The Impact of Age on Human Face Matching Performance with Images of Children

Eden Clothier

School of Psychology

University of Adelaide

October 2019

This thesis is submitted in partial fulfilment of the Honours degree of Psychological Science (Honours)

Word Count: 9385

Table of Contents

List of Figures

Abstract

The ability to accurately conduct facial comparisons with images of children is instrumental for various applied purposes, such as the prevention of child trafficking. Despite this, previous research has shown that one-to-one face matching is especially challenging on images of young children and those which show significant age-related facial changes. However, limited research has tested performance on more operationally challenging face matching tasks (one-to-eight) using images of children. This study used a one-to-eight task to explore the extent that performance varied across three childhood age groups (0-5, 5-10 and 10-15) with a 5-year age variation between target and comparison images. Participants (N = 42) completed 120 randomised face matching trials and their accuracy and confidence ratings were analysed. Results found the worst performance for the 0-5-year age group (16% accuracy), compared to 5-10 (26%) and 10-15 (30%) groups, suggesting that performance increased with age. Additionally, no significant differences were found between target-present and target-absent trials. The alarmingly high error rates found in all conditions highlights the importance of understanding and improving performance. Future research should continue to build upon these findings by testing generalisability to practitioner populations, exploring individual differences and evaluating ways to improve performance.

Declaration

This thesis contains no material which has been accepted for the award of any other degree of diploma in any University, and, to the best of my knowledge, this thesis contains no material previously published except where due reference is made. I give permission for the digital version of this thesis to be made available on the web, via the University of Adelaide's digital thesis repository, the Library Search and through web search engines, unless permission has been granted by the School to restrict access for a period of time

Eden Clothier

Oct 2019

Acknowledgements

First and foremost, I would like to thank my supervisors for their support this year, I couldn’t have asked for a better group of people to work with. To my primary supervisor, Dr. Dana Michalski, your feedback and encouragement has been instrumental to getting me to the finish line. The passion you put into your work is infectious and has kept me motivated even at the most stressful times. To my internal supervisor, A/Prof Carolyn Semmler, I truly admire your enthusiasm for research and statistics – I appreciate your support and patience, especially as I wrapped my head around the models. To Dr. Rebecca Heyer, thank you for igniting my interest in this area and leading me down this path. I value all your insights and the assistance you have provided throughout the year. I would also like to acknowledge the DST team for assisting with data collection and being so accommodating throughout the year.

Thank you to everyone else who has helped or supported me. To Britt and Gemma - your advice this year has been so valuable. Thank you for making time for me, sharing your previous experiences and providing perspective. To my lab family; Carlos, Athina and Carly, thank you for adopting me as an honorary lab member and all the laughs down on level 2; and for starting the thesis countdown that made me feel a sense of impending doom and kept me on my toes. To my graduate entry family; Meg – your wisdom and advice has kept me grounded. Thank you for always knowing the right things to say. Bart – your confidence, humour and drive (read: competitiveness) both stresses me out and pushes me to do better. You both motivate me, challenge me and inspire me to be my best self every day.

Finally, to Shaun – without you, I’m not sure this would have been possible. I appreciate your endless confidence and understanding. Thank you for being there for me, making me coffee, reading drafts at all hours, listening to me vent, and knowing when I just needed a break. Thank you.

CHAPTER 1:

Introduction

1.1 Rationale

Humans perform facial comparisons effortlessly every day to recognise friends, family, and other well-known figures (Garcia-Zurdo, Frowd & Manzanero, 2018). In a range of applied settings, these comparisons can also be used to verify and review identities they may not be familiar with. While there is extensive research evaluating human and algorithm performance on facial comparison tasks using images of adults, there is increasing interest in understanding performance using images of children (Michalski, 2017; Srinivas, Ricanek, Michalski, Bolme & King, 2019). Children exhibit markedly different facial changes than adults, making the ability to establish the identity of children more challenging in operational settings, such as general passport control, identification of missing children and radicalised minors, and prevention of child trafficking and exploitation (Jain, Nandakumar & Ross, 2016; Lander, Bruce & Bindemann, 2018; McCaffery, Robertson, Young & Burton, 2018).

An estimated 400,000 children are trafficked across international borders annually (U.S. Department of State, 2006). Effort has been made to combat this by requiring infants and children to have their own passport for travel purposes. However, this relies on the assumption that practitioners can make accurate identifications of children (Kramer, Mulgrew & Reynolds, 2018a). Despite the prevalence of using the face as biometric identification, research has consistently found poor performance on facial comparison tasks (Bruce et al., 1999; Lander et al., 2018; Megreya & Burton, 2008; Meissner, Susa & Ross, 2013; White Kemp, Jenkin, Matheson & Burton, 2014), especially for images that show age-related facial changes (Matthews & Mondloch, 2018). These findings have informed recent interest in the effects of ageing on face matching performance using images of children (Deb, Nain & Jain, 2018; Lander et al., 2018, Michalski, 2017). Due to the implications for child protection, it is essential that robust and accurate facial recognition systems and training are in place (Deb et al., 2018; Kramer et al., 2018a; White, Kemp, Jenkins & Burton, 2014). However, limited research means agencies are not yet equipped with empirical evidence to make informed decisions surrounding deficits to performance.

1.2 Overview of Face Matching

Face matching refers to the task of comparing two or more faces to determine whether they match a target identity (Facial Identification Scientific Working Group [FISWG], 2012). Faces have become one of the most common biometric modalities used in person identification; appearing on official documents, such as licenses, security passes and passports (Jain et al., 2016; Heyer, Semmler & Hendrickson, 2018). The use of faces as identification assumes that practitioners are able to make accurate judgements on face matching tasks. However, research shows that this is a difficult and error-prone task, even for passport officers (Burton, White & McNeill, 2010; White et al., 2014). These findings are often surprising and suggest that while humans exhibit expertise with identifying faces, this drops as familiarity with a person decreases and the number of faces for comparison increases (Graves et al., 2011; Heyer et al., 2018).

1.2.1 Familiar and unfamiliar faces. As the inherent human ability to conduct facial comparisons is strongly related to our familiarity with the faces, face matching performance can be split into two distinct categories; familiar and unfamiliar (Johnson, Dzirawiec, Ellis & Morton, 1991).

Familiar face matching refers to the comparison of faces we are experienced with, such as friends, family and other acquaintances; as well as notable figures around the world and in popular culture (Megreya & Burton, 2006; O'Toole, Phillips & Narvekar, 2008). People are exceptionally good at recognising these faces quickly and with high accuracy, even when faced with variations in image quality, illumination, and pose (Burton, Miller, Bruce, Hancock & Henderson, 2001; Phillips & O'Toole, 2014; Tistarelli, Yadav, Vatsu & Singh, 2013).

Unfamiliar face matching refers to the comparison of faces we are less experienced with and may have never seen. This is the type of face matching most often performed in passport control and law enforcement (FISWG, 2012).

Research has consistently shown higher performance when matching familiar faces (Kramer, Young & Burton, 2018; Richie et al., 2015). However, due to the ease of familiar face matching in our daily lives, it can be hard to appreciate the error-prone nature of unfamiliar face matching (McCaffery et al., 2018; Ritchie et al., 2015; White et al., 2014). A difference in cognitive and perceptual processes used for each type of matching may account for some of the difficulty we have when comparing unfamiliar faces (Megreya & Burton, 2006; Papesh, 2018). For example, people may rely on a holistic view of the face and focus on internal features, such as the eyes and nose, when matching familiar faces. Unfamiliar face matching may be more influenced by external features, such as hair and face shape, which are susceptible to change with time (Garcia-Zurdo et al., 2018; Kemp, Caon, Howard & Brooks, 2016; Young, Hay, McWeeny, Flude & Ellis, 1985). Additionally, our perceptions of what changes to appearance might be considered 'possible and permissible' based on our prior experience with faces can negatively influence accuracy on face matching trials both when there is and is not a match (Bruce, 1994; Michalski, 2017).

1.2.2 Target-present and target-absent face matching trials. Face matching tasks can be broken down into target-present and target-absent trials. When a face is compared to a watch-list of pre-defined persons of interest or checked for identity fraud, the goal is to confirm that none of the comparison images are of the applicant (Kemp et al., 2016). If there is no match to the target, the trial is target-absent. Alternatively, if a comparison image correctly matches the target identity, the trial is target-present.

Research has indicated that the type of trial has an impact on performance. For example, past research has found that participants find target-absent trials particularly challenging, finding an 11% drop in accuracy between target-present (88%) and target-absent (77%) trials (Megreya & Burton, 2006). Target-absent trials can be difficult due to the likelihood of images exhibiting high visual similarity to the target (Kemp et al., 2016). When a target is absent, what we believe to be 'possible and permissible' facial changes becomes broader and can be used to explain variations between images (Bruce, 1994; Michalski, 2017). Change to the proportion of target-present trials may also affect accuracy, as visual search studies suggest that when target stimuli (in this case, identity matches) occur infrequently, people are more likely to miss them (Stephens, Semmler & Sauer, 2017; Wolfe, Horowitz & Kenner, 2005; Wolfe et al., 2007). Finally, performance on both types of trials can also be influenced by the number of faces being compared, with performance being shown to decrease when more faces are added to the comparison process (Bruce et al., 1999; Heyer et al., 2018; White et al., 2015).

1.2.3 One-to-one and one-to-many face matching. Unfamiliar face matching can be further broken into two types, one-to-one and one-to-many; distinct in the number of images presented for comparison.

One-to-one tasks involve verifying the identity of a target by comparing it to another face to decide whether two photos (or a photo and a live face) depict the same person (FISWG, 2012; Jain et al., 2016). For passport control, this often involves either comparing a submitted applicant photo against a previous passport image, or comparing a live face to their passport image (Kemp et al., 2016; Zeng Ling, Latecki, Fitzhugh & Guo, 2012).

One-to-many tasks are more complex and involve determining an identity by comparing it against a group (known as a candidate list) of possible matches (Graves et al., 2011; Heyer, Chong & Semmler, 2019; Jain et al., 2016). One-to-many trials are used operationally to prevent identity fraud or identify persons of interest (White, Dunn, Schmid & Kemp, 2015). Put simply, one-to-one tasks are used to verify identity, while one-to-many are used to identify a target individual by reviewing a range of possibilities.

Candidate lists often consist of images chosen by a facial recognition algorithm that uses a probe image of a target identity to search and compare to a larger database, filtering out ‘non-matches’ based on various thresholds and parameters, and leading to an output of images that closely resemble the target identity (Graves et al., 2011). The generated candidate lists can be likened to and treated like a visual search array, which human practitioners must search to check for the presence or absence of a match to the target identity (Heyer et al., 2018). The complexity of one-to-many tasks and high pressure, fast-paced operational environments can lead to an increase in error rates and processing time (Heyer et al., 2018; Phillips, 2011).

1.3 Face Matching Operations and Performance

1.3.1 Facial recognition algorithms. The increasing use of automated facial recognition systems in both private and government industries has led to substantial advances in algorithm performance and a wider prevalence in face matching literature (Heyer & Semmler, 2013; White et al., 2015).

The highest performing algorithms now perform better than humans in face matching tasks that are of easy or intermediate difficulty (Fysh & Bindemann, 2018). Even for controlled front-facing images with variations in appearance, like those seen in a passport context, these systems are now comparable to humans (O’Toole et al., 2008; Phillips & O’Toole, 2014). However, while studies have found that one-to-one face matching systems have improved significantly; one-to-many matching still needs improvement (Bone & Blackburn, 2003; Grother, Quinn & Phillips, 2011).

Even with significant advancements, these systems are not infallible. Non-ideal conditions can challenge algorithms, and even the highest performing algorithms are susceptible to some variations to appearance (Lander et al., 2018). Consequently, humans remain integral to the identification process to verify and interpret the output of these systems (Graves et al., 2011; Heyer et al., 2018).

1.3.2 The human-computer interaction. Algorithms are often used to augment the comparison process (Heyer, MacLeod, Carter, Semmler & Ma-Wyatt, 2017; Michalski, Yiu & Malec, 2018). However, evaluations of algorithms often discount the effect of human error, despite human operators being the ultimate decision-makers (Heyer & Semmler, 2013; Lander et al., 2018). If human performance is poor, there is a higher rate of error, even when algorithms generate accurate, high-quality candidate lists. Error rates in target-absent trials suggest that candidate lists selected by algorithms are easily mistaken to be containing the target identity. Therefore, it is likely that the capacity of automated systems to search large databases comes at the cost of increasing difficulty for human practitioners (White et al., 2015).

Humans and systems work together to maintain the highest efficiency of facial recognition. A skill that humans currently have over algorithms is their affinity for detecting expressions and emotions and thus, the ability to detect duress that systems cannot (Michalski, 2017). This skill provides a significant advantage for the prevention of trafficking, as victims may exhibit signs of duress during the identification process. Humans are also more flexible than systems, making them better equipped to respond to unforeseen issues and changes compared with systems alone (Graves et al., 2011). On the other hand, face matching decisions require a high level of cognitive ability and working memory, and the use of facial recognition algorithms alleviates some of this cognitive workload and streamlines the process.

Together, these considerations indicate that practitioners and systems working together can maximise the effectiveness of the overall system by minimising both human and computer error (Graves et al., 2011). However, while great strides have been made in understanding and improving algorithms, there is still progress to be made regarding the human practitioner.

1.3.3 Human performance on face matching tasks. In a key study of one-to-many face matching, Bruce et al. (1999) showed participants a 1-to-10 candidate list and instructed them that the target may not be present. The results of this task showed surprisingly high error rates of around 30% on both target-present and target-absent trials. Even a follow-up study where all trials were target-present and forced choice showed results of only 79% accuracy (Bruce et al., 2001). These results have been replicated extensively (Bruce, Henderson, Newman & Burton, 2001; Burton et al., 2010; Megreya & Burton, 2006).

Error rates are high, even with similar images taken under optimal conditions (Lander et al., 2018; Mileva & Burton, 2018). For example, in a task using high-quality images of unfamiliar faces, upper bound performance levels of around 90% were found in a novice population (Burton et al., 2010). However, performance drops as conditions worsen and begin to resemble those found in operational settings, such as when matches do not perfectly resemble the target (Kramer et al., 2018a). Previous research has found that facial comparison accuracy can decline to 58%-70% when comparing photos taken just months apart (Lander et al., 2018; Megreya, Sandford & Burton, 2013), and declines further when images depict younger targets (Michalski, 2017; Yadav, Sigh, Vatsa & Noore, 2014). Overall, this suggests that face matching accuracy decreases as the time gap between target and comparison images increases.

1.4 The Impact of Age on Unfamiliar Face Matching

Deficits in performance over longer time gaps are a problem for facial recognition algorithms and human practitioners (Megreya et al., 2013; Michalski et al., 2018). A survey of practitioners found that 60% of respondents agreed that changes in appearance between images can negatively impact decision-making, and the most problematic changes included those caused by ageing (Heyer et al., 2017). Human faces can vary drastically between a passport photo being taken and presented for comparison, which has led to recent interest in ageing effects on face matching for images of both adults and children (Ling, Soatto, Ramanthan & Jacobs, 2010; Michalski, 2017; Tistarelli et al., 2013).

Despite operational implications, studies exploring face matching across ages have been limited due to lack of suitable datasets (Ling et al., 2010). Additionally, while studies exploring age changes with adult faces have found higher error rates (White et al., 2014), there has been less consideration of how difficult this task may be when matching images of children (Kramer et al., 2018a; Michalski, 2017; Yadav et al., 2014). Given that research shows deficits in human performance with adult images taken only a short time apart, it is intuitive that performance may be worse with images of children, who undergo substantial facial changes during childhood.

1.4.1 Age-related facial changes. Facial changes can be categorised into two types; texture and shape changes. Shape changes alter the structure of the face, such as those caused by bone growth. Texture changes can refer to the development of wrinkles, loss of elasticity and changes in pigmentation (Tistarelli et al., 2013). During childhood, faces have major shape changes and minimal textural changes, while texture changes are much more prominent in adulthood (Ramanathan & Chellappa, 2006; Tistarelli et al., 2013; Yadav et al., 2014).

From birth, a child's face undergoes significant craniofacial changes that are markedly different from those which occur during adulthood (Mahalingham & Kambhamettu, 2012), with some of these changes being stable across individuals, regardless of gender and ethnicity (Ricanek, Mahalingham, Albert & Vorder Bruegge, 2013). Infant faces begin to change substantially around the age of three or four. During childhood, faces change proportions; growing downwards and elongating, primarily due to growth in the maxilla and mandible to accommodate the teeth (Ricanek et al., 2013; Taylor, 2001). These facial changes continue until the age of 14 when growth is almost completely relative to the adult head (Ricanek et al., 2013).

Due to this, craniofacial shape cues are often more pronounced following puberty, while young children often look similar to one another (Kramer et al., 2018a). This has implications for performance, as with less information to distinguish between identities, facial comparison decisions become harder (Kramer et al., 2018a).

1.4.2 Face matching performance with images of children. In Australia, each child must have a passport to use as identification. Passports for children under the age of 16 are typically valid for five years and can be used for critical real-life applications, such as identifying missing people and preventing child trafficking (Michalski, 2017; Zhang, 2007). These operational applications make it vital to determine how effectively children can be identified using these images (Kramer et al., 2018a).

Facial comparisons with images of children have been shown to be more challenging than with images of adults (Deb et al., 2018; Michalski, 2017; Srinivas et al., 2019; Yadav et al., 2014; Zeng et al., 2012). Age gaps, like the five-year passport validity period, are likely to cause further difficulties due to substantial changes occurring throughout childhood (Kramer et al., 2018a). Previous research shows that facial ageing can increase error rates by 20% when images are separated by more than a year (White et al., 2015).

White et al. (2015) compared one-to-many face matching performance on images of children (aged 6-13), adolescents (aged 14-22) and adults (aged 40-47), and found poor performance on all age groups. Results found the poorest performance for the child age group, with only 39% accuracy. Images in this condition depicted a child aged between six to thirteen being compared to an image taken an average of six years earlier. Adolescent trials were similarly difficult. Conversely, while still poor (45%), accuracy on adult faces was statistically superior in comparison to the other two groups, despite there being an average of ten years between adult images (White et al., 2015). While this study used an operationally valid database and was designed to resemble an applied setting, it was limited by exploring performance on a single child age group and a lack of images depicting children between the ages of 0-5. As a substantial amount of growth occurs in this period, it is important to explore performance across all ages of childhood. Furthermore, poor performance in this study may have been impacted by both fatigue effects due to a large number of trials (300) and imposing a deadline on participants.

Kramer et al. (2018) examined how accurately people performed on a one-to-one task focused on infant faces and found 72% accuracy overall. Importantly, images for this task were taken within the same year and image pairs were manually selected with no attempt to ensure visual similarity between images, indicating that their results were at the upper bounds of human performance. Results also found a substantial drop in performance (64%) when trials depicted an operationally valid five-year age gap, by pairing an image of an infant with an image of a child aged around 5 years old (Kramer et al., 2018a). However, limitations of this study include its use of greyscale images and celebrity children; using celebrity children can be problematic for unfamiliar face matching tasks as performance can be influenced by potential familiarity from the media.

Michalski (2017) evaluated the performance of practitioners on several studies. One of these used a one-to-one task to examine performance with both child and adult images. Lower accuracy was found for child faces (73.9%) than adult faces (92.1%).

Further analysis exploring performance using a heat map data matrix across age (0-17 years) and age variations between images ranging from 0-10 years, found that accuracy was greater for images of older children. Furthermore, for each age tested, performance decreased as age variation increased. Both accuracy and confidence were poorest with infants regardless of the age variation depicted (59-70%). This study also found differences between performance on match (86.97%) and non-match (76.66%) trials. These results show that both facial changes throughout childhood and the type of trial presented have a significant impact on practitioner performance.

1.4.3 Limitations and implications of child face matching performance. Although results show that performance is variable across different ages of childhood and is lowest for infants and toddlers, a majority of studies have explored children at the group level rather than over individual ages (White et al., 2015). Grouping children as one overall age group gives little opportunity to explore how changes over childhood impact face matching performance. Additionally, a vast majority of the literature has focused on one-to-one face matching. Results from these studies cannot be extrapolated to a one-to-many context as one-to-many tasks have been shown to be more challenging, due to both the addition of extra faces for comparison and the chance of selecting a false match from the candidate list (Heyer et al., 2018).

The type of trial presented and the error made can have distinctive consequences in operational settings (Michalski, 2017). For example, the prevalence of target-absent trials is assumed to be much smaller for passport control than for investigative agencies. Investigative agencies are often tasked with identifying victims and finding missing people. Error rates on target-absent trials can mean using time and resources to follow false leads, while error rates on target-present trials pose a risk of victims going unidentified. For passport control, the goal is to confirm that a target identity is who they claim to be. In this context, there is a higher prevalence of target-present trials, with target-absent cases occurring through error and fraud. Passports are often valid for between 5-10 years, changes to appearance in this time can cause errors on target-present scenarios. These errors cause a false alarm and can lead to dissatisfied customers and a strain on resources. Conversely, errors on target-absent cases in this context can lead to the acceptance of fraudulent images, posing severe consequences, such as allowing trafficking victims through official checkpoints (Zhang, 2007).

Overall, studying how performance is impacted by both the age of the child being identified and the type of trial presented can begin to provide empirical evidence to help agencies understand and address detriments to performance, allowing them to determine where more training or alternative methods may be needed.

1.5 The Present Study

The current study aims to determine how human face matching performance varies depending on the age of a child and the type of trial presented. The study will use a one-to-many (1:8) face matching design which requires comparing images of children divided into three age groups (0-5, 5-10 and 10-15) with an age variation of 5 years between target and candidate list images to replicate the validity period of child passports.

Based on previous research, two research questions are explored:

(a) To what extent does participant performance differ when conducting face matching across three childhood age groups?

(b) To what extent does participant performance vary depending on the type of trial presented (target-absent or target-present)?

CHAPTER 2:

Method

2.1 Ethics Statement

This computer-based study was conducted by the University of Adelaide School of Psychology in conjunction with the Defence Science and Technology (DST) Group, Edinburgh, South Australia. Ethics approval was granted by the University of Adelaide Human Research Ethics Sub-Committee (HREC 19/57) and the DST Ethics Review Panel (ISSD 03-19).

2.2 Study Design

A within-subjects repeated measures design was used for this study. This meant that performance measures were taken across different levels of the independent variables with each participant tested in all conditions. For this study, the age groups (0-5, 5-10 and 10-15) and trial type (target-absent and target-present) were the independent variables. The performance measures for each participant (accuracy and confidence) served as the dependent variables. As each participant viewed the same trials, images were randomly presented and counterbalanced between participants to control for order effects.

2.3 Participants

Participants comprised a convenience sample of DST employees from Edinburgh, South Australia (N = 42) recruited on a voluntary basis. Participants were limited to DST employees as prior research has indicated that DST staff are more motivated than student populations and can show similar levels of performance to face matching practitioners (Heyer et al., 2018). Other inclusion criteria included being above the age of 18, proficient in English, and having normal or corrected-to-normal vision (i.e., through the use of glasses or contact lenses). The sample included 19 females and 23 males, ranging from 22-64 years old

(*M* = 42.2 years). A majority of participants identified as Caucasian (88%), with a minority identifying as Asian (12%). Participants had no formal training or experience conducting facial comparisons.

Participants were recruited through the placement of posters (Appendix A) around DST, online advertisement on the DST intranet site (SATURN), and word-of-mouth. Participation was completely voluntary, with participants made aware that they could withdraw at any time. No incentives were offered for participation besides light refreshments for the duration of the experiment and the opportunity to receive results if desired. Recruitment took place between the 31st of July and the 16th of August.

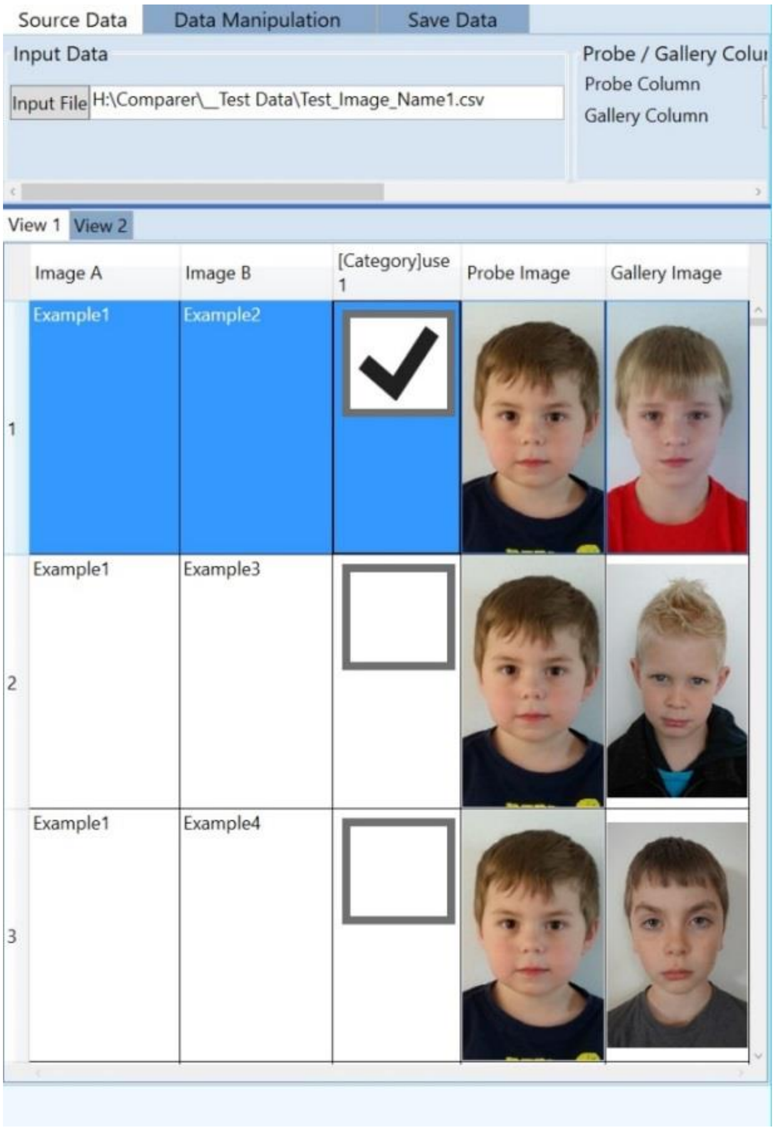
2.4 Materials

This study was computer-based and materials included the use of a controlled operational database of images, an image comparison software tool (Comparer), an experimental application and the use of DST computers.

2.4.1 Images. The primary material used for this study were operational images of children sourced from an operationally valid database.

2.4.1.1 Image database. The database used in the study was supplied by a government agency for research purposes and consisted of controlled images of faces. It has also been used in other related research (Michalski, 2017; Pearce, 2018; Snyder, 2018). Images were front facing and were consistent in size and quality. Target identities (aged 5, 10 or 15) from the database were used as probe images by an algorithm to search for a selection of the closest matches. Each probe image for target-present and target-absent trials was five years older than the images displayed in the corresponding candidate list. This age variation was selected to provide ecological validity as passports for children are usually valid for five years. Due to restrictions on this database, images cannot be provided within this thesis.

2.4.1.2 Image selection. Images for the study were selected from the database using an image comparison software tool called Comparer, designed by DST (Hole et al., 2015). This software presents the target images and a selection of their highest scoring potential matches returned by a facial recognition algorithm as side-by-side pairs. From there, a checkbox can be ticked to indicate selection for the dataset. This was used to select the 8 images that would be used for candidate list images for each target. Figure 1 provides a screenshot of this software.



Input File H:\Comparer_Test Data\Test_Image_Name1.csv

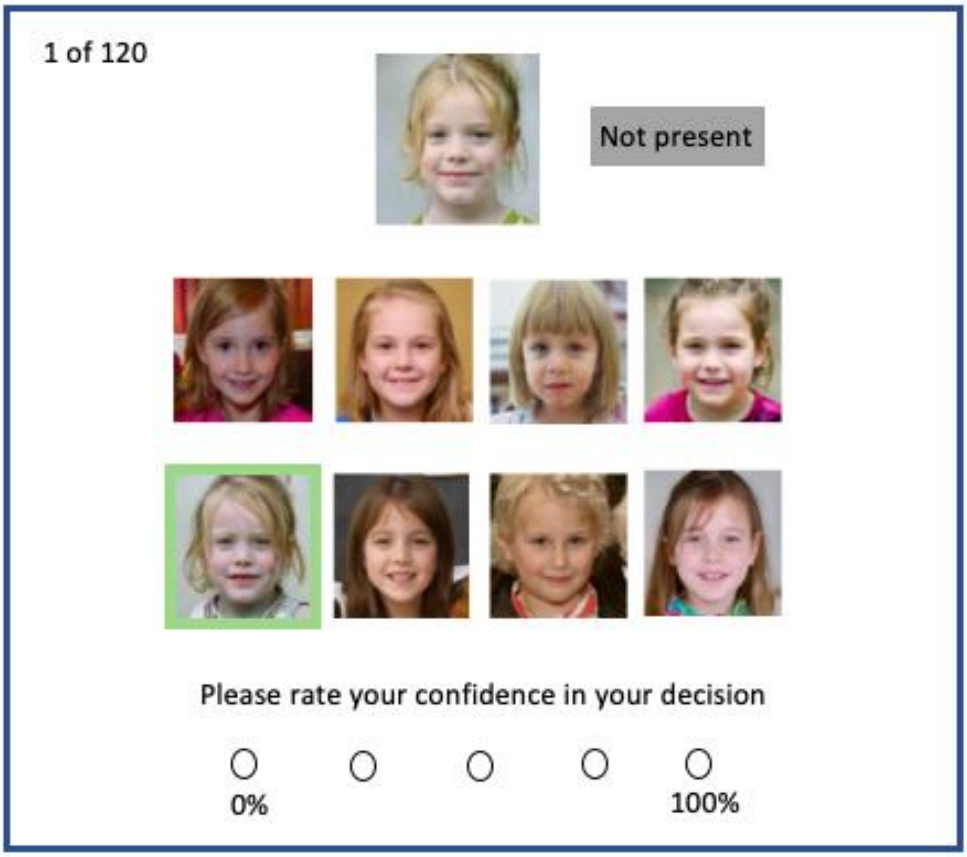
	Image A	Image B	[Category]use 1	Probe Image	Gallery Image
1	Example1	Example2	☑		
2	Example1	Example3	□		
3	Example1	Example4	□		

Figure 1. Screenshot of the Comparer software tool. Images are for illustration purposes only

A total of 120 target images and 960 candidate list images were selected, evenly divided into the age groups; 0-5 years, 5-10 years and 10-15 years, the number of target-absent and target-present trials and by gender. The end result was a dataset consisting of 40 trials for each age group, with 20 target-present and 20 target-absent trials. Additionally, 18 more images were selected to accommodate two practice trials. All images within the candidate lists were five years younger than the target.

Images were manually selected to produce a sample of images representative of the Australian population, check for any errors and ensure consistency across the dataset.

2.4.2 Experimental application. To conduct the experiment, an experimental application was custom designed by the researcher and programmed by DST staff based on the design specification (Appendix B). The application resembled the operational layout used in previous research (White et al., 2015). Figure 2 provides a diagram of the application.



1 of 120

Please rate your confidence in your decision

○ ○ ○ ○ ○

Figure 2. A diagram of a trial from the experimental application. Images are for illustration purposes only.

Due to restrictions of the operational database, secure computers were used to run the experiment. These computers had 23-inch monitors with 1920x1200 screen resolution. In addition to running the experiment, the application also collected and stored consent, demographic information, experimental data and registration for results.

2.5 Procedure

Prior to commencing the experiment, participants were provided with a copy of the Information Sheet and Consent Form (Appendix C), and DST Volunteer Guidelines (Appendix D). Participants received a verbal briefing (Appendix E) where they were guided through the task requirements, told how their data would be managed and were asked to read over and sign the provided documents. Participants also provided informed consent by indicating so on the first screen of the experiment; allowing them to be allocated a unique identifier which was used to manage their data. Participants completed a short demographic questionnaire where they indicated their age, gender and ethnicity before being asked to ensure they were wearing any required vision correction. Participants were then provided with further instructions on screen (see specification, Appendix B) and completed two practice trials to familiarise themselves with the task. Following the practice trials, participants were prompted to ask any questions and begin the study by pressing the ‘start’ button. Each session was conducted in-person and took approximately 1.5 hours to complete.

Participants completed a total of 120 trials. Each trial involved the participants deciding whether a target identity was present and if so, which image was the ‘match’. Participants could select a ‘not present’ option where they felt a match could not be made. Once a decision was made, participants were required to rate their confidence on a 5-point Likert-scale, where 1 indicated 0% confident and 5 indicated 100% confident. Each age group had an even split between target-absent and target-present trials and participants were reminded that the target may not always be present within the candidate list. This was to provide a basis for exploring the second research question.

Upon completion of the experiment, the final screen allowed participants to input an email address to receive their results. While each session was expected to take an hour and participants were instructed to work as quickly and accurately as possible, there was no time restriction placed on participants – allowing them to use more or less time as required.

CHAPTER 3:

Results

To explore the research questions, accuracy and confidence data were analysed for the three child age groups, 0-5, 5-10 and 10-15, as well as for each of the two trial types, target-absent or target-present.

3.1. Data Screening and Assumption Checking

Prior to conducting analyses, data were checked for errors and assessed for normality. A small number of outliers were found within the accuracy data; upon inspection they were determined to be legitimate cases of variation and remained in the dataset. Each condition was then assessed for normality using a combination of Shapiro-Wilk tests and visual inspection using Q-Q plots and histograms. A portion of the data were found to be significantly skewed and non-parametric tests were chosen for analysis. The use of non-parametric tests was favoured over data transformation as face-matching data has previously demonstrated skewness, indicating valid representation of performance (Michalski, 2017). When comparisons were made for more than two groups, Friedman tests were used. When results were significant, post-hoc pairwise comparisons were conducted using the Wilcoxon Signed-Ranks test with Bonferroni adjustment to protect against inflated Type-I error rates. Effect sizes were calculated using the formula: $r = z/\sqrt{N}$ (dividing the test statistic by the square root of the number of observations), and were interpreted as indicating a large (>0.5), medium (>0.3) or small (>0.1) effect size (Cohen, 1988). Due to non-parametric data, the median (*Mdn*) will be reported as a descriptive statistic for each condition. The mean (*M*) will appear alongside the median to allow direct comparisons with previous studies using parametric tests.

3.4. Performance Across the Three Age Groups

The first research question asked to what extent performance varied depending on the age of the child being identified. This was explored by analysing accuracy and confidence measures across age groups.

3.4.1 Accuracy. Visual inspection of the descriptive statistics in Figure 3 suggests an upwards trend where participants were the least accurate for the 0-5-year age group (*M* = 16%, *Mdn* = 10%) and most accurate for the 10-15-year age group (*M* = 30%, *Mdn* = 30%).

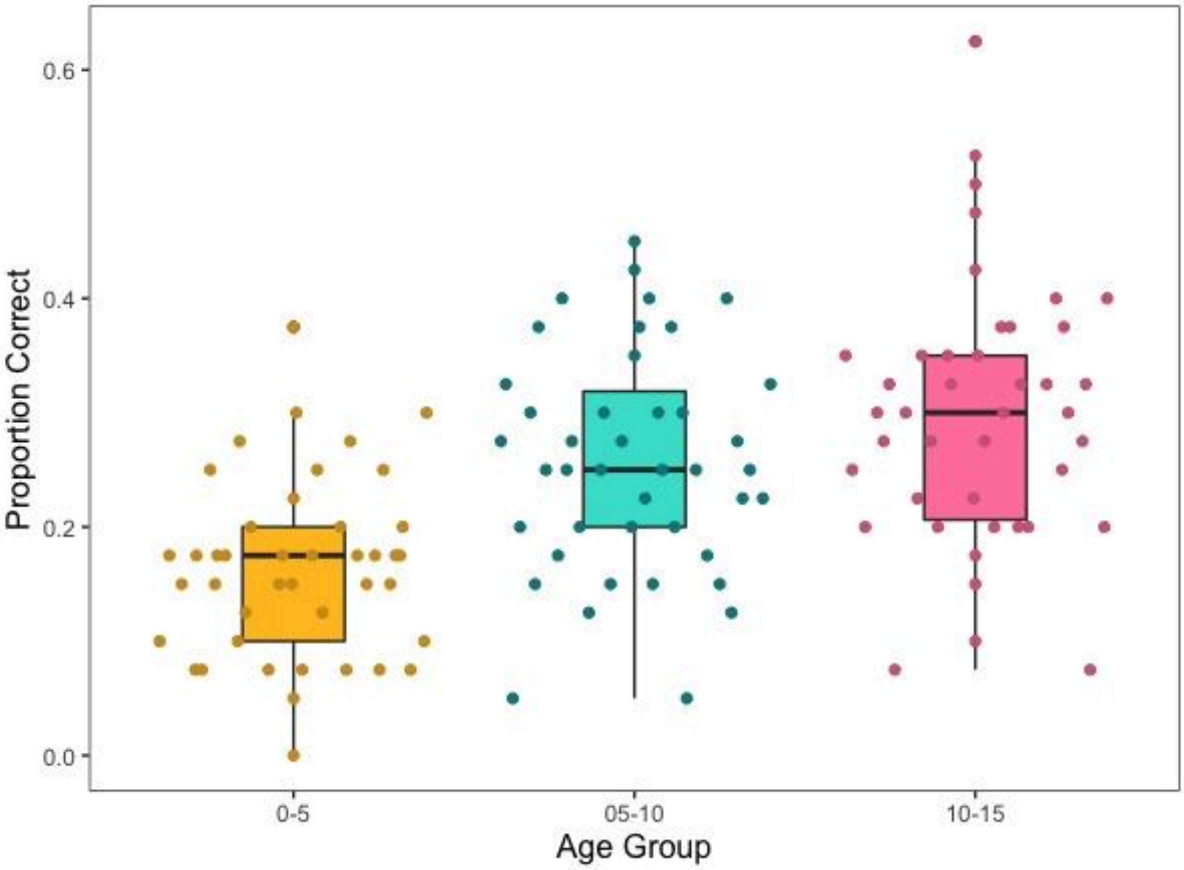


Figure 3. Descriptive statistic boxplots of overall accuracy (expressed as proportion correct) for each group, from left to right: 0-5, 5-10, 10-15-year age groups. Dots indicate individual performance.

A Friedman test showed statistically significant differences in accuracy based on the age of the child being identified ($\chi^2(2) = 44.73$, p <.001). Post-hoc tests were conducted to analyse these differences. A Bonferroni adjustment led to a significance criterion of $p < .016$. Wilcoxon Signed Rank tests revealed a significant decrease in accuracy for the 0-5-year age group compared to both the 5-10-year (*M* = 26%, *Mdn* = 25%), $z = -4.95$, $p <.001$, $r = .31$, and 10-15-year age groups, $z = -5.87$, $p <.001$, $r = .37$, with moderate effect sizes. A significant difference was also found between the 5-10-year and 10-15-year age group, $z = -2.61$, $p = .008$, $r = .16$.

3.4.2. Confidence. Figure 4 depicts the overall descriptive statistics for confidence in each age group.

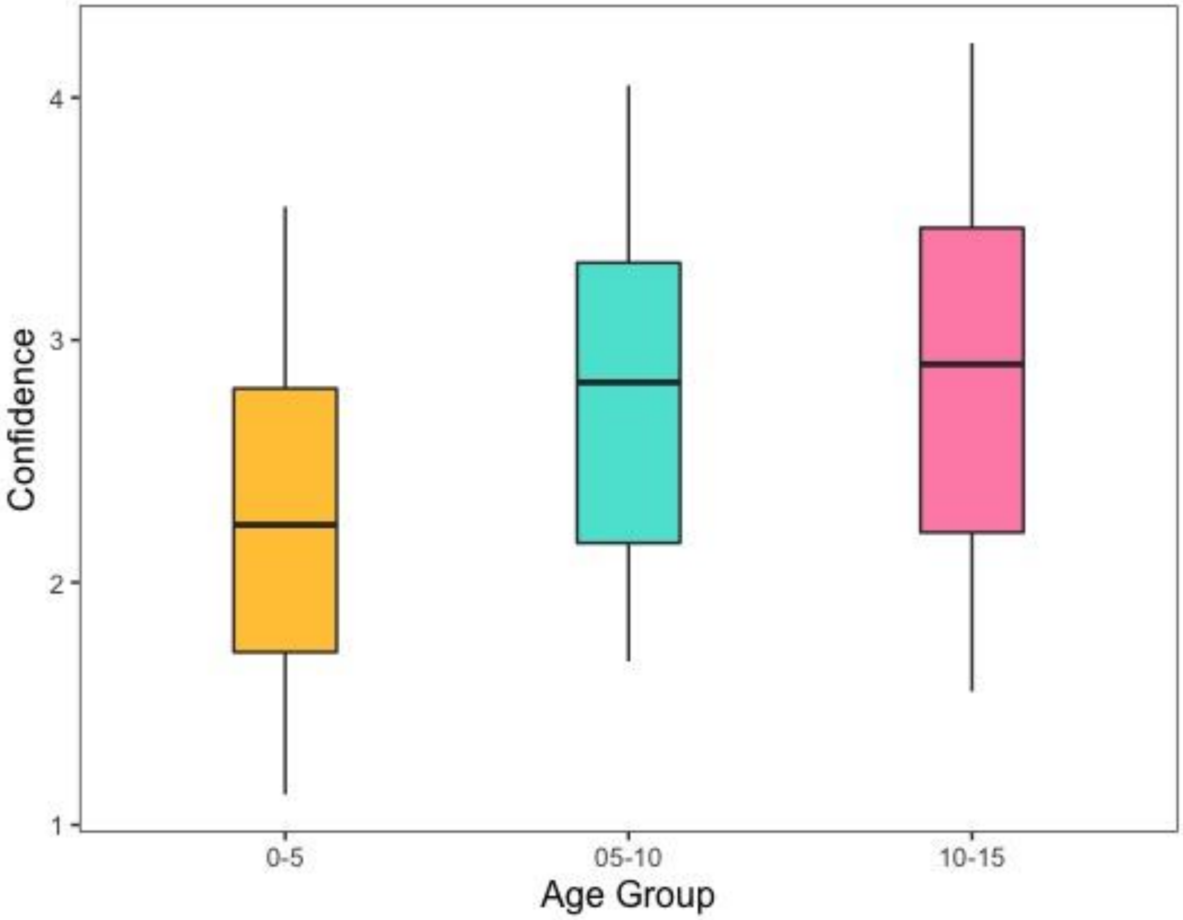


Figure 4. Boxplots for confidence in each age group (on a five-point Likert scale, where 1 indicates minimal confidence, and 5 indicates complete confidence).

Inspection of descriptive statistics suggests that as age group increases, mean level of confidence may also increase. A Friedman test found a statistically significant difference in confidence over the age groups ($\chi^2(2) = 60.04$, $p <.001$). Post-hoc analysis with Wilcoxon Signed Rank tests revealed a significant decrease in confidence for the 0-5-year age group ($M = 2.24$, $Mdn = 2$) compared to the 5-10-year ($M = 2.73$, $Mdn = 3$), $z = -6.15$, $p <.001$, $r = .67$, and 10-15-year age group ($M = 2.8$, $Mdn = 3$), $z = -6.57$, $p <.001$, $r = .72$. However, there were no significant differences between the 5-10-year and 10-15-year groups, $z = -1.48$, $p = 0.14$, $r = .16$).

3.5. Performance on Target-Present and Target-Absent Trials

The second research question queries the extent to which performance varies depending on the type of trial. This was analysed by comparing accuracy and confidence data across both types of trials.

3.5.1 Accuracy.

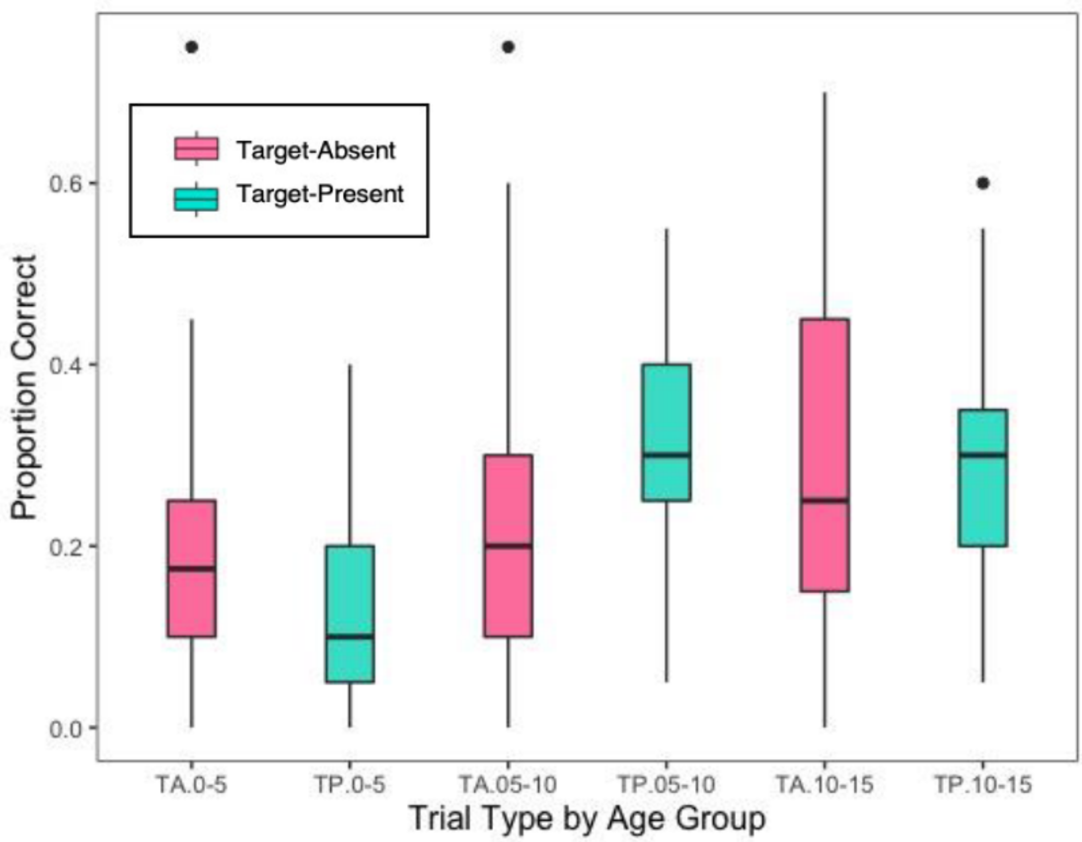


Figure 5. Descriptive boxplots for each of the three age groups based on trial type.

The Friedman test indicated that there was a statistically significance difference in accuracy based on the type of trial presented, $\chi^2(5) = 54.558$, $p <.001$.

Post-hoc analysis using Wilcoxon Signed Ranks tests was then conducted, with an adjusted significance criterion of .005.

3.5.1.1 Target-present trials across age groups. Comparison of target-present trials found that participants were significantly less accurate with target-present trials from the 0-5-year group (M = 14%, Mdn = 10%) compared to target-present trials for both the 5-10-year (M = 29%, Mdn = 30%), z = -5.19, p < .001, r = .8, and 10-15-year groups (M = 30%, Mdn = 30%), z = -5.08, p < .001, r = .78, both of these relationships show strong effect sizes. No significant differences were found between accuracy on target-present trials with the 5-10-year and 10-15-year age groups, z = -0.47, p = 0.64, r = .07.

3.5.1.2 Target-absent trials across age groups. Participants demonstrated a significant decrease in accuracy for target-absent trials from the 0-5-year group (M = 19%, Mdn = 18%) compared to target-absent trials from the 10-15-year group (M = 29%, Mdn = 25%), z = -3.14, p < .005, r = .48, and for target-absent trials from the 5-10-year group (M = 22%, Mdn = 20%) compared to the 10-15-year group, z = -3.28, p < .005, r = .51. No significant difference was found between accuracy on target-absent trials between the 0-5-year and 5-10-year age groups, z = -1.34, p = 0.18, r = .21.

3.5.1.3 Comparison of target-present and target-absent trials. Finally, no significant differences in accuracy were found between target-present and target-absent trials for any of the three age groups; 0-5-year, z = -1.65, p = 0.10, r = .25, 5-10-year, z = -2.39, p = 0.02, r = .37, and 10-15-year, z = -0.5, p = 0.62, r = .08).

3.5.2 Confidence.

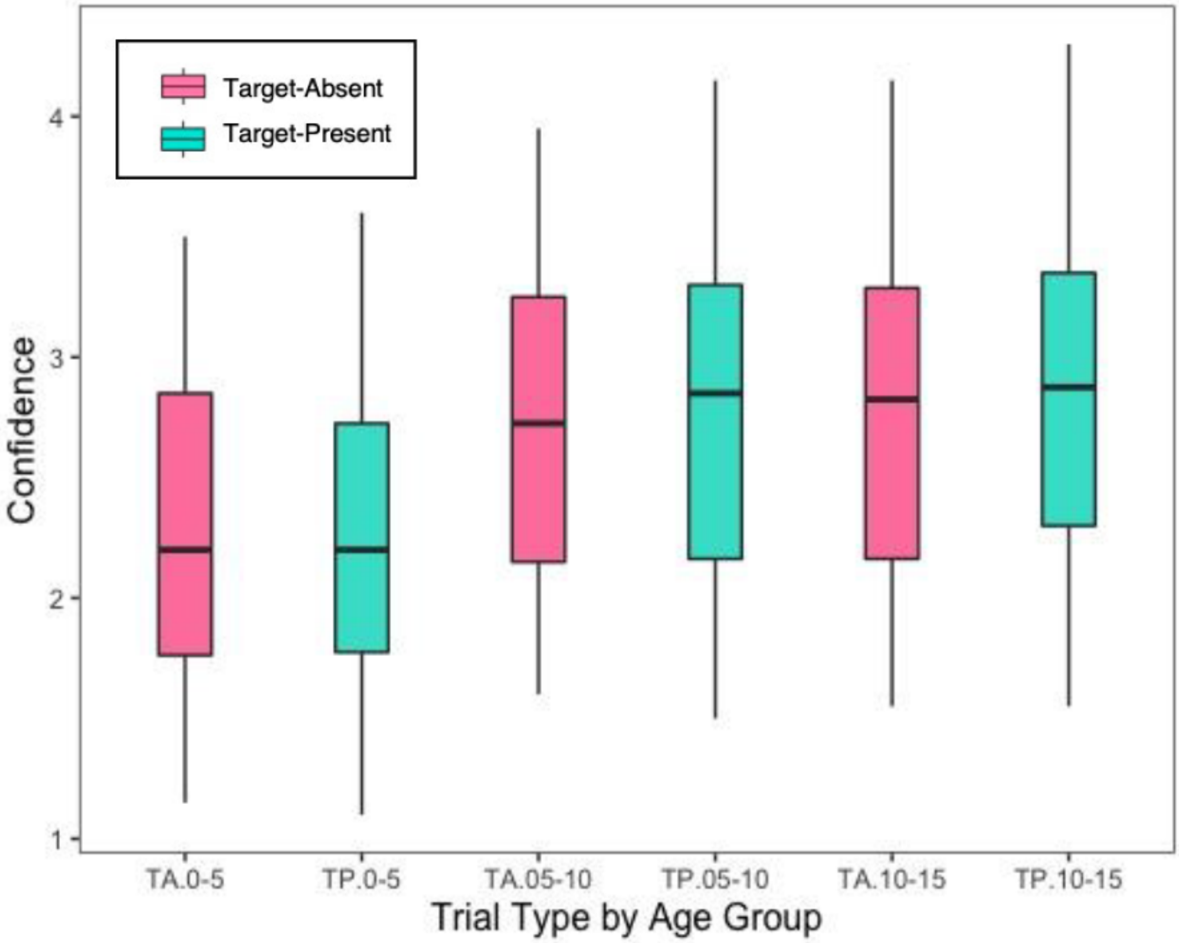


Figure 6. Descriptive statistics for confidence levels across both trial types and age group conditions

The Friedman test indicated that there were statistically significant differences in confidence based on trial type ($\chi^2(5) = 129.96$, $p < .001$).

Post-hoc analysis using Wilcoxon Signed Rank tests with a Bonferroni corrected significance criterion of .005 were conducted to explore relationships between groups.

3.5.2.1 Target-present trials across age groups. The results from these tests showed that there was a statistically significant decrease in confidence between target-present trials with the 0-5-year group ($M = 2.21$, $Mdn = 2.2$) compared to target-present trials for both the 5-10-year ($M = 2.79$, $Mdn = 2.85$), $z = -5.57$, $p < .001$, $r = .86$, and 10-15-year age groups (M

= 2.86, *Mdn* = 2.88), *z* = -5.51, *p* < .001, *r* = .85, and no significant differences for target-present trials between the 5-10-year and 10-15-year groups, *z* = -1.54, *p* = 0.12, *r* = .24.

3.5.2.1 Target-absent trials across age groups. Analysis of target-absent trials also found a significant decrease in confidence with images from the 0-5-year (*M* = 2.27, *Mdn* = 2.2) group compared to the 5-10-year (*M* = 2.68, *Mdn* = 2.73), *z* = -5.13, *p* <.001, *r* = .79, and 10-15-year (*M* = 2.75, *Mdn* = 2.83), *z* = -5.36, *p* <.001, *r* = .83, both with large effect sizes. A non-significant decrease in confidence was found for target-absent trials for the 5-10-year group, compared to the 10-15-year group, *z* = -1.46, *p* = 0.14, *r* = .23.

3.5.2.3 Comparison of target-present and target-absent trials. Finally, analysis of confidence levels on target-present and target-absent trials across age groups found a significant decrease in confidence for target-absent trials in the 5-10-year group, *z* = -2.85, *p* <.005, *r* = .44. Slight non-significant decreases in confidence were found for both the 0-5-year, *z* = -1.37, *p* = 0.17, *r* = .21, and 10-15-year age group, *z* = -2.75, *p* = .006, *r* = .42. These effect sizes are considered moderate.

3.2. Equal Variance Signal Detection (EVSD) Model

A limitation of using proportion correct to analyse performance is that it confounds criteria shifts with discriminability. Accordingly, a 6-parameter equal variance signal detection (EVSD) model with a max rule (Semmler, Kaesler & Dunn, 2018) was fit to the observed data to obtain estimates of *d'* and 5 criteria. Fitting a model can help describe and explain observed data, allowing observation of how criteria are placed across conditions and providing a nuanced picture of performance. To fit the model, data matrices consisting of frequency counts across the confidence bands and response types (Target Detection, False Alarm and Target Identification) were produced for each condition and overall data (Appendix F), which were used by the model to predict a set of parameters that best capture the observed data, using maximum likelihood estimation (Dunn, 2010). Despite finding a set of parameters at a maximum likelihood, the model did not adequately capture the data in any condition (Appendix G). Figure 7 depicts an overall plot of the Receiver operator characteristic (ROC), showing the performance of the sample, along with the recovered values of the best set of parameters from the EVSD model, illustrating the poor fit ($G^2(4) = 1735.9, p < .001$).

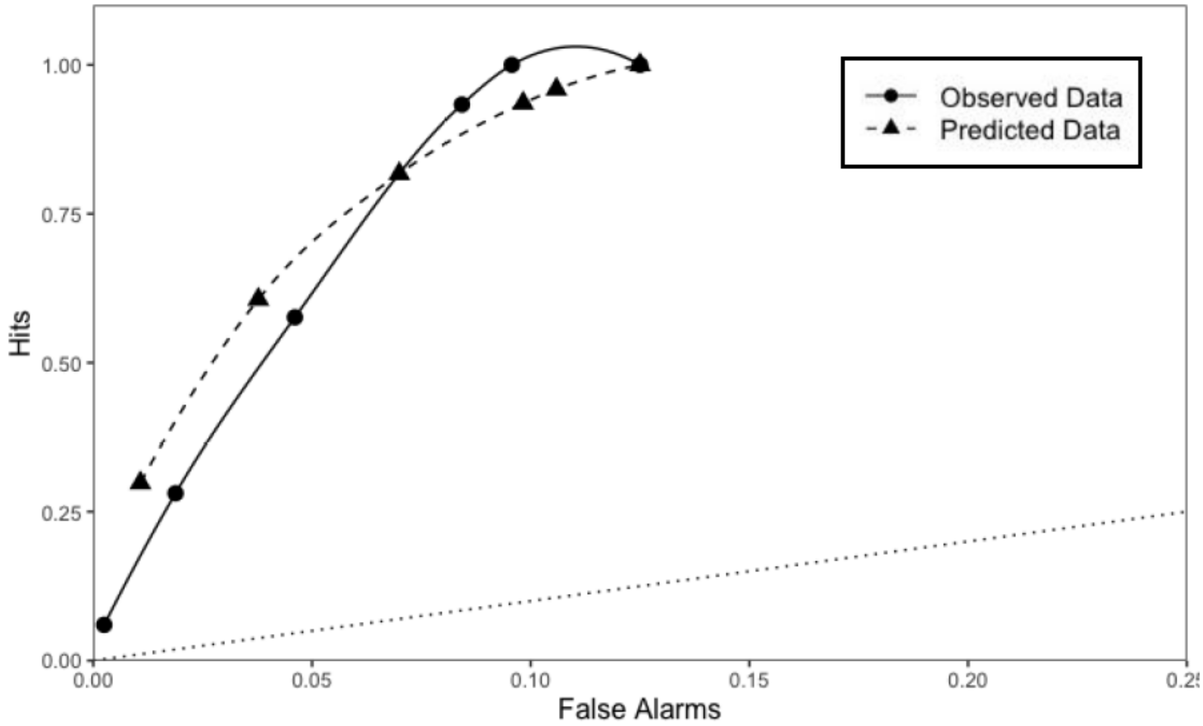


Figure 7. Receiver operator characteristic (ROC) curve showing actual (solid line) performance compared with the expected (dashed line) performance fitting the EVSD model using maximum likelihood estimation.

A 7-parameter unequal variance model was also fit and exhibited a poor ability to capture the data ($G^2(3) = 25.9, p < .001$). The set of parameters produced by the models differed substantially from the observed data (Appendix H). This was especially true around criteria 1 and 2, where the observed data were too conservative compared to what was predicted, indicating that participants were not placing their criteria in the way the model expected. This will be addressed in the discussion (section 4.3). We did not use non-parametric signal detection measures as these also require unmet assumptions regarding the signal and noise distributions (i.e., equal variance distributions).

3.3. Confidence-Accuracy Characteristic (CAC) Analysis

In order to assess the relationship between confidence and accuracy, we plotted a confidence-accuracy characteristic (CAC) curve, depicted in Figure 8.

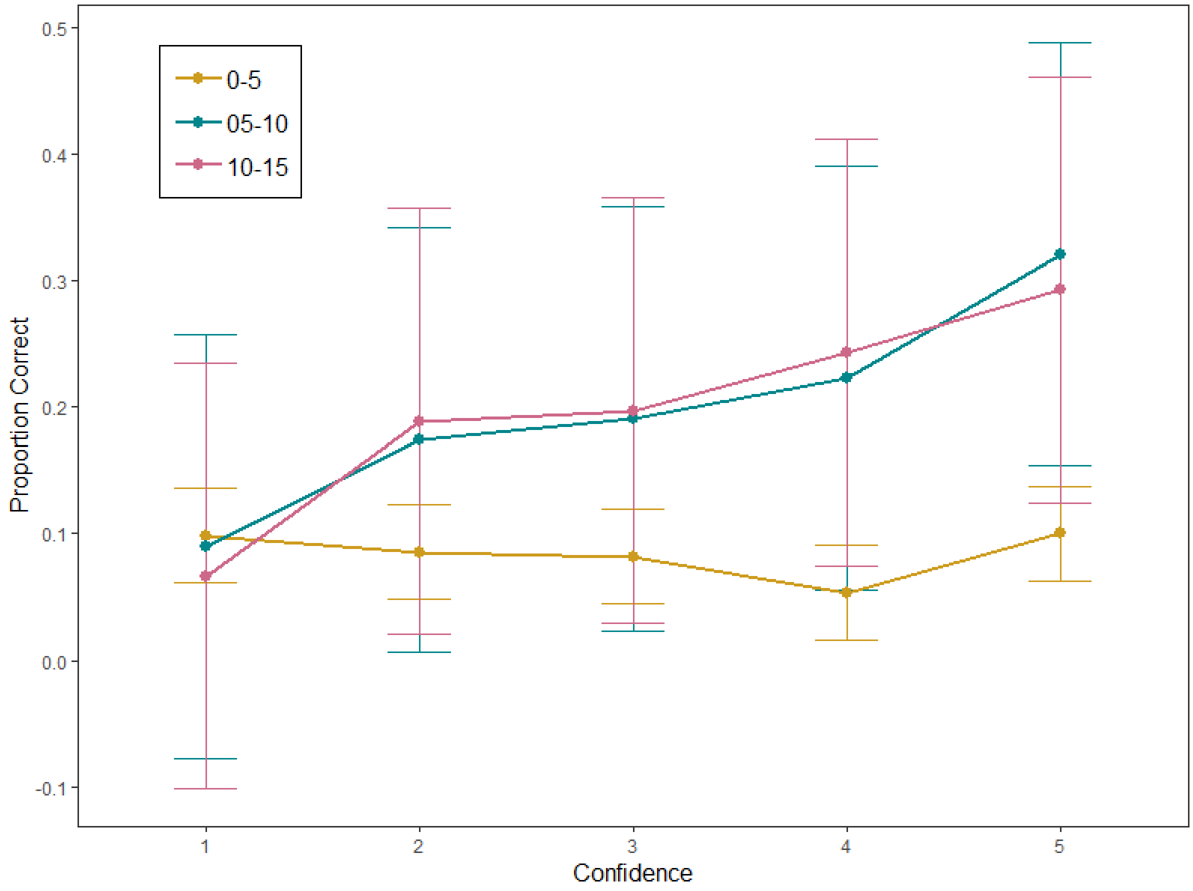


Figure 8. CAC curve for the three child age group conditions, depicting the proportion of correct responses at each level of reported confidence (on a 1-5 scale, aggregated across participants). The bars represent standard error bars.

At a minimum, it was expected that level of confidence should have some relationship with proportion correct. The CAC curve was produced using one of the approaches outlined in Mickes (2015) where all errors are counted, including misidentification errors, regardless of whether the trial was target-present or target-absent. This method was selected as there were no ‘known innocent’ identities within the candidate lists, with all options theoretically able to be the target. The CAC shows high variability within the data with overall trends indicating poor performance and almost no relationship between confidence and accuracy for the 0-5-year age group. Both the 5-10-year and 10-15-year groups show a small rise in accuracy as confidence increases. An over/under-confidence statistic (O/U) was computed for the overall data. The O/U statistic captures a responder’s overall tendency to report confidence that is higher or lower than is warranted by accuracy (Stephens et al., 2017). Values range from -1 (extreme under-confidence) to 1 (extreme over-confidence). The O/U statistic for the overall dataset was 0.16, implying that participants generally erred on the side of overconfidence. The O/U for individual groups also clustered around this value. Taken together, this suggests that the sample was slightly overconfident. However, due to high variance in the data, it is difficult to make any real inferences about the relationship across conditions.

3.6 Summary

Results indicated that participants were significantly less accurate and confident when matching images of children from the 0-5-year age group, in both target-present and target-absent conditions. Differences between the 5-10-year and 10-15-year age groups were non-significant but generally showed a slight increase in performance for the 10-15-year age group. This implies that performance increases as age increases. Overall performance across all conditions was poor, implying that the task was difficult in all conditions.

CHAPTER 4:

Discussion

This study aimed to evaluate the extent to which performance varied depending the age of a child being identified and the type of trial presented. To do this, a one-to-eight image comparison task was used with three child age groups (0-5-year, 5-10-year, and 10-15-year) and an age variation of 5 years between target and candidate list images, to mimic the current child passport validity period. Overall, results indicated that the 0-5-year age group was the hardest to identify. There were no major differences between target-present and target-absent trials for any of the age groups. The following section discusses and interprets these results based on the two research questions, and will consider limitations, implications and future directions for this research.

4.1 Overview of Performance on the Three Age Groups

The first research question aimed to explore how performance varied depending on the age of the child being identified. Results found accuracies of 16%, 26% and 30% for the three groups, indicating that performance was poor regardless of age group. Participants were significantly less accurate and confident for the 0-5-year age group, and the most accurate for the 10-15-year age group. These results were in line with poor performance found within another one-to-many study which analysed the difficulty of unfamiliar face matching with child images (White et al., 2015) and supported the notion that younger identities are hardest to identify (Lui et al., 2009; Michalski, 2017). Additionally, results were worse than those found on one-to-one studies (Kramer et al., 2018a; Michalski, 2017) and those which focused on adult faces (Bruce et al., 1999; Megreya et al., 2013; White et al., 2014). This supports the claims that one-to-many tasks are more difficult and that matching of child faces is harder than for adult faces. The poor performance for younger identities is likely due to the extensive craniofacial growth which occurs from an early age (Mahalingham &

Kambhamettu, 2012; Ricanek et al., 2013). Younger children also have fewer distinguishing features and less between-face variability, leading to less information to distinguish between identities (Kramer et al., 2018a; White et al., 2015).

The CAC analysis found high variability and minimal relationship between confidence and accuracy for all three age groups - the worst, once again, being for the 0-5-year group. High variability poses a problem for understanding the confidence-accuracy relationship and limits inferences that could be made about the data. These results were contrary to previous research which presents confidence as a good indicator of accuracy (Stephens et al., 2017; Lander et al., 2018). However, this study used a one-to-many task and focused on images of children and is the first study to examine the confidence-accuracy relationship in this context. Therefore, it is possible that the lack of relationship could be due to both the difficulty of one-to-many matching and limited practice matching faces of children. A lack of experience may mean participants are less aware of changes in faces over time and thus, may not appreciate the difficulty of matching the faces of children (Bobak, Mileva & Hancock, 2018). Further analysis is warranted and future research might consider adding extra trials so individual CACs can be plotted, rather than averaging over the sample.

Overall, results for the first research question supported the notion that comparing child faces is a highly difficult task, especially for younger identities. Performance increases as age increases, however, it seems that potentially from the age of 5, these increases are far less pronounced.

4.2 Overview of Performance on Target-Absent or Target-Present Trials

The second research question queried the extent to which performance varied depending on trial type for each age group. Namely, was it easier for participants to tell faces together or to tell faces apart across the age groups? The trial types were considered separately as previous research has found differences in accuracies, and different trial types have different functions (Megreya & Burton, 2006). Overall, performance was better for the 5-10- and 10-15-year age groups and much worse for the 0-5-year age group. Many of the relationships found were non-significant. For target-present trials, confidence and accuracy were significantly worse for the 0-5-year group (14% accuracy) compared to both other groups (29% and 30%). This converged with reports that participants are less likely to find a younger target when present in a candidate list (White et al., 2015). This could be because the most significant craniofacial growth happens during early childhood, causing these identities to exhibit higher within-subject variability over the 5-year age variation; combined with less between-person variability due to homogeneity of younger faces (Kramer et al., 2018a).

For target-absent trials, accuracy and confidence were significantly worse for both the 0-5 and 5-10-year age groups compared to the 10-15-year group. As the majority of growth has already taken place by this age (Ricanek et al., 2013) it is possible that there is less within-person variability in the older group, making it easier to determine whether the target is absent from the candidate list. Poor performance on the other groups could be because the faces have fewer distinguishing features, making them harder to tell apart (Michalski, 2017).

Within age groups, it was found that accuracy was not significantly different between target-present and target-absent trials, the only significant relationship being a decrease in confidence between target-absent and target-present trials for 5-10-year-olds. This lack of significant difference was inconsistent with reports from previous research which has found trial type to have a significant impact on performance (White et al., 2014; Megreya & Burton, 2006), but was similar to results from Bruce et al. (1999) where error rates of 30% were found for both target-present and target-absent conditions. Lack of significant differences might be because of the challenging nature of both child face matching and one-to-many tasks, that is, it may be due to a floor effect.

4.3 Evaluating Model Fit (Or Lack Thereof)

Both the EVSD and UVSD Models failed to adequately capture or explain the observed data, meaning that the set of parameters produced by the model differed substantially from the observed data. Despite maximum likelihood estimation minimising the discrepancy between predictions and data (Lewandowsky & Farrell, 2011), it appeared that the model tended to overestimate criteria 1 and 2 (in this case indicating high confidence) and underestimate 4 and 5 (indicating low confidence). This suggests participants were not using the confidence scale the way the model predicted and were too conservative in their confidence ratings. It is possible that this result was influenced by methodological factors, such as the lack of deadline imposed on participants, leaving them more time to analyse before making decisions, and time to exhibit response hesitancy. Additionally, participants were instructed to work as quickly and accurately as possible and be ‘sure’ before making decisions, which might have made participants compensate for difficulty by rating confidence conservatively. Another consideration is the decision rule used to fit the model. In this study, the maximum-of-outputs (max) rule was used as it is the simplest rule to implement. However, some research indicates that it is not the best choice (see Ma, Shen, Dziugaite, & van den Berg, 2015). An alternative to this rule is the ideal-observer (or Bayes-optimal) rule. Optimal rules are distinct from the max rule as they are Bayesian and probabilistic. They are also often more complex (Ma et al., 2015). Future research might benefit from exploring model fit using this rule.

It is possible that the models are not yet able to generalise to applied face-matching tasks with free conditions and may need substantial adjustments to account for this. Due to time constraints, this was not possible for this study. Instead, future work should test other models to find one with a better fit that can account for observed performance. Furthermore, advantage could be taken of the within-subjects nature of the data – with fitting of hierarchical models that can account for item and participant variability in the model parameters (Rouder & Lu, 2005).

4.4 Limitations

The current study is not without limitations, a selection of which will now be outlined. Firstly, the image dataset used in the experiment was representative of the Australian population and included images of various racial identities. This was chosen for operational validity, as practitioners will come into contact with a diverse range of people and are required to identify them accurately irrespective of ethnicity (Lander et al., 2018). However, this study did not explore any cross-race effects or own-race biases (Meissner et al., 2013). Previous research has suggested an own-race bias where matchers are more likely to make a correct identification on faces from their own race (Meissner et al., 2013). Considering that a vast majority of participants identified as Caucasian, this could have impacted on error rates. However, the effect itself has been challenged by studies that have either found no real difference or that non-Caucasian matches were actually easier (White et al., 2015).

This study also aggregated data across individuals and ignored any individual differences. This can be problematic as individual differences are known to be quite prominent on face matching tasks, with previous studies consistently showing a wide distribution of ability in novice and specialised population (Heyer et al., 2018; McCaffery et al., 2018; White et al., 2014). Individual differences are suggested to be one of the largest documented potential sources of error in face matching (Lander et al., 2018).

Finally, the novice sample used in this study is not representative of the total population of novices in reality. Their performance may have also been impacted by motivation level. Compared to practitioners, novices do not have much incentive to do well, which may have impacted their reporting and decisions. Although previous research has indicated that DST staff are more motivated than student populations and can even perform comparably to practitioner samples (Heyer et al., 2018), this may not necessarily apply to the current sample, and testing a novice population may not necessarily generalise to a practitioner population.

4.5 Strengths

Despite these limitations, this study boasted numerous strengths by addressing previous methodological concerns. For example, this study did not impose a deadline on participants as previous studies have (White et al., 2015). This was a strength as deadlines have been shown to negatively impact accuracy and may not be representative of real-world tasks.

Additionally, many past studies have grouped children into one overall ‘child’ age group to compare against performance on adult faces, limiting the inferences that can be made about how child age and age variation impacts on performance. This study focused exclusively on children and discriminated between more specific age ranges, exploring performance over three age groups from the ages of 0-15. This is important because of the significant amount of facial change during childhood (Ricanek et al., 2013; White et al., 2015). This allowed comparisons of how performance varied across childhood; finding that the faces of 0-5-year-olds are much harder to match than older children, and providing a basis for further exploration about whether some ages or age ranges within childhood specifically impact performance more than others. This can provide insight into human ability in unfamiliar face matching and can be used in future implementation of training.

Another strength was the use of an operationally valid database. Past studies have been limited by the lack of operational datasets containing child faces and have instead used face stimuli such as greyscale images and images of celebrities’ children due to online availability (Kramer et al., 2018a). Not only were the current study’s images operationally valid but candidate list images were selected by a state-of-the-art facial recognition algorithm. This added to operational validity and produced candidate lists representative of those seen in operational settings (Heyer et al., 2018). Conversely, previous studies have manually selected images on the basis of human similarity rating, limiting generalisability (Kramer et al., 2018a; Megreya & Burton, 2006).

Finally, a majority of studies looking at performance on child images have used a one-to-one task to evaluate performance, making this one of the first studies to use a one-to-many operationally valid experimental paradigm to evaluate novice performance on images across childhood.

4.6 Implications

This study shows that novices perform very poorly on one-to-many tasks when comparing images of children with a 5-year age variation. This poor accuracy decreases further as child age decreases, with 0-5-year-olds being exceptionally difficult to identify. The results confirm that matching child faces is an operationally difficult task, hindered by age-related facial changes and homogeneity between young child faces.

It is important to consider whether these results generalise to practitioners and how they may impact critical identification processes if so. Face matching plays an important role in identifying missing persons, general passport control and the prevention of child trafficking. Extremely low levels of accuracy are a problematic finding when the operational implications for child safety are considered. In the case of child trafficking, low matching accuracy can lead to more young victims being successfully trafficked with fraudulent passports (Michalski, 2017) and potentially never identified. To aid in the prevention of child trafficking, it is essential to understand performance on images of children across childhood and develop effective training programs and guidelines to address the fact that younger children are harder to identify. Results from this study can help contribute to empirical evidence to inform these endeavours.

Additionally, these results pose implications for the human-algorithm relationship. In systems implementing one-to-many checks, it is human users that are required to make final identity decisions, and efficiency is highly constrained by human accuracy (White et al., 2015). Current results imply that candidate list images are easily mistaken for the target, suggesting the capability for algorithms to search large databases of images comes at the cost of significantly increasing task difficulty for practitioners (White et al., 2015).

Aside from the practical standpoint, this research also has theoretical implications, such as contributing to the understanding and building of cognitive models of face recognition and matching (Lander et al., 2018). Finally, these results inform future research to further evaluate performance on images of children across the childhood years.

4.7 Future Directions

The results of this study highlight the need for further research in this area. While unfamiliar face matching with child images has started to gain traction in research circles, there are still great strides to be made in understanding and evaluating performance on these tasks.

Firstly, future research should dig deeper into face matching on child images across childhood and evaluate individual differences. Despite research consistently showing large individual differences between individuals on measures of speed, confidence and accuracy (Heyer et al., 2018; Towler et al., 2019; White et al., 2015; White 2014), many studies use an approach where performance is averaged across participants (Lander et al., 2018). Taking an alternative individual differences approach can be informative and may provide insight into individual abilities, and both potential training and recruitment processes for face matching practitioners.

As this research was conducted with a novice population, it is important to test practitioners to see whether the poor performance found generalises to applied contexts. Previous reports are mixed, with one study finding that practitioners performed comparatively to student populations with no training or experience (White et al., 2014) and others finding that practitioners outperformed novice populations (White et al., 2015). If the current results generalise to an expert population, it is important to explore how we may address poor performance. Therefore, evaluating the effectiveness of training courses and exploring alternative training strategies could be a vital research avenue, helping to ensure that training techniques are producing lasting improvement and practitioners are making optimal decisions. While guidelines for training do exist, these do not reference empirical research to support recommendations (FISWG, 2012; Towler et al., 2019).

Analysing the stability of facial features across childhood and broader lifespan could be a valuable research focus to explore what practitioners might benefit from focusing on when making face matching decisions. For example, it has been suggested that 'super-recognisers' (individuals with superior face matching skills) spend more time fixating on the centre of faces (Kramer et al., 2018a; Towler, White & Kemp, 2017). A survey of practitioners found that the eyes, nose and ears were the most popular features used in decision-making (Heyer et al., 2017). Future research could benefit from exploring whether this generalises to child faces and what facial features might best inform decisions.

4.8 Conclusion

This study aimed to evaluate performance changes across childhood age with an operationally valid 5-year age variation and found that performance for child images was very poor overall, with the worst performance found for the 0-5-year age group. This converges with findings that face matching is an error-prone task and that accuracy decreases as age decreases. This was the case regardless of trial type, with exploration of performance between target-present and target-absent trials finding no significant differences. The ability to make accurate face matching decisions with images of children is vital for a range of applied security settings, such as the prevention of child trafficking. This makes the high error rates found in this study alarming and highlights the need to continue work in this area, by testing whether results generalise to practitioners and whether we can effectively train practitioners to improve their performance on child images. Future research should evaluate this by testing practitioners and evaluating various training techniques.

References

Bobak, A. K., Mileva, V. R., & Hancock, P. J. B. (2018). Facing the facts: Naïve participants have only moderate insight into their face recognition and face perception abilities. *Quarterly Journal of Experimental Psychology, 72*(4), 872-881.

Bone, J., & Blackburn, D. (2003). Biometrics for narcoterrorist watch list applications (Tech. Rep.). Technical report, Crane Division, Naval Surface Warfare Center and DoD.

Bruce, V. (1994). Stability from variation: The case of face recognition. The M.D. Vernon memorial lecture. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology. 47A,* 5-28

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*(3), 207-218.

Burton, A. M., Miller, P., Bruce, V., Hancock, P. J. B., & Henderson, Z. (2001). Human and automatic face recognition: A comparison across image format. Vision Research, 41, 3185–319

Burton, A. M., White, D., & McNeill, A. (2010). The Glasgow Face Matching Test. *Behaviour Research Methods, 42*(1), 286-291.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioural Sciences* (2nd ed.). Hillsdale,New Jersey: Lawrence Erlbaum Associates.

Deb, D., Nain, N., & Jain, A. K. (2018). Longitudinal Study of Child Face Recognition. *2018 International Conference on Biometrics (ICB),* Gold Coast, QLD. Pp. 225-232.

Dunn, J. C. (2010). How to fit models of recognition memory data using maximum likelihood. *International Journal of Psychological Research, 3*(1), 140-149.

FISWG (2012). *Guidelines and recommendation for facial comparison methods* (Version 1.0). Retrieved from https://fiswg.org/FISWG_GuidelinesforFacialComparisonMethods_v1.0_2012_02_02 .pdf

Fysh, M. C., & Bindemann, M. (2018). Human-Computer Interaction in Face Matching. *Cognitive Science, 42*, 1714-1732.

García-Zurdo, R., Frowd, C. D., & Manzanero, A. L. (2018). Effects of facial periphery on unfamiliar face recognition. *Current Psychology,* 1-7.

Graves, I., Butavicius, M., MacLeod, V., Heyer, R., Parsons, K., Kuester, N., & Johnson, R. (2011). The role of the human operator in image-based airport security technologies. In E. A. L. Jain, & C. Abeynakake (Eds.), *Innovations in Defence Support Systems* (pp. 147- 181). Heidelberg, Germany: Springer.

Grother, P., Quinn, G. W., Phillips, P., J. (2011). Report on the Evaluation of 2D Still-Image Face Recognition Algorithms. NIST, Gaithersburg, MD, USA. Rep. 7709.

Heyer, R., & Semmler, C. (2013). Forensic confirmation bias: The case of facial image comparison. *Journal of Applied Research in Memory and Cognition, 2,* 68-70.

Heyer, R., Chong, C., & Semmler, C. (2019) Facial image comparisons of morphed facial imagery, Australian Journal of Forensic Sciences, 51:sup1, S5-S9.

Heyer, R., MacLeod, V., Carter, L., Semmler, C., & Ma-Wyatt, A. (2017). *Profiling the Facial Comparison Practitioner in Australia*. DST-GD-XXXX, DST Edinburgh, South Australia.

Heyer, R., Semmler, C., & Hendrickson, A. T. (2018). Humans and algorithms for facial recognition: the effects of candidate list length and experience on performance. *Journal of Applied Research in Memory and Cognition, 7*, 597-609.

Hole, M., McLindin, B., Hanton, K., Malec, C., Yiu, S.Y., & Hanly, G. (2015). An Overview of a DSTO Developed Human Operator Image Comparison Software Tool – Comparer. DSTO-GD-0855. Defence Science and Technology Organisation, Edinburgh.

Jain, A. K., Nandakumar, K., & Ross, A. (2016). 50 years of biometric research: Accomplishments, challenges, and opportunities. *Pattern Recognition Letters, 79*, 80–105.

Johnson, M. H., Dziurawiec, S., Ellis, H., & Morton, J. (1991). Newborns' preferential tracking of face-like stimuli and its subsequent decline. *Cognition, 40*(1-2), 1-19.

Kemp, R. I., Caon, A., Howard, M., & Brooks, K. R. (2016). Improving unfamiliar face matching by masking the external facial features. *Applied Cognitive Psychology, 30*(4).

Kramer, R. S. S., Mulgrew, J., & Reynolds, M. G. (2018a). Unfamiliar face matching with photographs of infants and children. *PeerJ*, *6*, e5010.

Kramer, R. S. S., Young, A. W., & Burton, A. M. (2018b). Understanding face familiarity. *Cognition, 172*, 46-58.

Lander, K., Bruce, V., & Bindemann, M. (2018). Use-inspired basic research on individual differences in face identification: implications for criminal investigation and security. *Cognitive Research: Principles and Implications, 3*:26.

Lewandowsky, S., & Farrell, S. (2011). *Computational Modeling in Cognition: Principles and Practice.* Thousand Oaks, CA: SAGE Publications, Inc.

Ling, H., Soatto, S., Ramanathan, N., & Jacobs, D. W. (2010). Face verification across age progression using discriminative methods. *IEEE Transactions on Information Forensics and Security, 5*(1), 82–91.

Lui, Y. M., Bolme, D., Draper, B. A., Beveridge, J. R., Givens, G., & Phillips, P. J. (2009). A meta-analysis of face recognition covariates. *Proceedings of the Third International Conference on Biometrics: Theory, Applications, and Systems*, 139–146.

Ma, W. J., Shen, S., Dziugaite, G., & van den Berg, R. (2015). Requiem for the max rule? *Vision Research, 116,* 179-193.

Mahalingham, G., & Kambhamettu, C. (2012). Face verification of age separated images under the influence of internal and external factors. *Image and Vision Computing, 30*, 1052-1061.

Matthews, C. M., & Mondloch, C. J. (2018). Improving Identity Matching of Newly Encountered Faces: Effects of Multi-image Training. *Journal of Applied Research in Memory and Cognition, 7*, 280-290.

McCaffery, J. M., Robertson, D. J., Young, A. W., Burton, A. M. (2018). Individual differences in face identity processing. *Cognitive Research: Principles and Implications, 3*:21.

Megreya, A. M., & Burton, A. M. (2006). Unfamiliar faces are not faces: Evidence from a matching task. *Memory & Cognition, 34*(4), 865–876.

Megreya, A. M., & Burton, A. M. (2008). Matching faces to photographs: Poor performance in eyewitness memory (without the memory). *Journal of Experimental Psychology: Applied, 14*(4), 364-372.

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*(6), 700–706.

Meissner, C. A., Susa, K. J., & Ross, A. B. (2013). Can I see your passport please? Perceptual discrimination of own- and other-race faces. *Visual Cognition, 21*(9-10), 1287-1305.

Michalski, D. J. (2017). *The impact of age-related variables on facial comparisons with images of children: algorithm and practitioner performance* (Doctoral dissertation). The University of Adelaide, Australia.

Michalski, D., Yiu, S. Y., & Malec, C. (2018). The Impact of Age and Threshold Variation on Facial Recognition Algorithm Performance using Images of Children. *2018 International Conference on Biometrics (ICB),* Gold Coast, QLD.

Mickes, L. (2015). Receiver operating characteristic analysis and confidence-accuracy characteristic analysis in investigations of system variables and estimator variables that affect eyewitness memory. *Journal of Applied Research in Memory and Cognition, 4*, 93-102.

Mileva, M., & Burton, A. M. (2018). Smiles in face matching: Idiosyncratic information revealed through a smile improves unfamiliar face matching performance. *British Journal of Psychology, 109,* 799-811.

O'Toole, A. J., Phillips, P. J., & Narvekar, A. (2008). Humans versus algorithms: Comparisons from the Face Recognition Vendor Test 2006. *8th IEEE International Conference on Automatic Face & Gesture Recognition,* Amsterdam, pp. 1-6.

Papesh, M. H. (2018). Photo ID verification remains challenging despite years of practice. *Cognitive Research: Principles and Implications, 3*:19.

Pearce, T. (2018). *The performance of novices with image pairs of children compared to adults on a face matching task.* (Honours dissertation). University of Adelaide, Australia.

Phillips, P. J., & O'Toole, A. J. (2014). Comparison of human and computer performance across face recognition experiments. *Image and Vision Computing, 32*(1), 74-85.

Phillips, P. J., & O'Toole, A. J. (2014). Comparison of human and computer performance across face recognition experiments. *Image and Vision Computing, 32,* 74-85.

Ramanathan, N., & Chellappa, R. (2006). Face verification across age progression. *IEEE Transactions on Image Processing, 15*(11), 3349–3361.

Ricanek, K., Mahalingam, G., Albert, A. M., & Vorder Bruegge, R. W. (2013). Human face ageing: A perspective analysis from anthropometry and biometrics. In M. Fairhurst (Ed.), *Age factors in biometric processing* (pp. 93–116). London, UK: The Institution of Engineering and Technology.

Ritchie, K. L., Smith, F. G., Jenkins, R., Bindemann, M., White, D., & Burton, A. M. (2015). Viewers base estimates of face matching accuracy on their own familiarity: Explaining the photo-ID paradox. *Cognition, 141*, 161-169.

Rouder, J. N., & Lu, J. (2005). An introduction to Bayesian hierarchical models with an application in the theory of signal detection. *Psychonomic Bulletin & Review*, *12*(4), 573-604.

Semmler, C., Kaesler, M., Dunn, J. (February, 2018). *An Introduction to Maximum Likelihood Estimation for Signal Detection Theory Applications to Eyewitness Identification Data*. Workshop presented at the SARMAC regional meeting. Adelaide, South Australia.

Snyder, G. (2018). *Human age estimation performance based on facial images: Potential implications for refugee processing* (Honours dissertation). University of Adelaide, Australia.

Srinivas, N., Ricanek, K., Michalski, D., Bolme, D. S., King, M. (2019). *Face Recognition Algorithm Bias: Performance Differences on Images of Children and Adults.* Paper presented at the 2019 IEEE Conference on Computer Vision and Pattern Recognition Workshops, Long Beach, CA.

Stephens, R. G., Semmler, C., & Sauer, J. D. (2017). The effect of the proportion of mismatching trials and task orientation on the confidence-accuracy relationship in unfamiliar face matching. *Journal of Experimental Psychology: Applied, 23*(3), 336-353.

Taylor, K. T. (2001). *Forensic Art and Illustration*. Boca Raton, FL: CRC Press.

Tistarelli, M., Yadav, D., Vatsu, M., & Singh, R. (2013). Short- and long-time ageing effects in face recognition. In M. Fairhurst (Ed.), *Age factors in biometric processing* (pp. 253–275). London, UK: The Institution of Engineering and Technology.

Towler, A., Kemp, R. I., Burton, A. M., Dunn, J. D., Wayne, T., Moreton, R., White, D. (2019). Do professional facial image comparison training courses work? *PLoS ONE, 14*(2): e0211037.

Towler, A., White, D., Kemp, R. I. (2017). Evaluating the feature comparison strategy for forensic face identification. *Journal of Experimental Psychology: Applied, 23*(1), 47-58.

U. S. Department of State (2006). Trafficking in persons report. Retrieved from: https://2009-2017.state.gov/documents/organization/66086.pdf

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

White, D., Kemp, R., Jenkins, R., & Burton, A. M. (2014). Feedback training for facial image comparison. *Psychometric Bulletin & Review, 21*(1), 100-106.

White, D., Kemp, R., Jenkins, R., Matheson, M., & Burton, A. M. (2014). Passport officers' errors in face matching. *PLoS ONE, 9*(8), 1-6.

Wolfe, J. M., Horowitz, T. S., & Kenner, N. M. (2005). Cognitive psychology: 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*(4), 623-638.

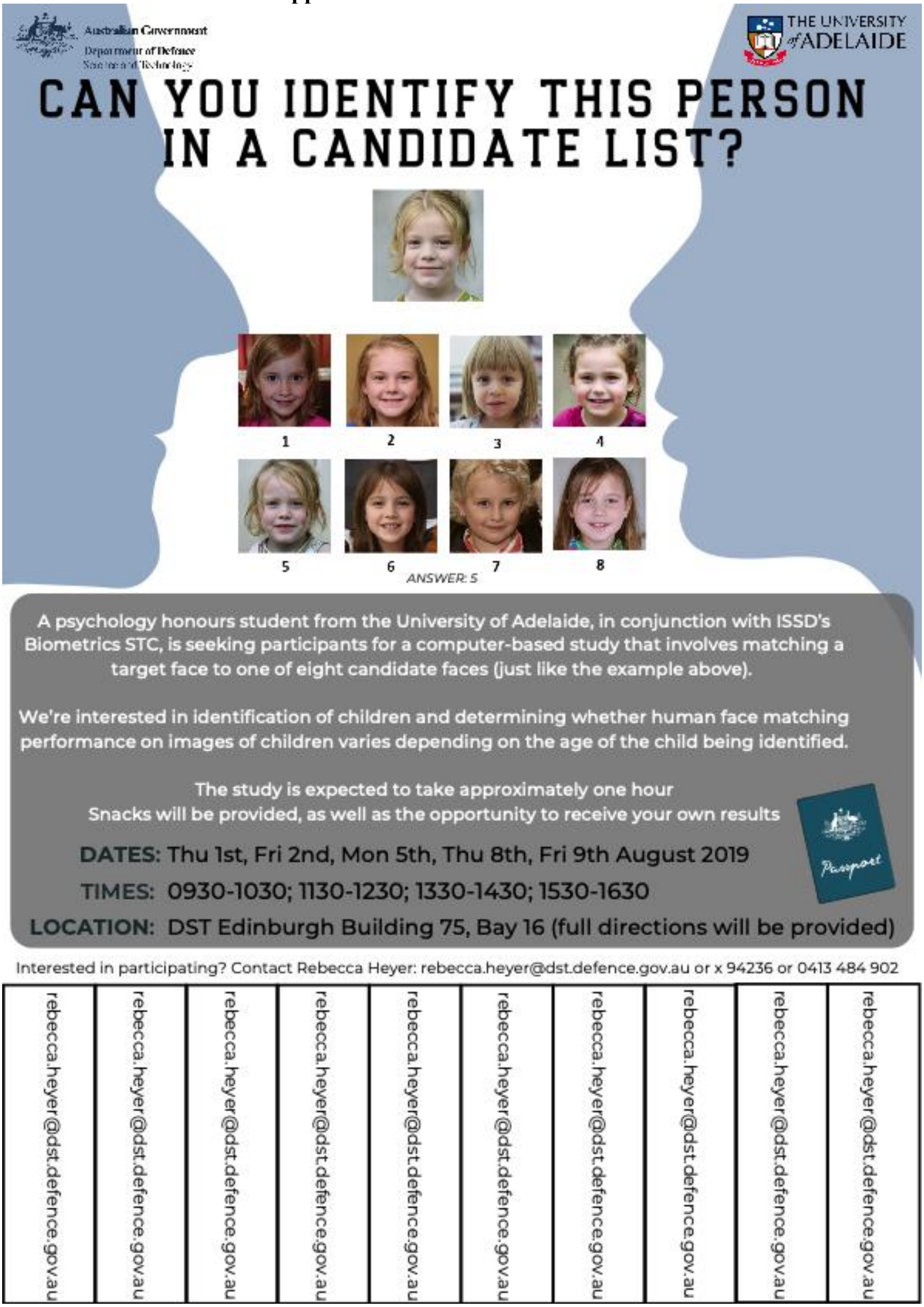
Yadav, D., Singh, R., Vatsa, M., & Noore, A. (2014). Recognising age-separated face images: Humans and Machines. *PLoS ONE, 9*(12), e112234.

Young, A. W., Hay, D. C., McWeeny, K. H., Flude, B. M., & Ellis, A. W. (1985). Matching familiar and unfamiliar faces on internal and external features. *Perception, 14*(6), 737-46.

Zeng, J., Ling, H., Latecki, L. J., Fitzhugh, S., & Guo, G., (2012). Analysis of facial images across age progression by humans. *ISRN Machine Vision*, Article ID 505974.

Zhang, S. X. (2007). *Smuggling and human trafficking of human beings: All roads lead to America.* Westport, CT: Praegar Publishers.

Appendix A: Recruitment Poster



Australian Government
Department of Defence
Science and Technology

THE UNIVERSITY of ADELAIDE

CAN YOU IDENTIFY THIS PERSON IN A CANDIDATE LIST?

ANSWER: 5

A psychology honours student from the University of Adelaide, in conjunction with ISSD's Biometrics STC, is seeking participants for a computer-based study that involves matching a target face to one of eight candidate faces (just like the example above).

We're interested in identification of children and determining whether human face matching performance on images of children varies depending on the age of the child being identified.

The study is expected to take approximately one hour
Snacks will be provided, as well as the opportunity to receive your own results

DATES: Thu 1st, Fri 2nd, Mon 5th, Thu 8th, Fri 9th August 2019

TIMES: 0930-1030; 1130-1230; 1330-1430; 1530-1630

LOCATION: DST Edinburgh Building 75, Bay 16 (full directions will be provided)

Passport

Interested in participating? Contact Rebecca Heyer: rebecca.heyer@dst.defence.gov.au or x 94236 or 0413 484 902

rebecca.heyer@dst.defence.gov.au
rebecca.heyer@dst.defence.gov.au
rebecca.heyer@dst.defence.gov.au
rebecca.heyer@dst.defence.gov.au
rebecca.heyer@dst.defence.gov.au
rebecca.heyer@dst.defence.gov.au
rebecca.heyer@dst.defence.gov.au
rebecca.heyer@dst.defence.gov.au
rebecca.heyer@dst.defence.gov.au
rebecca.heyer@dst.defence.gov.au

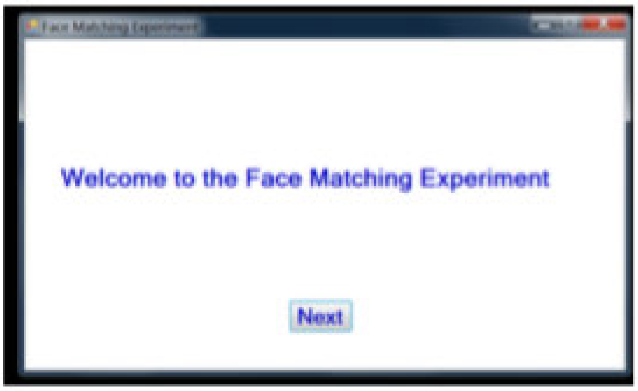
Appendix B – Design Specification

Child Face Matching Honours Project 2019

Experiment Overview

The aim of this experience is to determine whether a target face is present within a candidate list of faces five years younger than the target.

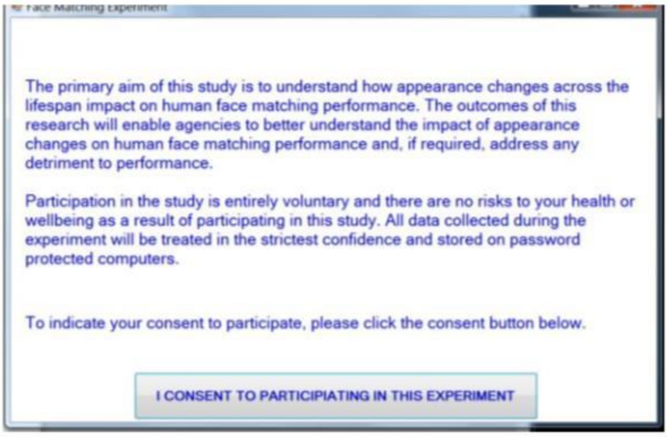
The experiment should start with a welcome screen. See example below:



Welcome to the Face Matching Experiment

For this experiment the 'Welcome to the Face Matching Experiment' title would be appropriate to keep.

Participants then click next to take them to an introduction and consent screen – containing a basic overview of the experiment and asking participants to click a button to indicate their consent to participate. See example below:



The primary aim of this study is to understand how appearance changes across the lifespan impact on human face matching performance. The outcomes of this research will enable agencies to better understand the impact of appearance changes on human face matching performance and, if required, address any detriment to performance.

Participation in the study is entirely voluntary and there are no risks to your health or wellbeing as a result of participating in this study. All data collected during the experiment will be treated in the strictest confidence and stored on password protected computers.

To indicate your consent to participate, please click the consent button below.

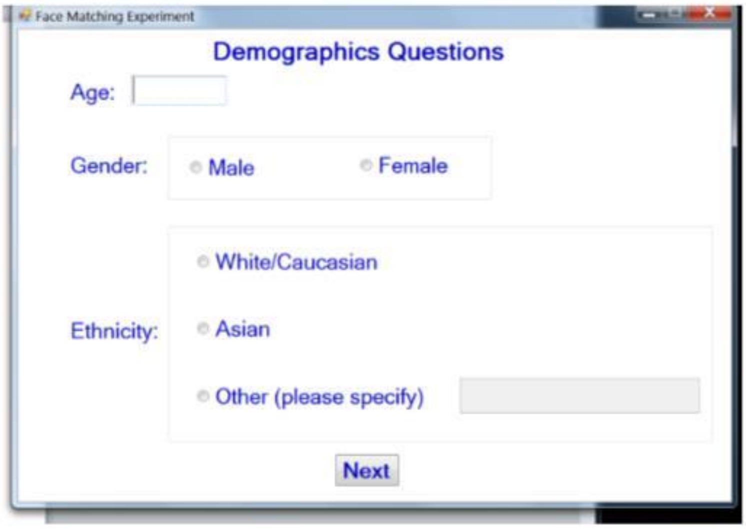
For this experiment, the basic overview paragraph will be changed to…

"The primary aim of this study is to understand whether changes to the face that occur throughout childhood impact on human face matching performance when people are presented with a candidate list of potential matches that are five years younger than the target identity. The outcomes of this research will enable agencies to better understand the impact of childhood age-related changes on human face matching performance and, if required, address any detriment to performance.

Participation in the study is entirely voluntary and there are no risks to your health or wellbeing as a result of participating in this study. All data collected during the experiment will be treated in the strictest confidence and stored on password protected computers.

To indicate your consent to participate, please click on the consent button below."

The following two screens collect basic demographic information (age, gender, ethnicity), as shown in the examples below, but with "Demographics Questions" changed to "Demographic Information (for statistical purposes)"



Demographics Questions

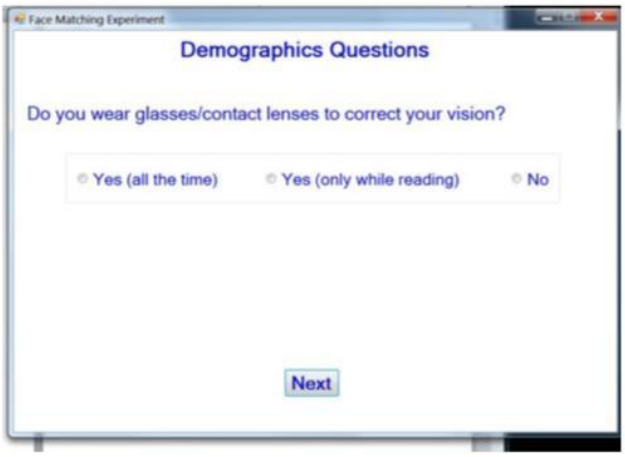
Age:

Gender: ○ Male ○ Female

Ethnicity:
○ White/Caucasian
○ Asian
○ Other (please specify)

It would be best to incorporate an error check for age >0 and <100. All questions must be answered to progress. (grey out next until all questions answered if possible).

The next page should ask whether participants wear glasses/contact lenses to correct their vision, see example below. This should be like the example but with the title "Vision Correction" instead of "Demographic Questions". Please also include a note to say "please ensure you are wearing your glasses or contact lenses throughout the experiment if required.

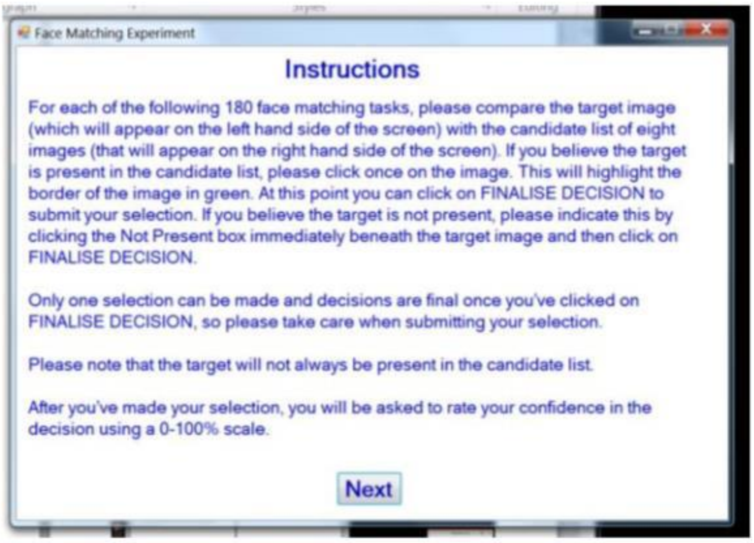


Demographics Questions

Do you wear glasses/contact lenses to correct your vision?

○ Yes (all the time) ○ Yes (only while reading) ○ No

The experimental portion of the application begins with an instruction screen, see example below.



Instructions

For each of the following 180 face matching tasks, please compare the target image (which will appear on the left hand side of the screen) with the candidate list of eight images (that will appear on the right hand side of the screen). If you believe the target is present in the candidate list, please click once on the image. This will highlight the border of the image in green. At this point you can click on FINALISE DECISION to submit your selection. If you believe the target is not present, please indicate this by clicking the Not Present box immediately beneath the target image and then click on FINALISE DECISION.

Only one selection can be made and decisions are final once you've clicked on FINALISE DECISION, so please take care when submitting your selection.

Please note that the target will not always be present in the candidate list.

After you've made your selection, you will be asked to rate your confidence in the decision using a 0-100% scale.

For this experiment, the instruction will be:

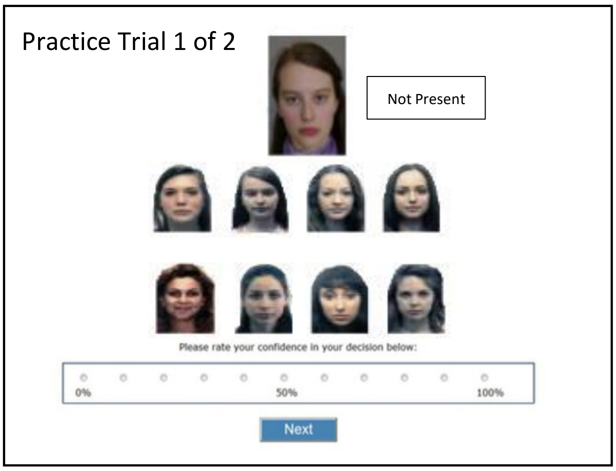
“In this experiment you will be asked to compare a target image (which will appear on top of the screen) with a candidate list of eight images (which will appear in two rows of four below the target image). If you believe the target is present in the candidate list, please click the image in the candidate list you believe is a ‘match’ to the target. This will highlight the border of the image in green. If you believe the target is not present in the candidate list, please indicate this by clicking the ‘Not Present’ box next to the target image. You will then be required to rate your confidence on a 0-100 percent scale below the candidate list.

Only one selection can be made and decisions are final once you click NEXT, so please be sure to take care when selecting your decision. Please note that the target may not always be present in the candidate list.

Once you have selected either a face or 'not present' and rated your confidence, you can click on NEXT to submit your selection.

You will now have the opportunity to complete a couple of practice trials before the experiment begins"

2 practice trials should then be included, the page should not be able to progress until a selection and confidence rating has been made. Selecting a face should lead to [a green box around the chosen face or a checkbox ticked underneath - CHECK] so participants can see their decision. Please see trial example below (some more space between the face and "practice trial 1 of 2" would be ideal)



Practice Trial 1 of 2

Not Present

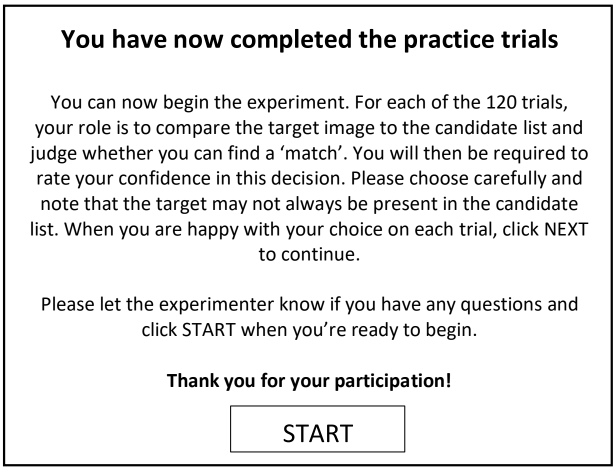
Please rate your confidence in your decision below:

○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○

0% 50% 100%

Next

Once practice trials are complete, a screen should come up with further instructions exactly as below:



You have now completed the practice trials

You can now begin the experiment. For each of the 120 trials, your role is to compare the target image to the candidate list and judge whether you can find a 'match'. You will then be required to rate your confidence in this decision. Please choose carefully and note that the target may not always be present in the candidate list. When you are happy with your choice on each trial, click NEXT to continue.

Please let the experimenter know if you have any questions and click START when you're ready to begin.

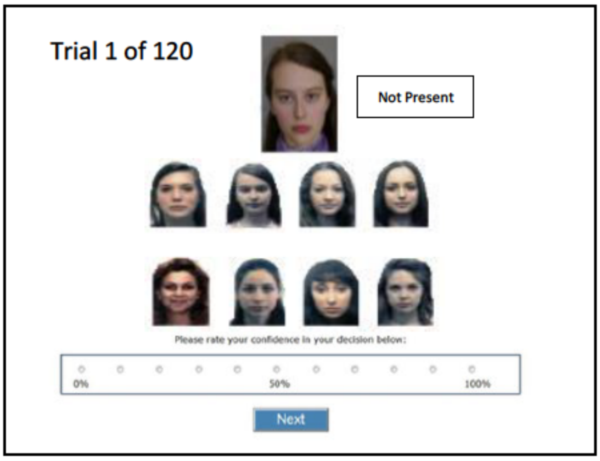
Thank you for your participation!

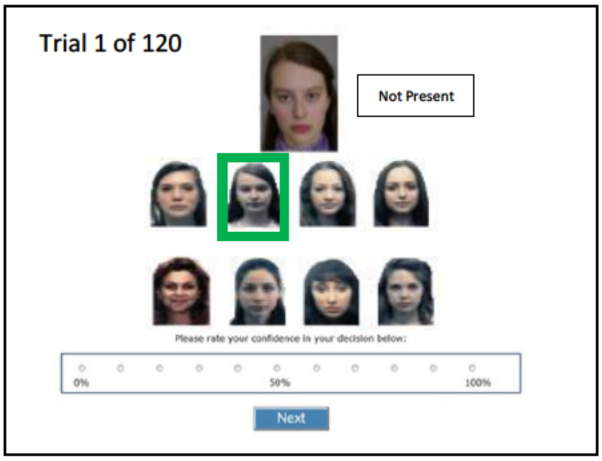
START

120 trials will be presented on the screen (one at a time) – with a target face on the top of the screen and a candidate list of 8 face images appearing in 2 rows of 4 below. When participants select a face (or the not present button) It should highlight the border of the image in green or highlight a checkbox underneath. See example below

There should also be a confidence rating scale below the candidate list images (see below) – please make this a five-point scale (where 1 is not at all confident and 5 is completely confident).

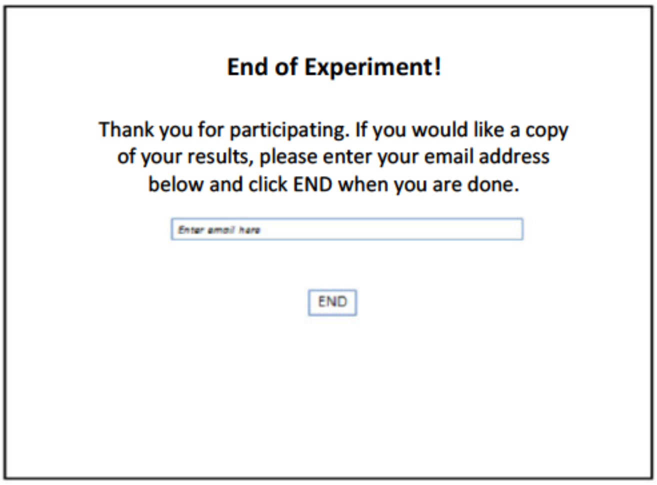
An indication of progress should also be visible on the screen – either in the top corner or middle of the screen – see below.





The 120 trials should be presented to participants in randomised order.

The final screen will then contain a register for results (see below)

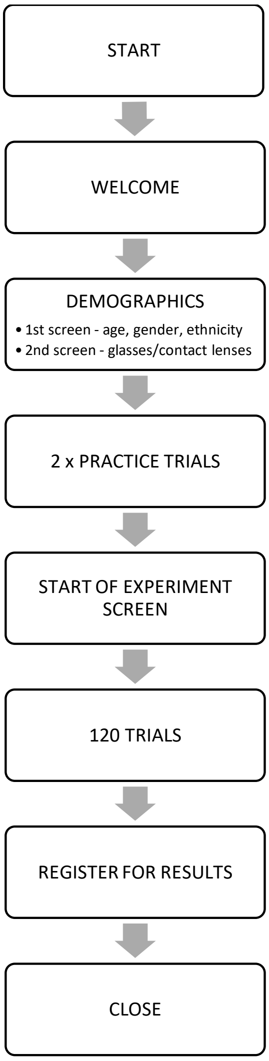


End of Experiment!

Thank you for participating. If you would like a copy of your results, please enter your email address below and click END when you are done.

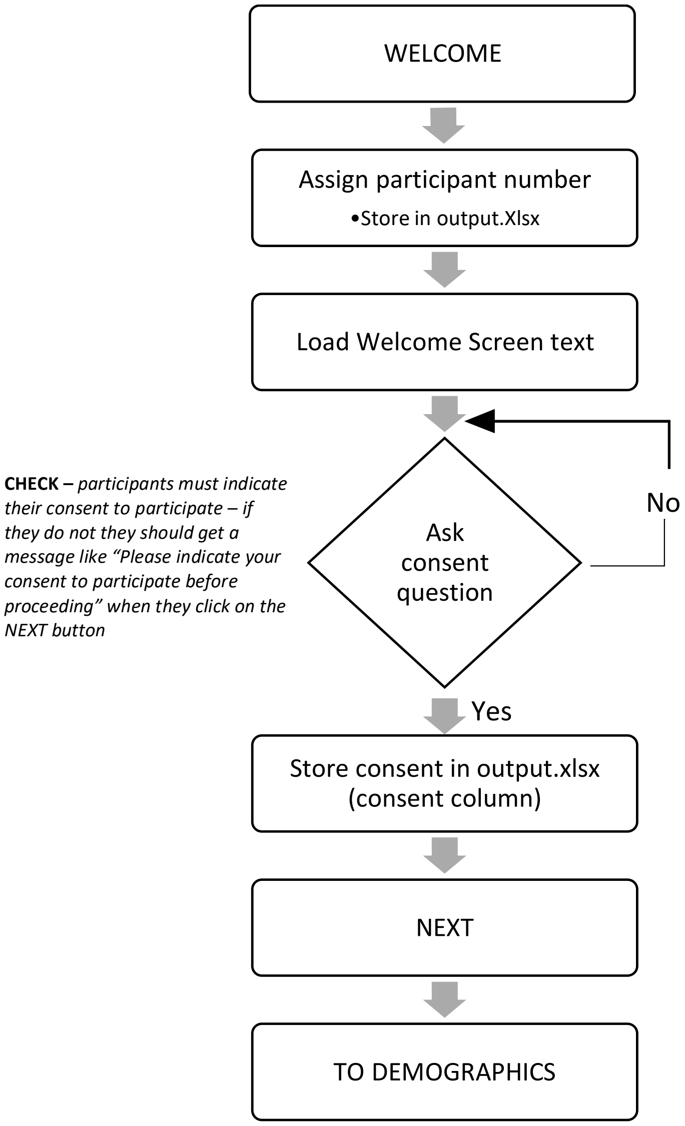
Participants should be able to click END whether they input an email address or not

FLOW CHART OF OVERALL EXPERIMENT


START
WELCOME
DEMOGRAPHICS
• 1st screen - age, gender, ethnicity
• 2nd screen - glasses/contact lenses
2 x PRACTICE TRIALS
START OF EXPERIMENT SCREEN
120 TRIALS
REGISTER FOR RESULTS
CLOSE

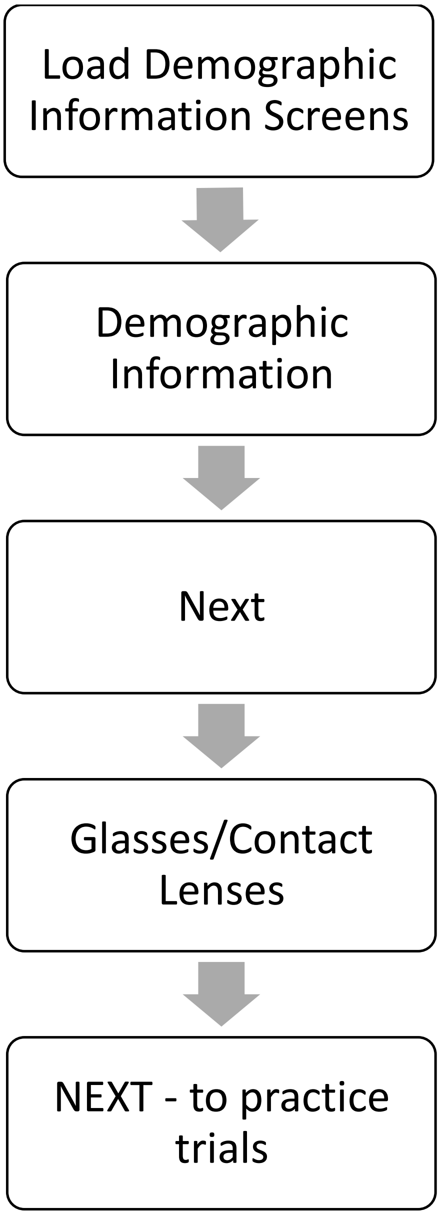
FLOWCHARTS FOR EACH PART OF THE EXPERIMENT

ome:

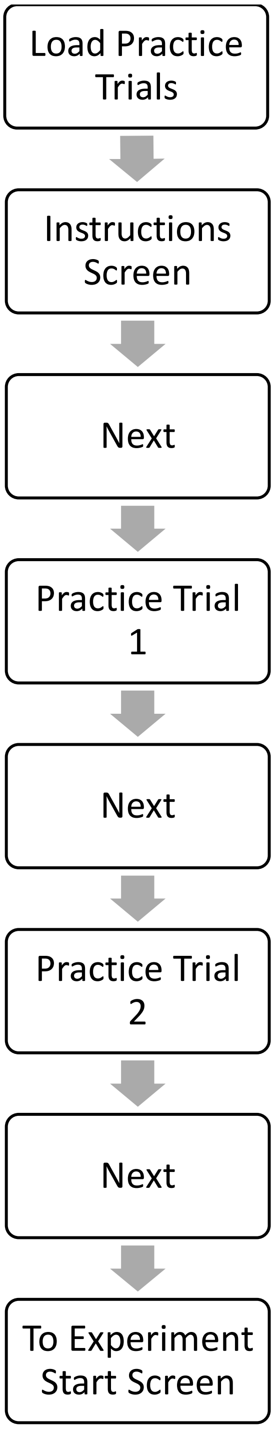


CHECK – *participants must indicate their consent to participate – if they do not they should get a message like "Please indicate your consent to participate before proceeding" when they click on the NEXT button*

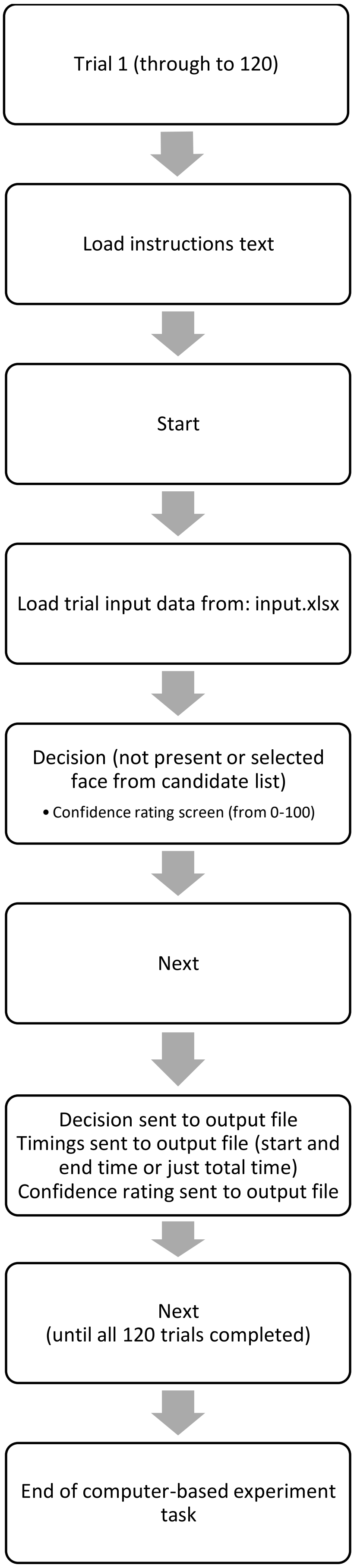
Demographics:


Load Demographic Information Screens
Demographic Information
Next
Glasses/Contact Lenses
NEXT - to practice trials

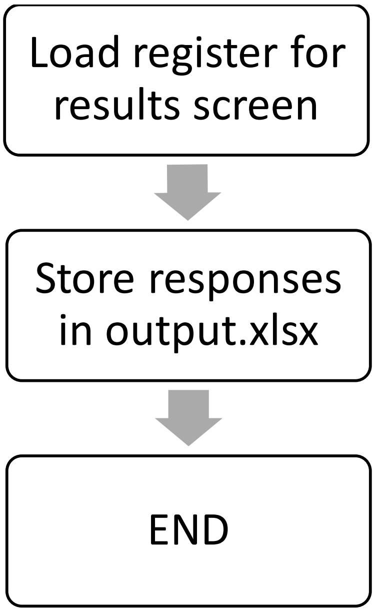
Practice Trials:



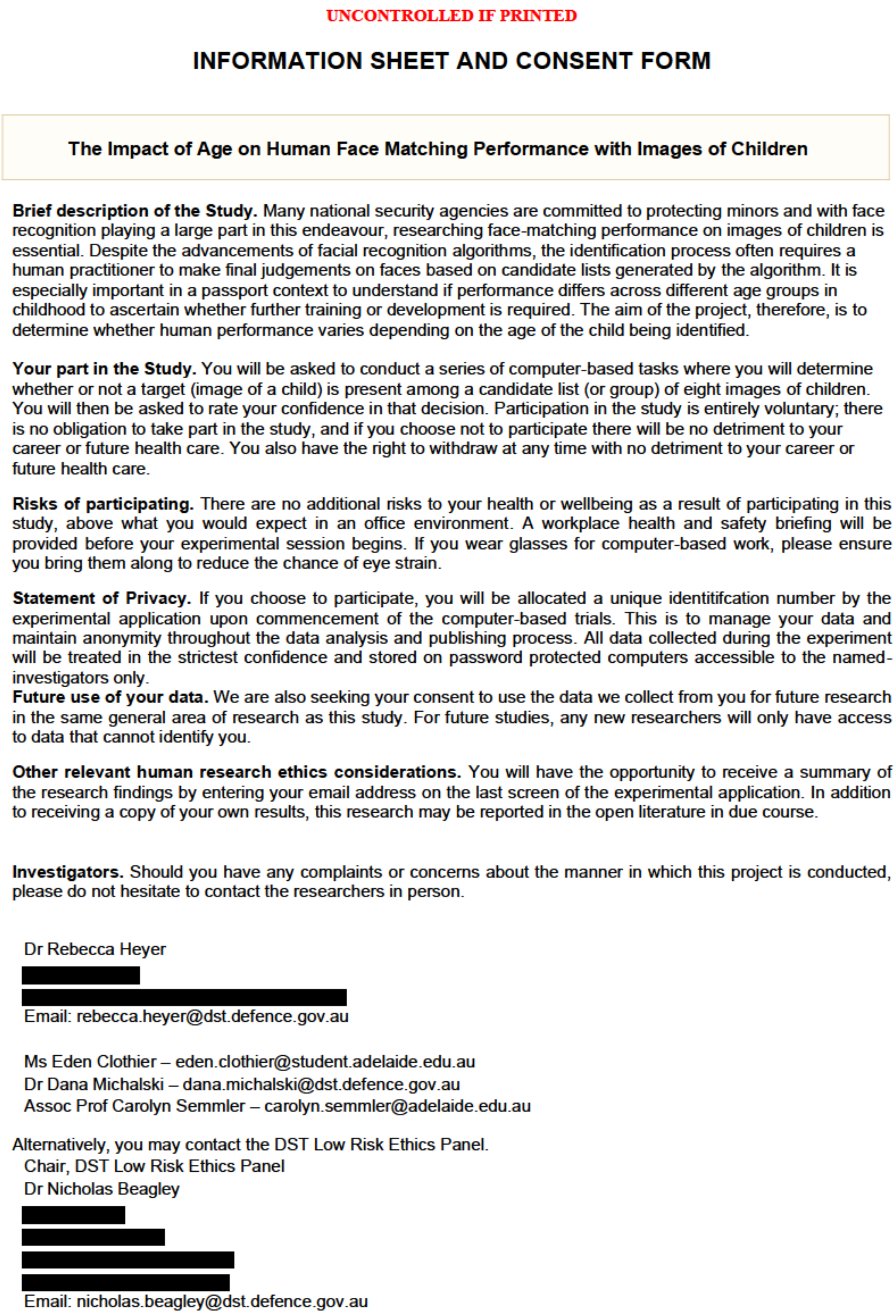
Experimental Trials:


Trial 1 (through to 120)
Load instructions text
Start
Load trial input data from: input.xlsx
Decision (not present or selected face from candidate list)
• Confidence rating screen (from 0-100)
Next
Decision sent to output file
Timings sent to output file (start and end time or just total time)
Confidence rating sent to output file
Next
(until all 120 trials completed)
End of computer-based experiment task

Final Screen:


Load register for results screen
Store responses in output.xlsx
END

Appendix C – Information Sheet and Consent Form



UNCONTROLLED IF PRINTED

INFORMATION SHEET AND CONSENT FORM

The Impact of Age on Human Face Matching Performance with Images of Children

Brief description of the Study. Many national security agencies are committed to protecting minors and with face recognition playing a large part in this endeavour, researching face-matching performance on images of children is essential. Despite the advancements of facial recognition algorithms, the identification process often requires a human practitioner to make final judgements on faces based on candidate lists generated by the algorithm. It is especially important in a passport context to understand if performance differs across different age groups in childhood to ascertain whether further training or development is required. The aim of the project, therefore, is to determine whether human performance varies depending on the age of the child being identified.

Your part in the Study. You will be asked to conduct a series of computer-based tasks where you will determine whether or not a target (image of a child) is present among a candidate list (or group) of eight images of children. You will then be asked to rate your confidence in that decision. Participation in the study is entirely voluntary; there is no obligation to take part in the study, and if you choose not to participate there will be no detriment to your career or future health care. You also have the right to withdraw at any time with no detriment to your career or future health care.

Risks of participating. There are no additional risks to your health or wellbeing as a result of participating in this study, above what you would expect in an office environment. A workplace health and safety briefing will be provided before your experimental session begins. If you wear glasses for computer-based work, please ensure you bring them along to reduce the chance of eye strain.

Statement of Privacy. If you choose to participate, you will be allocated a unique identitifcation number by the experimental application upon commencement of the computer-based trials. This is to manage your data and maintain anonymity throughout the data analysis and publishing process. All data collected during the experiment will be treated in the strictest confidence and stored on password protected computers accessible to the named-investigators only.

Future use of your data. We are also seeking your consent to use the data we collect from you for future research in the same general area of research as this study. For future studies, any new researchers will only have access to data that cannot identify you.

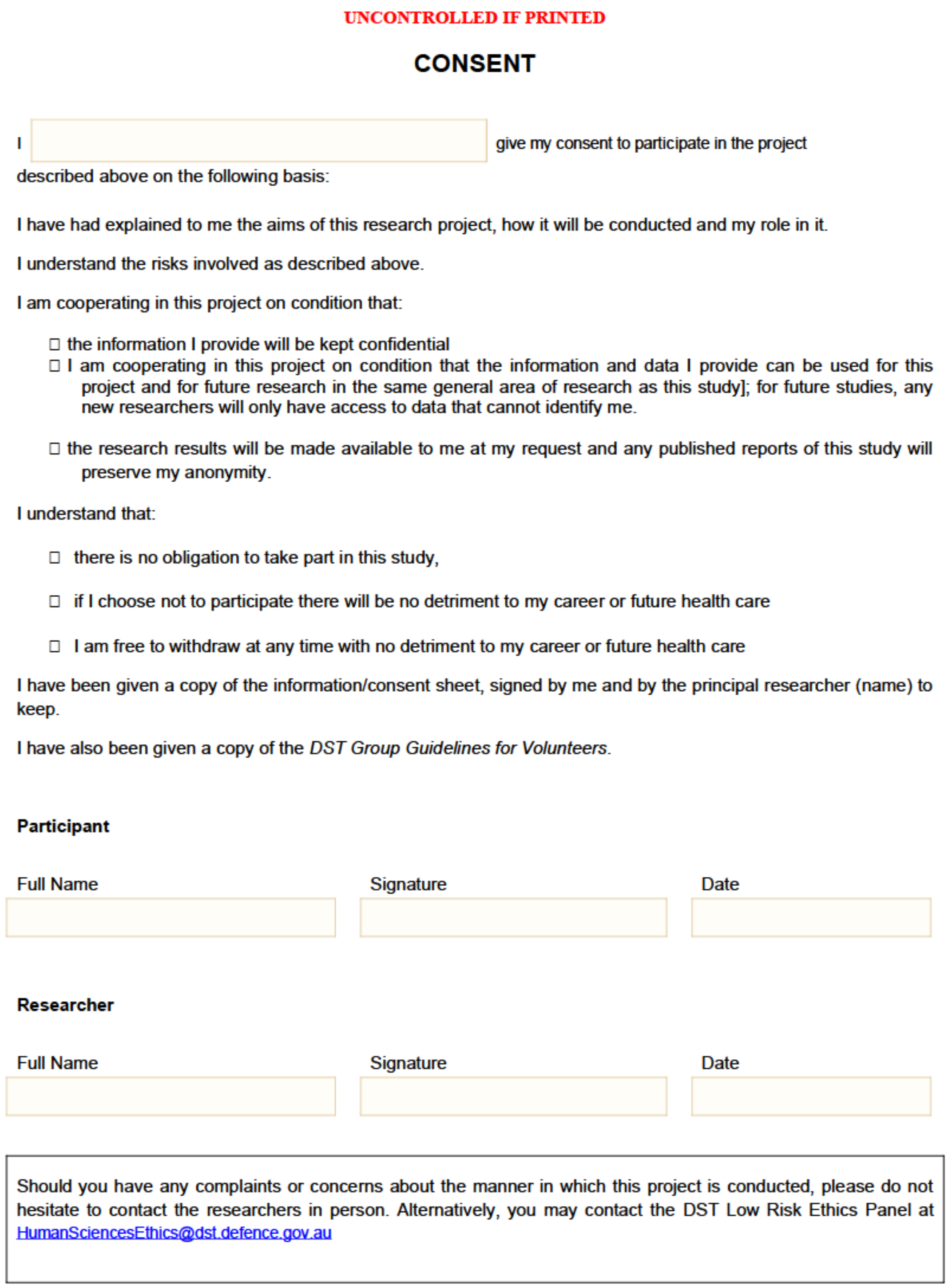
Other relevant human research ethics considerations. You will have the opportunity to receive a summary of the research findings by entering your email address on the last screen of the experimental application. In addition to receiving a copy of your own results, this research may be reported in the open literature in due course.

Investigators. Should you have any complaints or concerns about the manner in which this project is conducted, please do not hesitate to contact the researchers in person.

Dr Rebecca Heyer
[redacted]
Email: rebecca.heyer@dst.defence.gov.au

Ms Eden Clothier – eden.clothier@student.adelaide.edu.au
Dr Dana Michalski – dana.michalski@dst.defence.gov.au
Assoc Prof Carolyn Semmler – carolyn.semmler@adelaide.edu.au

Alternatively, you may contact the DST Low Risk Ethics Panel.
Chair, DST Low Risk Ethics Panel
Dr Nicholas Beagley
[redacted]
Email: nicholas.beagley@dst.defence.gov.au



UNCONTROLLED IF PRINTED

CONSENT

I ______ give my consent to participate in the project described above on the following basis:

I have had explained to me the aims of this research project, how it will be conducted and my role in it.

I understand the risks involved as described above.

I am cooperating in this project on condition that:

- □ the information I provide will be kept confidential
- □ I am cooperating in this project on condition that the information and data I provide can be used for this project and for future research in the same general area of research as this study]; for future studies, any new researchers will only have access to data that cannot identify me.
- □ the research results will be made available to me at my request and any published reports of this study will preserve my anonymity.

I understand that:

- □ there is no obligation to take part in this study,
- □ if I choose not to participate there will be no detriment to my career or future health care
- □ I am free to withdraw at any time with no detriment to my career or future health care

I have been given a copy of the information/consent sheet, signed by me and by the principal researcher (name) to keep.

I have also been given a copy of the *DST Group Guidelines for Volunteers*.

Participant

Full Name	Signature	Date

Researcher

Full Name	Signature	Date

Should you have any complaints or concerns about the manner in which this project is conducted, please do not hesitate to contact the researchers in person. Alternatively, you may contact the DST Low Risk Ethics Panel at HumanSciencesEthics@dst.defence.gov.au

Appendix D – DST Volunteer Guidelines

DST GUIDELINES FOR VOLUNTEERS

Thank you for taking part in Defence Science and Technology (DST) Group Research. Your involvement is much appreciated. This pamphlet explains your rights as a volunteer.

DST ethics review process

- DST Group has developed an approval process for low-risk research to ensure that human research complies with the requirements of the NHMRC (2007) *National Statement on Ethical Conduct in Human Research* and the Department of Defence (2007) *Health Manual Volume 23 Human Research in Defence – Instructions for Researchers*.
- If you are told that the project has DST ethics approval, this means that the Chief of Division or the DST Low Risk Ethics Panel has reviewed the research proposal and has agreed that the research is low-risk and is ethical. Ethical clearance through the Department of Defence and Veteran Affairs Human Research Ethics Committee (DDVA HREC) is not required for low-risk research.
- DST approval does not imply any obligation on commanders to order or encourage their service personnel to participate or to release troops from their usual workplace to participate. Obviously, the use of any particular personnel must have clearance from their commanders but commanders should not use DST Group approval to pressure personnel into volunteering.

Voluntary participation

- As you are a volunteer for this research project, you are under **no obligation** to participate or continue to participate. You may withdraw from the project **at any time** without detriment to your military career or to your medical care.
- At no time must you feel pressured to participate or to continue if you do not wish to do so.
- If you do not wish to continue, it would be useful to the researcher to know why, but you are under no obligation to give reasons for not wanting to continue.

Informed consent

- Before commencing the project you will have been given an information sheet which explains the project, your role in it and any risks to which you may be exposed.
- You must be sure that you understand the information given to you and that you ask the researchers about anything of which you are not sure.
- You should ensure you are satisfied that you understand the information sheet and agree to participate, and keep a copy.
- Before you participate in the project you should also have been given a consent form to sign. You must be happy that the consent form is easy to understand and spells out what you are agreeing to. If you are happy you should sign the consent form and keep an un-signed copy of the consent form.

Complaints

- If at any time during your participation in the project you are worried about how the project is being run or how you are being treated, then you should speak to the researchers.
- Alternatively, you can contact the Chair of the DST Low Risk Ethics Panel. Contact details are:

Chair, DST Group ERP
Dr Nicholas Beagley
[redacted]
Email: nicholas beagley@dst.defence.gov au

Appendix E – Verbal Briefing

VERBAL SPIEL FOR CHILD FACE MATCHING STUDY

1. Hi everyone, thank you for coming today! I'm Eden.

2. Before we get started I just want to let you know that in the case of an emergency please follow the emergency wardens to the assembly area which is on the grass by the BBQ area at the southern end of 75L. The men's toilets are adjacent this lab and the women are further down the corridor towards the northern end of the building, on your right.

3. Today you're participating in a computer-based study of face matching performance. A little background on today's study: Many security agencies are committed to protecting minors and with face recognition playing a large part in this endeavour, researching performance using images of children is vital. Despite advancements of recognition algorithms, the identification process often requires a human practitioner to make final judgements on faces based on candidate lists generated by the algorithm.

4. It is especially important in a passport context to understand if performance differs across different age groups in childhood to ascertain whether further training or development is required. Based on this, the aim of this project is to determine whether human performance varies depending on the age of the child being identified.

5. Within this study, you will be asked to conduct a series of computer-based tasks where you will determine whether or not a target image is present among a group of eight images of children. You will then be asked to rate your confidence. Participation is completely voluntary and you are free to withdraw at any time during the experimental session.

6. The study consists of three elements:
    a. Demographic questions
    b. Face matching study (including a couple of examples for practice)
    c. Register for a copy of results (prompt to put email address in)

7. Before we start, have you all read the information sheet and guidelines for volunteers provided to you in email? If not, please read through them as quickly as possible.

8. Any questions before starting?
9. If you're happy to participate please click NEXT to continue (this records your consent to participate)

1. Demographic Questions

Please complete the demographic questions in front of you. Let me know if you have any questions and please stop when you get to the Instruction screen.

[check that everyone has finished and then proceed to face matching]

2. **Face Matching**

Please take the time to now read through the instructions presented on the screen.

"In this experiment you will be asked to compare a target image (which will appear on top of the screen) with a candidate list of eight images (which will appear in two rows of four below the target image). If you believe the target is present in the candidate list, please click the image in the candidate list you believe is a 'match' to the target. This will highlight the border of the image in green. If you believe the target is not present in the candidate list, please indicate this by clicking the 'Not Present' box next to the target image. You will then be required to rate your confidence on a 5-point scale below the candidate list. Only one selection can be made and decisions are final once you click NEXT, so please be sure to take care when selecting your decision. Please note that the target may not always be present in the candidate list. Once you have selected either a face or 'not present' and rated your confidence, you can click on NEXT to submit your selection. You will now have the opportunity to complete a couple of practice trials before the experiment begins"

Please work through the two examples on your own and stop after the second example.

[check that everyone has finished the examples]

Within those examples, you saw one instance where the target was present and one where they were absent. Within the experiment, there will also be trials of both instances – please remember that the target may not always be in the candidate list. You will also have noticed that there is an age gap between the target and candidate list images. Each trial will present an age gap of five years over five age groups. Therefore, it is important to remember that you're not looking for identical images of a person (there will always be an age gap)

Any questions? Everyone happy they know what they're doing?

You will now complete 120 tasks like the two examples you've just completed. You will be timed and your accuracy and confidence will be measured. Please work as fast and as accurately as you can.

3. **Register for results**

Remember at the end of the experiment you can register for your results if you'd like them.

And please let me know when you're done.

Thanks and enjoy.

Appendix F – Observed SD Data Matrices

Overall Observed Data

	C1	C2	C3	C4	C5	C6 (No Choice)
TID	37	136	182	220	41	0
TD	78	355	579	759	218	531
FA	50	329	550	770	230	591

0-5-Year Age Group Observed Data

	C1	C2	C3	C4	C5	C6 (No Choice)
TID	1	7	27	51	28	0
TD	3	60	169	293	148	167
FA	7	71	161	304	136	161

5-10-Year Age Group Observed Data

	C1	C2	C3	C4	C5	C6 (No Choice)
TID	17	60	78	83	8	0
TD	36	137	198	229	40	200
FA	17	132	210	247	49	185

10-15-Year Age Group Observed Data

	C1	C2	C3	C4	C5	C6 (No Choice)
TID	19	69	77	86	5	0
TD	39	158	212	237	30	164
FA	26	126	179	219	45	245

These are the matrices used to fit the SD Models to the data (definitions and explanations below).

- C1 – C5: Confidence level in each decision
- C6: All times where participants selected "not present"
  *Values in C6 for TD are misses. Values in C6 for FA are correct rejections

- TID = Target Identification (or in SD terms, "Hit"): choosing the correct match on a target-present trial
- TD = Target Detected: Times where a match was identified, whether or not they were correct.
- FA = False Alarm: identifying a match on a target-absent trial.
- Miss: participant selected "not present" when there was a match
- Correct Rejection: participant selected "not present" correctly when no match was present

Appendix G – EVSD Model Outputs

				Model Parameters					
					Confidence Criterion				
Condition	χ^2	(df)	*p*	*d'*	c_1	c_2	c_3	c_4	c_5
Overall	1735.90	(4)	<.001	1.20	2.28	1.71	1.30	0.93	0.81
0-5	725.50	(4)	<.001	1.03	2.64	1.81	1.33	0.94	0.70
5-10	520.97	(4)	<.001	1.24	2.23	1.69	1.29	0.92	0.84
10-15	645.71	(4)	<.001	1.33	2.23	1.66	1.27	0.94	0.87

Overall Predicted Data – EVSD

	C1	C2	C3	C4	C5	C6 (No choice)
TID	338.8959	350.3538	239.6962	134.4960	26.65011	46.54999
TD	349.2252	419.6400	391.6539	630.2527	210.42436	518.80387
FA	216.9242	543.6375	649.6843	570.9707	153.35158	385.43180

0-5-Year Age Group Predicted Data – EVSD

	C1	C2	C3	C4	C5	C6 (No choice)
TID	44.54292	121.6938	88.88035	50.41449	15.29380	11.27747
TD	45.07765	137.5211	137.01913	233.30349	119.63771	167.44088
FA	27.41712	181.1754	240.01647	209.76451	88.32494	93.30153

5-10-Year Age Group Predicted Data – EVSD

	C1	C2	C3	C4	C5	C6 (No choice)
TID	131.07002	116.0920	78.29155	44.07127	5.755313	17.30889
TD	135.53449	140.7027	129.42410	208.85652	45.660400	179.82182
FA	81.81257	180.1639	212.61551	189.77557	33.174468	142.45795

10-15-Year Age Group Predicted Data – EVSD

	C1	C2	C3	C4	C5	C6 (No choice)
TID	148.49697	128.2159	76.37856	40.49726	4.671883	18.27438
TD	153.41583	156.2378	128.32357	187.09985	37.221703	177.70130
FA	81.69349	191.3051	208.54175	177.15564	27.110899	154.19314

Appendix H – UVSD Model Outputs

					Model Parameters					
						Confidence Criterion				
Condition	χ^2	(df)	p	d'	s	c_1	c_2	c_3	c_4	c_5
Overall	25.91	(3)	<.001	0.66	0.99	2.81	2.09	1.63	1.16	1.01
0-5	6.39	(3)	<.001	0.31	0.87	3.13	2.29	1.75	1.19	0.91
5-10	39.13	(3)	<.001	0.83	0.93	2.75	2.03	1.57	1.14	1.04
10-15	7.18	(3)	<.001	0.96	0.85	2.67	1.99	1.57	1.15	1.07

Overall Predicted Data – UVSD

	C1	C2	C3	C4	C5	C6 (No choice)
TID	36.24874	138.6263	177.4783	195.0255	49.52358	104.7677
TD	78.80831	388.8263	597.7601	757.7335	210.61740	486.2543
FA	48.96607	296.5041	526.3396	767.2002	237.85515	643.1348

0-5-Year Age Group Predicted Data – UVSD

	C1	C2	C3	C4	C5	C6 (No choice)
TID	0.4920681	8.702409	26.12358	51.04353	24.6994	27.77883
TD	5.5794210	65.853783	167.31243	302.02347	141.6516	157.57929
FA	5.8131555	65.302001	162.04488	293.90622	142.0513	170.88244

5-10-Year Age Group Predicted Data – UVSD

	C1	C2	C3	C4	C5	C6 (No choice)
TID	16.22127	61.66982	72.84586	68.34438	11.10243	40.92506
TD	33.17701	155.59660	216.77316	231.57396	40.60117	162.27810
FA	19.55356	112.98232	187.47443	238.98497	47.56481	233.43991

10-15-Year Age Group Predicted Data – UVSD

	C1	C2	C3	C4	C5	C6 (No choice)
TID	18.40830	69.04741	79.07606	76.44662	11.20889	47.60015
TD	40.01483	166.16530	212.10197	226.06380	35.23245	160.42165
FA	24.92867	117.58018	176.94974	229.89899	41.66702	248.97539