored by fusing responses from varying numbers of participants to amplify strategic differences across groups. Results indicated that examiners use the internal face more effectively than untrained participants, but failed to exploit identity information in the external face and body. We also showed that accuracy measures for examiners and controls must include both same-identity verifications and different-identity rejections to understand the role of perceptual skill and response bias in performance differences across participant groups.

Work published in in ***(ref. #2)****:*

Hu, Y., Jackson, K. Yates, A. White, D. Phillips, P. J. & O'Toole, A. J. (2017). Person recognition: Qualitative differences in how forensic face examiners and untrained people rely on the face vs. the body for identification. *Visual Cognition,* 25 (4-6), 492-506. http://dx.doi.org/10.1080/13506285.2017.1297339

Collaborative Decision Making to Improve Face Identification Accuracy

Face identification is more accurate when people collaborate in social dyads than when they work alone (Dowsett & Burton, 2015, *Br. J. Psychol.*, *106*, 433). Identification accuracy is also increased when the responses of two people are averaged for each item to create a 'non-social' dyad (White, Burton, Kemp, & Jenkins, 2013, *Appl. Cogn. Psychol.*, *27*, 769; White *et al.*, 2015, *Proc. R. Soc. B Biol. Sci.*, *282*, 20151292). Does social collaboration add to the benefits of response averaging for face identification? We compared individuals, social dyads, and non-social dyads on an unfamiliar face identity-matching test. We also simulated non-social collaborations for larger groups of people. Individuals and social dyads judged whether face image pairs depicted the same- or different identities, responding on a 5-point certainty scale. Non-social dyads were constructed by averaging the responses of paired individuals. Both social and non-social dyads were more accurate than individuals. There was no advantage for social over non-social dyads. For larger non-social groups, performance peaked at near perfection with a crowd size of eight participants. We tested three computational models of social collaboration and found that social dyad performance was predicted by the decision of the more accurate partner. We conclude that social interaction does not bolster accuracy for unfamiliar face identity matching in dyads beyond what can be achieved by averaging judgements.

Work published in ***(ref. #3)****:*
Jeckeln, G. Hahn, C. A., Noyes, E. Cavazos, J. G., & O'Toole, A. J. (2018). Wisdom of the social versus non-social crowd in face identification. *British Journal of Psychology.*
https://doi.org/10.1111/bjop.12291

Individual Differences and Group Effects: Interpreting experts' variability

This resource was prepared by the author(s) using Federal funds provided by the U.S. Department of Justice. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

There are large individual differences in people's face recognition ability. These individual differences provide an opportunity to recruit the best face-recognisers into jobs that require accurate person identification, through the implementation of ability-screening tasks. To date, screening has focused exclusively on face recognition ability, however real-world identifications can involve the use of other person-recognition cues. Here we incorporated body and biological motion recognition as relevant skills for person identification. We tested whether performance on a standardised face-matching task (the GFMT) predicts performance on three other identity-matching tasks, based on faces, bodies, and biological motion. We examined the results from group versus individual analyses. We found stark differences between the conclusions one would make from group analyses versus analyses that retain information about individual differences. Specifically, correlations and analysis of variance (ANOVAs) suggested that face recognition ability was related to performance for all person identification tasks. These analyses were strikingly inconsistent with the individual differences data, which suggested that the screening task was related only to performance on the face task. This study highlights the importance of individual data in the interpretation of results of person identification ability.

Work published in ***(ref. #4)****::*
Noyes, E. Hill, M. Q., & O'Toole, A. J. (2018). Face recognition ability does not predict person identification performance: Using individual data in the interpretation of group results. *Special Issue on Individual differences in face perception and person recognition*. (Eds. V. Bruce & K. Lander, & M. Bindemann), *Cognitive Research: Principles and Implications. 3(1), 23*.

Super-recognizers: Skills and Potential for Improving Security

Super-recognisers are people who perform face recognition and face matching tests with very high levels of accuracy. We reviewed the small literature available to date on super-recognisers and provide a summary of the key findings. Based on what is currently known, we argued that super-recognisers are best understood as the top performers sampled from a distribution of normal facial-recognition skills—rather than as a distinct population of people with 'superior recognition capacity'. This conclusion is based on findings that indicate that although super-recognisers *as a group* outperform controls on tasks of face processing, this is not true at an *individual level.* Individual 'super-recognisers' do not consistently exceed the expected performance range for normal face recognition skills. Moreover, their performance is not always consistent across related face processing tasks. Given this perspective on super-recognisers, we list open issues that should be addressed in future research. From an applied perspective, we argued that jobs that require accurate face identification skills, (e.g., law enforcement), should be filled by people with the best skills for the job. Given the limited consistency of super-recogniser performance across tasks, this may require tests that are targeted to the exact face processing skills needed for the success at a particular job. In this sense, super-recognisers should be considered 'skilled labour' for professional face recognition jobs and should be pre-tested as is done for professionals of all sorts (e.g., athletes, engineers).

Work published in ***(ref. #5 & 6)****::*
Noyes, E. Phillips, P. J. & O'Toole (2017). What is a super-recogniser? In Eds. M. Bindemann & A. Megreya) *Face Processing: Systems, Disorders and Cultural Differences.* Nova Science Publishers, Inc. NY, USA.
Noyes, E. Phillips, P. J. & O'Toole (2017). Face recognition assessments used in the study of super-recognisers**. arXiv:1705.04739**

Face Recognition and the "Other-race effect"

People recognize faces of their own race more accurately than faces of other races—a phenomenon known as the "Other-Race Effect" (ORE). Previous studies show that training with multiple variable images improves face recognition. Building on multi-image training, we take a novel approach to improving own- and other-race face recognition by testing the role of learning context on accuracy. Learning context was either *contiguous*, with multiple images of each identity seen in sequence, or *distributed*, with multiple images of an identity randomly interspersed among different identities. In two experiments, East Asian and Caucasian participants learned own- and other-races faces either in a contiguous or distributed order. In Experiment 1, people learned each identity from four highly variable face images. In Experiment 2, identities were learned from one image, repeated four times. In both experiments we found a robust other-race effect. The effect of learning context, however, differed depending on the variability of the learned images. The distributed presentation yielded better recognition when people learned from single repeated images (Exp. 1), but not when they learned from multiple variable images (Exp. 2). Overall, performance was better with multiple-image training than repeated single image training. We conclude that multiple-image training and distributed learning can both improve recognition accuracy, but via distinct processes. The former broadens perceptual tolerance for image variation from a face, when there are diverse images available to learn. The latter effectively strengthens the representation of *differences* among similar faces, when there is only a single learning image.

Work published in in ***(ref. #7):***
Cavazos, G., Noyes, E. & O'Toole, A. J. (2018). Learning context and the other-race effect: Strategies for improving face recognition. *Vision Research*.
https://doi.org/10.1016/j.visres.2018.03.003

Human Factors in Forensic Face Identification

Facial identification by forensic examiners is a core component of criminal investigations and convictions. These identifications are often done in challenging circumstances that require experts to match identity across images and videos taken at a various camera distances, under different illumination conditions, and across a wide range of poses. Until recently, laboratory studies of human face identification have concentrated, almost exclusively, on face identification by untrained (naïve) observers, with only a handful of studies focusing directly on the accuracy of expert forensic facial examiners. Over the last two decades, DNA-based exonerations of convicted criminals in the United States have revealed weaknesses in the forensic identification process due to *human factors*. In this paper, we reviewed and analyzed the factors known to impact facial identification accuracy for both naïve participants and trained experts. Combined, these studies point to a set of challenges that impact accuracy for both groups of participants. They also lead to an understanding of the specific conditions under which forensic facial examiners can surpass naïve observers at the task of face identification. Finally, we considered the role that computer-based face recognition systems can play in the future of forensic facial identification. These systems have made remarkable strides in recent years, raising new questions about how human and machine strengths at face identification can be combined to achieve optimum accuracy.

Work published in in ***(ref.*** *#8):*

White, D., Norell, K. Phillips, P. J. & O'Toole. (2016). Human factors in facial forensic examination. In (Eds. M. Tistarelli & C. Champod) *Biometrics in Forensic Science*. Springer Verlag.

Experiments to Contribute to the Development of Proficiency Tests (No-Cost Extension)

The final round of experiments examined the transferability of identity matching methods in examiner casework to the use of triad tests that allow for norming the test face stimuli so that they would be usable in proficiency tests. In this work, we applied Item Response Theory to the measurement of stimuli used in the tests of examiner skills. The primary challenge is to be able to generate pools of stimulus items that can be sampled to create tests of equal difficulty using different stimulus items. These kinds of tests have applied value in being utilized pre- and post-training to determine if the training is effective. Correctly applied, the methods will help to control for practice effects and other performance benefits due to familiarization with the stimuli. The work indicates that these methods are potentially highly effective in across populations and across time, for achieving a normative test that can be tailored to the skills of a targeted population (e.g., experts, trainees, etc.).

Implications for Criminal Justice Policy and Practice in the United States

We believe that the results of this project will have wide impact on the fields of forensic psychology and on criminal justice policy and practice in the United States. Our work shows, for the first time, the high skill levels of forensic facial examiners and reviewers over control groups of forensically trained fingerprint examiners and the general population. This gives scientific backing for policy decisions that endorse the use of forensic facial examiners as face identifiers in criminal justice situations.

The work also shows that the performance of the state-of-the-art face recognition algorithms compares favorably with trained human experts. This gives scientific backing for policy decisions that endorse the use of machine-based face recognition as part of the criminal justice tool kit.

An important finding of the work is that it indicates that the combination of the best algorithms with the best humans produces the most accurate face recognition ability. This gives scientific backing for policy decisions that endorse the use of human-machine collaborations in criminal justice applications. These combinations can exploit the strengths of both "systems" to improve accuracy.

Analysis of the data from the follow-up experiments points to the need to further investigate the difficulties of face recognition for other-race faces, viewpoint changes, and disguise. We see deficits for both human experts and machines in all three cases. The findings of different accuracy as a function of these factors for experts has an impact on understanding and predicting accuracy in applied scenarios in law enforcement. The work should also have impact in computer vision in that it provides a human benchmarks for face recognition technology. It is beginning to have an impact as an excellent test case for the use of statistical fusion in maximizing recognition accuracy.

We think the policy implications of this work are important. The research should assure the general public of the high skill of professional examiners and should also reassure the public about the use of face recognition technology in law enforcement.