**CITATION**: Towler, A., Kemp, R. I., & White, D. (2017). Unfamiliar face matching systems in applied settings. In *Face Processing: Systems, Disorders and Cultural Difference,*. Bindemann, M. & Megreya, A.M. (Eds.), Nova Science.

# UNFAMILIAR FACE MATCHING SYSTEMS IN APPLIED SETTINGS

***Alice Towler[1,2], Richard I. Kemp[2] and David White[2,*]***

[1]Department of Psychology, University of York, England, UK

[2]School of Psychology, University of New South Wales, Australia

** Corresponding Author address: Dr. David White, School of Psychology, University of New South Wales, Sydney NSW 2052, Australia. Email: david.white@unsw.edu.au.*

## ABSTRACT

Many years of research has established that humans are poor at identifying unfamiliar faces. This poses a significant problem in applied settings that depend on accurate face matching to verify the identity of unfamiliar people, such as when issuing photo-ID documents and in forensic investigations. However, in these situations, face matching decisions are seldom made by a single person. Instead, they are made by face matching *systems*, whereby chains of humans and computers make a series of identity judgements in a predetermined sequence. Psychologists and computer scientists have made good progress studying how individual components of these systems perform. However, very little is known about how well the components work together. In this chapter, we draw on research spanning psychology, forensic science and computer science to provide an initial indication of face matching system performance, suggest improvements to system design, and highlight key areas for future research.


**Keywords**: facial image comparison, face recognition, facial recognition software, person identification accuracy, biometrics, forensic science

## INTRODUCTION

Deciding whether two images are of the same unfamiliar person is a key component of many identity verification systems, for example when identifying criminal suspects, processing financial transactions and issuing photo-ID documents. The cost of identification errors in these systems can be serious: resulting, for example, in the issuance of false passports to criminals and the wrongful imprisonment of innocent people. Society therefore relies heavily on the accuracy of face matching decisions.

This reliance is not in keeping with a very large psychological literature showing that people perform very poorly when identifying unfamiliar faces (e.g., Bruce, Henderson, Newman, & Burton, 2001; Burton, Wilson, Cowan, & Bruce, 1999; Kemp, Towell, & Pike, 1997). More recently, it has become clear that this problem is not limited to lab-based research using novice participant groups, but that poor performance extends to professionals who match unfamiliar faces in their daily work (e.g., White, Dunn, Schmid, & Kemp, 2015; White, Kemp, Jenkins, Matheson, & Burton, 2014).

In many applied settings, complex procedures have been developed to support face matching decisions. For instance, significant advances in biometric identification technology has seen the widespread introduction of facial recognition software. In addition, a number of organisations have started to form specialist teams of face matching experts. Increasingly, identification decisions in applied settings are not made by a single person. Instead, face matching decisions are distributed across multiple human and computer processes. In this chapter, we refer to these distributed networks of decision-making as face matching *systems*.

The development of face matching systems has caused a widening gap in our scientific understanding of face identification in applied settings. This is because theoretical knowledge of face identification is almost entirely based on studies that examine the accuracy of *individuals* (see Bruce & Young, 2012; Calder, Rhodes, Johnson, & Haxby, 2011; Hole & Bourne, 2010). Because participants have been tested in isolation from the broader context of the decision-making framework, very little is known about how the component processes in these systems interact to produce accurate face matching decisions.

In this chapter, we consolidate evidence from the psychological, forensic and computer sciences to gain a preliminary understanding of current system performance, explore methods of optimising system design, and identify key areas for future research. We argue that it is

necessary for researchers to address the significant knowledge gaps in this area, not just to reveal the current (and potential) operational accuracy of face matching systems, but to improve our understanding of the factors underpinning expertise in person identification.

## EXAMPLES OF FACE MATCHING SYSTEMS

Shortly after the invention of photography, practitioners began to identify unfamiliar people by comparing photographs of their face (e.g., Bertillon, 1890). In recent years however, this task has become increasingly common due to the widespread use of CCTV, digital imagery and social media in criminal proceedings. In response to these technological changes, organisations have developed complex face matching systems to verify identity. Figure 1 illustrates two examples of common unfamiliar face matching systems used to screen for identity fraud when issuing photo-ID documents (e.g., passports, driver's licences; A) and to identify criminals in forensic investigations (B).

In photo-ID document issuance procedures (Figure 1A), an applicant will typically submit an application in person to a facial reviewer, who will then conduct an initial identity check by comparing their appearance to the photograph provided in the application. The application photograph will then be used by a facial recognition algorithm to query a database containing images of current ID document holders. The software returns an array of face images that are most similar to the application photo, as determined by the algorithm. To ensure the application is genuine, a facial reviewer must inspect this array to check that the applicant does not appear amongst the returned faces (see Figure 2B). In cases where the person does appear in the array, this can indicate a fraudulent application, and so the reviewer may refer the case to a specialist team of facial examiners who then conduct a detailed examination of the images.

Similar face matching processes now underpin forensic investigations in many police jurisdictions (Figure 1B; see also Maurer, 2016). For example, in some systems an investigator can submit facial imagery to a centralised service, who then use facial recognition software to search for highly similar identities in national databases of mugshot and driver's licence photos (see perceptuallineup.com; Maurer, 2016). A facial reviewer then inspects the identities returned by the algorithm and sends potential matches to the investigator. If the investigator suspects a matching identity has been found, they may refer

the case to a specialist team of facial examiners to make a formal identity judgment. In cases where this evidence leads to criminal charges, a facial examiner may prepare a forensic report describing the identification process for courtroom proceedings.

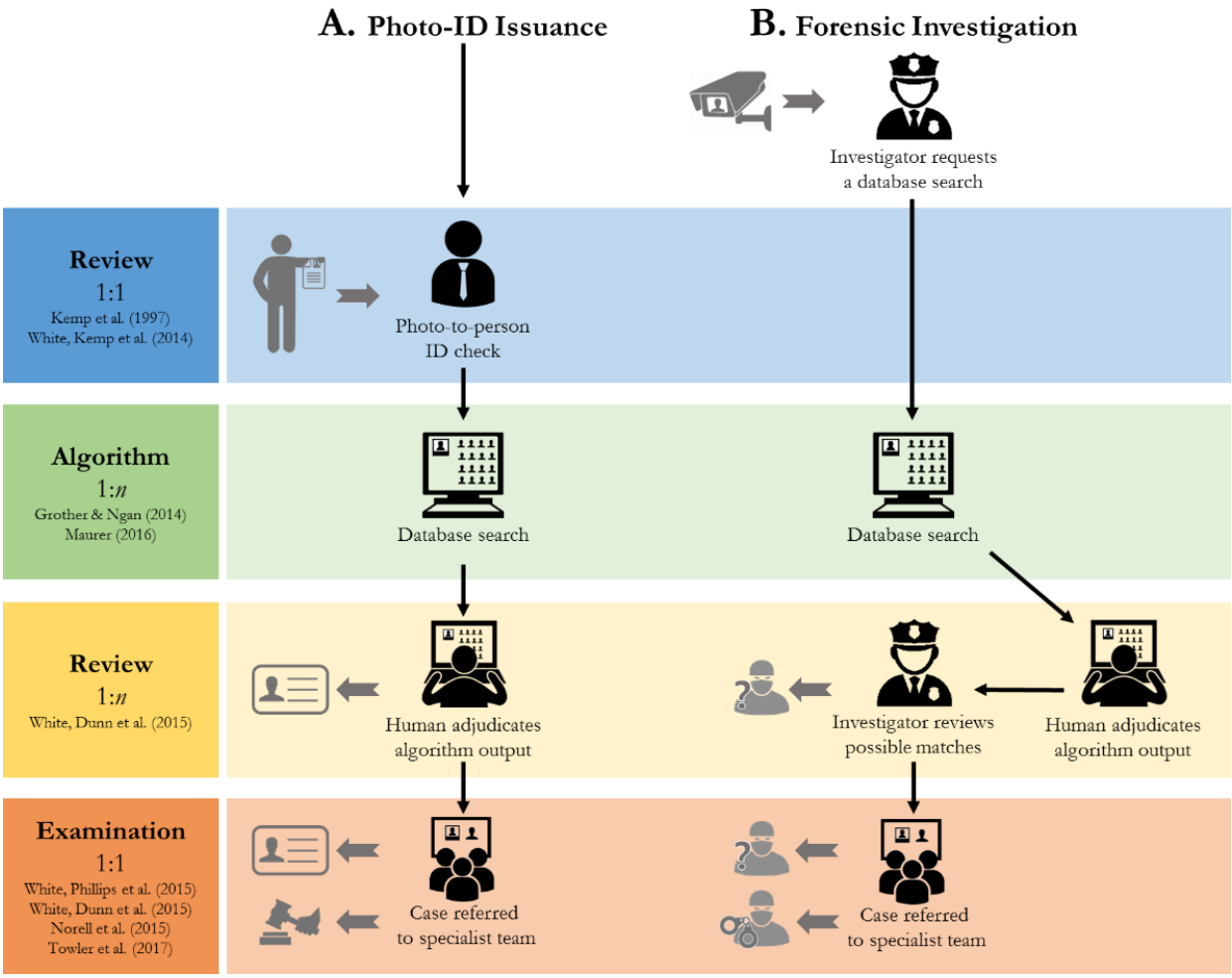


***Figure 1.*** *Example face matching systems used to screen for identity fraud when issuing photo-ID documents (A) and identify criminals in forensic investigations (B). Each system processes a photo of the identity document applicant or CCTV image of the offender and outputs a formal identity decision. See text for details. The figure 'Applied face matching systems' by Alice Towler is available at https://doi.org/10.6084/m9.figshare.4707052.v2 under a Creative Commons Attribution 2.0 licence (CC-BY 2.0).*

The success of face matching systems like these relies on the accuracy of multiple decisions made by humans and computers. Recent research has made significant progress in examining the performance of individual *parts* of these systems: for example, facial

reviewers (White, Dunn, et al., 2015; White, Kemp, Jenkins, Matheson, et al., 2014; White, Phillips, Hahn, Hill, & O'Toole, 2015), facial examiners (Norell et al., 2015; Towler, White, & Kemp, 2017; White, Dunn, et al., 2015; White, Phillips, et al., 2015), and the facial recognition algorithms staff use in their daily work (Grother & Ngan, 2014; Maurer, 2016). As a result, we have a relatively good understanding of how these components function in isolation, but it is still unclear how well these applied face matching systems function as a *whole*.

A field study conducted by the United States Government Accountability Office (Kutz, 2010) provides the only whole-of-system test of unfamiliar face matching performance. Experimenters applied for seven passports using counterfeit documents and information from fictitious or deceased people. Only two of the seven fraudulent applications were detected and refused. This study revealed significant flaws in the passport issuance system, including that facial recognition software was available to analysts during application screening procedures but was not routinely used. Therefore, although the evidence is limited, there appears to be significant scope for improving the face matching accuracy of these systems.

## HUMAN FACE MATCHING PERFORMANCE IN APPLIED SETTINGS

Researchers first tested human performance on unfamiliar face matching tasks in the 1990s. In early work, Kemp et al. (1997) tested the accuracy of supermarket cashiers in verifying the identity of shoppers from photo-ID documents. To the researchers' surprise, error rates were over 50%. Subsequent work using computerised lab-based procedures also found substantial error rates on these tasks, despite constructing tests that afforded participants optimal conditions for matching. For example, Bruce et al. (1999) constructed a one-to-many array task (see Figure 2A) using photos taken with high quality cameras on the same day and in good lighting. Despite these favourable conditions, participants still made 30% errors (see also Bruce et al., 2001; Burton, White, & McNeill, 2010).

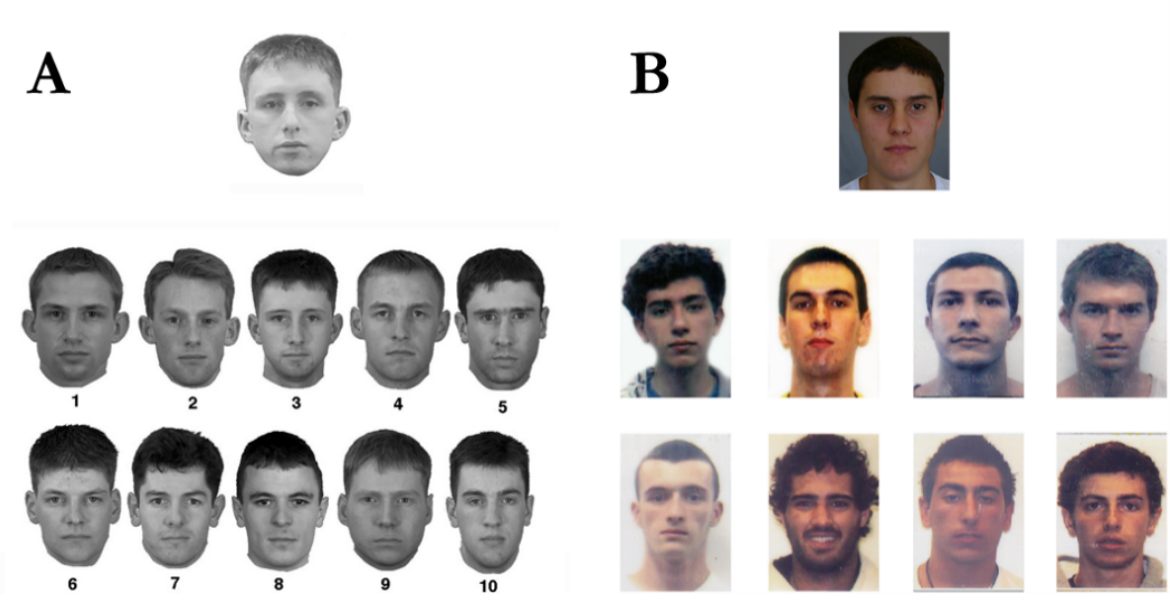


***Figure 2.*** *Examples of one-to-many face matching tasks showing the similarity between lab-based tasks devised in early studies (A; Bruce et al., 1999) and the tasks faced by users of facial recognition software (B; White, Dunn, et al., 2015). In both tasks, participants must decide if the target person pictured above the array is also pictured in the array, and if so to decide which image matches the target. For answers please see the Acknowledgments section at the end of this chapter.*

Importantly, optimal conditions are rarely encountered in applied identification tasks. Comparison images may be captured days, months or even years apart (Bahrick, Bahrick, & Wittlinger, 1975; Bruck, Cavanagh, & Ceci, 1991; Davis & Valentine, 2009; Megreya, Sandford, & Burton, 2013), and decisions may be made under significant time pressure (Bindemann, Fysh, Cross, & Watts, 2016; Wirth & Carbon, in press) and alongside additional biographical information checks (McCaffery & Burton, 2016). Images in forensic investigations are particularly challenging because they are usually captured in unconstrained conditions, such as on low-resolution security cameras and under poor lighting. All of these factors have been shown to impair unfamiliar face matching accuracy (see Hancock, Bruce, & Burton, 2000). This evidence converges on the rather unsettling conclusion that human identification decisions are likely to be particularly error-prone in applied settings, precisely where we need them to be most accurate.

Some studies of professional populations also show high rates of error. White, Kemp, Jenkins, Matheson, et al. (2014) found that Australian passport issuance officers, some of

whom had been working in the role for more than 20 years, fared no better on a test of face matching than a control group of inexperienced and untrained students. More recently, White, Dunn, et al. (2015) tested passport officers using the same one-to-many face matching task they perform in their daily work (Figure 2B), finding that these experienced passport officers made errors on around half of all trials.

Although this research points to high levels of error in operational environments, it is important to note that studies frequently report large individual differences in matching ability, including those of professional groups (see Figure 3; White, Dunn, et al., 2015; White, Kemp, Jenkins, Matheson, et al., 2014). Some people perform no better than chance whereas others perform with near perfect accuracy (e.g., Duchaine & Nakayama, 2006; Russell, Duchaine, & Nakayama, 2009). Recently, it has become clear that differences in face identification ability are stable across repeated testing, and have a large hereditary component (e.g., Shakeshaft & Plomin, 2015; Wilmer et al., 2010).

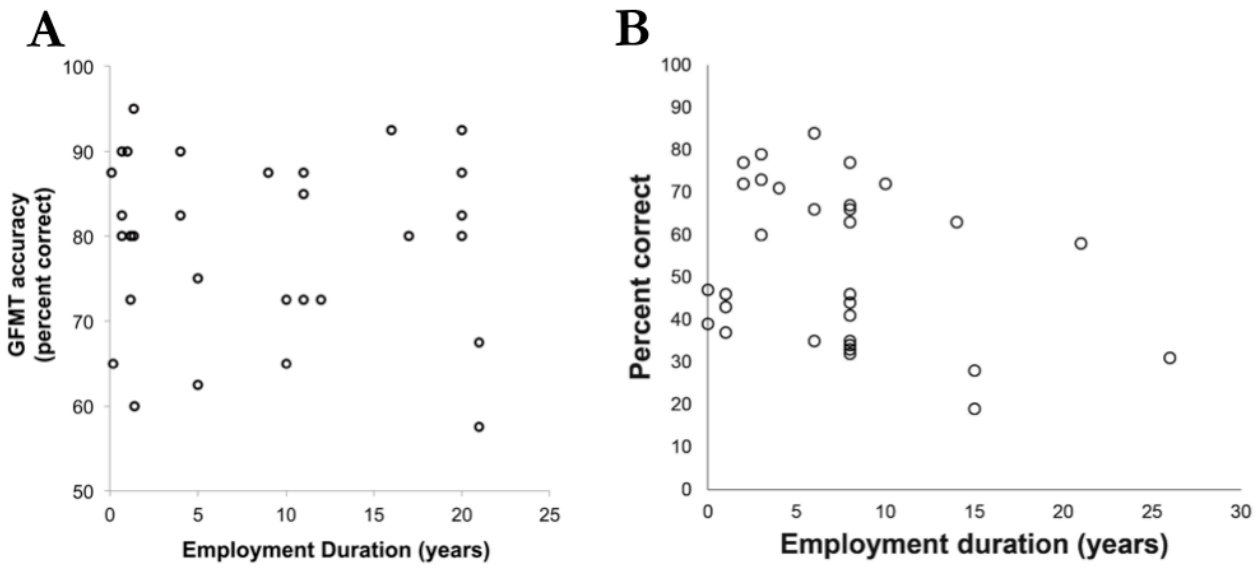


***Figure 3.*** *Individual differences in accuracy of passport officers in two tests of face matching ability. Some individuals perform very well, while others perform very poorly. Further, accuracy is unrelated to experience (B; White, Dunn, et al., 2015; A: White, Kemp, Jenkins, Matheson, et al., 2014).*

Together, this evidence suggests that deliberately recruiting highly skilled individuals for specialist face identification roles can help protect matching systems against poor human

accuracy observed in applied settings (Bobak, Dowsett, & Bate, 2016; White, Kemp, Jenkins, Matheson, et al., 2014). Indeed, some organisations have already adopted this approach, and tests suggest these groups are more accurate than control participants on face matching tests (e.g., Metropolitan Police and the Australian Passport Office; see Robertson, Noyes, Dowsett, Jenkins, & Burton, 2016; White, Dunn, et al., 2015 for details). These recent changes should be encouraging to researchers studying face identification, as they stand as clear evidence that current research is influencing policy and practice in real-world operational settings.

## VARIETIES OF EXPERTISE IN FACE IDENTIFICATION

Selection of specialist staff based on accuracy scores in face identification tests is a relatively new development in applied settings. Nevertheless, facial identification staff have been making identification decisions for many years. One important specialist cohort are forensic facial examiners, who are highly trained and have extensive experience conducting facial comparisons and preparing forensic reports for court. White, Phillips, et al. (2015) tested face matching performance in a group of internationally recognised facial examiners who regularly perform facial comparisons as part of their daily work. They found that the examiners consistently outperformed control subjects across three challenging identification tasks (see also Norell et al., 2015).

Overall however, research paints a rather inconsistent picture of error rates in professional populations. Some professional groups perform no better than untrained students (White, Dunn, et al., 2015; White, Kemp, Jenkins, Matheson, et al., 2014), while others appear to have superior abilities (Davis, Lander, Evans, & Jansari, 2016; Robertson et al., 2016; Towler et al., 2017; White, Dunn, et al., 2015; White, Phillips, et al., 2015; Wirth & Carbon, in press). It is beyond the scope of the present article to provide an overarching account of these findings (but see Noyes, Phillips & O'Toole, this volume). Nevertheless, it is important to note that the different populations in these studies differ in terms of experience, training and recruitment.

Their roles are also quite different. For example, the facial identification staff tested by White, Dunn, et al. (2015) were divided into two distinct groups: facial *reviewers* and facial

*examiners* (also, see Figure 1). This terminology originates in industry training proficiency guidelines (Facial Identification Scientific Working Group, 2011), and reflects a distinct separation of duties within applied face matching systems. Facial reviewers perform high volumes of quick, intuitive identification decisions (White, Dunn, et al., 2015; White, Kemp, Jenkins, Matheson, et al., 2014). By comparison, facial examiners conduct laborious, detail-driven examinations (e.g., in specialist units), and as a result tend to make far fewer decisions over much longer periods of time (Towler et al., 2017; White, Dunn, et al., 2015; White, Kemp, Jenkins, Matheson, et al., 2014; White, Phillips, et al., 2015).

Given the different job descriptions of facial reviewers and examiners, these roles may also require different recruitment procedures. This will be an especially important consideration if reviewer and examiner roles turn out to require different cognitive skills. Indeed, evidence suggests that forensic examiners may approach face matching tasks differently to those that have naturally high levels of accuracy. Specifically, forensic examiners' expertise appears to rely on an analytic, feature-based approach to comparison (Towler et al., 2017; White, Phillips, et al., 2015). This is *qualitatively* different to the enhanced holistic processing underpinning accurate face identification in the population at large (Bobak, Bennetts, Parris, Jansari, & Bate, 2016; Russell et al., 2009), perhaps pointing to separable routes to accurate face matching.

Novices also demonstrate considerable diversity in matching strategies. Many studies report a dissociation between match and non-match trial accuracy, whereby some people are very good at detecting matching identities but not good at detecting non-matching identities, and vice versa (e.g., Megreya & Burton, 2006; Megreya & Burton, 2007; Towler et al., 2017; White, Burton, Jenkins, & Kemp, 2014). Other work shows large individual differences in eye movement patterns to faces (Arizpe, Walsh, Yovel, & Baker, in press), even amongst those with naturally superior ability (Bobak, Parris, Gregory, Bennetts, & Bate, 2016).

Thus, it appears different sets of cognitive strategies can give rise to the same high levels of performance. Interestingly, this heterogeneity in cognitive profiles is also observed at the opposite end of the ability spectrum – in people with Developmental Prosopagnosia, who have specific difficulties in identifying faces (e.g., Dalrymple, Garrido, & Duchaine, 2014; White, Rivolta, Burton, Al-Janabi, & Palermo, 2017). These observations have important implications for optimising the design of face matching systems. For instance, it may not be enough to select individuals purely on the basis of quantitative measures of test performance.

Instead, assignment of individuals to expert teams and expert roles may require a more detailed, qualitative understanding of the perceptual and cognitive abilities of potential recruits. Critically, their abilities should be matched to the task they perform. For example, it may be important to assign some people to tasks that involve very quick decisions and others to tasks that require slower and more deliberative decisions.

Later in the chapter we return to how integrating diversity into system design might improve overall system accuracy. First, we consider the related question of how systems might balance the different strengths of humans and machines to improve overall system performance.

## HUMANS USING COMPUTERS

Forensic analysts have a long history of using digital technology to assist face identification decisions (see Gibelli et al., 2016 for a review of forensic comparison methods). For example, some digital tools automatically generate facial measurements (Bulut & Sevim, 2013; Halberstein, 2001; Ventura, Zacheo, & Pala, 2004), and others facilitate image superimposition (Stavrianos et al., 2012; Vanezis & Brierley, 1996). However, many of these methods show poor validity and reliability in empirical tests, and can actually *impair* identification accuracy (e.g., Davis, Valentine, & Davis, 2010; Kleinberg, Vanezis, & Burton, 2007; Moreton & Morely, 2011; Strathie & McNeill, 2016; Strathie, McNeill, & White, 2012). This stands as an important reminder that as modern artificial intelligence technologies bring further opportunities to assist forensic practice, they should only be adopted once their value has been empirically verified.

The most dramatic technology-led change in face matching systems over recent years has been the introduction of facial recognition algorithms. The accuracy of these algorithms has improved substantially in recent years, and so their utility in applied settings is now unquestioned. Indeed, the best algorithms now outperform novice humans in all but the most challenging conditions (O'Toole, An, Dunlop, & Natu, 2012; Phillips et al., 2011).

One of the key strengths of facial recognition algorithms is their ability to search databases containing millions of images very quickly, and return the most similar images to a human operator (see Figure 1). Although this expands the capability of face matching

systems far beyond what was previously possible, integrating human and computer decision-making in this way may also lead to new sources of identification errors. Demonstrating this, White, Dunn, et al. (2015) tested Australian passport issuance officers using precisely the same facial recognition system used to screen for identity fraud in their daily work. As is typical in these systems, the passport officers checked the output of an algorithm's database search to confirm that the applicant was not present in an array of eight possible matches (see Figure 1A and Figure 2B). On this difficult real-world task using real passport images, passport officers made errors on 50% of trials: considerably higher than the 20-30% errors typically observed in lab-based one-to-many tasks (e.g., Bruce et al., 1999). It is likely that this difference was at least partly due to non-matching faces in these arrays being selected from a very large national database, on the basis of their similarity to the probe image. The introduction of facial recognition algorithms may therefore increase the burden on humans to perform face matching tasks; both in terms of the volume and difficulty of the decisions being made (see also Graves et al., 2011).

Can matching systems be designed to combine human and computer decision-making more effectively? One possibility is to exploit the fact that algorithms and humans rely on different information to determine identity (O'Toole, Abdi, Jiang, & Phillips, 2007), and find different decisions challenging (O'Toole et al., 2012). For example, Rice, Phillips, Natu, An, and O'Toole (2013) showed that humans performed well above chance on a one-to-one matching task in which algorithms scored 100% *incorrect*. In this study, humans were able to find identifying information in the external facial features and other body information that were inaccessible to the algorithms used in the study (see also Kumar, Berg, Belhumeur, & Nayar, 2009).

The accuracy of human-computer decisions can be strengthened by exploiting this diversity in matching strategy through a process known as fusion. Fusion computationally aggregates independent judgments of humans and computers to form a single identification decision. O'Toole et al. (2007) showed that fusing human and algorithm face similarity ratings resulted in almost perfect discrimination on a challenging one-to-one matching task. This approach exploits the disparate strategies of humans and computers so that the strengths of one can help counteract the weaknesses of the other. Implementing fusion in applied settings could therefore go some way to alleviating the unacceptably high error-rate of current human-computer decision-making.

Another potential method for improving human-computer decision-making is to fuse similarity ratings at the level of individual features (i.e., 'feature-level fusion': see Ross & Govindarajan, 2005). For example, recent work by computer scientists has examined the extent to which algorithm-based comparison of individual facial features can be used to support forensic decision-making (e.g., Tome, Vera-Rodriguez, Fierrez, & Ortega-Garcia, 2015; Zeinstra, Veldhuis, & Spreeuwers, 2016). Parallel work has also examined the extent to which human ratings of feature similarity, made by novices and facial examiners, provide a reliable basis for identification decisions (Towler et al., 2017). Both of these feature-based approaches have helped to establish the discriminative value of individual facial features in face matching decisions (converging on the somewhat surprising result that ears are particularly useful for identification). In future work, it may be possible to combine human and algorithm feature-to-feature similarity judgments to improve face matching accuracy, perhaps by using automated annotation to support human decision-making (see Jain, Klare, & Park, 2012).

## HUMANS WORKING TOGETHER

In many applied settings, face-matching decisions are made by groups of individuals, either in sequence or in collaboration with one another. However, little is known about the accuracy of these group decisions. One way to estimate group performance is to use the fusion approach, as described in the previous section, to aggregate responses of people to form nominal groups.

We have used this approach in previous work to estimate the accuracy of both novices and facial examiners working in groups (see Figure 4; Towler et al., 2017; White, Burton, Kemp, & Jenkins, 2013; White, Phillips, et al., 2015). As can be seen in Figure 4, group performance increases as the responses of more individuals are aggregated. These ‘wisdom of the crowd’ effects (see Galton, 1907; Kerr & Tindale, 2004; Surowiecki, 2004) are even more impressive when forensic examiners’ decisions are aggregated. Both White, Phillips, et al. (2015) and Towler et al. (2017) show near ceiling performance by combining decisions of just 3 or 4 expert forensic examiners (see Figure 4B and 4C). This empirical work indicates

that when a few experts are each given the same case, and work on this case independently, their aggregate response is likely to be highly accurate.

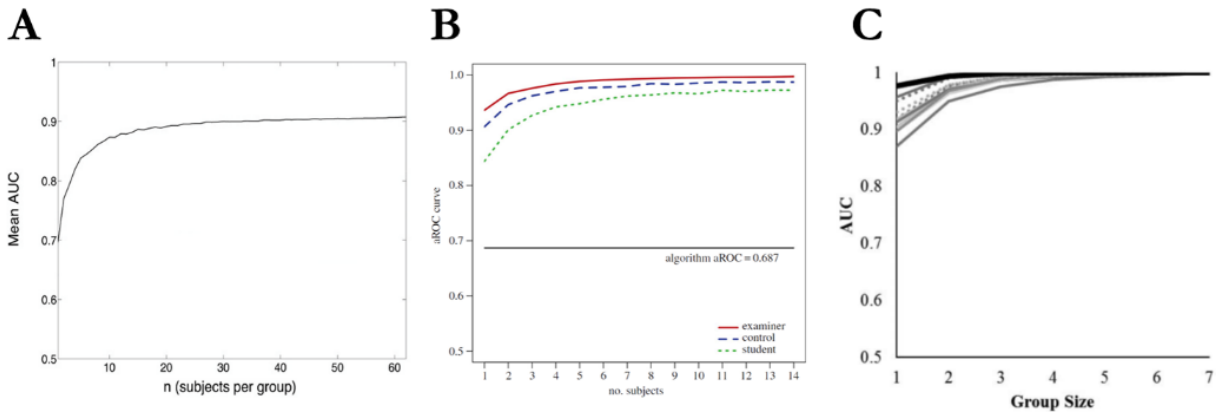


***Figure 4.*** *'Wisdom of the crowd' effects in unfamiliar face matching. Pairwise same/different face matching accuracy is improved by aggregating independent judgements from multiple people. Panel A shows large gains on the Glasgow Face Matching Test (see Burton et al., 2010) by grouping just 8 people (White et al., 2013). Panel B shows similar benefits when grouping decisions by facial examiners, control subjects and untrained students on a challenging test using images captured under unconstrained environmental conditions (White, Phillips, et al., 2015).* The thick dark line in Panel C shows almost perfect discrimination for groups of two or more facial examiners in a more recent study (Towler et al., 2017).

This research also points to a rather straightforward method of improving the accuracy of identification decisions made in applied settings. In many face matching systems, decisions are distributed across networks of individual users based in different physical locations. Simply aggregating the independent responses of these multiple users would likely result in significant improvements to identification accuracy.

However, in some applied situations, staff work together rather than independently to reach joint decisions. Fusion approaches do not adequately model this situation and so perhaps do not provide an indication of the accuracy of these group decisions. An elegant study by Dowsett and Burton (2014) addresses this question. Participants completed three face matching tests in which they decided if pairs of images showed the same person or different people. In the first test, participants worked individually. In the second test, these

participants worked in pairs and collaborated to reach a joint decision. These joint decisions were more accurate than those made in the individual phase and in some cases exceeded performance of the highest performing group member. This finding indicates that collaborative decisions made by teams of analysts in applied settings are likely to be more accurate than if single analysts made the identification decision on their own.

After the collaborative test, Dowsett and Burton again tested participants individually. Importantly, the benefit of paired decision-making carried over into this individual test, with the lower-performing member of each pair showing a significant improvement in this phase relative to the first individual test. The implication of this finding is that people with poor ability may learn successful strategies from those with higher levels of ability. This may be particularly relevant to organisations that currently employ groups of high performing face matchers (i.e., 'super-recognisers': see Davis et al., 2016; Robertson et al., 2016; White, Dunn, et al., 2015) suggesting that recruitment of high performers can provide benefits to performance of existing staff.

## TUNING THE SYSTEM

How can we improve the accuracy of real-world face matching systems? As discussed, current scientific knowledge shows that the largest gains are achieved through recruitment of high performers. Another very promising solution is to aggregate the face matching decisions of multiple individuals. We have recommended elsewhere that these approaches should be the first steps to improving performance in face matching systems (White et al., 2013; White, Kemp, Jenkins, Matheson, et al., 2014).

Nevertheless, it is also important to consider whether other changes can improve their reliability further. One promising avenue is to make best use of avaliabile imagery, by showing multiple images of a target identity to facial comparison staff where possible (Bindemann & Sandford, 2011; White, Burton, Jenkins, & Kemp, 2014; Ritchie & Burton, 2017). Another widely-adopted strategy is to train staff to use a feature-based approach to image comparison. Studies suggest that this can lead to increases in accuracy (e.g., Towler, White, & Kemp, 2017), insofar as participants allocate attention to features that reliably signal identity (Towler, White & Kemp, 2014).

As described in the previous section, system accuracy could also be improved by exploiting the diverse strengths of trained examiners, computers and novices, as diversity in matching strategies appears to generate large benefits in group performance (Hong & Page, 2004; O'Toole et al., 2007). Future research should examine whether recruitment and training of staff can help facilitate optimal combinations of diverse expert strategies. Compared to recruitment of high performers and response aggregation however, the gains from these methods are likely to be more nuanced, and are therefore well suited to 'tuning' face matching systems.

In designing optimal systems, it is also important to anticipate changes associated with assigning face matching experts to teams. If current trends continue, it is quite likely that many organisations will deploy specialist teams whose daily work will consist almost entirely of face identification decisions. However, research shows that university students' accuracy steadily declines when they perform face matching tasks for an extended period of time (Alenezi, Bindemann, Fysh, & Johnston, 2015).

One way to counter this is to provide feedback on decision accuracy. Feedback is critical for learning across a range of domains (see Hattie & Timperley, 2007), and has been shown to benefit face matching performance (White, Kemp, Jenkins, & Burton, 2014). Critically, feedback may help maintain vigilance on face matching tasks. Alenezi and Bindemann (2013) found that the provision of feedback maintained accuracy levels, preventing a decline observed when feedback was not provided. In many applied settings, it may prove difficult to provide accurate feedback on casework because ground truth in operational scenarios is typically unknown. Nevertheless, it may be possible to introduce feedback into these systems by using verified images that are already stored in the system. For example, a previous passport photo of the applicant could be inserted into the face array reviewed by a passport issuance officer (see Figure 2B). The system could then provide the officer with immediate feedback on the accuracy of their decision.

A further challenge in real-world settings is that staff rarely encounter critical cases. For example, identity fraud is estimated to occur in only 0.25% of passport applications (BBC, 2007). However, research indicates that rare targets are extremely difficult to detect and are often missed (Wolfe, Horowitz, & Kenner, 2005). Papesh and Goldinger (2014) investigated target prevalence effects in a face matching task by presenting non-matching identities on 10% or 50% of trials. They found that participants were twice as likely to miss the non-

matching identities in the low prevalence condition compared to the high prevalence condition, making almost 50% errors (cf. Bindemann, Avetisyan, & Blackwell, 2010). Considering that the prevalence rate of critical cases in applied settings is estimated to be much lower than the 10% used in this study, low target prevalence is likely to be a significant source of error in applied face matching tasks.

In addition to maintaining vigilance, feedback could also help counteract these low target prevalence effects. In a visual search task, Wolfe et al. (2007) demonstrated that short bursts of feedback training on high prevalence trials protected against a drop in detection on subsequent low prevalence tests where no feedback was provided. A similar approach, where casework is interleaved with short sessions of high prevalence feedback training, using cases for which the ground truth is known by system administrators, may minimise the low target prevalence effects observed in applied face matching tasks (see Papesh & Goldinger, 2014).

Finally, it will be important to consider the broader workflow of human analysts in future research. Should identity checks be separated from other tasks, such as authentication of signatures and birthdates? Indeed, research suggests that cues to fraud in biographical information are often missed when checked alongside identity (Kutz, 2010; McCaffery & Burton, 2016). Balancing the optimal distribution of work amongst human analysts will be an important challenge for researchers and practitioners in this area, and is a particularly relevant question given the move towards creating specialist teams of face matchers.

## CONCLUSIONS

Face matching systems that incorporate human and computer decision-making are used to verify the identity of unknown people in a variety of applied settings. In this chapter, we have focussed on face matching systems used to screen photo-ID document applications for identity fraud, and those used to support forensic investigations. However, similar systems are used in many other defence, security and surveillance operations.

A significant amount of work has investigated the accuracy of individual components of these systems. However, because whole-of-system evaluation studies are not avaliable, the accuracy of face identification processes, in general, are unknown. Evidence does show that facial recognition software commonly integrated into these systems is relatively accurate, at

least for high quality imagery. Critically however, the software itself does not make the final decision. Rather, humans are required to review the output of algorithms and make final identity judgments. Current evidence shows that human analysts are surprisingly error-prone on these tasks, including those with extensive experience working within these systems.

Based on the studies reviewed in this chapter, we conclude that the application of psychological research can lead to substantial improvements in face matching systems. Future research should consider the complex interactions between decision-makers in face matching systems, and seek to distribute decision-making in a way that optimises both system accuracy and efficiency. Achieving this aim will require collaboration between psychologists, computer scientists and system administrators. Unreliable face matching systems pose substantial risks: potentially leading to the issuance of fraudulent identity documents, or wrongful convictions of innocent people. Therefore, we hope that a system-level approach to this problem will help to promote safer and fairer societies in the future.

## ACKNOWLEDGMENTS

Preparation of this chapter was supported by Australian Research Council Linkage Project grants to RK and DW in partnership with the Australian Passport Office (LP130100702, LP160101523). The answers to face matching arrays in Figure 2 are: A = target number 3; B = target absent.

## REFERENCES

Alenezi, H. M., & Bindemann, M. (2013). The effect of feedback on face-matching accuracy. *Applied Cognitive Psychology, 27*, 735-753.

Alenezi, H. M., Bindemann, M., Fysh, M. C., & Johnston, R. A. (2015). Face matching in a long task: Enforced rest and desk-switching cannot maintain identification accuracy. *PeerJ, 3*, e1184.

Arizpe, J., Walsh, V., Yovel, G., & Baker, C. I. (in press). The categories, frequencies, and stability of idiosyncratic eye-movement patterns to faces. *Vision Research*.

Bahrick, H. P., Bahrick, P. O., & Wittlinger, R. P. (1975). Fifty years of memory for names and faces: A cross-sectional approach. *Journal of Experimental Psychology: General, 104*, 54-75.

BBC. (2007). 10,000 passports go to fraudsters. Retrieved from http://news.bbc.co.uk/1/hi/uk_politics/6470179.stm.

Bertillon, A. (1890). *La photographie judiciaire avec un appendice sur la classification et l'identification anthropométrique. Bibliothèque photographique*. Gauthier-Villars et fils: Paris.

Bindemann, M., Avetisyan, M., & Blackwell, K. (2010). Finding needles in haystacks: Identity mismatch frequency and facial identity verification. *Journal of Experimental Psychology: Applied, 16*, 378-386.

Bindemann, M., Fysh, M., Cross, K., & Watts, R. (2016). Matching Faces Against the Clock. *i-Perception, 7*, 2041669516672219.

Bindemann, M., & Sandford, A. (2011). Me, myself, and I: Different recognition rates for three photo-IDs of the same person. *Perception, 40*, 625-627.

Bobak, A. K., Bennetts, R. J., Parris, B. A., Jansari, A., & Bate, S. (2016). An in-depth cognitive examination of individuals with superior face recognition skills. *Cortex, 82*, 48-62.

Bobak, A. K., Dowsett, A. J., & Bate, S. (2016). Solving the border control problem: Evidence of enhanced face matching in individuals with extraordinary face recognition skills. *PLoS ONE, 11*, 1-13.

Bobak, A. K., Parris, B. A., Gregory, N. J., Bennetts, R. J., & Bate, S. (2016). Eye-movement strategies in developmental prosopagnosia and "super" face recognition. *The Quarterly Journal of Experimental Psychology, 70*, 201-217.

Bruce, V., Henderson, Z., Greenwood, K., Hancock, P. J. B., Burton, A. M., & Miller, P. (1999). Verification of face identities from images captured on video. *Journal of Experimental Psychology: Applied, 5*, 339-360.

Bruce, V., Henderson, Z., Newman, C., & Burton, A. M. (2001). Matching identities of familiar and unfamiliar faces caught on CCTV images. *Journal of Experimental Psychology: Applied, 7*, 207-218.

Bruce, V., & Young, A. W. (2012). *Face Perception*. Hove: Taylor and Francis Ltd.

Bruck, M., Cavanagh, P., & Ceci, S. J. (1991). Fortysomething: Recognizing faces at one's 25th reunion. *Memory and Cognition, 19*, 221-228.

Bulut, Ö., & Sevim, A. (2013). The efficiency of anthropological examinations in forensic facial analysis. *Turkish Journal of Police Studies, 15*, 139-158.

Burton, A. M., White, D., & McNeill, A. (2010). The Glasgow Face Matching Test. *Behavior Research Methods, 42*, 286-291.

Burton, A. M., Wilson, S., Cowan, M., & Bruce, V. (1999). Face recognition in poor-quality video: Evidence from security surveillance. *Psychological Science, 10*, 243-248. Calder, A. J., Rhodes, G., Johnson, M. H., & Haxby, J. V. (Eds.). (2011). *The Oxford Handbook of Face Perception*. Oxford: Oxford University Press.

Dalrymple, K. A., Garrido, L., & Duchaine, B. (2014). Dissociation between face perception and face memory in adults, but not children, with developmental prosopagnosia. *Developmental Cognitive Neuroscience, 10*, 10-20.

Davis, J. P., Lander, K., Evans, R., & Jansari, A. (2016). Investigating predictors of superior face recognition ability in police super-recognisers. *Applied Cognitive Psychology, 30*, 827-840.

Davis, J. P., & Valentine, T. (2009). CCTV on trial: Matching video images with the defendant in the dock. *Applied Cognitive Psychology, 23*, 482-505. Davis, J. P., Valentine, T., & Davis, R. E. (2010). Computer assisted photo-anthropometric analyses of full-face and profile facial images. *Forensic Science International, 200*, 165-176. Dowsett, A. J., & Burton, A. M. (2014). Unfamiliar face matching: Pairs out-perform individuals and provide a route to training. *British Journal of Psychology, 106*, 433-445.

Duchaine, B. C., & Nakayama, K. (2006). The Cambridge Face Memory Test: Results for neurologically intact individuals and an investigation of its validity using inverted face stimuli and prosopagnosic participants. *Neuropsychologia, 44*, 576-585.

Facial Identification Scientific Working Group. (2011). Guidelines and recommendations for facial comparison training to competency (Version 1.1). Retrieved from https://www.fiswg.org/document/viewDocument?id=22.

Galton, F. (1907). Vox populi. *Nature, 75*, 450-451.

Gibelli, D., Obertová, Z., Ritz-Timme, S., Gabriel, P., Arent, T., Ratnayake, M., . . . Cattaneo, C. (2016). The identification of living persons on images: A literature review. *Legal Medicine, 19*, 52-60.

Graves, I., Butavicius, M. A., MacLeod, V., Heyer, R., Parsons, K., Kuester, N., . . . Johnson, R. (2011). The role of the human operator in image-based airport security technologies. *Studies in Computational Intelligence, 338*, 147-181.

Grother, P., & Ngan, M. (2014). *Face Recognition Vendor Test (FRVT)*. Retrieved from Information Access Division, National Institute of Standards and Technology.

Halberstein, R. A. (2001). The application of anthropometric indices in forensic photography: three case studies. *Journal of Forensic Sciences, 46*, 1438-1441.

Hancock, P. J., Bruce, V., & Burton, A. M. (2000). Recognition of unfamiliar faces. *Trends in Cognitive Sciences, 4*, 330-337.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*, 81-112.

Hole, G. J., & Bourne, V. (2010). *Face Processing: Psychological, Neuropsychological, and Applied Perspectives*. Oxford: Oxford University Press.

Hong, L., & Page, S. E. (2004). Groups of diverse problem solvers can outperform groups of high-ability problem solvers. *Proceedings of the National Academy of Sciences of the United States of America, 101*, 16385-16389.

Jain, A. K., Klare, B., & Park, U. (2012). Face matching and retrieval in forensics applications. *IEEE MultiMedia, 19*, 20-28.

Kemp, R. I., Towell, N. A., & Pike, G. E. (1997). When seeing should not be believing: Photographs, credit cards and fraud. *Applied Cognitive Psychology, 11*, 211-222.

Kerr, N. L., & Tindale, R. S. (2004). Group performance and decision making. *Annual Review of Psychology, 55*, 623-655.

Kleinberg, K. F., Vanezis, P., & Burton, A. M. (2007). Failure of anthropometry as a facial identification technique using high-quality photographs. *Journal of Forensic Sciences, 52*, 779-783.

Kumar, N., Berg, A. C., Belhumeur, P. N., & Nayar, S. K. (2009). *Attribute and simile classifiers for face verification.* Paper presented at the IEEE 12th international conference on computer vision.

Kutz, G. (2010). *Undercover tests show passport issuance process remains vulnerable to fraud*. Testimony before the Subcommittee on Terrorism and Homeland Security, Committee on the Judiciary, U.S. Senate.

Maurer, D. C. (2016). *Face recognition technology: FBI should better ensure privacy and accuracy*. Report to the ranking member, subcommittee on privacy, technology and the law, committee on the judiciary, U.S. Senate.

McCaffery, J. M., & Burton, A. M. (2016). Passport checks: Interactions between matching faces and biographical details. *Applied Cognitive Psychology, 30*, 925-933.

Megreya, A. M., & Burton, A. M. (2006). Unfamiliar faces are not faces: Evidence from a matching task. *Memory and Cognition, 34*, 865-876.

Megreya, A. M., & Burton, A. M. (2007). Hits and false positives in face matching: A familiarity-based dissociation. *Perception and Psychophysics, 69*, 1175-1184.

Megreya, A. M., Sandford, A., & Burton, A. M. (2013). Matching face images taken on the same day or months apart: The limitations of photo ID. *Applied Cognitive Psychology, 27*, 700-706.

Moreton, R., & Morely, J. (2011). Investigation into the use of photoanthropometry in facial image comparison. *Forensic Science International, 212*, 231-237.

Norell, K., Brorsson Lathen, K., Bergstrom, P., Rice, A., Natu, V., & O'Toole, A. J. (2015). The effect of image quality and forensic expertise in facial image comparisons. *Journal of Forensic Sciences, 60*, 331-340.

O'Toole, A. J., Abdi, H., Jiang, F., & Phillips, P. J. (2007). Fusing face-verification algorithms and humans. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, 37*, 1149-1155.

O'Toole, A. J., An, X., Dunlop, J., & Natu, V. (2012). Comparing face recognition algorithms to humans on challenging tasks. *ACM Transactions on Applied Perception*, 1-15.

Papesh, M. H., & Goldinger, S. D. (2014). Infrequent identity mismatches are frequently undetected. *Attention, Perception, and Psychophysics, 76*, 1335-1349.

Phillips, P. J., Beveridge, J. R., Draper, B. A., Givens, G., O'Toole, A. J., Bolme, D. S., . . . Weimer, S. (2011). *An introduction to the good, the bad, & the ugly face recognition challenge problem.* Paper presented at the IEEE international conference on automatic face and gesture recognition.

Rice, A., Phillips, P. J., Natu, V., An, X., & O'Toole, A. J. (2013). Unaware person recognition from the body when face identification fails. *Psychological Science, 24*, 2235-2243.

Ritchie, K. L., & Burton, A. M. (2017). Learning faces from variability. *The Quarterly Journal of Experimental Psychology*, *70*, 897-905.

Robertson, D. J., Noyes, E., Dowsett, A. J., Jenkins, R., & Burton, A. M. (2016). Face recognition by metropolitan police super-recognisers. *PLoS ONE, 11*.

Ross, A. A., & Govindarajan, R. (2005). Feature level fusion of hand and face biometrics. *Defense and Security* (pp. 196-204): International Society for Optics and Photonics.

Russell, R., Duchaine, B., & Nakayama, K. (2009). Super-recognizers: People with extraordinary face recognition ability. *Psychonomic Bulletin and Review, 16*, 252-257.

Shakeshaft, N. G., & Plomin, R. (2015). Genetic specificity of face recognition. *Proceedings of the National Academy of Sciences, 112*, 12887-12892.

Stavrianos, C., Zouloumis, L., Papadopoulos, C., Emmanouil, J., Petalotis, N., & Tsakmalis, P. (2012). Facial mapping: Review of current methods. *Research Journal of Medical Sciences, 6*, 77-82.

Strathie, A., & McNeill, A. (2016). Facial wipes don't wash: Facial image comparison by video superimposition reduces the accuracy of face matching decisions. *Applied Cognitive Psychology, 30*, 504-513.

Strathie, A., McNeill, A., & White, D. (2012). In the Dock: Chimeric Image Composites Reduce Identification Accuracy. *Applied Cognitive Psychology, 26*, 140-148.

Surowiecki, J. (2004). *The wisdom of crowds: Why the many are smarter than the few*. New York: Little Brown.

Tome, P., Vera-Rodriguez, R., Fierrez, J., & Ortega-Garcia, J. (2015). Facial soft biometric features for forensic face recognition. *Forensic Science International, 257*, 271-284.

Towler, A., White, D., & Kemp, R. I. (2014). Evaluating training methods for facial image comparison: The face shape strategy does not work. *Perception*, *43*, 214-218.

Towler, A., White, D., & Kemp, R. I. (2017). Evaluating the feature comparison strategy for forensic face identification. *Journal of Experimental Psychology: Applied, 23*, 47-58.

Vanezis, P., & Brierley, C. (1996). Facial image comparison of crime suspects using video superimposition. *Science & Justice, 36*, 27-33.

Ventura, F., Zacheo, A. V., A., & Pala, A. (2004). Computerised anthopomorphometric analysis of images: A case report. *Forensic Science International, 146*, S211-S213.

White, D., Burton, A. M., Jenkins, R., & Kemp, R. I. (2014). Redesigning photo-ID to improve unfamiliar face matching performance. *Journal of Experimental Psychology: Applied, 20*, 166-173.

White, D., Burton, A. M., Kemp, R. I., & Jenkins, R. (2013). Crowd effects in unfamiliar face matching. *Applied Cognitive Psychology, 27*, 769-777.

White, D., Dunn, J. D., Schmid, A. C., & Kemp, R. I. (2015). Error rates in users of automatic face recognition software. *PLoS ONE, 10*, 1-14.

White, D., Kemp, R. I., Jenkins, R., & Burton, A. M. (2014). Feedback training for facial image comparison. *Psychonomic Bulletin and Review, 21*, 100-106.

White, D., Kemp, R. I., Jenkins, R., Matheson, M., & Burton, A. M. (2014). Passport officers' errors in face matching. *PLoS ONE, 9*, 1-6.

White, D., Phillips, P. J., Hahn, C. A., Hill, M., & O'Toole, A. J. (2015). Perceptual expertise in forensic facial image comparison. *Proceedings of the Royal Society of London B: Biological Sciences, 282*, 1814-1822.

White, D., Rivolta, D., Burton, A. M., Al-Janabi, S., & Palermo, R. (2017). Face matching impairment in developmental prosopagnosia. *The Quarterly Journal of Experimental Psychology, 70*, 287-297.

Wilmer, J. B., Germine, L., Chabris, C. F., Chatterjee, G., Williams, M., Loken, E., . . . Duchaine, B. (2010). Human face recognition ability is specific and highly heritable. *Proceedings of the National Academy of Sciences, 107*, 5238-5241.

Wirth, B. E., & Carbon, C. C. (in press). An easy game for frauds? Effects of professional experience and time pressure on passport-matching performance. *Journal of Experimental Psychology: Applied*.

Wolfe, J. M., Horowitz, T. S., & Kenner, N. M. (2005). Rare items often missed in visual searches. *Nature, 435*, 439-440.

Wolfe, J. M., Horowitz, T. S., Van Wert, M. J., Kenner, N. M., Place, S. S., & Kibbi, N. (2007). Low target prevalence is a stubborn source of errors in visual search tasks. *Journal of Experimental Psychology: General, 136*, 623-638.

Zeinstra, C., Veldhuis, R., & Spreeuwers, L. (2016). *Discriminating power of FISWG characteristic descriptors under different forensic use cases*. Paper presented at the 2016 international conference of the biometrics special interest group (BIOSIG).