An Analysis of Face Matching Accuracy and Related Variables: A Meta-Analysis and Two Follow-Up Studies

by

Danielle M. Rumschik

A thesis submitted to the

School of Graduate and Postdoctoral Studies in partial

fulfillment of the requirements for the degree of

Doctor of Philosophy in Forensic Psychology

Faculty of Social Sciences and Humanities

University of Ontario Institute of Technology (Ontario Tech University)

Oshawa, Ontario, Canada

July 2022

THESIS EXAMINATION INFORMATION

Submitted by: **Danielle Rumschik**

Doctor of Philosophy in **Forensic Psychology**

Thesis title: An Analysis of Face Matching Accuracy and Related Variables: A Meta-Analysis and Two Follow-Up Studies

An oral defense of this thesis took place on July 13, 2022, in front of the following examining committee:

Examining Committee:

Chair of Examining Committee	Leigh Harkins
Research Supervisor	Amy May-Leach
Research Co-supervisor	
Examining Committee Member	Brian Cutler

Examining Committee Member	Matt Shane
University Examiner	Joseph Eastwood
External Examiner	Veronica Stinson, Saint Mary's University

The above committee determined that the thesis is acceptable in form and content and that a satisfactory knowledge of the field covered by the thesis was demonstrated by the candidate during an oral examination. A signed copy of the Certificate of Approval is available from the School of Graduate and Postdoctoral Studies.

ABSTRACT

Face matching is used in many different transportation and policing situations and is an important last line of defense in security contexts such as the TSA. Despite how important this task is in ensuring security, no meta-analyses have been performed to find the average effect sizes of previously researched variables. To rectify this issue, I conducted a series of studies aimed at clarifying the field with regard to the factors important for influencing face-matching accuracy. Study one was a meta-analysis; studies two and three tested specific questions that arose from the meta-analysis. As expected, view, expertise, and feedback were found to be important; however, inconsistent with multiple reports, base rates of mismatch did not reach significance. A second study further examined base rates of mismatch to see if awareness of the base rate is necessary for the known low prevalence effect to occur. Results found that awareness mitigated the low prevalence effect instead of inducing it. A final study examined how expressions affect face matching accuracy. Participants were most accurate when evaluating photos showing neutral stimuli as opposed to expressive stimuli. The results of these studies provide information that can be used to increase face matching accuracy which is crucial as many security decisions rely on face matching and face matching is increasingly involved in legal decision making.

Keywords: face matching; accuracy; base rates of mismatch; expression; security

AUTHOR'S DECLARATION

I hereby declare that this thesis consists of original work of which I have authored. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I authorize the University of Ontario Institute of Technology (Ontario Tech University) to lend this thesis to other institutions or individuals for the purpose of scholarly research. I further authorize University of Ontario Institute of Technology (Ontario Tech University) to reproduce this thesis by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research. I understand that my thesis will be made electronically available to the public.

The research work in this thesis that was performed in compliance with the regulations of Research Ethics Board under REB File #s 16491 & 16492

Danielle Rumschik

STATEMENT OF CONTRIBUTIONS

I hereby certify that I am the sole author of this thesis and that no part of this thesis has been published or submitted for publication. I have used standard referencing practices to acknowledge ideas, research techniques, or other materials that belong to others. Furthermore, I hereby certify that I am the sole source of the creative works and/or inventive knowledge described in this thesis.

ACKNOWLEDGEMENTS

I would like to thank Dr. Brian Cutler for his support throughout the process of my degree and this dissertation. Thank you to Dr. Amy Leach who helped to refine my research and make it much better than it would have been without her help. I would also like to thank Dr. Garrett Berman for his continued support throughout my education, along with thanking him for setting me up with a good foundation and skills with which to complete this dissertation. Thank you to my husband, Jake Rumschik, for supporting me and encouraging me even if he does not quite understand what I am looking at. Lastly, thank you to the Dr. Matt Shane, Dr. Joseph Eastwood, and Dr. Veronica Stinson for taking time to evaluate my dissertation and provide feedback to ensure my dissertation is high quality.

TABLE OF CONTENTS

LIST OF TABLES

CHAPTER 1

CHAPTER 2

CHAPTER 3

CHAPTER 4

LIST OF FIGURES

CHAPTER 1

CHAPTER 2

CHAPTER 3

CHAPTER 4

LIST OF ABBREVIATIONS AND SYMBOLS

IV – independent variables

SC – study characteristics

MD – mean difference

θ - mean difference

m – mean

SE – standard error

sd – standard deviation

n - sample size

w – weight

CFMT+ - Cambridge Face Matching Test+

GFMT – Glasgow Face Matching Test

LPE – low prevalence effect

Chapter 1. Introduction

1.1 Background & Significance

On January 6th, 2021, hundreds of rioters stormed the U.S. Capitol to prevent the certification of the 2020 presidential election results. Rioters disturbed Congressional proceedings, attacked Capitol police officers, and stole confidential documents. In the days following the attack, the FBI released photos and videos of the rioters to elicit tips and identifications from the public. Internet detectives focused in on a man with salt-and-pepper hair and a "CFD"-labeled beanie who threw a fire extinguisher at police officers (Kornfield, 2021). After the public identified the rioter and filed a tip with the FBI, his identity was released leading to people calling him incessantly, harassing his family, and lurking outside his home. Unfortunately, David Quintavalle, the man being harassed, was nowhere near the U.S. Capitol, instead he was 700 miles away grocery shopping in Chicago. Despite being a law-abiding citizen, Quintavalle was left defending himself and rebuilding his reputation while still being bombarded with hateful calls and messages. Amateur investigators compared the photos provided by the FBI to photos of Quintavalle available on social media and decided that he was a close enough match to the photo of the rioter to publicly accuse him of insurrection.

This process is known as face matching. Face matching is the task of deciding whether two, typically simultaneously, presented stimuli (photos, videos, or people) depict the same person. While viewing, deciders must scan one stimulus and determine the face shape, shape of facial features, and the relative distance between those features and compare them to the other stimulus to search for any discrepancies, like a "spot the difference" task (Sate et al., 2011). Millions of face matching decisions are made every day. For example, the United States Transportation Security Agency (TSA) makes decisions on whether to allow people onto flights

based on government-issued identification documents, such as passports, driver's licenses, and Trusted Traveler's documents. In October of 2019, seven million passengers traveled from LAX (Los Angeles World Airports, 2019). According to the U.S. Federal Aviation Administration, 2.6 million airline passengers fly each day. On top of this, people are routinely required to prove their identities to purchase alcohol, cigarettes, lottery tickets, and other controlled substances, and to show ID cards to enter schools, workplaces, and other protected environments.

Along with the traditional uses for face matching, the increasing use of algorithmic and artificial intelligence identification processes being used by state and government organizations around the world introduces a new task for face matchers. Algorithmic and artificial intelligence-guided face matching is used in at least 30 states in the USA and is commonly used in the UK to identify wanted persons (Garvie et al., 2016). After an algorithm or AI identifies a photo as a potential match to the wanted person, a human (usually a police officer or federal agent) must then decide if the returned photo actually matches the wanted person. The human decision-maker introduces errors into the process that can result in falsely accusing, arresting, and imprisoning people (e.g., Hill, 2020a; Hill, 2020b).

Despite the ubiquity of face matching tasks, some research has found poor accuracy rates (Bindemann, & Sandford, 2011; Henderson et al., 2001; Kemp, et al., 1997). Even under the best conditions, face matching accuracy rates range between 70% to 90% (e.g., Fysh & Bindemann, 2018; Norell et al., 2014; White et al., 2014; White et al., 2015). These general accuracy rates are influenced by the photos and stimuli used. For example, a person's appearance can change significantly over time (Megreya et al., 2013), and aspects of the photo such as lighting (Favelle, et al., 2017), photo quality (Bindemann et al., 2013), and angle (Bruce et al., 1999) can drastically change the representation of a person in a photo; two photos of the same person can

look very different and two photos of different people can look very similar (Buolamwini et al., 2020). Terrorism has been inextricably linked with ID theft or fraud as well (Sullivan, 2004), which, in some cases, should be detected through face matching. Therefore, the importance of ensuring correct decisions made at security points cannot be understated.

Unfamiliar face matching, as opposed to familiar face matching, is the type of matching most often employed in face matching decision-making (White et al., 2021). Unfamiliar face matching entails making matching decisions about stimuli that show a person that the decision-maker does not have any familiarity with. Familiarity can come from knowing a person in real life or from seeing a person frequently through media (Clutterbuck & Johnston, 2010). Accuracy rates are higher for familiar face matching, around 90% (Megreya & Burton, 2007). Some research has attempted to "familiarize" face matching decision-makers to improve accuracy rates (e.g., Clutterbuck & Johnston, 2005; Osborne & Stevenage, 2008) with small, but promising, results. However, the large numbers of people that pass through security checkpoints and other places where face matching is used makes familiarization as a means of improving accuracy impossible. Thus, it is necessary to examine unfamiliar face matching decisions and improve accuracy with unfamiliar people.

Unfamiliar face matching has been the subject of research for 30 years and consists of hundreds of experiments. However, up until this point, only qualitative reviews of the research have been conducted (e.g., Bindemann, 2021; Fysh & Bindemann, 2017). Therefore, I conducted a meta-analysis to provide empirically based estimates of importance for variables that affect face matching accuracy and lay a framework for expert testimony and decision making related to face matching. Guided by the results of this meta-analysis, I designed and conducted two additional experiments that aimed to test the generalizability of face matching

findings. Specifically, in Study 2, I manipulated base rates to evaluate whether awareness of base rates of mismatch is necessary for the expected reduction in accuracy to occur, and in Study 3 I examined different types of expressions in order to better align results with the type of real-world contexts that may be encountered when one needs to match unfamiliar faces.

1.1.1 Base Rates of Mismatch

In many laboratory-based face matching studies, 50% of trials are designed as mismatch trials (see Bindemann, et al., 2010) in order to ensure an equal number of match/mismatch trials. However, Bindemann, et al. (2010) suggest that this 50%/50% split between matching and mismatching photo pairs may not be representative of real life, which may in turn result in inflated accuracy rates. The effect of base rates of mismatch in research has been coined the low prevalence effect (LPE; e.g., Papesh & Goldinger, 2014). The LPE suggests that under realistic viewing conditions, low prevalence of mismatches is a large, persistent source of errors; specifically, a face matcher is more likely to incorrectly classify face pairs when a small number of true mismatches are present in the overall sample (e.g., Papesh et al., 2018; Weatherford et al., 2020). However, research has been mixed with regard to whether the LPE actually exists in face matching research. A recent study (Davis et al., 2021) found that participants did not exhibit the expected LPE unless they were aware of the low base rate of mismatch condition. Based on this body of research, I designed and conducted an experiment (Study 2) further assessing if the LPE is affected by participants' awareness of the base rate condition.

1.1.2 Expression

Photo identifications and travel documents in the United States usually require the photographed subject to have a neutral expression. However, neutral expressions may not be the best representation of a person in a day-to-day situation since people likely do not have a neutral

face when presenting their identity documents. Previous face matching research has examined the influence of smiles on face matching accuracy (e.g., Bruce et al., 1999; Mileva & Burton, 2018) and found that matching smiling face pairs results in higher accuracy than neutral face pairs. But, identification cards and travel documents use neutral, non-expressive photos while people can present in a face matching situation with any expression. Therefore, in order to test more ecologically valid conditions, I conducted a study using different expression photo pairs - one neutral photo and one expressive photo - and expanded the types of expressions examined.

1.2 Thesis Objectives

The goals of this dissertation were as follows:

1) To quantify the effect sizes for variables related to face matching accuracy
2) To examine and investigate unexpected results found in the process of a meta-analysis
3) To explore the effect of base rates of mismatch and how it interacted with awareness to change accuracy, sensitivity, decision making criteria, and response bias
4) To explore the effects of incongruous expressions on face matching accuracy

Chapter 2. Meta-Analysis

2.1 Background & Significance

Unfamiliar face matching has been the subject of psychological research for over 30 years, with the first known study with the goal of assessing face matching accuracy taking place in 1986 (Young et al., 1986). Typically, face matching research employs a basic experimental structure: participants are randomly assigned to a condition and then asked to make face matching decisions. Face matching stimuli are presented as a one-to-one pair of stimuli or a one-to-array (more akin to a lineup), participants view the stimuli and then indicate whether the

stimuli show a match or a mismatch or if the probe photo is in the array. For this meta-analysis, I focused on the one-to-one pair of stimuli, as it is more similar to the tasks used at security checkpoints.

Laboratory research is particularly appropriate for improving our understanding of the psychological mechanisms underlying face matching due to the ability to control the environment of the study and the specifics of the variables being manipulated. Research conducted in the laboratory typically involves presenting photo pairs to amateur participants (e.g., students) through a computer. This basic structure can then be manipulated to test many different variables such as age of the photo, the photo quality, or the base rate of mismatches (e.g., Bindemann & Sandford., 2011; Bindemann et al., 2013; Weatherford et al., 2020). For example, in a study by Bruce et al. (1999), amateur participants were shown unfamiliar faces taken from high-quality video against arrays of photos. The photos were taken on the same day and shown in a frontal view. Participants were asked to decide whether the photo arrays contained the target photo. Other studies have aimed to more closely approximate real-word conditions by using ecologically valid stimuli, such as photos taken from two separate cameras at different times (e.g., Bindemann & Sandford, 2011), or by carrying out the research in more ecological settings. For example, Kemp, Towell, and Pike (1997) conducted one of the earliest field studies within which they manipulated credit card identification photos to determine if customer use of credit cards with affixed photos would reduce fraud. Their study was conducted in a supermarket, with authentic credit cards, employed cashiers, and real transactions with confederates posing as customers (Kemp et al., 1997).

Recent qualitative reviews (e.g., Bindemann, 2021 Fysh & Bindemann, 2017; White & Kemp, 2019) provide summaries and syntheses of various specific aspects of the face matching

research. These reviews offer in-depth evaluations of previous research and estimates of the importance of moderating variables, but they cannot produce exact measurements, a task better suited for meta-analyses. A meta-analysis of the face matching research is needed; research has been accumulating for the past 30 years, but so far only qualitative reviews have been done. A meta-analysis can provide insight into effect sizes and statistical importance, help to reconcile or explain inconsistent findings in the general research, and help to chart future research questions by looking at variables not examined in previous studies. Finally, a meta-analysis can provide important context and general information that can be used for expert testimony.

I compiled face matching research and conducted a meta-analysis to estimate face matching accuracy and determine the overall effects of independent variables on face matching accuracy. I collected 39 different studies consisting of 78 different tests (total sample size of 8,205 participants) of face matching accuracy and coded them for 17 different variables including study characteristics, independent variables, and dependent variables. Below I list the variables that I examined in the meta-analysis and summarized the relevant face-matching research about those variables.

It is my intention that this meta-analysis provide useful applied data, so I based my variable decisions on the most often investigated variables in the face matching research. I also included variables known to affect facial recognition technology (e.g., Buolamwini et al., 2020; Garvie et al., 2016; Lui et al., 2009) and facial recognition (e.g., Herlitz & Loven, 2013), and variables known to affect eyewitness identifications (e.g., Wells et al., 2020). The psychological process of making an eyewitness identification and the process of an unfamiliar face matching identification are very similar; the only thing missing in a face matching decision is the absence of a long-term memory trace. Unfamiliar face matching decisions involve some working

memory and requires the decision maker to keep a picture of one of the comparison stimuli in their minds when looking at the other comparison stimulus. A face matching decision is especially similar to a show-up procedure where the witness is asked to identify the suspect from single person "lineup" or photo array. Therefore, it is assumed that variables that affect eyewitness identification will also be important in face matching decision making. In the following sections, I introduce the variables considered for inclusion in the meta-analysis and briefly discuss themes arising from research in those areas.

2.1.1 Photo Color

Face matching research typically uses face pairs or other stimuli that are either both in color or both in greyscale, but in real-world conditions the mugshot or identification photo may be in greyscale while the person standing in front of the matcher is in full color. Research examining the influence of photo color on face matching accuracy is nascent, but raises an important, real-world question in face matching accuracy. Bobak, Mileva, and Hancock (2019) tested the difference between congruent (e.g., both color or both greyscale) and mixed (one color photo and one greyscale photo) stimuli and its effect on face matching accuracy. Overall, participants showed a more conservative bias in the greyscale condition than the other conditions, meaning that they were more likely to reject a true pair as a mismatch. The researchers suggest that color incongruency in face matching might disguise subtle differences between faces and impair piecemeal processing, resulting in lower face matching accuracy.

2.1.2 Photo Quality

Poor image quality has been repeatedly linked to reduced performance in facial identifications and facial matches from surveillance and CCTV video (e.g., Bindemann et al., 2013; Burton et al., 1999; Henderson et al., 2001; Lee et al., 2009; Lie et al., 2003). With high

resolution photos, accuracy can be as high as 85%-90%, but pixilation of any kind reduces accuracy significantly down to 60%-66% (Bindemann et al., 2013). The dramatic reduction in identification accuracy illustrates that image pixilation exerts a strong effect on unfamiliar face matching. Face matching is already an error-prone task; in simple identification task with high quality images, accuracy can be poor, and participants can be inappropriately confident in their ability to match images of the same person (Henderson, Bruce, & Burton, 2001). Introducing a pixelated or low-quality image may further reduces the accuracy and the degree of pixilation might determine the likelihood that a correct identification can be made (Bindemann et al., 2013).

2.1.3 Base Rates of Mismatch

Many face matching studies use 50% mismatch trials, meaning that half of the trials are matches and half of the trials are mismatches (see Bindemann et al.,2010). However, it is likely that the incidence of mismatches (e.g., imposters, fake IDs, etc.) in applied contexts is significantly lower. In passport identification, for example, the mismatch rate could be as a low as 0.0075% (Hetter & Cripps, 2014). In applied contexts, the base rate of mismatches likely depends on the context of the decision. For examples, bars in college towns probably have a high rate of fake IDs, but fake passports or IDs at security are likely rare. Additionally, not all face matching decisions have the same consequences; the consequences of missing a mismatch at a bar are much smaller than the consequences of missing a mismatch at an airport.

Bindemann et al. (2010) tested whether low mismatch prevalence resulted in increased or decreased accuracy across four experiments. In all the experiments, participants were informed of their prevalence condition (low vs. high) and the mismatches in the low prevalence condition were always presented at the end of the trial block. Across four experiments, more mismatches

were detected under the low prevalence condition than in the high prevalence condition. Based on this evidence researchers suggested that low mismatch frequency did not impair mismatch detection in unfamiliar face matching. The lack of a pure low prevalence effect is also seen in Weatherford et al. (2020). However, other researchers (e.g., Papesh & Goldfinger, 2014; Susa et al., 2019) suggest that the lack of an effect is due to participants being aware of the condition they were in and the mismatches always being at the end of trial blocks.

Papesh and Goldinger (2014) tested low (10%) vs. high (50%) mismatch prevalence conditions and found inflated miss rates under low prevalence conditions. In the low condition mismatch errors rates were around 45%, roughly doubling the rate for the high condition (i.e., 20% errors). This low prevalence effect on accuracy persisted when participants were allowed to correct their initial judgements, when they were asked to give certainty judgements after their decision, and when they were permitted second looks at the face pairs. Papesh and Goldinger's findings suggest that under realistic viewing conditions, low prevalence of mismatches is a large, persistence source of errors in face matching. Similar results are seen in Papesh, Heisick, and Warner (2018) and Susa et al. (2019). Overall, the cumulative results suggest that low mismatch prevalence reduces face matching accuracy, but some studies (e.g., Bindemann et al., 2010; Weatherford et al., 2020) have not found that effect.

2.1.4 Expertise

Expertise in face matching research is most often related to employment or the length of time spent regularly making face matching decisions, regardless of whether any formal training is received in face matching decision-making and techniques. Face matching experts have been conceptualized as passport officers (White et al., 2014), police officers (Wirth & Carbon, 2017), notaries and bank workers (Papesh, 2018). It seems intuitive that observers with more experience

would be more accurate when making face matching decisions. Research, however, does not always bear this out. Papesh (2018) examined photo identification verification between student groups and professional groups to determine if experience moderates face matching accuracy rates. Notaries, bank workers, and undergraduate students were exposed to 30 photo pairs. Both professional groups performed similarly to the student group and the only reliable individual difference predictor of accuracy was age, not experience; older participants performed more accurately than younger participants White et al. (2014) tested passport officers' ability to make same or different identity judgements in person-photo pairs and photo-photo pairs. Overall, officers averaged 10% errors on person to photo tests. Approximately 6% of valid photos were wrongly rejected and 14% of fraudulent photos were wrongly accepted as depicting the person in front of the officer. In the photo-to-photo experiment match accuracy was 70.9% and mismatch accuracy was 89.4%. The results demonstrated by the passport officers were did not significantly differ from students' results, suggesting that experience alone does not improve accuracy on face matching tasks.

In contrast, Wirth and Carbon (2017) found that German federal police officers significantly outperformed novices, but still had a high ratio of missed frauds. There were large individual differences in matching accuracy, ranging from 59.5% to 98.4% accurate. Overall, high true acceptance rates were found (M = 94.6%), but false acceptance rates were also high (M = 23.7%). Matches were more reliably detected than mismatches. There was a significant difference between the novice sample and the professional sample with the professional sample performing more accurately than the novice sample. Within the professional sample, the less experienced police officers performed better than the more experienced police officers, in direct

contrast with what would be expected. Therefore, there is conflict in the research about whether expertise alone increases face matching accuracy.

2.1.5 View Variation

View and angle are some of the variables with the strongest effect on the ability to match faces. This is especially important to consider when using photos or video taken from CCTV and other surveillance videos because the systems are usually placed high above a space, meaning that the video captured will not show the person in a full face-on view, the person will usually be shown from at least a slight angle. When using upright, front-facing photos matching accuracy ranges from 70% to 98% (Bruce et al., 1999; Favelle, Hill, & Claes, 2017; Kramer & Reynolds, 2018; Megreya & Bindemann, 2015; Towler, White, & Kemp, 2017). However, when the photo shows the person from any angle other than straight on, accuracy is reduced to anywhere from 48% to 85% (Bruce et al., 1999; Favelle et al., 2017; Kramer & Reynolds, 2018; Megreya & Bindemann, 2015; Towler et al., 2017).

Some of the research on view and angle variation looks at inverted faces, which is a common stimulus used in visual processing literature (e.g., Rossion, 2008; Tanaka & Farah, 1993; Young, Hellawell, & Hay, 1987). Megreya and Bindemann (2015) examined one-to-one face matching with upright and inverted faces. They found significantly reduced accuracy for inverted face pairs with 76% correct IDs for upright pairs and 67% correct for inverted pairs. When comparing upright and inverted face views Towler and colleagues (2017) found similar results: participants were more accurate with upright stimuli (87%) than inverted stimuli (72%).

Other view-focused research focused on the differences between front-facing face stimuli and profile face stimuli. Favelle and colleagues (2017) looked at the interaction between lighting and viewpoint by rotating both the camera and the lighting around the y-axis and the x-axis of a

computer-generated face. Frontal views resulted in the highest accuracy, that changes of view across either axis resulted in similarly reduced accuracy, and that the pattern of results was similar regardless of whether view variation occurred as a result of head or camera movement. When using a one-to-many face matching procedure, Bruce et al. (1999) found greater accuracy for full-front-facing stimuli (70%) than 30-degree profile view stimuli (61%). Front views resulted in more hits and correct rejections than the rotated condition. On the other hand, Kramer and Reynolds (2018) found no difference in performance across frontal and profile views. They also tested presenting both frontal and profile views at the same time and found no accuracy benefit when both views were presented together.

2.1.6 Time Pressure

Airport and border service security agents are expected to accurately and quickly perform face matching tasks throughout their shifts. Australian and UK passport officers are expected to process about 90% of passengers in a queue within 30 minutes of arriving on shift (Fysh & Bindemann, 2017). This expectation can be problematic because it reduces the amount of time that the agent can view the photo, and accuracy rates rise and fall with photo presentation duration (Chiller-Glaus, Schwaninger, & Hofer, 2007). Reports from airport security personnel reveal that they usually only have a few seconds to assess travel documents including the verification of the document photo plus other factors of authenticity, such as the official seals and personal information. Wirth and Carbon (2017) further examined time pressure by comparing passport-matching accuracy in novices under no time pressure, under a local time limit (i.e. fixed presentation times), or under a global time limit (25 minutes for all trials). They found a match bias, with participants performing worse on mismatch trials throughout the

experiment. This match bias was particularly detrimental in the global time limit condition, resulting in a 12.2% fall in accuracy rates.

Fysh and Bindemann (2017) also manipulated time pressure using two different onscreen displays. Participants were placed under increasing time pressure (from 10s per trial to 2s per trial) or decreasing time pressure (from 2s per trial to 10s per trial). The onscreen displays were updated in real time to reflect the participant's average response time and the number of trials remaining. Displays also indicated whether the participant was on track to complete the experiment within the required timeframe. Time pressure appeared to specifically impact performance on mismatch trials, with accuracy deteriorating as the average time target per trials was reduced. Accuracy on match trials improved over the course of the experiment from 66% to 81%, while mismatch accuracy decreased from 74% to 53%. In a second experiment using the same stimuli, Fysh and Bindemann (2017) had participants complete four blocks of face matching trials where time pressure varied from eight to two seconds. Again, accuracy was higher for match trials and under strict time pressure observers exhibited a bias to classify more faces as matches. Over both experiments, the match bias that emerged accounted for up to 21% of errors on mismatch trials. This bias was also seen in Bindemann, Fysh, Cross, and Watts' (2016) experiment concerning time pressure and face matching accuracy. These studies represent the time pressure in face matching research in general and suggest that an increase in time pressure may lead to a reduction in face matching accuracy.

2.1.7 Exposure Time

Exposure time refers to the amount of time that a stimulus pair is present for a participant to view. Once the exposure time has expired, participants are able to take as long as they want to decide whether the pair was a match or mismatch, but the photo pair will not be available for

them to view. Exposure time and time pressure are similar, but exposure time is concerned with the amount of time available to view the stimuli pair while time pressure is the amount of time available to view the photo pair and make a decision before being forced to move on to the next photo pair. Previous research has shown that maximal ID accuracy can be reached with exposure duration of only 90ms to a face (Veres-Injac & Schwaninger, 2009). Özbek and Bindemann (2011) tested display time in face matching. Accuracy increased with exposure duration and peaked in the 2000ms condition. Match accuracy was at chance levels with 200ms display time but improved with increased display time. Mismatch accuracy was higher at 200ms, but accuracy decreased with increased display time. This suggests a possible mismatch response bias at short exposure times. Chiller-Glaus and colleagues (2007) manipulated display duration and the presence of an additional task to determine whether the rushed conditions that airport security works under affects face matching accuracy. Photo pairs were displayed for either one second, four seconds, or self-paced. ID verification accuracy was significantly better with increased display duration. Therefore, research overall suggests that the strict timelines enforced by some security organizations can have a detrimental effect on accuracy, increasing the chance for unnecessarily inconveniencing travelers or unknowingly accepting fraudulent documents.

2.1.8 Training

Training in face matching research refers to whether participants receive any type of formal training or guidelines concerning how to make face matching decisions. Training guidelines set by forensic institutes in Europe help forensic facial examiners to make decisions based on morphological-anthropological facial features (Ali et al., 2015). When comparing facial photos, forensic facial examiners are trained to focus on the shape of facial features, such as the mouth, eyes, nose, ears, and eyebrows; to examine the relative distance among different relevant

facial features; to focus on the contour of the cheek- and chin-lines; and finally, to focus on any lines, moles, wrinkles, or scars on the face. Many of the trainings used by face matching professionals are based on what makes sense intuitively without much research to back the efficacy of the program. Thus, Towler et al. (2017) set out to evaluate the feature comparison strategy used by forensic facial examiners. In the first experiment, untrained students were asked to rate the similarity of facial features in image pairs prior to making same/different judgments. Rating the similarity of facial features improved matching accuracy on match trials, but not mismatch trials. When comparing forensic facial examiners to novices, examiners performed more accurately than students and their accuracy was not affected by stimulus orientation. This suggests that the similarity training received by forensic facial examiners may increase accuracy, but that it takes some time to learn how to use it and is not effective immediately. Towler et al. (2017) provides a single example of the general overview of training in the face matching literature: training seems to improve accuracy, but the transfer of training to on-the-job performance is poor and requires repeated use before it can be reliably used to improve accuracy.

2.1.9 Cross-race Matching

The Cross-Race Effect or Other-Race Effect (ORE) is a well-known phenomenon in memory tasks (e.g., Brigham, Bennet, Meissner, & Mitchell, 2007; Meissner & Brigham, 2001) but also persists in face matching, which requires no memory. Megreya, White, and Burton (2011) presented British and Egyptian students with simultaneous one-to-many face matching decisions. Trials consisted of both British and Egyptian face pairs. Both UK and Egyptian participants performed better overall with own-race arrays than with other race arrays. Similar results were seen in both one-to-one face matching and one-to-many face matching: participants were more

accurate when matching same-race faces than other-race faces (Kokje, Bindemann, & Megreya, 2018; Susa et al., 2019).

2.1.10 Quantity of Photos

Photo identifications present a single photo of an unfamiliar person that is then used to compare to the person presenting the identification. However, identifications are often valid for many years, and a person's appearance can change drastically in that time. Also, a single photo may not provide sufficient information about a person's face so that an observer can assess all the possible ways a person may present (White et al., 2014). To remedy this, researchers have proposed presenting multiple photos of the same person (arrays) or a single photo depicting a face-average, which is an array of photos that has been condensed into a single photo. White and colleagues (2014) consistently found an accuracy advantage for face-averages and photo arrays over single photos. Ritchie et al. (2020) found opposite results. Taking a different approach, Gentry and Bindemann (2019) provided examples of matches and mismatches alongside to-be-decided face pairs and found that examples improved overall face matching accuracy. Most of the improvement in accuracy, however, was due to improved accuracy for observers who were least accurate in the baseline block. Therefore, examples seemed to help individuals with lower natural face matching abilities and did not improve accuracy for every observer (e.g., Ritchie et al., 2020).

2.1.11 Feedback

Face matching decision makers rarely are told anything about whether the decisions they make are correct or incorrect, especially in security situations. In the lab providing feedback about the accuracy of decisions can help inform decision makers about how to better identify mismatches and how to improve accuracy (e.g., Alenezi & Bindemann, 2013); thus, providing

feedback for face matching decisions has potential utility as a training tool. Participants were tasked with making face matching decisions and either given immediate, trial-by-trial feedback or no feedback at all (Alenezi & Bindemann, 2013). While accuracy on mismatch trials gradually declined throughout the matching task for the no feedback group, the group that received feedback maintained their initial level of accuracy (M = 90%) throughout the experiment. The feedback effect also transferred to new, not previously seen stimuli. However, the maintenance effect seen with trial-by-trial feedback does not occur when overall outcome feedback is used. Overall, face matching research suggests that specific trainings that provide trial-by-trial feedback could be useful in maintaining accuracy levels and reducing the match bias that is often seen in long and tedious face matching tasks. However, the influence of feedback has not been tested outside of the lab.

2.1.12 Hypotheses

Based on previous research, I expected significant effects of trial type, base rates of mismatch, expertise, feedback, cross-race, photo amount, exposure time, time pressure, and view and pose. I hypothesized that photo color, photo illumination, base rates of mismatch, and view of the stimuli would significantly predict the effect estimate in the meta regression. The other variables that were included in my analyses were included for exploratory purposes as this was the first meta-analysis in a relatively new area of research. Thus, I made no *a priori* predictions concerning those variables. Since I conducted a meta-analysis using the mean difference effect size based on the same value scale (0.0 – 1.0), it was possible to compare effect sizes across variables to determine the relative importance of variables in the field of face matching; however, comparisons across variables was not a goal for this analysis, therefore I made no *a priori* predictions concerning relative importance.

2.2 Method

I searched multiple research databases for all face matching articles between the years of 1960 and 2020. I chose this range of years because modern face matching research was not conducted before 1960, and an ending date of 2020 gave me the most recent research when I started this project in 2020. I searched Google Scholar, PsycInfo, PsycArticles, Scholars Portal Journals, Scopus, JSTOR, ProQuest Science Journals, PubMed, and Web of Science. The search terms used were "face matching," "person matching," "facial comparison," "forensic facial comparison," and "imposter identification." Theses, dissertations, and unpublished manuscripts were excluded in favor of peer-reviewed, English publications so that the results of the meta-analysis will meet Daubert criteria (*Daubert v. Merrell Dow Pharmaceuticals Inc.*, 1993) and can be used to inform court decisions. As research on human face matching ability has grown, so has research concerning face and person matching algorithms. Articles included in the analysis concerned human face matching only (no algorithm or algorithm and human interaction) with typically developed adults with typical face matching skills (no super-recognizers). Articles specifically researching the abilities of super-recognizers (individuals with naturally exceptional face matching skills (Bobak et al., 2015)) were excluded from the meta-analysis, but participants with different levels of expertise were still included. Therefore, there is a chance that some individual participants included in the studies would qualify for super-recognizer status as natural face matching abilities change based on the individual. However, expertise in this meta-analysis was defined based on job/employment status (e.g., passport officers, notary workers, etc.) as opposed to face matching abilities. Articles containing brain imaging were excluded. Articles included in the final list consisted of at least one unique simultaneous, one-to-one, face matching experiment.

The database search resulted in 477 articles. After a cursory title search, 77 articles were removed for not fitting the criteria, leaving 400 articles. Next, the list of articles was uploaded into Rayyan (https://rayyan.qcri.org/welcome), an open-source abstract filtering software. Rayyan filtered for duplicate ($N = 105$) and unpublished ($N = 10$) articles. The articles were then filtered based on the established exclusion criteria resulting in a final article list of 52 articles and 123 experiments. Ultimately, thirteen experiments were removed from the analyses due to missing dependent variables and effect sizes resulting in the final sample of 39 studies and 78 experiments. See Figure 2.1 for a PRISMA chart of the included studies.

Figure 1.
PRISMA chart detailing the reduction of included variable in the meta-analysis

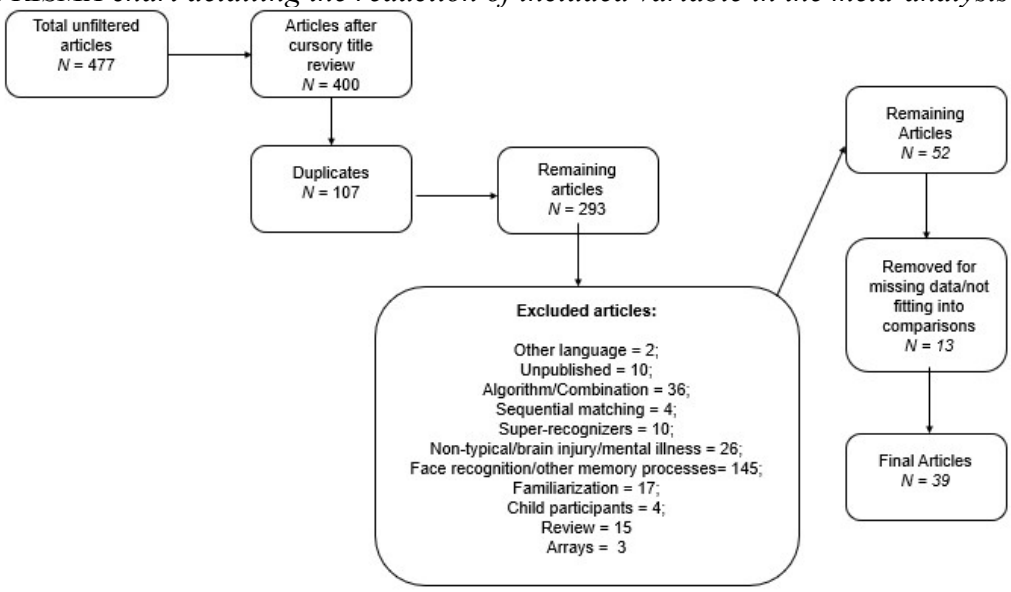


Figure 2.1. Prisma chart detailing variables included in analysis

2.2.1 Types of Variables

Three different kinds of variables were coded for each study: study characteristics, independent variables, and dependent variables. Study characteristics are variables in a research

methodology that are held constant in a single study but varied across studies (Shapiro & Penrod, 1986). These characteristics can be integrated to analyze the influence of specific variables and the relative importance of study factors on performance across studies. Independent variables are factors manipulated by the researcher in each study. Study characteristics and independent variables are not mutually exclusive, and variables can appear on both lists. In the meta-analysis, comparisons were analyzed with the independent variables, and study characteristics were used as predictor variables in the meta-regression. Three different types of dependent variables were coded, all forms of accuracy: overall accuracy, match accuracy, and mismatch accuracy. Accuracy for each group was taken directly as reported in the articles. If overall accuracy was not reported, it was estimated using data from figures and tables, or calculated using reports of match and mismatch accuracy. See Table 2.1 for a complete list of coded variables.

Table 2.1.

List of meta-analysis coding variables

Study Characteristics Variables (SC)	Independent Variables (IV)	Dependent Variables
Age of Observer	Cross-race	Overall Accuracy
Percent Women	Photo Amount	Match Accuracy
Photo Illumination	Expression	Mismatch Accuracy
Photo Color	Feedback	
Photo Quality	Base Rates of Mismatch	
Base Rates of Mismatch	Expertise	
Expertise	View Variation	
View Variation	Time Pressure	
Time Pressure	Exposure Time	
Exposure Time	Training	
Training		

Table 2.1. List of meta-analysis coding variables

2.2.2 Variable Definitions

The following list of definitions clarifies how the coding variables were conceptualized and whether they were study characteristics, independent variables, or both.

1. **Age of observer (SC)** – the average age of participants in a sample, coded as a continuous variable. If no average age was reported and the average age could not be calculated, the variable was left blank. Two values were coded for each study in a comparison: the average age for the two extremes of the IV. For example, if a study was included in the cross-race comparison, the average age of the observer was coded for participants in the same race condition and participants in the other race condition. Values coded for age of observer ranged from 18.76 to 50.40.
2. **Percent women (SC)** – percentage of female participants in a sample, coded as a continuous variable. Two values were coded for each study in a comparison, one for the percentage of women in the two most extreme categories of the IV. For example, if a study was included in the base rates of mismatch comparison, then one value was coded for the low base rate condition and one for the high base rate condition. Values coded for percent women ranged from 5% to 97%.
3. **Photo illumination (SC)** – the lighting of the photos (e.g., well-lit, dimly lit), self-reported by the authors of the study. Values for photo illumination were coded dichotomously such that 1 = well-lit and 2 = dimly-lit. Two values were coded for each test in a comparison – one for each level of the IV.
4. **Photo Color (SC)** – whether the photos were presented in greyscale and/or color. Values were coded based on a nominal coding scale with 1 = greyscale, 2 = color, 3 = both/varied. Values were coded based on what was reported by authors of the study. Two values were coded for each test in a comparison – one for each level of the IV.

5. **Photo quality (SC)** – the quality of the comparison photos, coded nominally (1 = high quality, 2 = low quality, 3 = both/varied). Values were coded based on what the authors reported about their stimuli. Two values were coded for each test in a comparison.
6. **Base Rates of Mismatch (SC/IV)** – the percentage of mismatches presented to participants throughout the experiment. This variable was considered a continuous variable and coded as a percentage of how many of the trials contained a mismatch. This value was reported by authors in descriptions of their study design and procedure. Values ranged from 2% to 90%
7. **Expertise (SC/IV)** – the self-reported amount of experience that the participants have, based on how it was reported in each article, coded dichotomously (amateur vs. expert)
8. **View Variation (SC/IV)** – how the subject in the photos were presented for viewing, coded using a nominal coding scale (1 = upright. front-view, 2 = other, 3 = both)
9. **Time Pressure (SC/IV)**– the amount of time the participants had to make a decision about whether the stimuli match or not. This variable was coded continuously based on the time allotted for the decision that the authors reported in seconds. Values ranged from 0 (no time pressure) to 1,500 seconds (25 minutes).
10. **Exposure Time (SC/IV)** – the amount of time the photo is visible to the participants when they are making a decision about whether the stimuli match or not, from 200ms to unlimited (no restriction).
11. **Training (SC/IV)** – whether participants have participated in any type of training course whether provided by the researchers or an outside source, coded dichotomously (training vs. no training)

12. **Cross-Race (IV)** – whether participants are the same race as the subjects in the stimuli or different races, coded dichotomously (same race vs. other race) based on the conditions reported by the study authors.
13. **Photo Amount (IV)** - the amount of comparison photos presented, coded dichotomously (single photo vs. array of photos)
14. **Feedback (IV)** – whether participants received any type of feedback about the accuracy of their decisions, coded dichotomously (feedback vs. no feedback). If any feedback was given about accuracy, regardless of when the feedback was given (e.g., trial-by-trial vs. global) it was coded as feedback present.
15. **Overall Accuracy (DV)** – the total percentage of correct decisions made throughout the experiment.
16. **Match Accuracy (DV)** – the percentage of correct decisions made when the stimuli showed the same person; "hits"
17. **Mismatch Accuracy (DV)** – the percentage of correct decisions made when the stimuli showed two different people.

2.2.3 Coding

I coded each study for study characteristics, independent variables, and dependent variables. Only one coder was used because the information being coded was objective and not a matter of opinion. If a study had more than two levels of an independent variable, then I coded the two most extreme levels. For continuous independent variables the effect size analyses were based on an artificial dichotomous low-high distinction and records only the most extreme values. Each study was coded for methodological data, sample size, mean data and standard

deviations. The mean difference effect was used. Effect sizes were calculated using RevMan5 (Higgins et al., 2021).

Missing data were treated conservatively. If outcome data were missing, then they were estimated from tables and graphs using WebPlotDigitizer (Rohatgi, 2020). WebPlotDigitzier uses the height and length of pixels along with the extreme values of the x- and y-axes of a graph to determine the value of a selected spot on the graph. Graphs from 49 different experiments were introduced to WebPlotDigitizer to determine values. Unless the value of an independent variable or study characteristics variable could be taken directly from the article, computed, estimated from a table or graph, or was received from correspondence with the author, it was left blank. Thirty-three tests were missing data necessary for the meta-analysis. The open-source framework (OSF.io) and any supplemental materials were consulted before reaching out to authors for data. Thirteen authors were emailed concerning 43 different tests. Responses were received from most of the authors that were contacted who provided either the necessary data or data that could be used to calculate the necessary data. Ultimately nine experiments were removed from the analyses due to missing dependent variables and effect sizes resulting in the final sample of 39 studies. If only the standard deviation was missing for a study, the values were imputed based on the recommendations established by Cochrane Reviews (Higgins et al., 2021). See Table 2.2 for a complete list of included articles and the comparisons the articles were used in.

Table 2.2

Summary of included studies, sample size, and comparisons

Study	# of Tests Included	Sample Size Included	Analyses
Alenezi & Bindemann, 2013	5		
Exp 1		50	Trial Type, Feedback, Avg. Accuracy

	Exp 2		50	Trial Type, Feedback, Avg. Accuracy
	Exp 3		50	Trial Type, Feedback, Avg. Accuracy
	Exp 4		50	Trial Type, Feedback
	Exp 5		50	Trial Type, Feedback
Bindemann, Attard, Leach, & Johnston, 2013		3		
	Exp 1		20	Trial Type, Avg. Accuracy
	Exp 2		20	Trial Type, Avg. Accuracy
	Exp 4		20	Trial Type, View & Pose, Avg. Accuracy
Bindemann, Avetisyan, & Blackwell, 2010		4		
	Exp 1		54	Trial Type, Base Rate, Avg. Accuracy
	Exp 2		18	Trial Type, Base Rate, Avg. Accuracy
	Exp 3		36	Trial Type, Base Rate, Avg. Accuracy
	Exp 4		36	Trial Type, Base Rate, Avg. Accuracy
Bindemann, Fysh, Cross, & Watts, 2016		2		
	Exp 1		40	Time Pressure
	Exp 2		20	Time Pressure
Bobak, Mileva, & Hancock, 2019		2		
	Exp 1		42	Trial Type , Avg. Accuracy
	Exp 2		52	Trial Type , Avg. Accuracy
Davis & Valentine, 2008		3		
	Exp 1		198	Trial Type, Avg. Accuracy
	Exp 2		200	Trial Type, Avg. Accuracy
	Exp 3		376	Trial Type, Avg. Accuracy
Dowsett & Burton, 2015		3		
	Exp 1		20	Trial Type, Avg. Accuracy
	Exp 2		36	Trial Type, Avg. Accuracy
	Exp 3a		36	Trial Type, Avg. Accuracy
Edmonds & Lewis, 2007		1	20	View & Pose
Favelle, Hill, & Claes, 2017		1	67	View & Pose
Fletcher, Butavicius & Lee, 2008		1	48	Exposure Time
Fysh & Bindemann, 2017		2		

	Exp 1		80	Time Pressure
	Exp 2		60	Time Pressure
Gentry & Bindemann, 2019		1	90	Trial Type, Photo Amount , Avg. Accuracy
Kemp, Caon, Howard, & Brooks, 2016		1	180	Trial Type, Avg. Accuracy
Kokje, Bindemann, & Megreya, 2018		1	74	Trial Type, Cross-Race, Avg. Accuracy
Kramer & Reynolds, 2018		3		
	Exp 1		54	Trial Type, Avg. Accuracy
	Exp 2		33	Photo Amount
	Exp 3		51	Trial Type, Photo Amount, View & Pose, Avg. Accuracy
Kramer & Ritchie, 2016		1	52	Trial Type, Avg. Accuracy
McCaffery & Burton, 2016		2		
	Exp 1		80	Trial Type, Avg. Accuracy
	Exp 3		96	Trial Type, Avg. Accuracy
Megreya & Bindemann, 2018		2		
	Exp 1		60	Trial Type, Avg. Accuracy
	Exp 2		32	Trial Type, Avg. Accuracy
Megreya & Burton, 2008		1	50	Trial Type, Avg. Accuracy
Megreya & Burton, 2006		1	30	Trial Type, View & Pose, Avg. Accuracy
Megreya, Bindemann, & Harvard, 2011		1	45	Trial Type, Avg. Accuracy
Megreya, Sandford, & Burton, 2013		2		
	Exp 1		80	
	Exp 2		80	Trial Type, Avg. Accuracy
Mileva & Burton, 2018		2		
	Exp 1		40	Trial Type, Avg. Accuracy
	Exp 2		60	Trial Type, Avg. Accuracy
Moore & Johnston, 2013		2		
	Exp 1		40	Trial Type, Avg. Accuracy
	Exp 2		28	Trial Type, Avg. Accuracy
Ozbek & Bindemann, 2011		1	30	Trial Type, Exposure Time, Avg. Accuracy
Papesh & Goldinger, 2014		4		

	Exp 1		61	Trial Type, Base Rate, Avg. Accuracy
	Exp 2		83	Trial Type, Base Rate, Avg. Accuracy
	Exp 3		40	Trial Type, Base Rate, Avg. Accuracy
	Exp 4		53	Trial Type, Base Rate, Avg. Accuracy
Papesh, 2018		1	883	Trial Type, Expertise, Avg. Accuracy
Papesh, Heisick, & Warner		5		
	Exp 1		118	Trial Type, Base Rate, Feedback, Avg. Accuracy
	Exp 2		58	Trial Type, Base Rate, Avg. Accuracy
	Exp 3		103	Trial Type, Base Rate, Avg. Accuracy
	Exp 4		76	Trial Type, Base Rate, Avg. Accuracy
	Exp 5		76	Trial Type, Base Rate, Avg. Accuracy
Ritchie, Mireku, & Kramer, 2020		2		
	Exp 1		959	Trial Type, Photo Amount, Avg. Accuracy
	Exp 2		1040	Trial Type, Photo Amount, Avg. Accuracy
Stephens, Semmler, & Sauer, 2017		2		
	Exp 1		95	Trial Type, Base Rate, Avg. Accuracy
	Exp 2		60	Trial Type, Base Rate, Avg. Accuracy
Susa, Michael, Dessenberger, & Meissner, 2019		2		
	Exp 1		200	Trial Type, Base Rate, Cross-Race, Avg. Accuracy
	Exp 2		68	Trial Type, Base Rate, Cross-Race, Avg. Accuracy
Towler, White, & Kemp, 2017		2		
	Exp 1		47	Trial Type, Expertise, View & Pose, Avg. Accuracy
	Exp 2		102	Trial Type, Avg. Accuracy
Weatherford, Erickson, Thomas, Walker, & Schein, 2020		3		

	Exp 1		91	Trial Type, Base Rate, Avg. Accuracy
	Exp 2		83	Trial Type, Base Rate, Feedback, Avg. Accuracy
	Exp 3		85	Trial Type, Base Rate, Feedback, Avg. Accuracy
White, Burton, Jenkins, & Kemp, 2014		3		
	Exp 1		44	Trial Type, Photo Amount, Avg. Accuracy
	Exp 2		72	Trial Type, Exposure Time, Photo Amount, Avg. Accuracy
	Exp 3		56	Photo Amount
White, Burton, Kemp, & Jenkins, 2013		1	300	Trial Type, Avg. Accuracy
White, Kemp, Jenkins, & Burton, 2014		2		
	Exp 1		42	Trial Type, Feedback, Avg. Accuracy
	Exp 2		56	Feedback
White, Kemp, Jenkins, Matheson, & Burton, 2014		1	65	Trial Type, Expertise, Avg. Accuracy
White, Phillips, Hahn, Hill, & O'Toole, 2015		1	59	Expertise
Wirth & Carbon, 2017		1	96	Trial Type, Expertise, Time Pressure, Avg. Accuracy
Number of studies (k)		**39**		
Number of tests (N)		**78**		
Number of participants (n)		**8205**		

Table 2.2. Summary of included studies, sample size, and comparisons

2.2.4 Accuracy Analyses

Average accuracy rates were calculated for overall accuracy, control group match accuracy, and control group mismatch accuracy using the accuracy rates reported or calculated from all articles included. Average weighted accuracy rates were also calculated to take the sample size of each experiment into account.

2.2.5 Effect Size Analyses

Once all the articles had been coded, 10 different comparisons were conducted. Ultimately, three of the comparisons could not be completed due to too small of a sample size (*N*

< 5): training, exposure time, and cross-race. The outcome variable for all comparisons was overall accuracy, and each comparison contained subgroups for match and mismatch accuracy. All data was input into Review Manager 5.4 (Higgins et al., 2021) and summary effects were calculated using the mean difference. Mean difference was chosen because all the outcome measurements were made on the same scale (0-100% accurate; expressed as a number between 0 - 1); the mean difference also allows for the combination of between- and within-subject (Higgins et al., 2021). Individual study estimates were calculated using mean differences. Difference in means was calculated in Review Manager 5.4 using the pre-programed statistics (Deeks & Higgins, 2010); the mean difference is given by

$$\boldsymbol{MD_i = m_{1i} - m_{2i}}$$

with standard error

$$\boldsymbol{SE\{MD_i\} = \sqrt{\frac{sd_{1i}^2}{n_{1i}} + \frac{sd_{2i}^2}{n_{2i}}}}$$

An inverse-variance method was used for all comparisons. The inverse-variance method pools all mean differences and individual effect sizes are weighted according to the reciprocal of their variance which is calculated as the square of the standard error. The intervention effect estimate (mean difference) is denoted by $\boldsymbol{\theta_i}$. Weights are calculated as follows:

$$\boldsymbol{w_i = \frac{1}{(SE\{\theta_i\})^2}}$$

The summary effect is then calculated:

$$\boldsymbol{\theta_{IV} = \frac{\sum w_i \theta_i}{\sum w_i}}$$

with

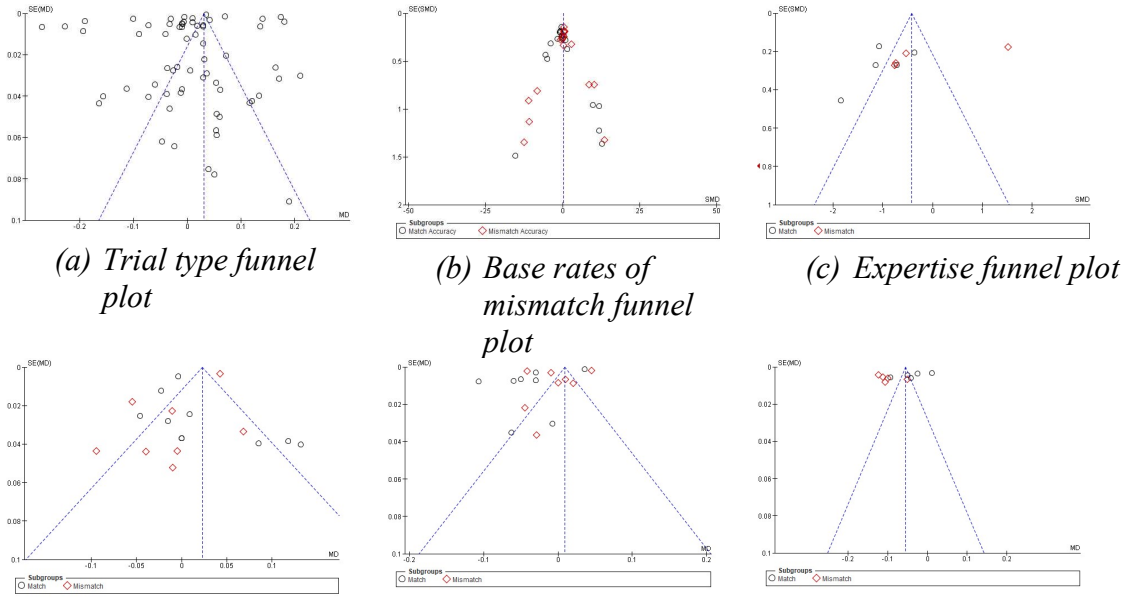
$$\boldsymbol{SE\{\theta_{IV}\} = \frac{1}{\sqrt{\sum w_i}}}$$

All statistics for formulae were taken from Deeks & Higgins (2010).

For comparisons that contained greater than 10 studies (trial type & base rate) both fixed effects and a random effects models were run, and the random effects model was presented if there was no evidence of funnel plot asymmetry, as recommended in the Cochrane Review Handbook (Higgins et al., 2021). Funnel plot asymmetry indicates bias in the sample of effect sizes included in the meta-analytic comparison. Bias may come from sample heterogeneity or publication bias. If a comparison contained fewer than 10 studies then only a fixed effects model was conducted as fixed effects are recommended for comparisons with smaller numbers of studies (Higgins et al., 2021). Therefore, the base rate comparison was the only comparison that is presented using the random effects model; all other comparisons use the fixed effects model. See Figure 2.2 for the funnel plots for each comparison. The funnel plots for expertise, photo amount, and time pressure show evidence of publication bias; this is likely a consequence of excluding unpublished manuscripts to ensure our results meet Daubert criteria.

Figure 2.2

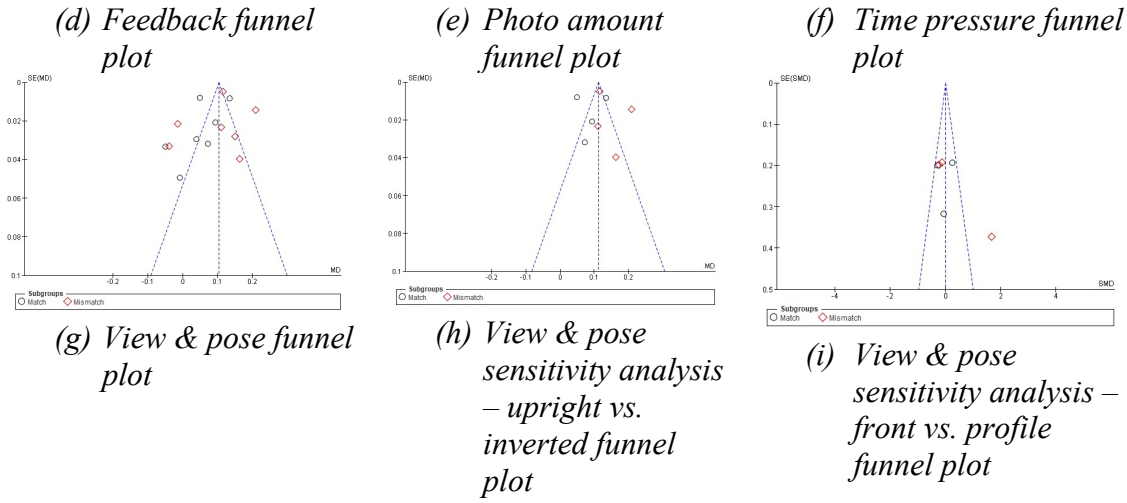
Funnel plots for all comparisons



(a) Trial type funnel plot

(b) Base rates of mismatch funnel plot

(c) Expertise funnel plot



(d) Feedback funnel plot

(e) Photo amount funnel plot

(f) Time pressure funnel plot

(g) View & pose funnel plot

(h) View & pose sensitivity analysis – upright vs. inverted funnel plot

(i) View & pose sensitivity analysis – front vs. profile funnel plot

Figure 2.2. Funnel plots for all comparisons

2.2.6 Study Characteristic Analysis

The relative importance of study characteristics was assessed using multiple regressions for each comparison with the MD as the predicted variable and all coded study characteristics as predictor variables. This analysis was exploratory as this was the first meta-analysis of face-matching research with a small number of studies and a limited amount of variability in the studies. Therefore, no specific hypotheses were created to direct the analysis and all predictor variables were added in the first block.

All data have been made publicly available at the Open-Source Framework and can be accessed at https://osf.io/23kfx/?view_only=e4cade430d2c4ffaa0e17db24946ff48.

2.3 Results

2.3.1 Average Accuracy

The average overall accuracy for control participants was calculated (see Table 2.3). Only control participants were considered in order to provide a useful baseline accuracy rate for expert testimony. Therefore, any experiment without a control was removed from the analysis. A control was considered either a baseline measurement or the measurement from a group of participants that did not receive any of the experimental conditions. The resulting dataset

contained 65 studies and 5,511 participants. The average overall accuracy for control participants was 80%. Match and mismatch accuracy and weighted averages were also calculated.

Table 2.3

List of computed average accuracy rates for included articles.

Type of Accuracy	*M*	*SD*
Overall accuracy	80.00%	7.74%
Weighted overall accuracy	80.20%	8.18%
Control match accuracy	80.50%	9.08%
Weighted control match accuracy	82.20%	8.90%
Control mismatch accuracy	79.50%	6.39%
Weighted control mismatch accuracy	78.30%	7.47%

Table 2.3. List of computed average accuracy rates for included articles

2.3.2 Effect Size Analyses

A summary table of effect size results from the meta-analytic comparisons can be found in Table 2.4. Forest plots for each comparison can be found in Appendix 1.

Table 2.4

Summary of meta-analysis statistics for all comparisons

Comparison	*k*	*n*	*MD*	Lower CI	Upper CI	*Z*	Effect size *p*	Tau^2	Chi^2	df	Heterogeneity *p*	I^2
Trial Type	64	10,842	0.03	0.03	0.03	64.73	0.000**		21,687.30	53	0.000**	100%
Base Rate	20	3,232	-0.01	-0.06	0.04	0.52	0.600	0.03	25,399	39	0.000**	100%
Expertise	5	2,300	-0.02	-0.02	-0.02	10.73	0.000**		365.16	9	0.000**	98%
Feedback	10	1,254	0.02	0.02	0.03	8.86	0.000**		175.53	19	0.000**	89%
Photo Amount	8	9,438	0.01	0.01	0.01	9.85	0.000**		1,831.99	15	0.000**	99%
Time Pressure	5	896	-0.05	-0.06	-0.05	36.17	0.000**		997.71	9	0.000**	99%
View & Pose	8	1,182	0.10	0.10	0.11	31.93	0.000**		208.40	15	0.000**	93%

Note: *k* = number of tests in the comparison; *n* = number of participants in the comparison; * = significant at the $p = 0.05$ level, ** significant at the $p < 0.001$ level, *** significant at the $p < 0.0001$ level; Tau^2 = estimate of between-study variance in the random-effect analysis; Chi^2 = estimate of whether observed differences in results are compatible with change alone; I^2 = amount of variance due to heterogeneity in the sample

Table 2 4. Summary of meta-analysis statistics for all comparisons

I^2 assess the amount of variance due to heterogeneity in the sample. Each comparison and most of the subgroup analyses revealed large I^2 values suggesting a significant amount of

variance due to heterogeneity. There was some evidence that I^2 values are biased with smaller meta-analyses (von Hippel, 2015), sometimes showing 12-point differences between expected and observed values. Regardless of the true values of I^2, heterogeneity needs to be considered when interpreting the results. Thus, it is important to remember this limitation when using these results in policy decision making and legal decision making and to use the presented 95% confidence interval for the effect sizes to ensure a more accurate estimate. The meta-regression (see section 2.3.4) provides some insight into the heterogeneity within the comparisons. Additionally, the model used for the comparison can affect the heterogeneity and interpretation of results which will be discussed more in depth in the discussion section.

All comparisons were significant other than the base rates of mismatch comparison, $Z = 0.52$, $p = 0.600$. The confidence interval of the estimated effect, $MD = -0.01$ [-0.06, 0.04], crossed zero, suggesting that the effect of base rates was not a reliable effect in the studies included in this meta-analysis. This finding did not support my hypothesis about base rates.

The trial effect comparison found an effect estimate of $MD = 0.03$ [0.03, 0.03] suggesting that participants were more accurate in match trials than mismatch trials, supporting my hypothesis.

For expertise I found a significant effect estimate of $MD = -0.02$ [-0.02, -0.02], $Z = 10.73$, $p < 0.000$ indicating that experts outperformed novice decision makers and supporting my hypothesis.

The photo amount comparison was significant ($MD = 0.01$, [0.01, 0.01]) suggesting that single comparison photos lead to more accurate decisions than multiple comparison photos, $Z = 9.85$, $p = 0.00$. I did hypothesize a significant effect for photo amount, but this result is unexpected because it is assumed that providing more comparison photos allows for a more

complete understanding of all the different ways a person might look, leading to more accurate classification (e.g., Bruce 1982; Longmore et al., 2008). This result seems to be driven by one study with an overinflated weight due to the large sample size and small standard deviation (Ritchie et al., 2020 exp 2, $SD_{overall}$ = 0.032, n = 1,040).

The time pressure comparison revealed a significant negative effect (MD = -0.05 [-0.06, -0.05], Z = 36.17, p < 0.000) showing that participants are more accurate when under low time pressure.. This supports my original hypothesis.

Finally, I found a significant positive effect of view showing that participants were more accurate when using upright, front facing as opposed to profile or inverted views, MD = 0.10 [0.10, 0.11], Z = 31.93, p < 0.000. These results support my hypothesis. The view comparison contained both upright, front facing views compared to both profile and inverted views, so a sensitivity analysis was conducted to determine whether the observed effect estimate is due the upright vs. inverted comparison or the front vs. profile comparison alone. The first analysis I conducted was the front vs. profile view comparison (Appendix 1 for forest plots for sensitivity analysis). The front vs. profile view comparison consisted of three tests and 500 participants. This comparison resulted in a non-significant effect size of MD = 0.01 [-0.17, 0.19], Z = 0.11, p = 0.91, I^2 = 81%. The upright vs. inverted view comparison revealed a large, significant, positive effect, MD = 0.11 [0.10, 0.12], Z = 31.71, p < 0.000, I^2 = 94%. Therefore, the effect seen in the original view comparison is driven by the upright vs. inverted view comparison.

2.3.3 Subgroup Analyses

Subgroup analyses examined match vs. mismatch accuracy separately in order to explore the impact of the underlying heterogeneity that was seen in the overall comparisons. (See Table 2.5).

Subgroup analyses revealed quantitative interactions for all the comparisons that showed significant subgroup differences. Specifically, interventions affect match and mismatch decisions to different degrees; an equal number of significant effects were seen for both match and mismatch decisions, but mismatch decisions were affected more strongly. The photo amount comparison showed statistically significant subgroup differences but the difference in magnitude between match and mismatch differences was so small that the statistical difference may be an artifact of the low power of the χ^2 test and the small number of studies included in the comparison.

Table 2.5

Summary of tests of subgroup differences.Comparison	*MD*	Lower CI	Upper CI	*Z*	*P*	*Chi²*	df	*p*	I^2
Base Rate						0.26	1	0.61	0%
Match	-0.03	-0.09	0.03	0.97	0.33	10580.3	19	0.000**	100%
Mismatch	0.00	-0.1	0.11	0.05	0.96	14402.24	19	0.000**	100%
Expertise						17.55	1	0.000**	94%
Match	-0.05	-0.06	-0.03	7.36	0.000**	6.59	4	0.16	39%
Mismatch	-0.02	-0.02	-0.01	8.86	0.000**	341.02	4	0.000**	99%
Feedback						60.77	1	0.000**	98%
Match	0.00	-0.01	0.01	0.74	0.46	32.23	9	0.000**	72%
Mismatch	0.04	0.03	0.04	11.78	0.000**	82.52	9	0.000**	89%
Photo Amount						49.22	1	0.000**	98%
Match	0.01	0.01	0.02	11.9	0.000**	925.93	7	0.000**	99%
Mismatch	0.00	0.00	0.00	1.62	0.10	856.84	7	0.000**	99%
Time Pressure						613.69	1	0.000**	100%
Match	-0.03	-0.03	-0.02	14.09	0.000**	14.09	4	0.000**	99%
Mismatch	-0.11	-0.11	-0.1	41.52	0.000**	85.07	4	0.000**	95%
View & Pose - Original						21.33	1	0.22	95%
Match	0.09	0.08	0.10	17.59	0.000**	79.09	7	0.000**	91%
Mismatch	0.12	0.11	0.12	27.04	0.000**	107.97	7	0.000**	94%
View & Pose - Upright vs. Inverted						23.47	1	0.000**	96%
Match	0.09	0.08	0.10	16.57	0.000**	55.16	3	0.000**	92%
Mismatch	0.13	0.12	0.13	27.47	0.000**	39.98	3	0.000**	92%

Note: $**p < 0.000$; Chi^2 = estimate of whether observed differences in results are compatible with change alone; I^2 = amount of variance due to heterogeneity in the sample

Table 2.5. Summary of tests of subgroup difference.

2.3.4 Meta Regression Results

A meta-regression was planned *a priori* to further explore heterogeneity and the influence of study characteristics. A meta-regression is an extension of a meta-analysis and creates a model of the linear relationship between study-level covariates and effect sizes (Bornstein et al., 2009). Meta-regressions can provide explanations for the heterogeneity seen in a meta-analysis by indicating if there are any individual predictors accounting for a large amount of heterogeneity. Also, meta-regressions allow for the examination of methodological variables across studies. I planned to undertake a meta-regression for each comparison, but the small number of studies included in the comparisons coupled with issues of missing data meant that I was only able to complete meta-regressions for the trial type comparison and the base rate comparison. The base rate comparison did not elicit any significant effects; however, the trial type comparison did, and are reported below.

The meta-regression was performed with the outcome variable of trial type effect size (*MD*), the standard error for each study in the trial type comparison and year published were entered as covariates. Sample size, percentage of female participants, average age of participant, base rate, photo illumination, photo color, photo quality, expertise of participants, view, time pressure, and exposure time entered as predictor variables in a single block. Most of the study characteristic values were the same across both levels of comparison variable, so only one of the coded values was used for each study. If the study characteristics had two different values based on the level of the comparison value, then the values were averaged (e.g., percent women and age of observer), or the value thought to affect the meta-regression the most was used. Thus, for the base rate predictor variable, the lowest base rate value for each test was used for this analysis;

for expertise, the highest level of expertise was entered into the analysis; for time pressure, the highest level of time pressure was entered into the analysis; and, for exposure time the lowest level of exposure time was entered into the analysis. Ultimately, photo illumination and time pressure were removed from the analysis because they lacked variability. The meta-regression was a random-effect meta-regression with the restricted maximum likelihood method. The model was marginally significant, $\chi^2(16) = 24.021$, $p = 0.089$. Photo color was the only significant, individual predictor. See Table 2.6 for a summary of the meta-regression findings.

Table 2.6
Summary of meta-regression findings for trial type analysis.

	χ^2	df	p	τ^2	I^2	H^2	R^2
Model Summary	24.021	16	0.089	0.007	99.50%	200.78	20.90%

Variable	Beta	*SE*	*t*	*P*	Lower CI	Upper CI
Year = 2010	0.109	0.132	0.825	0.419	-0.167	0.385
Year = 2013	0.09	0.112	0.801	0.433	-0.145	0.325
Year = 2014	-0.096	0.097	-0.996	0.332	-0.298	0.106
Year = 2016	-0.325	0.233	-1.396	0.179	-0.813	0.162
Year = 2017	0.024	0.1	0.24	0.813	-0.185	0.233
Year = 2018	-0.026	0.1	-0.259	0.799	-0.235	0.183
Year = 2019	0.046	0.128	0.362	0.721	-0.222	0.315
Sample Size	0.000	0.000	0.145	0.886	0.000	0.000
% Women	-0.014	0.109	-0.125	0.902	-0.242	0.215
Avg. Age	0.003	0.007	0.467	0.646	-0.011	0.018
Base Rate	0.036	0.159	0.226	0.823	-0.296	0.368
Photo Color	0.162	0.064	2.528	0.021*	0.028	0.295
Photo Quality	-0.037	0.039	-0.945	0.357	-0.119	0.045
Expertise	-0.12	0.075	-1.593	0.128	-0.277	0.038
Exposure Time	0.004	0.044	0.092	0.928	-0.089	0.097

Note: *$p < 0.05$

Table 2.6. Summary of meta-regression findings for trial type analysis

2.4 Discussion

The purpose of this meta-analysis was to quantitatively summarize the face matching literature and to identify areas that may benefit from more research. Results suggest that while a good amount of research on face matching has been conducted, there is still more work that needs to be done. My analyses showed that the variables that impact face matching ability include trial type, expertise, feedback, photo amount, time pressure, and view. View and exposure time were the two factors that were most stable and had the largest effects on face matching accuracy. Base rates were the only non-significant variable.

The investigation into base rates included the most studies and is the variable that has been researched the most. Base rates were the only variable in my analysis that did not show a significant, reliable effect, which was the opposite of my predictions. This is a surprising result as base rates are thought to highly affect face matching accuracy (Susa et al., 2019; Weatherford et al., 2020). The lack of significant results likely stems from the nine non-significant results in the Bindemann et al. (2010) and Weatherford et al. (2020) studies. In terms of the Weatherford et al. (2020) experiments, feedback and within stimulus variability were also manipulated, and interactions between these variables and the base rates manipulation were found. The results of the Weatherford et al. experiments found some evidence of low prevalence effect such as reduced mismatch accuracy in low prevalence conditions, but the significance did not rise to the level of a main effect. Therefore, the effect of base rates was lower in the meta-analysis and contributing to a non-significant result. One potential reason for the lack of significant results in the Bindemann et al. (2010) study is that participants were aware of which condition they were in, which may have allowed them to adjust their decision-making criteria based on that knowledge in order to avoid the low prevalence effect. Tversky and Kahneman (1973) found that

people often did not employ base rates appropriately when making decisions. Overall, participants may not be able to conceptualize the small number of base rates that occur in face matching research in order to use that information to make decisions. Or, participants may need explicit information about the base rates to make appropriate decisions; Davis et al., (2021) found similar results where a low prevalence effect did not occur until the participants were aware of the low number of mismatches in the stimuli. While my results do not preclude the effect of base rates, it does suggest that base rates have a much more variable and study- or sample-dependent effect.

Higher levels of accuracy were associated with more expertise (Towler, White, & Kemp, 2017; White Phillips, Hahn, Hill, & O'Toole, 2015). These results support my hypothesis; however, the observed effect size was small. Certain types of careers may lead to better face matching accuracy due to either the tasks involved, or the types of training given to those in the career. This idea is supported by a recent review of expertise in face matching by White et al. (2020). White et al. (2020) reviewed 29 tests of face matching accuracy in professional versus novice groups and found that 12 of those tests showed no difference, but facial examiners consistently outperformed novices. This suggests that simply doing the task repeatedly is not enough to increase accuracy, but that training, practice, and feedback may all enhance performance for professional face matchers. The small effect size seen in this meta-analysis may be due to the types of experts involved in the studies. For example, Papesh (2018) found no differences between students, notaries, and bank workers, but Wirth and Carbon (2017) used police officers and found that they outperformed novices.

Providing feedback to participants was also associated with increased face-matching ability. This finding supports my hypothesis. Accuracy feedback is rarely, if ever, provided for

professional face matchers. Feedback is important because it allows the matchers to assess and change their criterion for matches and mismatches in order to increase their accuracy and reduce misidentifications. All different types of feedback were collapsed into a single group in this analysis, so there is a chance that the effect could be higher if only a specific type of feedback was examined. For example, White, Kemp, Jenkins, and Burton (2014) found that trial-by-trial feedback was more effective than overall feedback after each block. At this time, not enough studies examined trial-by-trial feedback for a comparison between trial-by-trial and overall feedback to be conducted.

Increasing the number of photos used when making matching decisions also resulted in increased accuracy, which matched what I predicted. An increased number of photos helps to reveal details to the face that may be hidden by the lighting, angle, or camera artifacts from single photos and can help to reveal idiosyncrasies in faces (Burton, 2013; Burton, Kramer, Ritchie, & Jenkins, 2016). The larger number of photos can also help face matchers to assess their criterion for matches and mismatches by having a larger number of photos to compare the target photo to.

Higher time pressure was associated with lower overall accuracy. This was a predicted result, as higher time pressure means that the matcher has less time to devote to making the decision and may feel pressure that then leads them to make an incorrect decision. This analysis collapsed both types of time pressure (global vs. local) into one variable, so larger effects may result when using local time pressure as seen in Wirth and Carbon (2017). Currently, not enough studies exist examining global vs. local time pressure to conduct a meta-analytic comparison.

Finally, upright frontal views in photos resulted in higher accuracy than any other pose. This finding supported my hypothesis. This effect was one of the largest and most reliable effects

in the entire analysis. Results of the sensitivity analysis revealed that the upright vs. inverted studies were driving the observed effect size in the view comparison. View is an extremely important factor to consider when creating face matching stimuli or making face matching decisions. While the differences seen between upright vs. inverted views make sense given face processing literature showing that inverted faces are not processed the same way as upright faces (Farah et al., 1998; McKone et al., 2007; Robbins & McKone, 2003; Rossion, 2008; Tsao & Livingstone, 2008; Yovel & Kanwisher, 2008), it is surprising that significant differences in pose were not seen. Previous research on pose has shown that changing the location of the camera when taking a photo or even moving ones face even a few degrees from center can totally change the way the person looks (Favelle et al., 2017). View and pose are such important variables in unfamiliar face matching because the two photos or the photo and the person presented are the first time you are seeing the subject and are the only representations of the person you have to assess whether the stimuli match. Stimuli of unfamiliar people provide limited information about how someone looks. This limited amount of information is very difficult to use to identify someone when they do not look exactly like the single representation of the person that you have been given (Burton, 2013; Burton, Kramer, Ritchie, & Jenkins, 2016). Therefore, view changes between stimuli make it even more difficult to identify someone because the limited amount of information you have about the way the person looks does not line up with the new view of the person. While this distinction between view and pose and view driving the effect seen in the meta-analysis makes sense and is theoretically useful, it does not provide much practical utility. There are almost no situations where decision makers would be asked to use inverted stimuli to make a face matching decision.

There are some limitations that should be considered when interpreting the results of my meta-analysis. First, the use of a single coder for the entire coding process can introduce error. Typically, multiple coders are used for meta-analyses to reduce bias and ensure more objective coding of the included articles (Higgins et al., 2021). The current study only employed a single coder because most of the coding was for objective values that were not open to interpretation; however, the use of single coder is a limitation and may have introduced error and bias into the meta-analysis. For example, as the coder, I was aware of the hypotheses of the study and may have coded in a way more likely to support those hypotheses.

Dichotomizing independent variables and study characteristics into extreme values may have led to overestimations of estimated effect sizes. The decision to dichotomize independent variables and study characteristics was based on methodology employed by Shapiro and Penrod (1986). Dichotomizing variables into extremes leads to a low versus high comparison that allows for the inclusion of more studies since in most instances, too few studies used the same operationalization and levels of a variable.

On top of this, the large I^2 statistics seen in the meta-analysis is something to be concerned about. Large I^2 values indicate substantial amounts of heterogeneity within the studies included in the comparisons. This may suggest that the studies are too different and should not be meta-analyzed together or may be due the natural heterogeneity seen between study samples. I attempted to explore the heterogeneity using subgroup analyses and meta-regression and found little clarity. I did find qualitative interactions in many of the subgroup analyses, but not a significant amount to explain the large amounts of heterogeneity seen in the overall comparison. Additionally, the meta regression model did not reach significance. The inclusion of the most impactful values in the meta regression for study characteristics with two different values based

on IV level (e.g., base rates of mismatch, time pressure, exposure time, and expertise) likely led to an over-estimated regression statistic, R^2, and beta coefficients for those predictors. However, the meta-regression did not reach significance and no individual predictors were significant, so that limitation is less important in the interpretation of the findings. For the comparisons that utilized a fixed-effect model (all except the base rates comparison) we can effectively ignore heterogeneity; however, the large amount of unaccounted for heterogeneity may mean that the effect estimate found in this analysis may represent an effect that does not actually occur in real-life.

Furthermore, most of the comparisons contained less than 10 studies. The small number of studies may mean that larger studies are over-weighted in the analyses due to the larger weighting assigned to studies with larger numbers of participants and smaller standard errors. The majority of the meta-analyses and reviews conducted using the Cochrane protocol contained fewer than seven studies (Higgins et al., 2021). So, while it is important to consider this limitation when interpreting the results of the meta-analysis, the protocol used has been designed to limit effects of outlier studies in small meta-analyses.

Despite these limitations, this meta-analysis helps to advance the literature over qualitative reviews in three different ways. Firstly, while this meta-analysis uses much of the same data as recent qualitative reviews, the statistical analysis employed here provides more actual data about the state of face matching research and variables that are important in the assessment of face matching accuracy. Specifically, while qualitative reviews give synthesized estimates of the effect of variables that moderate face matching accuracy, my meta-analysis reveals exact measurements of effect sizes and statistical importance. Secondly, this meta-analysis helps to reconcile inconsistent findings in the general research so that forthcoming

research can explore new avenues. Finally, my meta-analysis helps to inform future research based on study characteristics and unexpected results.

The results of my meta-analysis can be used by other researchers to develop training programs for face matching professionals by advising researchers of what variables are most important to include. Also, as face matching is increasingly used with identification technology for arrests and testimony (e.g., Buolamwini et al., 2020; Garvie et al., 2016), my results provide information for courts, judges, and lawyers about face matching and the accuracy of these types of decisions, which is necessary information when determining the admissibility of testimony related to any new technology. The results of this meta-analysis consolidate statistical information for experts to use in courts.

The current meta-analysis provides a wealth of information about the state of face matching research but is limited by the studies included in the analysis. The studies I included were limited to a subset of the available research; specifically, unaided, unfamiliar face matching with participants with average face matching abilities. This subset of research encompasses the majority of face matching decisions and the majority of the type of people making those decisions. However, familiarity with the targets can significantly influence the accuracy of decisions and as face matching is increasingly used in policing scenarios (e.g., Buolamwini et al., 2020) it is more likely that the decision makers could be at least somewhat familiar with some of the targets. There's also the relatively unexplored issue of AI-aided face matching, which is also used in policing scenarios. These limitations lead to potential directions for future research.

As opportunities for face matching – including the use of facial identification software and face matching algorithms -- become more common among police and state organizations it is essential to understand the variables that can affect the final matching decision made by face

matchers. A misidentification or missed mismatch can result in wrongful imprisonment and revoking of an innocent person's rights. Therefore, organizations need to be aware of the fallibility of these types of decisions and how they can be improved. Additionally, future research should focus on trying to explain the unexpected results from this meta-analysis, specifically concerning base rates of mismatch.

Chapter 3. Base Rate Study

3.1 Background & Significance

Many face matching studies use 50% mismatch trials (see Bindemann, Avetisyan, & Blackwell, 2010). However, it is likely that the incidence of mismatches (e.g., imposters, fake IDs, etc.) in applied contexts is significantly lower. In passport identification, for example, the mismatch rate could be as a low as 0.0075% (Hetter & Cripps, 2014). Bindemann et al. (2010) recognized that the 50% split between matching and mismatching photo pairs typically seen in laboratory research is not representative of real life and may inflate accuracy rates. The effect of different match/mismatch base rates was examined by including either a low base rate (2% mismatch) or a high base rate (50% mismatch). Participants were informed of their base rate condition (low vs. high) and the mismatches in the low base rate condition were always presented at the end of the trial block. Low base rates resulted in participants detecting more mismatches than high base rates. However, this increase in mismatch accuracy came with a higher false positive rate. Based on this evidence, the researchers concluded that low mismatch frequency did not impair mismatch detection in unfamiliar face matching.

Other researchers (e.g., Papesh & Goldfinger, 2014; Susa et al., 2019) suggest that the lack of a low prevalence effect is due to participants being aware of the condition they were in and the mismatches always being at the end of trial blocks. Papesh and Goldinger (2014) tested low

(10%) vs. high (50%) mismatch prevalence conditions with unaware participants and found inflated miss rates under low prevalence conditions. In the low condition, mismatch errors rates were around 45%, roughly doubling the rate for the high condition (i.e., 20% errors). This low prevalence effect on accuracy persisted when participants were allowed to correct their initial judgements, when they were asked to give certainty judgements after their decision, and when they were permitted second looks at the face pairs. Papesh and Goldinger's findings suggest that under more realistic viewing conditions, low prevalence of mismatches is a large, persistent source of errors in face matching. Similar results are seen in Papesh, Heisick, and Warner (2018) and Susa et al. (2019). Overall, the cumulative results suggest that low mismatch prevalence reduces face matching accuracy. However, the meta-analysis I conducted found no overall effect of low base rates of mismatch.

The meta-analysis collapsed participants across aware and unaware participants. This may have washed out a potential effect of low base rates. In support of this, a recent study found, unexpectedly, that control subjects showed higher accuracy in low prevalence conditions when, no mismatch prevalence information was provided (Davis et al., 2021). In a second experiment, the expected low prevalence effects were observed when top-end-of-typical range ability participants (participants with scores between 84-94 on the Cambridge Face Memory Test: Extended (CFMT+) and 37-39 on the Glasgow Face Matching Test (GFMT)) were informed of their condition. This study suggests that participants only display a low prevalence effect if they are aware of prevalence in advance, which is opposite from what Bindemann et al. (2010) found and what other researchers (e.g., Papesh & Goldinger, 2014) have suggested.

Therefore, the existing findings in the literature are mixed, and more research is required to clarify the issue. To this end, the current study seeks to examine whether prior knowledge of

mismatch prevalence is necessary for the LPE to occur in face matching research. The results of this study will also provide insight into the unexpected results of the meta-analysis. Additionally, the current study uses a lower mismatch base (10%) which is a closer approximation of real-world face matching contexts where mismatches are rare.

3.1.1 Hypotheses

Based on the previous research, I hypothesized that awareness is not necessary for the LPE to occur and that participants who are aware of their low prevalence condition will not show any differences in match and mismatch accuracy. Additionally, I hypothesized that there would be no main effect of prevalence, based on the results found in the meta-analysis. I also hypothesized that participants who were aware of their condition would be more accurate than not aware participants.

3.2 Method

This study employed a 3 (prevalence: 10% mismatch, 50% mismatch, 90% mismatch) x 2 (awareness: aware vs. unaware) x 2 (trial: match vs. mismatch) repeated measures ANOVA with prevalence and awareness as between-subjects measures and trial as the repeated measure.

3.2.1 Participants

One hundred ninety-eight Ontario Tech University students were recruited for this experiment. Participants were compensated with 0.5 university research credits One participant was removed for incomplete data leaving a final sample of 197. A G*Power 3.1 sensitivity analysis revealed that with $N = 197$, $\alpha = 0.05$, and 80% power, the minimum effect size that this study can reliably detect is $f = 0.129$ or $\eta_p^2 = 0.016$ (Faul et al., 2007). The final sample contained 63% ($N = 124$) females, 32% ($N = 64$) males, 4% ($N = 8$) gender non-conforming

individuals, and 2% ($N = 4$) who preferred not to answer. The majority of the sample identified as White (40%), 25% of the sample identified as South Asian, and 9% of the sample identified as Black. The remaining approximately 25% of the sample included Arab, Asian, East Asian, First Nations, Latinx, and Pacific Islander participants. Participants aged in range from 18 to 57 years old with an average age of 20.55 years old.

3.2.2 Materials

When planning this study, I planned to use 1%, 50%, and 99% prevalence conditions, but actually tested 10%, 50%, and 90% prevalence conditions. This mistake was not caught until after all data had been collected. Therefore, participants were presented with information stating that they were in the 1%, 50%, or 99% condition, but the conditions were really 10%, 50%, and 90%.

3.2.2.1 Face Matching Stimuli. Stimuli were taken from the Glasgow Face Matching Test (GFMT; Burton et al., 2010). The GFMT consists of premade facial pairs. The pairs are made up of black and white, front facing photos that have been cropped to show only the face. Photo pairs were created using volunteers from The University of Glasgow; 304 individuals were photographed (132 women) with an average age of 22.9 for men and 23.2 for women. The final GFMT is comprised of 168 pairs of faces with full-face poses and a neutral expression taken from two different cameras roughly 15 minutes apart. Of the total face pairs, half (84) are same-face trials depicting two photos of the same person. The same people are also used in different-face trials so that one of the person's images is presented alongside a similar face from the database. Non-matching pairs were created based on pairwise similarity measures generated using a sorting technique (see Bruce et al., 1999). The foils for nonmatching trials were the faces most similar to each of the target identities. No information about the racial composition of

individuals used in the GFMT was provided. There are no obviously non-White individuals included in the set. The GMFT has been validated against other tests of face matching ability (Burton et al., 2010).

I presented 40 trials to participants in all three conditions. The 50% prevalence condition included 20 match and 20 mismatches; the 10% condition included 36 matches and four mismatches; the 90% prevalence condition included four matches and 36 mismatches. All pairs in the 50% condition were sampled randomly from the GFMT. For the 10% prevalence condition, four randomly chosen mismatches and the 20 matches used in the equal prevalence condition were used, along with an additional 16 randomly sampled matches from the GFMT to create the 40 face pairs. The 90% prevalence condition was created using the 20 mismatch pairs from the 50% condition, along with an additional 16 randomly sampled mismatched pairs, and four matched pairs randomly chosen from the matches used in the 50% condition. Each prevalence condition included an equal number of male and female pairs across match and mismatch trial types. Facial pairs and all other stimuli for the study can be found in Appendix B.

3.2.2.2 Awareness Statement. Participants were shown an awareness statement. In the aware conditions, participants were shown the following statement: "This set of stimuli contains [1%, 50%, or 99%] mismatches. Please keep that in mind when answering the questions while trying to be as quick and accurate as possible." The statement included no information about matches. In Davis et al. (2021) and other previous research, awareness statements only included information about one type of trial. Base rate information is typically given with only information about one condition (e.g., Tversky & Kahneman, 1973). In the unaware conditions, participants were shown this statement: "This set of stimuli contains both matches and

mismatches. Please keep that in mind when answering the questions while trying to be as quick and accurate as possible."

3.2.2.3 Manipulation Check. In order to assess whether participants were truly aware of their mismatch prevalence condition, participants were exposed to a manipulation check after completing all the face matching pairs. Participants were asked, "Which mismatch condition were you told you were in?" Participants selected a response from either "1%," "50%," "99%," or "I was not informed of my mismatch condition."

3.2.3 Procedure

Participants were recruited from the Ontario Tech participant pool using SONA. All participation took place online through a survey hosted on Qualtrics. Participants were randomly assigned to one of the six between-subjects conditions. Participants first saw the instructions for the survey. The instructions reminded participants that features may change over time and encouraged them to be as quick and accurate as possible when answering questions. The instructions were the same for all conditions. Before viewing any pairs, participants were shown an awareness statement. After the awareness statement participants viewed the face pairs in a random order. Each pair was presented with the question "Do the two photos show a match or a mismatch?" After completing all 40 trials participants answered a few demographics questions and then were thanked and debriefed. Participants were also exposed to a manipulation check asking them which percentage of mismatches they were exposed to so that I could ensure all participants were aware of their condition. On average, the whole process took 15 minutes.

3.3 Results

In the aware conditions, 60% (58/96) of the participants did not pass the manipulation check. In the unaware conditions, only two participants failed the manipulation check.

Participants who failed the manipulation check averaged a study completion time of 7 minutes, less than half of the amount of time taken by the overall sample. Participants may have failed the participant check due to the amount of time participants completed the study in or due to a lack of salience of the awareness statement employed. Therefore, *t*-tests were conducted to determine whether accuracy rates were significantly different between the group of participants that passed the manipulation check and those that did not. No significant differences were seen between the sets of participants for any of the types of accuracy (mismatch, $t(195) = -1.045$, $p = 0.297$; match, $t(195) = 0.482$, $p = 0.630$; overall, $t(195) = 0.453$, $p = 0.651$ However, this analysis is likely underpowered; out of an abundance of caution I am presenting only the results for the subset of participants that passed the manipulation check. A summary of results for the overall sample can be found in Appendix C for comparison.

A sensitivity analysis was conducted with the sample that passed the manipulation check and revealed that with $N = 137$, $\alpha = 0.05$, and 80% power, the minimum effect size that this study can reliably detect is $f = 0.15$ or $\eta_p^2 = 0.022$.

3.3.2. Participants who Passed Manipulation Check

The participants that passed the manipulation check ($N = 137$) were evenly distributed across the prevalence conditions (10% $N = 47$; 50% $N = 50$; 90% $N = 40$), but were over-represented in the "not aware" condition (Not aware $N = 99$; Aware $N = 38$). All presented within-subjects results employ sphericity assumed *F*-tests and all presented post-hoc comparisons use the Bonferroni correction provided in SPSS. The current SPSS package does not provide an adjusted alpha when using the Bonferroni correction, instead it performs the mathematically equivalent procedure of taking the observed, uncorrected *p*-value and multiplying it by the number of comparisons made. Then, instead of using a new alpha value to

assess significance, the multiplied *p*-value is compared to the *a priori* alpha level of 0.05 (IBM, 2020).

A 3(prevalence) x 2 (awareness) x 2 (trial type) repeated measures ANOVA with prevalence and awareness as between-subjects measures and trial type as the repeated measure was conducted. Results showed a significant main effect of trial, $F(1) = 8.059$, $p = 0.005$, $\eta_p^2 = 0.058$, suggesting that mismatches elicited greater accuracy than matches.

No main effects were seen for prevalence ($F(2) = 1.363$, $p = 0.259$, $\eta_p^2 = 0.02$) or awareness ($F(1) = 1.978$, $p = 0.162$, $\eta_p^2 = 0.015$). A two-way trial x prevalence interaction was found, $F(2) = 6.983$, $p = 0.001$, $\eta_p^2 = 0.096$, but was qualified by a higher-order interaction. No interaction was found for trial x awareness ($F(1) = 3.045$, $p = 0.083$, $\eta_p^2 = 0.023$) or prevalence x awareness ($F(2) = 0.919$, $p = 0.402$ $\eta_p^2 = 0.014$). Finally, a three-way trial x prevalence x awareness interaction was seen, $F(2) = 7.977$, $p < 0.001$, $\eta_p^2 = 0.109$. In the 10% condition, unaware participants were more accurate on mismatch (vs. match) trials. In the 90% condition, unaware participants were more accurate on match (vs. mismatch) trials. No significant differences in terms of trial type were seen for participants in the 50% condition.

Figure 3.1
Summary of means for the trial x prevalence x awareness interaction

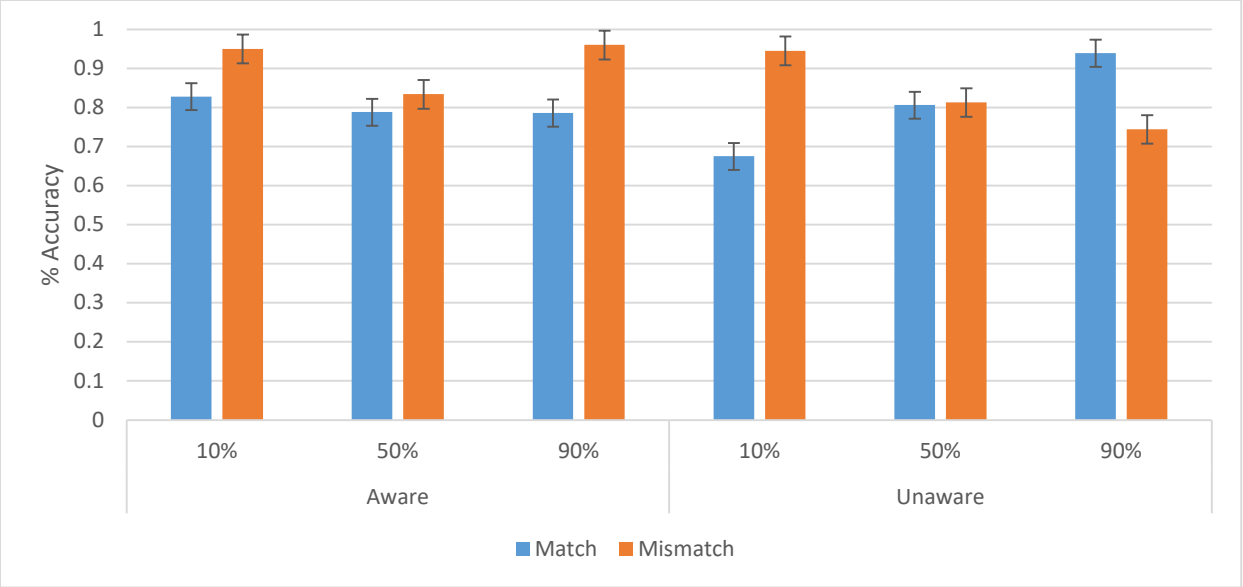


Figure 3. 1. Summary of means for the trial x prevalence x awareness interaction

3.3.3 Signal Detection Analyses

Accuracy rates only provide part of the answer when looking at face matching decisions. Accuracy rates do not delineate between discrimination and bias in decision making. Discrimination (*d'*) and bias (*β)* are important to consider because they provide information about how participants are making decisions and they also give a fuller picture of what might be done to improve decision making overall. Discrimination indicates the participants' sensitivity (i.e., their abilities to differentiate between matches and mismatches); *d'* is a measure of performance independent of bias (*β*). Bias describes the pattern of responses that a participant exhibits and whether a participant is more likely to answer match, mismatch, or answers both evenly. Therefore, I conducted signal detection analyses (SDT) to learn more about the discrimination and bias of the participants in this experiment. Match accuracy for hit rates and false alarms was calculated by subtracting mismatch rates from 1. For hit rates of 1 or false alarm rates of 0, the standard correction was applied, where N = is the maximum number of trials for the respective measure (i.e., lures or targets):

$$False\ Alarm\ Correction = \frac{1}{2N}$$

$$Hit\ Correction = 1 - \frac{1}{2N}$$

(MacMillan & Kaplan, 1985)

Discrimination (*d'*) is calculated as follows with Φ indicating a *Z*-score (MacMillan, 1993):

$$d' = \Phi_{hits} - \Phi_{false\ alarms}$$

β was used as the measure of bias and was calculated as follows (Stanislaw & Todorov, 1999):

$$\beta = e^{\left\{\frac{[\Phi_{false\ alarms}]^2 - [\Phi_{hits}]}{2}\right\}^2}$$

I planned to conduct signal detection analyses for both the sample of participants that passed the manipulation check and those who did not pass the manipulation check to further examine if a difference exists between those samples, however, the distribution of the participants in the sample that did not pass the manipulation check meant that there were not enough participants in the not aware condition to make meaningful comparisons. Thus, only the results for the sample who passed the manipulation check are presented.

3.3.4 SDT for Sample who Passed the Manipulation Check

A 3 (prevalence) x 2 (awareness) ANOVA was conducted for *d'*. No significant difference was found for prevalence, $F(2) = 2.067$, $p = 0.131$, $\eta_p^2 = 0.031$. A significant main effect was seen for awareness, $F(1) = 5.075$, $p = 0.026$, $\eta_p^2 = 0.038$. Aware participants exhibited higher discrimination than not aware participants

Finally, a significant interaction was found between prevalence and awareness, $F(2) = 5.228$, $p = 0.007$, $\eta_p^2 = 0.075$. Figure 3.2. Significant differences in sensitivity across prevalence were only seen for participants in the not aware condition. Unaware participants in the 50%

mismatch condition were significantly more sensitive than unaware participants in the 10% condition ($p < 0.001$) and the 90% condition ($p = 0.005$). No significant difference was seen between the 10% and 90% conditions ($p = 0.714$).

Figure 3.2
Prevalence x awareness interaction for d'

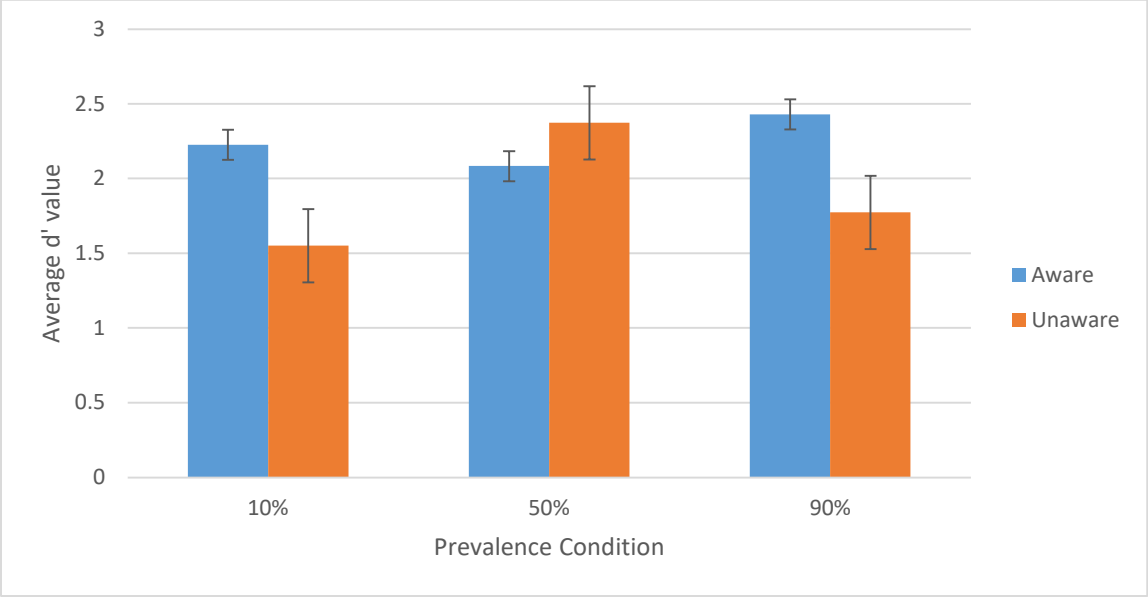


Figure 3. 2. Prevalence x awareness interaction for d'

A 3 (prevalence) x 2 (awareness) ANOVA was conducted for β. A significant main effect was found for prevalence, $F(2) = 5.449$, $p = 0.005$, $\eta_p^2 = 0.078$. Participants in the 90% condition were significantly more likely to classify something as a mismatch than participants in the 10% condition ($p = 0.004$). No significant differences were seen between response bias in the 10% condition and the 50% ($p = 0.351$) or the 50% condition and the 90% condition ($p = 0.148$). A significant main effect was found for awareness as well, $F(1) = 6.645$, $p = 0.011$, $\eta_p^2 = 0.049$. Participants in the aware condition were more likely to respond mismatch than participants in the unaware condition. Finally, a significant interaction was found between prevalence and awareness, $F(2) = 7.911$, $p < 0.001$, $\eta_p^2 = 0.109$. See Figure 3.3. Significant differences across prevalence conditions were only seen in the aware condition. Aware participants in the 90%

condition showed the highest value of *β,* demonstrating that they were the most likely to respond with "mismatch," significantly higher than participants in the 10% condition ($p < 0.001$) and the 50% condition ($p = 0.025$).

Figure 3.3.
Prevalence x awareness interaction for β

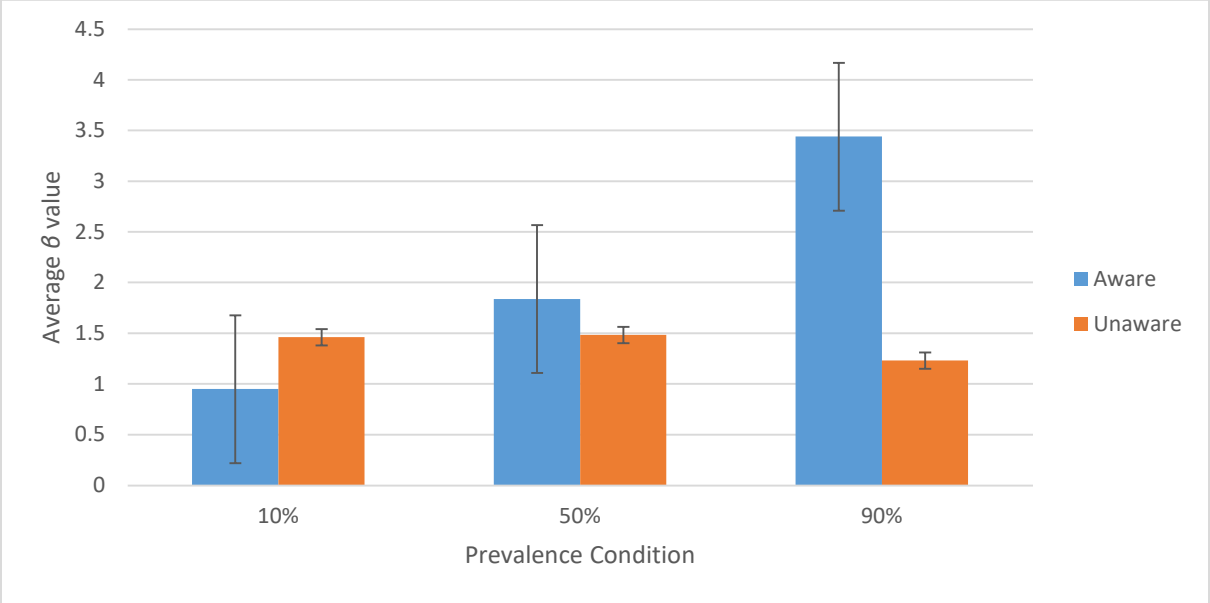


Figure 3. 3. Prevalence x awareness interaction for β

One-sample *t*-tests were conducted for each condition to see if participants' responses differed from chance (no sensitivity or no bias). The test statistic for *d'* was 0 (indicating no sensitivity) and 1 for *β* (indicating no bias). In terms of *d'*, results of the *t*-tests suggest that participants discriminated between matches and mismatches significantly better than chance, regardless of condition. For *β*, the tests indicated a bias to respond "mismatch" in the not aware 10% condition and the aware 90% condition, but mostly participants did not differ from chance, suggesting the lack of bias in most conditions. See Table 3.1 for a summary of *t*-test statistics.

Table 3.1

Summary of one-sample t-test statistics

SDT statistic	Awareness	Prevalence	*t*	*df*	*p*

d'		10%	11.99	14	<.001**
		50%	7.92	15	<.001**
	Aware	90%	7.46	6	<.001**
		10%	15.45	31	<.001**
		50%	14.69	31	<.001**
	Not Aware	90%	18.27	32	<.001**
β		10%	-.29	14	.770
		50%	1.95	15	.070
	Aware	90%	4.29	6	.005*
		10%	6.05	31	<.001**
		50%	1.88	31	.069
	Not Aware	90%	.79	32	.43

Note:* indicates a significance at the $p < 0.05$ level; ** indicates a significance at the $p < 0.001$ level

Table 3. 1. Summary of one-sample t-test statistics.

3.4 Discussion

Awareness did not induce a low prevalence effect and accuracy was high in the 10% mismatch condition regardless of awareness. Davis et al. (2021) posited that the unexpected results in low prevalence participants (increased accuracy instead of decreased accuracy) seen in their first experiment was because those participants were given no information about mismatch prevalence. They then tested this idea in their second experiment where they found support for the idea that awareness of the base rates of mismatch was necessary for the low prevalence effect to occur. Ultimately, the results of the current study are more in line with the findings of Bindemann, et al. (2010), with participants showing increased mismatch accuracy with awareness of low mismatch rates.

Accuracy is the outcome variable most used in previous research and allows for comparisons with said research, whereas signal detection theory provides additional detail and value by revealing details about how the manipulations affect decision-making. Discrimination did not differ across prevalence conditions without the presence of awareness, but the average discrimination was rather high ($d' > 1.5$). Comparisons to chance revealed that matches and mismatches were discriminable regardless of mismatch prevalence condition. Prevalence and

awareness interacted to produce the highest level of sensitivity for not aware participants in the 50% condition. This finding suggests that any change from the 50-50 split makes it more difficult for participants to tell the difference between matches and mismatches. Reductions in discrimination outside the 50-50 split may be due to a participant's expectations of the composition of the stimuli. Unaware participants may expect an even split between mismatches and matches based on the representative heuristic (Kahneman & Tversky, 1973). However, when participants were informed of their prevalence condition, sensitivity did not significantly differ by prevalence condition. This suggests that awareness may modify sensitivity, but not in the expected direction. Post hoc tests revealed that aware participants in the 10% condition exhibited higher sensitivity than unaware participants in the 10% condition, suggesting that awareness did not induce a low prevalence effect in the 10% condition, but increased discriminability, protecting against a low prevalence effect.

Participants' response bias across prevalence conditions only differed in the aware condition. Participants exposed to the 90% condition were more hesitant to classify something as a match. This suggests that aware participants were able to accurately adjust their responses. Aware 10% conditions should be more willing to classify trials as matches since they know that only 10% of the trials will contain a mismatch, but these participants exhibited no response bias. The lack of a response bias means that participants in the 10% aware condition were not able to integrate the base rate information they received to correctly adjust their responses, a result supported by previous research concerning base rates (e.g., Tversky & Kahneman, 1973). This condition, the 10% aware condition, is most analogous to what happens in real-world settings – most of the identity documents a face matching decision-maker encounters will be matches, and they know that, yet the current study suggests that even then decision-makers do not default to a

match bias. Often, emphasis is placed on increasing accuracy and improving decision-making to detect mismatches therefore ensuring that a threat does not pass security checkpoints. The potential consequences of missing a mismatch at security checkpoint, such as TSA or a border crossing, are large and may even be catastrophic. Thus, it may be beneficial for decision-makers to stay vigilant and suspicious even though they know most of the documents will be legitimate, even at the expense of efficiency. However, expending cognitive resources by scrutinizing all the documents they see could lead to a depletion of cognitive resources that may allow a fraudulent document to pass through (Baumeister et al., 1998; Muraven & Baumeister, 2000).

The current study did not find support for the idea that awareness is necessary for the low prevalence effect to occur. A large portion of the sample were unable to pass a manipulation check, so there is a chance that an effect of awareness could occur with a more balanced sample. There is a potential lack of power in this study because a large portion of the participants failed the manipulation check. However, the sensitivity analysis I conducted for the sample that passed the manipulation check ($N = 137$) showed that an effect size of $\eta_p^2 = 0.022$ could be reliably detected in this study. All of the effect sizes found here are larger than $\eta_p^2 = 0.022$. Additionally, findings are similar when looking at the full sample and when looking at the sample that passed the manipulation check. So, while the power of this study was reduced, the pattern of results and the sensitivity analysis suggest that findings may be robust. Still, future research should continue to examine awareness and its interaction with the LPE with a more explicit warning

Results of the current study suggest that informing professional face matchers of how many mismatches they should expect when conducting their jobs could help to change how they make decisions, but only if a large proportion of mismatches are expected. Also, findings suggest that awareness (or specifically, the lack thereof) can affect decision-maker's actual abilities to

discriminate between matches and mismatches. Future research should examine whether the unexpected results seen here are artifacts of the warning used or if there is something about low prevalence awareness that effects participants' accuracy counterintuitively.

Chapter 4. Expression Study

4.1 Background & Significance

Photo identifications and passports in the United States typically require the photo subject to have a neutral expression. However, neutral expressions may not be the best representation of what a person looks like in a day-to-day situation. Originally, I intended to examine expression in the meta-analysis, but expression was not explored enough in previous research to have enough studies to conduct a comparison in the meta-analysis. Therefore, I wanted to explore the effect of expression on face matching accuracy.

Bruce et al. (1999) asked participants to match a neutral probe photo or a smiling probe photo to full simultaneous lineups. All participants in the lineup had neutral expressions. Results showed no difference in accuracy between neutral and smiling photos. On the other hand, Mileva and Burton (2018) manipulated whether the face pairs depicted neutral faces, open-mouth smiling faces, or closed-mouth smiling faces. Face matching accuracy was higher when photos were of smiling (vs. neutral) faces. Open-mouth smiles increased accuracy more than closed-mouth smiles, suggesting that open-mouth smiles change the face in an idiosyncratic way that aids in face matching decisions. Results similar to Mileva and Burton (2018) have been seen in facial identification research (Lander, Christie, & Bruce, 1999; Lander, Chuang, & Wickham, 2006). Chen et al. (2011) examined how facial expressions affected response times for sequential facial identification and found that happy, sad, and neutral faces were matched more quickly than fearful, angry, and disgusted faces. Studies in facial recognition have shown that certain emotion

stimuli are more efficiently encoded, consolidated, and retrieved than neutral stimuli – specifically, those related to happiness (Kensiger, 2004; La Bar & Cabeza, 2006; Righi et al., 2012). Chen et al. (2015) found that happy faces were more accurately recognized than angry faces. Additionally, emotional expression affected both encoding and retrieval processes in identity recognition.

Identification cards are usually required to have a neutral, non-expressive photo while the person presenting that identification can present with any expression. Previous research (e.g., Mileva & Burton, 2018) has examined matched emotional pairs in face recognition and face matching but has yet to examine mismatched expressions in face matching. Therefore, the current study uses mismatched expression face pairs to create a more ecologically valid situation for face matching decision-makers.

The current study examines whether different types of expressions aid or hinder face matching accuracy. Pairs in the current study will use mismatched expressions: a neutral face paired with an emotional expression. The expressions I intended to test in the study were neutral, happiness, sadness, fear, disgust, and anger. These expressions were chosen because they are considered universal emotions (Ekman, 1999). However, based on the results of the pilot study and the number of believable mismatched face pairs that could be created, the included expressions had to be reduced to neutral, happiness, and anger.

4.1.1 Hypotheses

I hypothesize that happiness will result in the highest accuracy overall (e.g., Kensiger, 2004). Due to the limited previous research and exploratory nature of the included expressions, I make no specific predictions about the other types of expressions.

4.2 Pilot Study

A pilot study was conducted to determine how many believable mismatch facial pairs could be made using the FACES database (Ebner et al., 2010) based on similarity ratings. FACES is a set of color images of 171 women and men of varying ages displaying six facial expressions. All ages and both men and women were used as stimuli because my meta-analysis revealed that age and gender did not cause a different in accuracy rates. The similarity ratings from the pilot test dictated what photos were used for stimuli for the full study as well as how many different expressions we were able to test.

4.2.1 Participants

49 participants were recruited from Ontario Tech University through SONA. Participants were compensated with 0.5 university research credits for participation. No demographic information was collected for the pilot study.

4.2.2 Materials

The FACES database consists of two sets of photos per person per facial expression. Photos are separated by age with young, middle-aged, and older categories. The database consists of photos all taken on the same day and shows the hairstyle of all the subjects. For the pilot test, all photo pairs were shown only with neutral expressions in order to get the purest similarity ratings. Mismatches used in the pilot study were created based on my judgements of which faces looked most similar using the age, gender, eye color, hair color, and hairstyle of the subjects in the photos. The final set of stimuli consisted of 50 mismatches.

4.2.3 Procedure

Participants were recruited through the Ontario Tech participant pool using SONA. All participation took place online through a survey hosted on Qualtrics. Participants were first shown the instructions for the survey. The instructions were as follows: “In the next section you

will be shown pairs of faces and asked how similar you think the faces look to each other. You will rate the similarity on a scale from 1 to 5 with 1 = not at all similar and 5 = extremely similar. Please take your time and try to answer as accurately as possible." After the instructions, participants saw the face pairs in a random order. Each pair was presented with the question "How similar are the two faces shown?" On average, participant took 8 minutes.

4.2.4 Pilot Study Results

The average similarity rating for facial pairs was 2.14 (*SD* = 0.42). Mismatches were chosen to be used in the full study if they were judged to be at least 50% similar; if a facial pair had an average similarity rating of 2.5 or higher it was included. The 50% criterion was chosen to ensure that the mismatches were at least somewhat believable as potential matches. Ultimately, only 12 of the face matching pairs achieved at least a 50% similarity rating with an average rating of *M* = 2.72, *SD* = 0.13. The 12 face pairs that achieved this cut-off were 50% female face pairs. Originally, I intended to examine six expressions, but with only 12 believable mismatched face pairs that would mean there would only be 2 mismatch pairs per expression without repeating faces in the stimulus set. Thus, the full study was modified to contain only three expression categories: neutrality, happiness, and anger. These expressions were chosen because previous research suggests that the largest differences in accuracy will exist between these expressions (e.g., Chen et al., 2011; Mileva & Burton, 2018). I also thought that it would be more likely that a person would present at a security checkpoint with an angry expression than an expression of disgust or fear and was the most logical choice for an expression that may result in lower accuracy based on Chen et al.'s (2011) results. Each expression category contained four mismatched face pairs.

4.3 Full Study Method

A within-subjects 2 (trial: match vs. mismatch) x 3 (expression: neutrality, happiness, anger) design was used for this study.

4.3.1. Participants

45 Ontario Tech University students were recruited for this study. According to a G*Power 3.1 sensitivity analysis, with $N = 45$, $\alpha = 0.05$, and 80% power, the smallest effect size that can be reliably detected in this study is $\eta_p^2 = 0.045$, *Cohen's f* = 0.21. Participants were compensated with 0.5 university research credits for their participation. No participants were removed from the sample. Ages of participants ranged from 18 to 48 with an average age of 22.24. Seventy-three percent of the participants identified as female. Almost 50% of the participants identified as white and approximately 30% of the sample identified as South Asian.

4.3.2 Materials

The final set of stimuli consisted of 12 matches and 12 mismatches, with four mismatches for each of the three expressions. Each expression condition was randomly allotted 4 mismatched face pairs from the pairs that achieved 50% similarity in the pilot study. Matches were chosen randomly from the remaining FACES database. All photos were taken on the same day and show the hairstyle of all the subjects. Each expression set contained four female face pairs and four male face pairs, with two mismatched face pairs for each sex. Photo pairs consist of one expressive face compared to one neutral face.

Figure 4.1

Example of stimuli for each expression condition

a) Example of match stimuli in the anger condition

b) Example of a mismatch in the happiness condition

c) Example of match stimuli in the neutral condition

Figure 4. 1. Example of stimuli for each expression condition

4.3.3 Procedure

Participants were recruited using the Ontario Tech participant pool using SONA. All participation took place online through a survey hosted on Qualtrics. Participants first saw the instructions for the survey. The instructions reminded the participants that features may change over time and encouraged them to be as quick and accurate as possible when answering questions. After the instructions, participants saw the face pairs in a random order. Each pair was presented with the question "Do the two photos show a match or a mismatch?" All participants completed all 24 trials in a randomized order. Participants were also randomly exposed to an attention check randomly during the face-matching trials. The attention check was a simple math

equation (2+2 =?) to ensure that participants were reading the questions. After completing all trials participants answered a few demographics questions and were then thanked and debriefed.

4.4 Results

All participants passed the attention check and no participants needed to be removed from the sample.

A within-subjects 2 (trial: match vs. mismatch) x 3 (expression: neutrality, happiness, anger) ANOVA was used for this study with accuracy as the dependent variable. All presented within-subjects results employ sphericity assumed *F*-tests and all presented post-hoc comparisons use the Bonferroni correction provided in SPSS. Results revealed a marginally significant main effect of trial type, $F(1) = 3.342$, $p = 0.074$, $\eta_p^2 = 0.074$. Participants were more accurate on match trials ($M = 0.909$, $SD = 0.030$) than on mismatch trials ($M = 0.841$, $SD = 0.022$). A main effect of expression was also found, $F(2) = 5.476$, $p = 0.006$, $\eta_p^2 = 0.11$, such that participants were significantly more accurate on neutral trials ($M = 0.917$, $SD = 0.021$) than on anger trials ($M = 0.858$, $SD = 0.023$, $p = 0.016$) and happiness trials ($M = 0.850$, $SD = 0.23$, $p = 0.005$). There was no significant difference between accuracy on happiness and anger trials, $p = 0.673$. No interaction was found between trial type and expression, $F(2) = 1.617$, $p = 0.204$, $\eta_p^2 = 0.35$.

Figure 4.1

Accuracy rates by expression

Figure 4. 2 Accuracy rates by expression type

Performance in all conditions was significantly better than chance. See Table 4.1 for a summary of one-sample *t*-test results.

Table 4.1

Summary of one-sample t-test results

					95% Confidence Interval of the Difference	
	t	df	*p*	Mean Difference	Lower	Upper
Anger Match	-1117.875	44	<.001*	-49.13333	-49.2219	-49.0448
Anger Mismatch	-1832.098	44	<.001*	-49.15000	-49.2041	-49.0959
Happiness Match	-1155.590	44	<.001*	-49.12222	-49.2079	-49.0366
Happiness Mismatch	-1518.105	44	<.001*	-49.17778	-49.2431	-49.1125
Neutral Match	-5213.836	44	<.001*	-49.01667	-49.0356	-48.9977
Neutral Mismatch	-1181.751	44	<.001*	-49.15000	-49.2338	-49.0662

Note: * indicates a significant difference at the $p < 0.05$ level

Table 4. 1. Summary of one-sample t-test results.

4.4.1 Signal Detection Analyses

As in Study 2, I conducted a signal detection analysis on the data. The same formulas as described above were used to calculate *d'* and *β*.

A repeated measures ANOVA, with expression as the within-participants variable, was conducted on *d'* scores. Results found a main effect for *d'*, $F(2) = 5.689$, $p = 0.005$ $\eta_p^2 = 0.127$. Discrimination was highest in the neutral condition ($M = 2.090$) and significantly different from discrimination in the anger condition ($M = 1.866$, $p = 0.021$) and happiness condition ($M = 1.789$, $p = 0.002$). No significant difference was seen between the anger and happiness conditions ($p = 0.425$).

I also conducted a repeated measures ANOVA on β. There was no significant effect of expression, $F(2) = 0.260$, $p = 0.772$, $\eta_p^2 = 0.01$.

Signal detection measures were compared to chance using one-sample *t*-test. All *d'* values were significantly different from chance. Neutral and happiness conditions showed biased responding significantly different from chance while the anger condition did not (See Table 4.2). Both neutral and happiness conditions revealed a match bias.

Table 4.2

Summary of one-sample t-tests for d' and β.

SDT statistic	Expression	*t*	*df*	*p*
d'	Happiness	19.02	41	<0.001**
	Neutral	35.42	41	<0.001**
	Anger	24.66	40	<0.001**
β	Happiness	-2.67	41	0.01*
	Neutral	-2.037	41	0.05*
	Anger	-1.75	40	0.089

Note: test statistic for $\beta = 1$; test statistic for $d' = 0$; *denotes a significant difference at the $p = 0.05$ level** denotes a significant difference at $p < 0.001$

Table 4. 2. Summary of one-sample t-tests for d' and β

4.5 Discussion

Participants were most accurate, and able to discriminate between matches and mismatches, on trials of neutral pairs. There were no differences between happiness and anger trials in terms of performance. These results did not support my hypothesis that happiness would result in the highest accuracy. Happiness resulting in lower accuracy is contrary to the results seen in previous research (e.g., Lander et al., 1999; Mileva & Burton, 2018). Mileva and Burton (2018) found that smiles resulted in higher accuracy than neutral trial pairs; however, the stimuli they used showed the targets smiling in both photos in the pair. In the current study, the stimuli showed only one happy photo and one neutral photo to better approximate real-world conditions. It may be that smiling reveals helpful idiosyncrasies in the face when comparing the face to another smiling photo or person, but those same idiosyncrasies become a hinderance when comparing a smiling photo or person to a neutral photo.

In all expression conditions, sensitivity was significantly higher than chance suggesting that participants were able to discriminate between matches and mismatches reliably. Sensitivity was highest in the neutral condition with no difference between the happiness and anger conditions – matching what was seen in the accuracy analysis. No significant differences were seen between the conditions in terms of response bias (β). However, both the neutral and happiness conditions showed a pattern of biased responding towards matches that was significantly different from chance.

Participants in the neutral condition were the only participants to see matched expressions. That confound might account for the expression effect. While the mismatch expressions may have resulted in results contrary from previous research, the current study was modeled on the confounded conditions that exist in the real world; examining this confound was

one of the goals of the study. In the U.S., most IDs are required to have a neutral photo, but sometimes the person presenting the ID will have a neutral expression and sometimes they will have a different expression. My results suggest that the mismatch in expressions between a person and their ID seen in real-world settings may negatively affect face matching accuracy. Therefore, maybe people should have to adopt a neutral expression when they show ID in order to prevent a reduction in accuracy. Further studies could help to unpack the confound and examine whether it was the use of a neutral target expression or the similarity in expressions that produced the neutral face advantage seen in this study.

The stimuli used also only had to pass a 50% similarity threshold, likely making the mismatch pairs artificially easy resulting in increased accuracy across conditions. A more difficult stimulus set would likely lead to a similar pattern of results but with lower levels of accuracy. Face matching in the real-world is likely harder than the task used here as most people using fake IDs would try to use an ID with a photo that looks as similar to them as possible. Future research in this area should try to use a more ecologically valid and difficult stimuli set.

The study presented here had to reduce the types of expressions used in the stimuli from six to three due to the available believable mismatch stimuli. Consequently, future research should explore more types of expressions. For example, fear has been shown to reduce facial recognition (e.g., Righi et al., 2012). Fear, sadness, and other emotions that a person may present within a face matching situation should be examined. As face matching becomes more common in legal and criminal spheres, the types of facial expressions depicted in probe photos and face matching stimuli is less likely to be the neutral stimuli that is used in laboratory research, so research should make attempts to examine different, emotional stimuli to increase the applicability of results to real-world situations.

Performance in this study was significantly higher than chance in all conditions. Pairs with neutral expressions lead to a slightly better performance. However, this study used stimuli that was likely artificially easy; additional research needs to be done to correct this limitation. Security decisions often rely on face matching decisions, so it is crucial that researchers continue this line of research to evaluate how much expressions effect face matching accuracy and ways to mediate their effect.

Chapter 5. General Discussion

The goals of this dissertation were to quantify the effect sizes for variables related to face matching decision making, to further investigate variables important to face matching decision making, to examine unexpected results found in the process, and to advance knowledge concerning face matching and improving face matching accuracy. To this end, three studies were conducted that each sought to clarify important features of the face matching process. Study 1 accomplished this through a meta-analysis that provided a high-level summary of the work to date. Studies 2 and 3 followed up on identified gaps in the meta, and targeted specific under-investigated features of the face matching literature.

Face matching research has been conducted for over 30 years, but only qualitative reviews have been conducted thus far. Face matching is employed as a means of security and is increasingly being used by law enforcement as a means of identifying suspects. Previous research has identified different variables that effect face matching accuracy such as photo quality, base rates of mismatch, expertise, view variation, cross-race matching, training, and feedback (for a review see Bindemann, 2021). However, research has yet to provide empirical estimates of average effect sizes of these moderating variables.

Therefore, a meta-analysis was conducted to quantify effect sizes of variables important to face matching decision making. Results found that face matching accuracy is affected by the view of the target, feedback provided to the decision makers, expertise, the amount of comparison photos, and time pressure. Surprisingly, base rates of mismatch were not shown to affect face matching accuracy. Base rates of mismatch are a widely examined aspect of face matching and the LPE that results from low base rates of mismatch is thought to be a reliable and potent source of increased error in face matching decision making. Recent research suggests that awareness of mismatch prevalence may be necessary for the LPE to occur (Davis et al., 2021). Therefore, I conducted an experiment testing that suggestion.

Base rates of mismatch (prevalence) were manipulated along with participants' awareness of what prevalence condition they were in. Results found a lack of a pure low prevalence effect as participants were most accurate in the low prevalence conditions. Overall, awareness did help participants to be more accurate and altered participants' response bias. Thus, participants are able to change how they make decisions based on the information that they learn about the stimuli set. The results of this study provide important information for face matching decision makers and the organizations that employ face matching decision makers: it is possible to change response bias and sensitivity by providing a rather small amount of information about what decision-makers should expect. Future research should continue to investigate these changes in bias and sensitivity to determine their longevity and what information is necessary to increase face matching accuracy across all conditions.

Finally, the third study examined expressions in face matching. I hypothesized that participants would be most accurate when matching happy photos and least accurate when matching angry photos. My hypothesis was not supported. Participants were most accurate when

matching neutral faces and showed no differences in accuracy when matching happy or angry faces. These results suggest that errors may increase when the person presenting the document does not also present with a neutral face.

Use of signal detection theory in the face matching research is emerging. Previous research has reported signal detection measures (e.g., Davis et al., 2021), but have chosen to focus on the typical outcome of accuracy. Signal detection theory provides details about decision-making; for example, in Study 2, SDT allowed me to uncover changes in sensitivity that were not captured in the accuracy analysis. So, by focusing on signal detection analyses, my work helps to move the face matching research forward by not only continuing the use of SDT, but also in terms of making new discoveries. My results revealed that face matching response biases can be changed but that it does not always happen uniformly or in expected ways.

There are some overall limitations to the experimental studies conducted in this dissertation that make the face matching decisions made in these studies easier than face matching in the field. First is the use of optimized stimuli that are not representative of the stimuli seen by real face matching decision makers. Especially as face matching is used more frequently to identify suspects from CCTV and other surveillance cameras it becomes less likely that two still photos will be used as face matching stimuli. Even when still photos are taken from video stimuli from CCTV and surveillance cameras, there is the issue of low-quality stimuli. Any type of pixelation (e.g., Bindemann et al., 2013) or changes in view (e.g., Bruce et al., 1999) reduces face matching accuracy. Secondly, participants are engaged in the task of face matching for a relatively short time – around 15 minutes or less, on average. Alenezi et al. (2015) found that face matching accuracy decreased over time. Real face matching decision-makers are likely involved in the task of face matching for longer than 15 minutes at a time, suggesting that

accuracy in real-world settings would be lower than the accuracy seen in the experimental studies I conducted. Finally, there is an artificially short delay between the pictures being taken in the stimuli we used. The U.S. Real ID requirement allows for a photo to be used for up to eight years (Department of Homeland Security, 2021). So, there can be years between the when the photo was taken and when the person presents the ID. The person's looks can change a lot with time, which can be compounded by changes in hair color, hair styles, glasses, and facial hair (e.g., Buolamwini et al., 2020). Therefore, the results of my studies likely show higher accuracy rates than would be seen in real-world face matching tasks.

Face matching performance in these studies was relatively robust: 84% for the Study 2, and 87.5% for the Study 3. This is further bolstered by sensitivity being greater than chance in all conditions for both studies. Discriminability between matches and mismatches seems to be relatively easy to achieve, but performance is complicated by biases. Biased responding was present in both studies, suggesting that it is rather easy to induce biases in face matching decision-makers. In Study 2, statistically significant mismatch biases were seen in the unaware 10% and the aware 90% condition. The mismatch bias makes sense and is a correct adjustment to the information given in the 90% condition. However, in the unaware 10% condition, the presence of a mismatch bias does not follow from the information given to participants. Logically, participants should exhibit a match bias in the 10% condition. It may be that participants expect mismatches to be present and become wary when they keep seeing matches, so they start second-guessing the pairs they believe to be matches. Study 3 did not provide participants with any information that should influence their pattern of responding, yet a match bias emerged in both the neutral and happiness conditions. Most of the biased responding in Studies 2 and 3 do not logically follow from the information and stimuli presented to

participants. This suggests that biases in face matching decision-making may be influenced by extraneous variables outside of the experiment and, maybe, the participants' own expectations of the stimuli.

In Study 2, there was no evidence for a low prevalence effect in face matching accuracy. However, discriminability was highest in the unaware 50% condition, suggesting that any departure from an equal ratio of matches and mismatch affects discriminability. This finding has implications for base rates face matching research as a whole. Most of the previous base rates research has used a 50-50 split as a comparison for experimental manipulations, but a 50% mismatch rate may artificially inflate face matching accuracy to levels not really seen in real-world conditions. Therefore, future research should further investigate this finding to see if it is replicable or if it is an artifact of our study.

I also found that neutral pairs led to the highest accuracy and discriminability when compared to mismatched expression pairs (Study 3). This result falls in line with previous research suggesting that any changes between the ID and the person presenting the ID results in reduced accuracy (e.g., Fysh & Bindemann, 2017). One way to increase face matching accuracy, then, is to include multiple photos on IDs (e.g., Study 1; White et al., 2014). Multiple comparison photos provide different views of the person at different times and in different situations which would increase the face matching decision-maker's ability to recognize the person. Future research should look at the intersection between using multiple photos and paraphernalia changes (e.g., hair style and color, jewelry, etc.) to establish the optimal number of photos on an ID that reduces face matching errors.

Overall, this group of studies shows that face matching accuracy can be relatively high with optimal stimuli but can be easily reduced based on variables related to the stimuli used (e.g.,

view), the stimulus set (e.g., expressions), information provided about the stimulus set (e.g., base rates), or the participant themselves (e.g., expertise, expectations). This reduction in accuracy can be driven by either reductions in sensitivity or biased responding and that participants' biased responding does not always follow the expected direction. Thus, it seems unwise to rely solely on face matching decision-makers in situations where the public's security is at stake.

Chapter 6. Reference

Alenezi, H., & Bindemann, M. (2013). The effect of feedback on face matching accuracy. *Applied Cognitive Psychology, 27*(6), 735-753. https://doi.org/10.1002/acp.2968

Alenezi, H. M., Bindemann, M., Fysh, M. C., & Johnston, R. A. (2015). Face matching in a long task: Enforced rest and desk-switching cannot maintain identification accuracy. *PeerJ*, *3*, 1-18. doi: 10.7717/peerj.1184

Baumeister, R. F., Bratslavsky, E., Muraven, M., & Tice, D. M. (1998). Ego depletion: Is the active self a limited resource? *Journal of Personality and Social Psychology, 74*, 1252-1265.

Bindemann, M. (Ed.). (2021). *Forensic face matching.* Oxford University Press.

Bindemann, M., Attard, J., Leach, A., & Johnston, R. A. (2013). The effect of image pixilation on unfamiliar face matching. *Applied Cognitive Psychology, 27*(6), 707-717. https://doi.org/10.1002/acp.2970

Bindemann, M., Avetisyan, M., & Blackwell, K. (2010). Finding needles in haystacks: Identity mismatch frequency and facial identification verification. *Journal of Experimental Psychology: Applied, 16*(4), 378-386. doi: 10.1037/a0021893

Bindemann, M., Fysh, M. C., Cross, K., & Watts, R. (2016). Matching faces against the clock. *i-Perception, 6*, 1-18. doi: 10.1177/2041669S16672219

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

Bobak, A. K., Hancock, P. J. B., & Bate, S. (2015). Super-recognisers in action: Evidence from face-matching and face memory tasks. *Applied Cognitive Psychology, 30*, 81-91.

Bobak, A. K., Mileva, V. R., & Hancock, P. J. B. (2019). A grey area: How does image hue affect unfamiliar face matching? *Cognitive Research: Principles and Implications, 4*(27), 2-10. https://doi.org/10.1186/s41235-019-0174-3

Bornstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. (2009). Introduction to meta-analysis. Chichester: John Wiley & Sons.

Bruce, V. (1982). Changing faces: Visual and non-visual coding processes in face recognition. *British Journal of Psychology, 73*, 105-116.

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*(4), 339-360.

Buolamwini, J., Ordonez, V., Morgenstern, J., & Learned-Miller, E. (2020). Facial recognition technologies: A primer. *The Algorithmic Justice League*. https://global-uploads.webflow.com/5e027ca188c99e3515b404b7/5ed1002058516c11edc66a14_FRTs PrimerMay2020.pdf

Burton, A. M. (2013). Why has research in face recognition progressed so slowly? The importance of variability. *Quarterly Journal of Experimental Psychology, 66*(8), 1467-1485. https://doi.org/10.1080/17470218.2013.800125.

Burton, A. M., Kramer, R. S.S., Ritchie, K. L., & Jenkins, R. (2016). Identity from variation: Representations of faces derived from multiple instances. *Cognitive Science, 40*, 202-223. https://doi.org/10.1111/cogs.12231.

Chen, W., Lander, K., & Liu, C. H. (2011). Matching faces with emotional expression. *Frontiers in Psychology, 2*(206), 1 – 10.

Chen, W., Liu, C. H., Li, H., Tong, K., Ren, N., & Fu, X. (2015). Facial expression at retrieval affects recognition of facial identity. *Frontiers in Psychology, 6*(780), 1- 9.

Clutterbuck, R., & Johnston, R. A. (2005). Demonstrating how unfamiliar faces become familiar using a face matching task. *European Journal of Cognitive Psychology, 17*, 97-116.

Daubert v. Merrell Dow Pharmaceuticals Inc., 509 U.S. 579 (1993).

Davis, J. P., Dray, C., Petroc, N., & Belanova, E. (2021). Low prevalence match and mismatch detection in simultaneous face matching: Influence of face recognition ability and feature focus guidance. *Attention, Perception, & Psychophysics, 83*(7), 2937-2954.

Davis, J. P., & Valentine, T. (2009). CCTV on trial: Matching video images with the defendant in the dock. *Applied Cognitive Psychology, 23*, 482-505. doi: 10.1002/acp.1490

Department of Homeland Security (2021). *DHS announces extension of REAL ID full enforcement deadline.* DHS.gov. https://www.dhs.gov/real-id/news/2021/04/27/dhs-announces-extension-real-id-full-enforcement-deadline

Dowsett, A. J., & Burton, A. M. (2015). Unfamiliar face matching: Pairs out-perform individuals and provide a route to training. *British Journal of Psychology, 106*(#), 433-445.

Edmonds, A. J., & Lewis, M. B. (2007). The effect of rotation on configural encoding in a face-matching task. *Perception, 36*, 446-460. doi: 10.1068/p5530

Ekman, P. (1999). Basic emotions. In T. Dalgleish and T. Power (Eds.). *The handbook of cognition and emotion* (pp. 45-60). New York, NY: John Wiley & Sons.

Favelle, S., Hill, H., & Claes, P. (2017). About face: Matching unfamiliar faces across rotations of view and lighting. *i-Perception, 8*(6), 1-20. doi: 10.117/204166951774421

Farah, M. J., Wilson, K. D., Drain, M., & Tanaka, J. N. (1998). What is "special" about face perception? *Psychological Review, 105*(3), 482-498.

Faul, F., Erdfelder, E., Lang, A., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*(2), 175-191.

Fysh, M. C., & Bindemann, M. (2017). Effects of time pressure and time passage on face-matching accuracy. *Royal Society Open Science, 4*, 1-13. doi: 10.1098/rsos.170249

Fysh, M. C., & Bindemann, M. (2017). Forensic face matching: A review. In M. Bindemann & A. M. Megreya (Eds.), *Face processing: Systems, disorders and cultural difference*s (pp. 1-20). Nova Science Publishing, Inc.

Garvie, C., Bedoya, A. M., & Frankle, J. (2016). The perpetual lineup: Unregulated police face recognition in America. *Georgetown Law Center on Privacy & Technology.* www.perpetuallineup.org.

Gentry, N. W., & Bindemann, M. (2019). Examples improve facial identity comparison. *Journal of Applied Research in Memory and Cognition, 8*, 376-385.

Hill, K. (2020a). Wrongfully accused by an algorithm. *The New York Times.* https://www.nytimes.com/2020/06/24/technology/facial-recognition-arrest.html.

Hill, K. (2020b). Another arrest, and jail time, due to a bad facial recognition match. *The New York Times.* https://www.nytimes.com/2020/12/29/technology/facial-recognition-misidentify-jail.html

IBM. (2020). The calculation of Bonferroni-adjusted p-values. *IBM.* https://www.ibm.com/support/pages/calculation-bonferroni-adjusted-p-values

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review, 80*(4), 237-251. doi: 10.1037/h0034747

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), 622-627.

Kensinger, E. A. (2004). Remembering emotional experiences: The contribution of valence and arousal. *Reviews in Neuroscience, 15*, 241-251.

Kokje, E., Bindemann, M., & Megreya, A. M. (2018). Cross-race correlations in the abilities to match unfamiliar faces. *Acta Psychologica, 185*, 13-21.

Kornfield, M. (2021). The wrong ID: Retired firefighter, comedian, and Chuck Norris falsely accused of being Capitol rioters. *The Washington Post.* https://www.washingtonpost.com/technology/2021/01/16/sleuths-falsely-identify-rioters/

Kramer, R. S. S., & Reynolds, M. G. (2018). Unfamiliar face matching with frontal and profile views. *Perception, 47*(4), 414-431. doi: 10.1177/0301006618756809

Kramer, R. S., & Ritchie, K. L. (2016). Disguising Superman: How glasses affect unfamiliar face matching. *Applied Cognitive Psychology, 30*(6), 841-845.

LaBar, K. S., & Cabeza, R. (2006). Cognitive neuroscience of emotional memory. *National Review of Neuroscience, 7*, 54-64.

Lander, K., Christie, F., & Bruce, V. (1999). The role of movement in the recognition of famous faces. *Memory & Cognition, 27*, 974-985.

Lander, K., Lewis, C., & Wickham, L. (2006). Recognizing face identity from natural and morphed smiles. *Quarterly Journal of Experimental Psychology, 59*(5), 801-808.

Longmore, C. A., Liu, C. H., Young, A. W. (2008). Learning faces from photographs. *Journal of Experimental Psychology: Human Perception and Performance, 34*, 77-100.

Los Angeles World Airports (2019). Traffic comparison (TCOM) Los Angeles International Airport. https://www.lawa.org/-/media/bb9842a7ecec43efa519a84ff7455ad7.pdf

MacMillan, N. A. (1993). Signal detection theory as data analysis method and psychological decision model. In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences: Methodological issues* (pp. 21-57). Hillsdale, NJ: Erlbaum.

MacMillan, N. A., & Kaplan, H. L. (1985). Detection theory analysis of group data: Estimating sensitivity from average hit and false-alarm rates. *Psychological Bulletin, 98*, 185-199.

McCaffery, J. M., & Burton, A. M. (2016). Passport checks: Interactions between matching faces and biographical details. *Applied Cognitive Psychology, 30*, 925-933. doi: 10.1002/acp.3281

McKone, E., Kanwisher, N., & Duchaine, B. C. (2007). Can generic expertise explain special processing for faces? *Trends in Cognitive Sciences, 11*, 8-15.

McIntyre, A. H., Hancock, P. J. B., Kittler, J., & Langton, S. R. H. (2013). Improving discrimination and face matching with caricature. *Applied Cognitive Psychology, 27*(6), 725-734.

Megreya, A. M., & Bindemann, M. (2015). Developmental improvement and age-related decline in unfamiliar face matching. *Perception, 44*, 5-22.

Megreya A. M., & Bindemann, M. (2018). Feature instructions improve face-matching accuracy. *PLoS ONE, 13*(3), 1-16. https://doi.org/10.1371/journal.pone.0193455.

Megreya, A. M., & Burton, A. M. (2007). Hits and false positives in face matching: A familiarity-based dissociation. *Perception & Psychophysics, 69*(7), 1175-1184.

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. doi: 10.1037/a0013464

Megreya, A. M., Bindemann, M., & Havard, C. (2011). Sex differences in unfamiliar face identification: Evidence from matching tasks. *Acta Psychologica, 137*, 83-89.

Megreya, A. M., Sandford, A., & Burton, A. M. (2013). Matching face images taken on the same day or months apart: The limitations of photo ID. *Applied Cognitive Psychology. 27*, 700-706. doi: 10.1002/acp.2965.

Megreya, A. M., White, D., & Burton, A. M. (2011). The other-race effect does not rely on memory: Evidence from a matching task. *The Quarterly Journal of Experimental Psychology, 64*, 1473-1483.

Meissner, C. A., & Brigham, J. C. (2001). Thirty years of investigating the own-race bias in memory for faces: A meta-analytic review. *Psychology, Public Policy, & Law, 7*, 3 – 35.

Meissner, C. A., Brigham, J. C., & Butz, D. A. (2005). Memory for own-and other-race faces: A dual process approach. *Applied Cognitive Psychology, 19*(5), 545-567.

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.

Moore, R. M., & Johsnton, R. A. (2013). Motivational incentives improve unfamiliar face matching accuracy. *Applied Cognitive Psychology, 27*, 754-760. doi: 10.1002/acp.2964.

Muraven, M. R., & Baumeister, R. F. (2000). Self-regulation and deletion of limited resources: Does self-control resemble a muscle? *Psychological Bulletin, 126*, 247-259.

Osborne, C. D., & Stevenage, S. V. (2008). Internal feature saliency as a marker of familiarity and configural processing. *Visual Cognition, 16*, 23-43.

Özbek, M., & Bindemann, M. (2011). Exploring the time course of face matching: Temporal constraints impair unfamiliar face identification under temporally unconstrained viewing. *Vision Research, 51*, 2145-2155.

Papesh, M. H. (2018). Photo ID verification remains challenging despite years of practice. *Cognitive Research: Principles and Implications, 3*(19), 1-9. https://doi.org/10.1186/s41235-018-01-10-y

Papesh, M. H., & Goldinger, S. D. (2014). Infrequent identity mismatches are frequently undetected. *Attention, Perception, & Psychophysics, 76*95), 1335-1349. doi: 10.3758/s13414-014-0630-6.

Papesh, M. H., Heisick, L. L., & Warner, K. A. (2018). The persistent low-prevalence effect in unfamiliar face-matching: The roles of feedback and criterion shifting. *Journal of Experimental Psychology: Applied, 24*(3), 416-430. http://dx.doi.org/10.1037/xap0000156.

Righi, S., Marzi, T., Toscani, M., Baldassi, S., Ottonello, S., & Viggiano, M. P. (2012). Fearful expressions enhance recognition memory: Electrophysiological evidence. *Acta Psychology, 139*, 7-18.

Ritchie, K. L., Mireku, M. O., & Kramer, R. S. S. (2020). Face averages and multiple images in a live matching task. *British Journal of Psychology, 111*, 92-102.

Robbins, R., & McKone, E. (2003). Can holistic processing be learned for inverted faces? *Cognition, 88*, 79-107.

Rossion, B. (2008). Picture-plane inversion leads to qualitative changes of face perception. *Acta Psychologica, 128*(2), 274-289.

Sato, T., Miyazaki, M., & Watanabe, T. (2011). Visual search skills in task of spot difference. *Bio Web of Conferences, 1*, 1-4.

Shapiro, P. N., & Penrod, S. (1986). Meta-analysis of facial identification studies. *Psychological Bulletin, 100*(2), 139-156.

Stanislaw, H., & Todorov, N. (1999). Calculation of signal detection theory measures. *Behavior Research Methods, Instruments, & Computers, 31*(1), 137-149.

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.

Sullivan, B. (2004). 9/11 report light on ID theft issues. *NBC.* http://nbcnews.com/id/wbna5594385

Susa, K. J., Michael, S. W., Dessenberger, S. J., & Meissner, C. A. (2019). Imposter identification in low prevalence environments. *Legal and Criminological Psychology, 24*, 179-193. doi: 10.1111/lcrp.12138

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), 1-17. https://doi.org/10.137/journal.pone.0211037

Towler, A., White, D., & Kemp, R. I. (2017). Evaluating the feature comparison strategy for forensic face identification. *Journal of Experimental Psychology: Applied, 23*, 47-58. https://doi.org/10.1037/xap0000108.

Towler, A., White, D., & Kemp, R. I. (2014). Evaluating training methods for facial image comparison: The face shape strategy does not work. *Perception, 43*, 214-218. doi: 10.1068/p7676

Tsao, D., Y., & Livingstone, M. S. (2008). Mechanisms of face perception. *Annual Review of Neuroscience, 31*, 411-437.

Tversky, A. & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology, 5*(2), 207-232.

Weatherford, D. R., Erickson, W. B., Thomas, J., Walker, M. E., & Schein, B. (2020). You shall not pass: How facial variability and feedback affect the detection of low-prevalence fake IDs. *Cognitive Research: Principles and Implications, 5*(3), 1-15. https://doi.org/10.1186/s41235-019-0204-1.

White, D., & Kemp, R. (2020). Identifying people from images. In N. Brewer & A. Bradfield Douglas (Eds.), *Improving the criminal justice system: Perspectives from psychological science,* Guilford Publications.

White, M., & Li, J. (2006). Matching faces and expression in pixelated and blurred photos. *The American Journal of Psychology, 119*, 21-28.

White, D., Burton, A. M., Jenkins, R., & Kemp, R. I. (2014). Redesigning photo-ID to improve unfamiliar face matching performance. *Journal of Experimental Psychology: Applied, 20*(2), 166-173. doi: 10.1037/xap00000009

White, D., Burton, A. M., Kemp, R. I., & Jenkins, R. (2013). Crowd effects in unfamiliar face matching. *Applied Cognitive Psychology, 27*, 769-777. doi: 10.1002/acp.2971.

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

White, D., Kemp, R. I., Jenkins, R., Matheson, M., & Burton, A. M. (2014). Passport offices' errors in face matching. *PLoS ONE, 9*(8), 1-6. doi: 10.137/journal.pone.0103510.

White, D., Phillips, P. J., Hahn, C. A., Hill, M., & O'Toole, A. J. (2015). Perceptual expertise in forensic facial image comparison. *Proceedings of the Royal Society B: Biological Sciences, 282*(1814), 1-8. http://dx.doi.org/10.1098/rspb.2015.1292.

White, D., Towler, A., & Kemp, R. (2021). Understanding professional expertise in unfamiliar face matching. *Forensic Face Matching: Research and Practice*, 62.

Wirth B. E., & Carbon, C. (2017). An easy game for frauds? Effects of professional experience and time pressure on passport-matching performance. *Journal Experimental Psychology: Applied, 23*(2), 138-157. http://dx.doi.org/10.1037/xap0000114.

Young, A. W., McWeeny, K. H., Hay, D. C., & Ellis, A. W. (1986). Matching familiar and unfamiliar faces on identity and expression. *Psychological Research, 48*, 63-68.

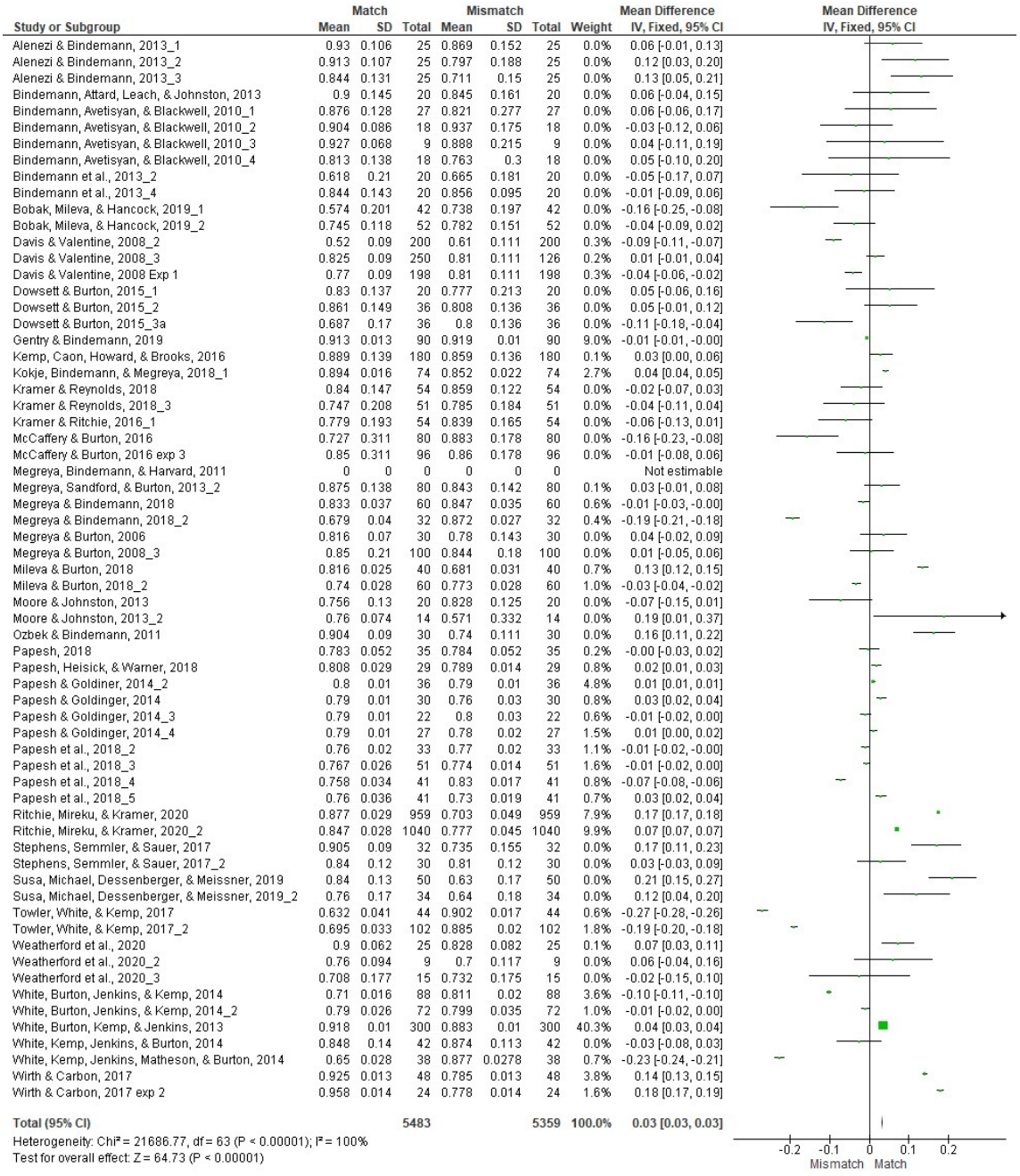
Yovel, G., & Kanwisher, N. (2008). The representations of spacing and part-based information are associated for upright faces by dissociated for objects: Evidence from individual differences. *Psychonomic Bulletin Review, 15*(5) 933-939.

Appendices

Appendix A.

Forest Plots for Study 1 (Meta-Analysis)

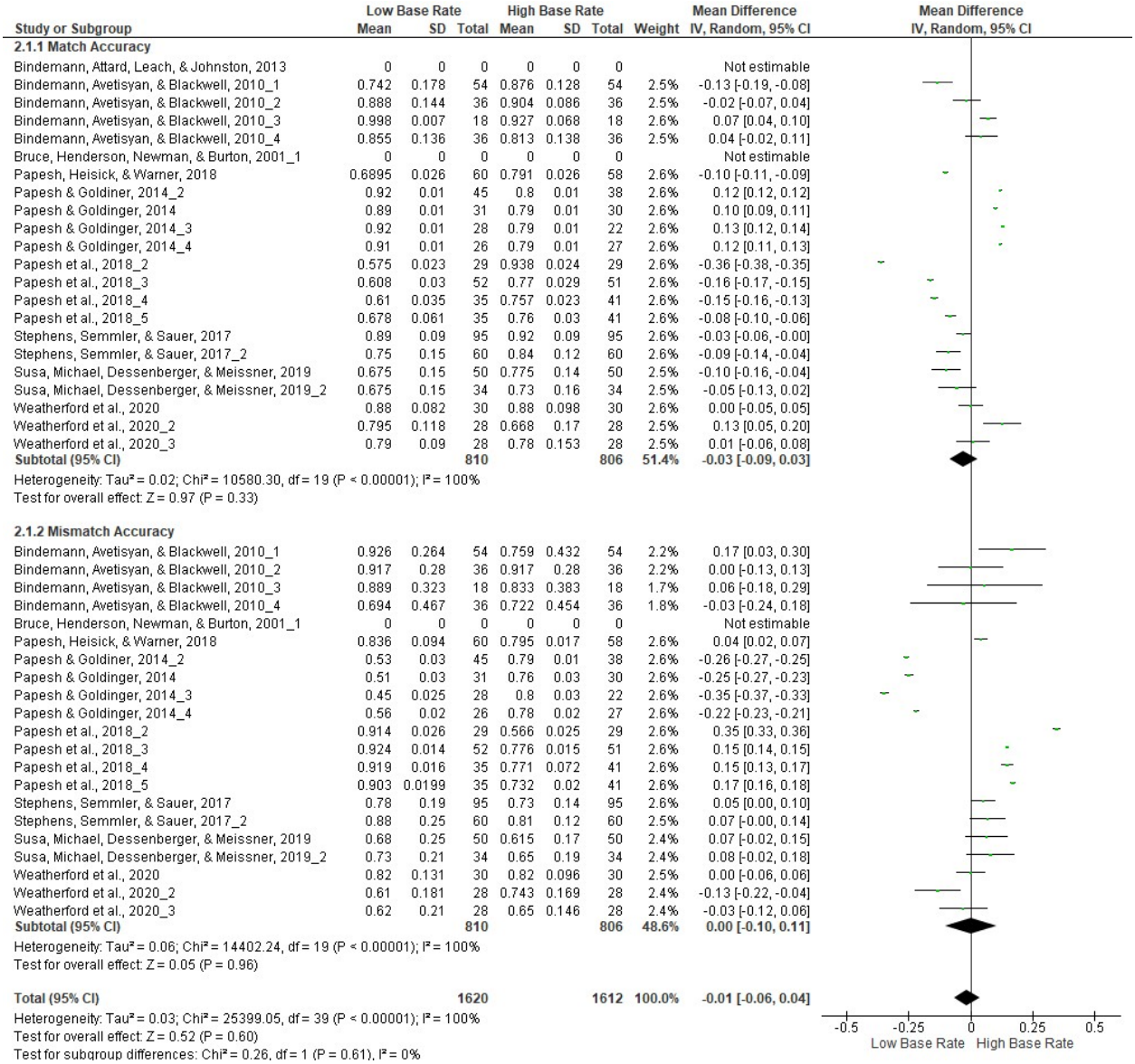
A1. Trial type forest plot



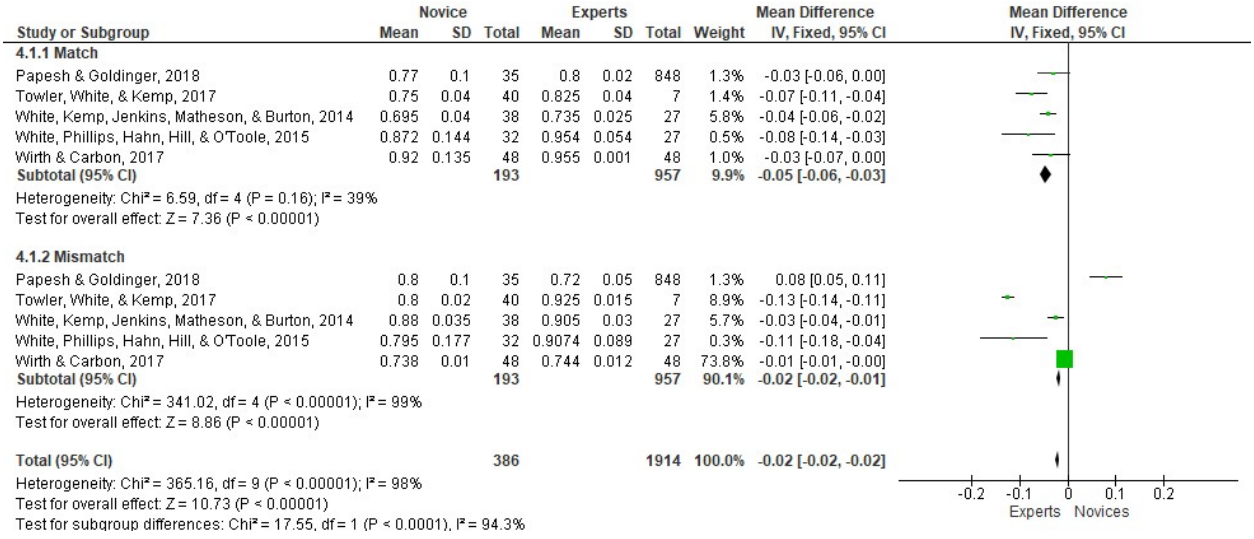
Study or Subgroup	Match Mean	Match SD	Match Total	Mismatch Mean	Mismatch SD	Mismatch Total	Weight	Mean Difference IV, Fixed, 95% CI
Alenezi & Bindemann, 2013_1	0.93	0.106	25	0.869	0.152	25	0.0%	0.06 [-0.01, 0.13]
Alenezi & Bindemann, 2013_2	0.913	0.107	25	0.797	0.188	25	0.0%	0.12 [0.03, 0.20]
Alenezi & Bindemann, 2013_3	0.844	0.131	25	0.711	0.15	25	0.0%	0.13 [0.05, 0.21]
Bindemann, Attard, Leach, & Johnston, 2013	0.9	0.145	20	0.845	0.161	20	0.0%	0.06 [-0.04, 0.15]
Bindemann, Avetisyan, & Blackwell, 2010_1	0.876	0.128	27	0.821	0.277	27	0.0%	0.06 [-0.06, 0.17]
Bindemann, Avetisyan, & Blackwell, 2010_2	0.904	0.086	18	0.937	0.175	18	0.0%	-0.03 [-0.12, 0.06]
Bindemann, Avetisyan, & Blackwell, 2010_3	0.927	0.068	9	0.888	0.215	9	0.0%	0.04 [-0.11, 0.19]
Bindemann, Avetisyan, & Blackwell, 2010_4	0.813	0.138	18	0.763	0.3	18	0.0%	0.05 [-0.10, 0.20]
Bindemann et al., 2013_2	0.618	0.21	20	0.665	0.181	20	0.0%	-0.05 [-0.17, 0.07]
Bindemann et al., 2013_4	0.844	0.143	20	0.856	0.095	20	0.0%	-0.01 [-0.09, 0.06]
Bobak, Mileva, & Hancock, 2019_1	0.574	0.201	42	0.738	0.197	42	0.0%	-0.16 [-0.25, -0.08]
Bobak, Mileva, & Hancock, 2019_2	0.745	0.118	52	0.782	0.151	52	0.0%	-0.04 [-0.09, 0.02]
Davis & Valentine, 2008_2	0.52	0.09	200	0.61	0.111	200	0.3%	-0.09 [-0.11, -0.07]
Davis & Valentine, 2008_3	0.825	0.09	250	0.81	0.111	126	0.2%	0.01 [-0.01, 0.04]
Davis & Valentine, 2008 Exp 1	0.77	0.09	198	0.81	0.111	198	0.3%	-0.04 [-0.06, -0.02]
Dowsett & Burton, 2015_1	0.83	0.137	20	0.777	0.213	20	0.0%	0.05 [-0.06, 0.16]
Dowsett & Burton, 2015_2	0.861	0.149	36	0.808	0.136	36	0.0%	0.05 [-0.01, 0.12]
Dowsett & Burton, 2015_3a	0.687	0.17	36	0.8	0.136	36	0.0%	-0.11 [-0.18, -0.04]
Gentry & Bindemann, 2019	0.913	0.013	90	0.919	0.01	90	9.0%	-0.01 [-0.01, -0.00]
Kemp, Caon, Howard, & Brooks, 2016	0.889	0.139	180	0.859	0.136	180	0.1%	0.03 [0.00, 0.06]
Kokje, Bindemann, & Megreya, 2018_1	0.894	0.016	74	0.852	0.022	74	2.7%	0.04 [0.04, 0.05]
Kramer & Reynolds, 2018	0.84	0.147	54	0.859	0.122	54	0.0%	-0.02 [-0.07, 0.03]
Kramer & Reynolds, 2018_3	0.747	0.208	51	0.785	0.184	51	0.0%	-0.04 [-0.11, 0.04]
Kramer & Ritchie, 2016_1	0.779	0.193	54	0.839	0.165	54	0.0%	-0.06 [-0.13, 0.01]
McCaffery & Burton, 2016	0.727	0.311	80	0.883	0.178	80	0.0%	-0.16 [-0.23, -0.08]
McCaffery & Burton, 2016 exp 3	0.85	0.311	96	0.86	0.178	96	0.0%	-0.01 [-0.08, 0.06]
Megreya, Bindemann, & Harvard, 2011	0	0	0	0	0	0		Not estimable
Megreya, Sandford, & Burton, 2013_2	0.875	0.138	80	0.843	0.142	80	0.1%	0.03 [-0.01, 0.08]
Megreya & Bindemann, 2018	0.833	0.037	60	0.847	0.035	60	0.6%	-0.01 [-0.03, -0.00]
Megreya & Bindemann, 2018_2	0.679	0.04	32	0.872	0.027	32	0.4%	-0.19 [-0.21, -0.18]
Megreya & Burton, 2006	0.816	0.07	30	0.78	0.143	30	0.0%	0.04 [-0.02, 0.09]
Megreya & Burton, 2008_3	0.85	0.21	100	0.844	0.18	100	0.0%	0.01 [-0.05, 0.06]
Mileva & Burton, 2018	0.816	0.025	40	0.681	0.031	40	0.7%	0.13 [0.12, 0.15]
Mileva & Burton, 2018_2	0.74	0.028	60	0.773	0.028	60	1.0%	-0.03 [-0.04, -0.02]
Moore & Johnston, 2013	0.756	0.13	20	0.828	0.125	20	0.0%	-0.07 [-0.15, 0.01]
Moore & Johnston, 2013_2	0.76	0.074	14	0.571	0.332	14	0.0%	0.19 [0.01, 0.37]
Ozbek & Bindemann, 2011	0.904	0.09	30	0.74	0.111	30	0.0%	0.16 [0.11, 0.22]
Papesh, 2018	0.783	0.052	35	0.784	0.052	35	0.2%	-0.00 [-0.03, 0.02]
Papesh, Heisick, & Warner, 2018	0.808	0.029	29	0.789	0.014	29	0.8%	0.02 [0.01, 0.03]
Papesh & Goldiner, 2014_2	0.8	0.01	36	0.79	0.01	36	4.8%	0.01 [0.01, 0.01]
Papesh & Goldinger, 2014	0.79	0.01	30	0.76	0.03	30	0.8%	0.03 [0.02, 0.04]
Papesh & Goldinger, 2014_3	0.79	0.01	22	0.8	0.03	22	0.6%	-0.01 [-0.02, 0.00]
Papesh & Goldinger, 2014_4	0.79	0.01	27	0.78	0.02	27	1.5%	0.01 [0.00, 0.02]
Papesh et al., 2018_2	0.76	0.02	33	0.77	0.02	33	1.1%	-0.01 [-0.02, -0.00]
Papesh et al., 2018_3	0.767	0.026	51	0.774	0.014	51	1.6%	-0.01 [-0.02, 0.00]
Papesh et al., 2018_4	0.758	0.034	41	0.83	0.017	41	0.8%	-0.07 [-0.08, -0.06]
Papesh et al., 2018_5	0.76	0.036	41	0.73	0.019	41	0.7%	0.03 [0.02, 0.04]
Ritchie, Mireku, & Kramer, 2020	0.877	0.029	959	0.703	0.049	959	7.9%	0.17 [0.17, 0.18]
Ritchie, Mireku, & Kramer, 2020_2	0.847	0.028	1040	0.777	0.045	1040	9.9%	0.07 [0.07, 0.07]
Stephens, Semmler, & Sauer, 2017	0.905	0.09	32	0.735	0.155	32	0.0%	0.17 [0.11, 0.23]
Stephens, Semmler, & Sauer, 2017_2	0.84	0.12	30	0.81	0.12	30	0.0%	0.03 [-0.03, 0.09]
Susa, Michael, Dessenberger, & Meissner, 2019	0.84	0.13	50	0.63	0.17	50	0.0%	0.21 [0.15, 0.27]
Susa, Michael, Dessenberger, & Meissner, 2019_2	0.76	0.17	34	0.64	0.18	34	0.0%	0.12 [0.04, 0.20]
Towler, White, & Kemp, 2017	0.632	0.041	44	0.902	0.017	44	0.6%	-0.27 [-0.28, -0.26]
Towler, White, & Kemp, 2017_2	0.695	0.033	102	0.885	0.02	102	1.8%	-0.19 [-0.20, -0.18]
Weatherford et al., 2020	0.9	0.062	25	0.828	0.082	25	0.1%	0.07 [0.03, 0.11]
Weatherford et al., 2020_2	0.76	0.094	9	0.7	0.117	9	0.0%	0.06 [-0.04, 0.16]
Weatherford et al., 2020_3	0.708	0.177	15	0.732	0.175	15	0.0%	-0.02 [-0.15, 0.10]
White, Burton, Jenkins, & Kemp, 2014	0.71	0.016	88	0.811	0.02	88	3.6%	-0.10 [-0.11, -0.10]
White, Burton, Jenkins, & Kemp, 2014_2	0.79	0.026	72	0.799	0.035	72	1.0%	-0.01 [-0.02, 0.00]
White, Burton, Kemp, & Jenkins, 2013	0.918	0.01	300	0.883	0.01	300	40.3%	0.04 [0.03, 0.04]
White, Kemp, Jenkins, & Burton, 2014	0.848	0.14	42	0.874	0.113	42	0.0%	-0.03 [-0.08, 0.03]
White, Kemp, Jenkins, Matheson, & Burton, 2014	0.65	0.028	38	0.877	0.0278	38	0.7%	-0.23 [-0.24, -0.21]
Wirth & Carbon, 2017	0.925	0.013	48	0.785	0.013	48	3.8%	0.14 [0.13, 0.15]
Wirth & Carbon, 2017 exp 2	0.958	0.014	24	0.778	0.014	24	1.6%	0.18 [0.17, 0.19]
Total (95% CI)			**5483**			**5359**	**100.0%**	**0.03 [0.03, 0.03]**

Heterogeneity: Chi² = 21686.77, df = 63 (P < 0.00001); I² = 100%
Test for overall effect: Z = 64.73 (P < 0.00001)

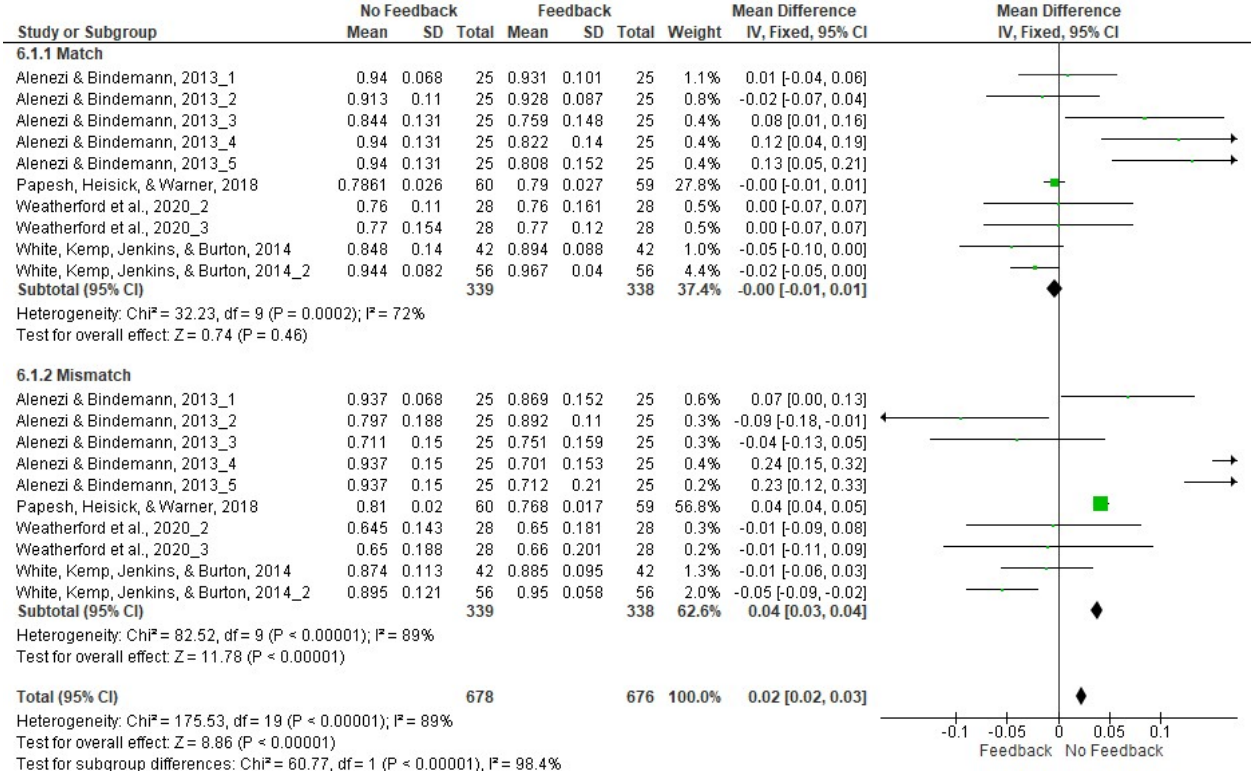
A2. Base rate of mismatch forest plot



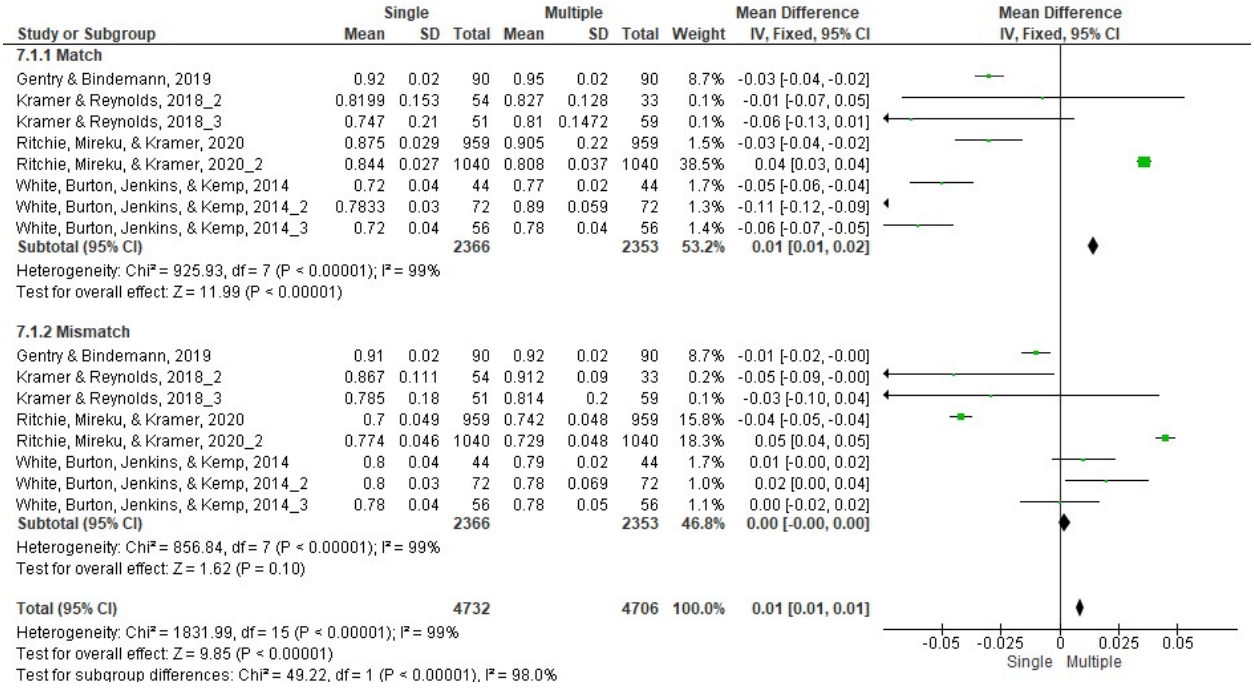
A3. Expertise forest plot



A4. Feedback forest plot

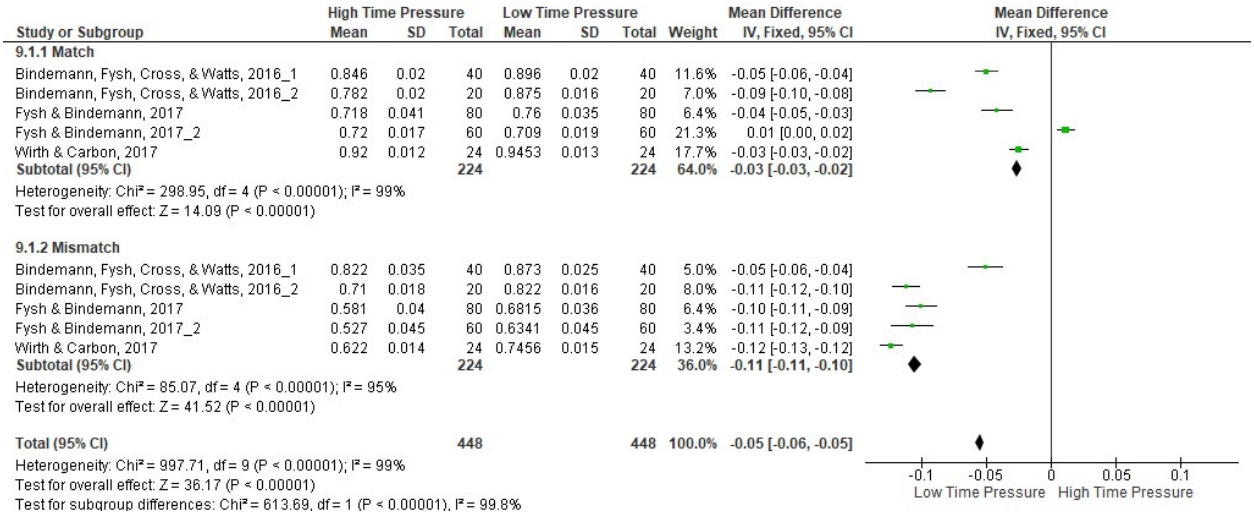


A5. Photo amount forest plot



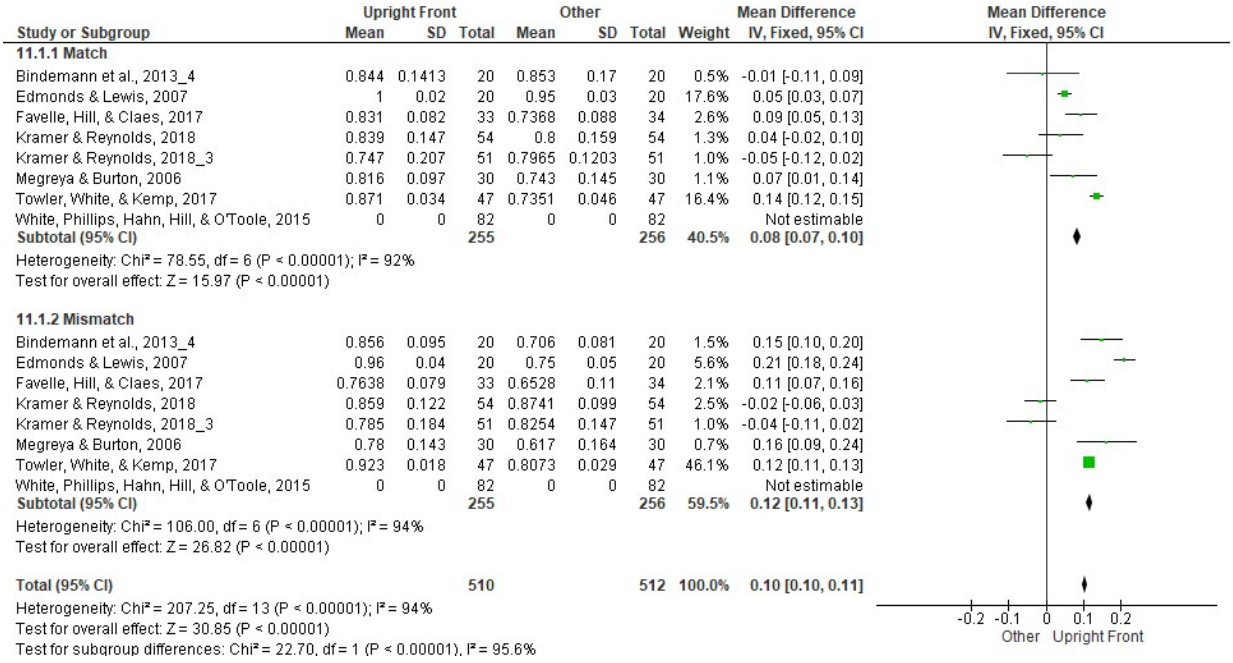
Study or Subgroup	Single Mean	Single SD	Single Total	Multiple Mean	Multiple SD	Multiple Total	Weight	Mean Difference IV, Fixed, 95% CI
7.1.1 Match								
Gentry & Bindemann, 2019	0.92	0.02	90	0.95	0.02	90	8.7%	-0.03 [-0.04, -0.02]
Kramer & Reynolds, 2018_2	0.8199	0.153	54	0.827	0.128	33	0.1%	-0.01 [-0.07, 0.05]
Kramer & Reynolds, 2018_3	0.747	0.21	51	0.81	0.1472	59	0.1%	-0.06 [-0.13, 0.01]
Ritchie, Mireku, & Kramer, 2020	0.875	0.029	959	0.905	0.22	959	1.5%	-0.03 [-0.04, -0.02]
Ritchie, Mireku, & Kramer, 2020_2	0.844	0.027	1040	0.808	0.037	1040	38.5%	0.04 [0.03, 0.04]
White, Burton, Jenkins, & Kemp, 2014	0.72	0.04	44	0.77	0.02	44	1.7%	-0.05 [-0.06, -0.04]
White, Burton, Jenkins, & Kemp, 2014_2	0.7833	0.03	72	0.89	0.059	72	1.3%	-0.11 [-0.12, -0.09]
White, Burton, Jenkins, & Kemp, 2014_3	0.72	0.04	56	0.78	0.04	56	1.4%	-0.06 [-0.07, -0.05]
Subtotal (95% CI)			**2366**			**2353**	**53.2%**	**0.01 [0.01, 0.02]**
Heterogeneity: Chi² = 925.93, df = 7 (P < 0.00001); I² = 99%								
Test for overall effect: Z = 11.99 (P < 0.00001)								
7.1.2 Mismatch								
Gentry & Bindemann, 2019	0.91	0.02	90	0.92	0.02	90	8.7%	-0.01 [-0.02, -0.00]
Kramer & Reynolds, 2018_2	0.867	0.111	54	0.912	0.09	33	0.2%	-0.05 [-0.09, -0.00]
Kramer & Reynolds, 2018_3	0.785	0.18	51	0.814	0.2	59	0.1%	-0.03 [-0.10, 0.04]
Ritchie, Mireku, & Kramer, 2020	0.7	0.049	959	0.742	0.048	959	15.8%	-0.04 [-0.05, -0.04]
Ritchie, Mireku, & Kramer, 2020_2	0.774	0.046	1040	0.729	0.048	1040	18.3%	0.05 [0.04, 0.05]
White, Burton, Jenkins, & Kemp, 2014	0.8	0.04	44	0.79	0.02	44	1.7%	0.01 [-0.00, 0.02]
White, Burton, Jenkins, & Kemp, 2014_2	0.8	0.03	72	0.78	0.069	72	1.0%	0.02 [0.00, 0.04]
White, Burton, Jenkins, & Kemp, 2014_3	0.78	0.04	56	0.78	0.05	56	1.1%	0.00 [-0.02, 0.02]
Subtotal (95% CI)			**2366**			**2353**	**46.8%**	**0.00 [-0.00, 0.00]**
Heterogeneity: Chi² = 856.84, df = 7 (P < 0.00001); I² = 99%								
Test for overall effect: Z = 1.62 (P = 0.10)								
Total (95% CI)			**4732**			**4706**	**100.0%**	**0.01 [0.01, 0.01]**
Heterogeneity: Chi² = 1831.99, df = 15 (P < 0.00001); I² = 99%								
Test for overall effect: Z = 9.85 (P < 0.00001)								
Test for subgroup differences: Chi² = 49.22, df = 1 (P < 0.00001), I² = 98.0%								

A6. Time pressure forest plot

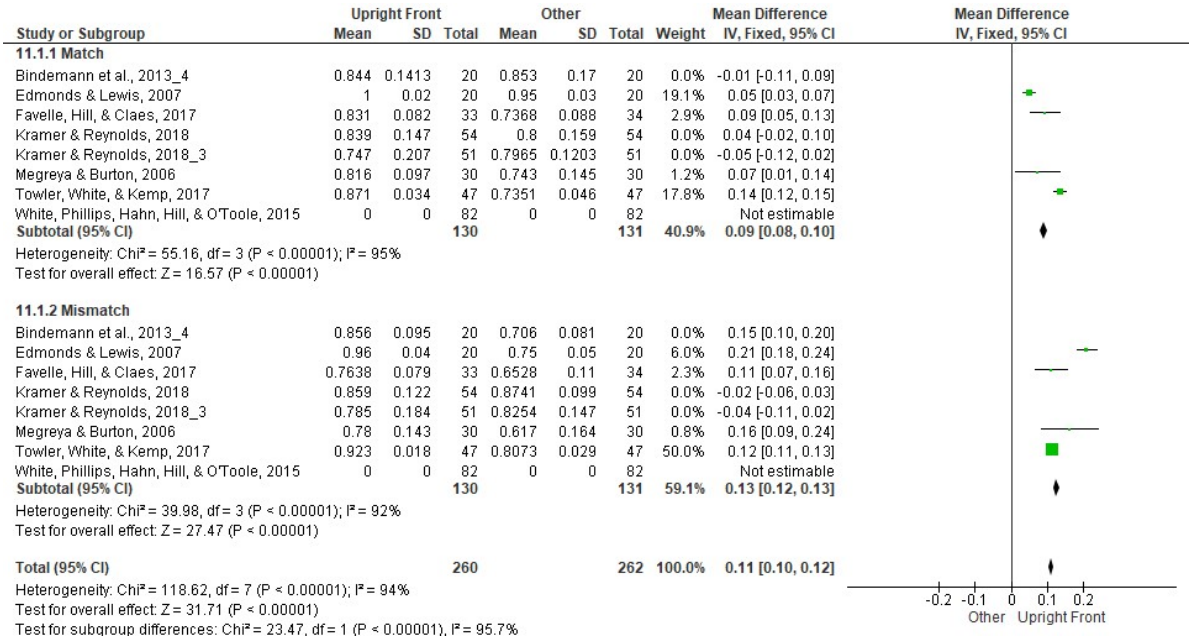


Study or Subgroup	High Time Pressure Mean	High Time Pressure SD	High Time Pressure Total	Low Time Pressure Mean	Low Time Pressure SD	Low Time Pressure Total	Weight	Mean Difference IV, Fixed, 95% CI
9.1.1 Match								
Bindemann, Fysh, Cross, & Watts, 2016_1	0.846	0.02	40	0.896	0.02	40	11.6%	-0.05 [-0.06, -0.04]
Bindemann, Fysh, Cross, & Watts, 2016_2	0.782	0.02	20	0.875	0.016	20	7.0%	-0.09 [-0.10, -0.08]
Fysh & Bindemann, 2017	0.718	0.041	80	0.76	0.035	80	6.4%	-0.04 [-0.05, -0.03]
Fysh & Bindemann, 2017_2	0.72	0.017	60	0.709	0.019	60	21.3%	0.01 [0.00, 0.02]
Wirth & Carbon, 2017	0.92	0.012	24	0.9453	0.013	24	17.7%	-0.03 [-0.03, -0.02]
Subtotal (95% CI)			**224**			**224**	**64.0%**	**-0.03 [-0.03, -0.02]**
Heterogeneity: Chi² = 298.95, df = 4 (P < 0.00001); I² = 99%								
Test for overall effect: Z = 14.09 (P < 0.00001)								
9.1.2 Mismatch								
Bindemann, Fysh, Cross, & Watts, 2016_1	0.822	0.035	40	0.873	0.025	40	5.0%	-0.05 [-0.06, -0.04]
Bindemann, Fysh, Cross, & Watts, 2016_2	0.71	0.018	20	0.822	0.016	20	8.0%	-0.11 [-0.12, -0.10]
Fysh & Bindemann, 2017	0.581	0.04	80	0.6815	0.036	80	6.4%	-0.10 [-0.11, -0.09]
Fysh & Bindemann, 2017_2	0.527	0.045	60	0.6341	0.045	60	3.4%	-0.11 [-0.12, -0.09]
Wirth & Carbon, 2017	0.622	0.014	24	0.7456	0.015	24	13.2%	-0.12 [-0.13, -0.12]
Subtotal (95% CI)			**224**			**224**	**36.0%**	**-0.11 [-0.11, -0.10]**
Heterogeneity: Chi² = 85.07, df = 4 (P < 0.00001); I² = 95%								
Test for overall effect: Z = 41.52 (P < 0.00001)								
Total (95% CI)			**448**			**448**	**100.0%**	**-0.05 [-0.06, -0.05]**
Heterogeneity: Chi² = 997.71, df = 9 (P < 0.00001); I² = 99%								
Test for overall effect: Z = 36.17 (P < 0.00001)								
Test for subgroup differences: Chi² = 613.69, df = 1 (P < 0.00001), I² = 99.8%								

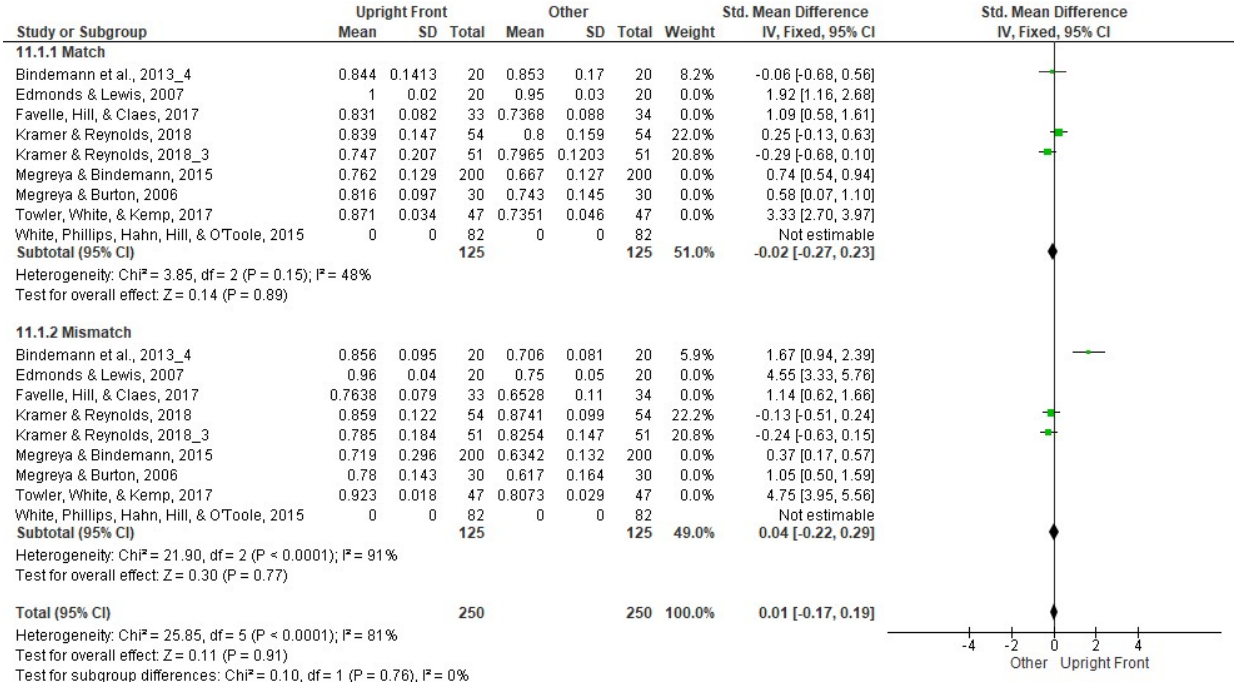
A7. View and pose forest plot – overall



A8. View and pose forest plot – upright vs. inverted



A9. View and pose forest plot – front vs. profile



Study or Subgroup	Upright Front Mean	SD	Total	Other Mean	SD	Total	Weight	Std. Mean Difference IV, Fixed, 95% CI
11.1.1 Match								
Bindemann et al., 2013_4	0.844	0.1413	20	0.853	0.17	20	8.2%	-0.06 [-0.68, 0.56]
Edmonds & Lewis, 2007	1	0.02	20	0.95	0.03	20	0.0%	1.92 [1.16, 2.68]
Favelle, Hill, & Claes, 2017	0.831	0.082	33	0.7368	0.088	34	0.0%	1.09 [0.58, 1.61]
Kramer & Reynolds, 2018	0.839	0.147	54	0.8	0.159	54	22.0%	0.25 [-0.13, 0.63]
Kramer & Reynolds, 2018_3	0.747	0.207	51	0.7965	0.1203	51	20.8%	-0.29 [-0.68, 0.10]
Megreya & Bindemann, 2015	0.762	0.129	200	0.667	0.127	200	0.0%	0.74 [0.54, 0.94]
Megreya & Burton, 2006	0.816	0.097	30	0.743	0.145	30	0.0%	0.58 [0.07, 1.10]
Towler, White, & Kemp, 2017	0.871	0.034	47	0.7351	0.046	47	0.0%	3.33 [2.70, 3.97]
White, Phillips, Hahn, Hill, & O'Toole, 2015	0	0	82	0	0	82		Not estimable
Subtotal (95% CI)			**125**			**125**	**51.0%**	**-0.02 [-0.27, 0.23]**
Heterogeneity: Chi² = 3.85, df = 2 (P = 0.15); I² = 48%								
Test for overall effect: Z = 0.14 (P = 0.89)								
11.1.2 Mismatch								
Bindemann et al., 2013_4	0.856	0.095	20	0.706	0.081	20	5.9%	1.67 [0.94, 2.39]
Edmonds & Lewis, 2007	0.96	0.04	20	0.75	0.05	20	0.0%	4.55 [3.33, 5.76]
Favelle, Hill, & Claes, 2017	0.7638	0.079	33	0.6528	0.11	34	0.0%	1.14 [0.62, 1.66]
Kramer & Reynolds, 2018	0.859	0.122	54	0.8741	0.099	54	22.2%	-0.13 [-0.51, 0.24]
Kramer & Reynolds, 2018_3	0.785	0.184	51	0.8254	0.147	51	20.8%	-0.24 [-0.63, 0.15]
Megreya & Bindemann, 2015	0.719	0.296	200	0.6342	0.132	200	0.0%	0.37 [0.17, 0.57]
Megreya & Burton, 2006	0.78	0.143	30	0.617	0.164	30	0.0%	1.05 [0.50, 1.59]
Towler, White, & Kemp, 2017	0.923	0.018	47	0.8073	0.029	47	0.0%	4.75 [3.95, 5.56]
White, Phillips, Hahn, Hill, & O'Toole, 2015	0	0	82	0	0	82		Not estimable
Subtotal (95% CI)			**125**			**125**	**49.0%**	**0.04 [-0.22, 0.29]**
Heterogeneity: Chi² = 21.90, df = 2 (P < 0.0001); I² = 91%								
Test for overall effect: Z = 0.30 (P = 0.77)								
Total (95% CI)			**250**			**250**	**100.0%**	**0.01 [-0.17, 0.19]**
Heterogeneity: Chi² = 25.85, df = 5 (P < 0.0001); I² = 81%								
Test for overall effect: Z = 0.11 (P = 0.91)								
Test for subgroup differences: Chi² = 0.10, df = 1 (P = 0.76), I² = 0%								

Appendix B.

Materials for Study 2 (Base Rates of Mismatch)

B1. Instructions (same for all conditions)

In the following section you will be asked to decide whether the presented photos show the same or different people. If you think the two photos show the same people, select the "Match" option. If you think the two photos show different people, select the "Mismatch" option. As you answer the questions, try to be as quick and as accurate as possible. Keep in mind that time may have passed between the two photos and that hairstyles, hair colors, and other aspects of the person's face, such as makeup, may have changed. Please do not zoom in while taking the survey.

B2. Awareness Warnings

B2.1. Aware Warning – 10% condition

This set of stimuli contains **1% mismatches**. Please keep that in mind when answering the questions while trying to be as quick and accurate as possible.

B2.2. Unaware Warning – 10% condition

This set of stimuli contains both matches and mismatches. Try to be as quick and accurate as possible while answering the following questions.

B2.3. Aware Warning – 50% condition

This set of stimuli contains **50% mismatches**. Please keep that in mind when answering the questions while trying to be as quick and accurate as possible.

B2.4. Unaware Warning – 50% condition

This set of stimuli contains both matches and mismatches. Please keep that in mind when answering the questions while trying to be as quick and accurate as possible.

B2.5. Aware Warning – 90% condition

This set of stimuli contains **99% mismatches**. Please keep that in mind when answering the questions while trying to be as quick and accurate as possible.

B2.6. Unaware Warning – 90% condition

This set of stimuli contains both matches and mismatches. Try to be as quick and accurate as possible while answering the following questions.

B3. Stimuli

B3.1. 10% Condition

Do the two photos show a match or a mismatch?

Match ○ Mismatch ○

Do the two photos show a match or a mismatch?

Match ○ Mismatch ○

Do the two photos show a match or a mismatch?

Match ○ Mismatch ○

Do the two photos show a match or a mismatch?

Match ○ Mismatch ○

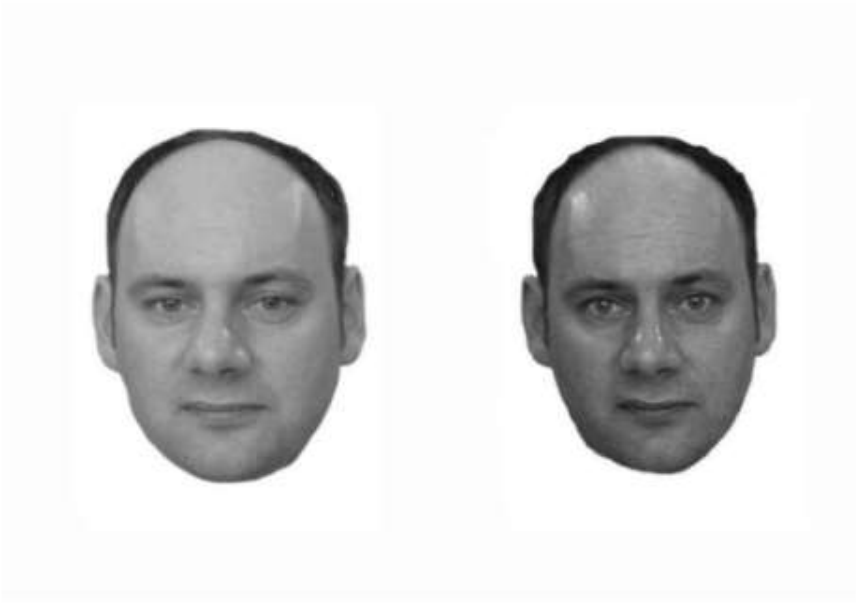
Do the two photos show a match or a mismatch?

Match ○ Mismatch ○

Do the two photos show a match or a mismatch?

Match ○ Mismatch ○

Do the two photos show a match or a mismatch?



Match

Mismatch

Do the two photos show a match or a mismatch?

Match

Mismatch

Do the two photos show a match or a mismatch?

Match ○ Mismatch ○

Do the two photos show a match or a mismatch?

Match ○ Mismatch ○

Do the two photos show a match or a mismatch?

Match ○ Mismatch ○

Do the two photos show a match or a mismatch?

Match ○ Mismatch ○

Do the two photos show a match or a mismatch?

Match ○ Mismatch ○

Do the two photos show a match or a mismatch?

Match ○ Mismatch ○

Do the two photos show a match or a mismatch?

Match ○ Mismatch ○

Do the two photos show a match or a mismatch?

Match ○ Mismatch ○

Do the two photos show a match or a mismatch?

Match ○ Mismatch ○

Do the two photos show a match or a mismatch?

Match ○ Mismatch ○

Do the two photos show a match or a mismatch?

Match ○ Mismatch ○

Do the two photos show a match or a mismatch?

Match ○ Mismatch ○

Do the two photos show a match or a mismatch?

Match ○ Mismatch ○

Do the two photos show a match or a mismatch?

Match ○ Mismatch ○

Do the two photos show a match or a mismatch?

Match ○ Mismatch ○

Do the two photos show a match or a mismatch?

Match ○ Mismatch ○

Do the two photos show a match or a mismatch?

Match ○ Mismatch ○

Do the two photos show a match or a mismatch?

Match ○ Mismatch ○

Do the two photos show a match or a mismatch?

Match ○ Mismatch ○

Do the two photos show a match or a mismatch?

Match ○ Mismatch ○

Do the two photos show a match or a mismatch?

Match ○ Mismatch ○

Do the two photos show a match or a mismatch?

Match ○ Mismatch ○

Do the two photos show a match or a mismatch?

Match ○ Mismatch ○

Do the two photos show a match or a mismatch?

Match ○ Mismatch ○

Do the two photos show a match or a mismatch?

Match ○ Mismatch ○

Do the two photos show a match or a mismatch?

Match ○ Mismatch ○

Do the two photos show a match or a mismatch?

Match ○ Mismatch ○

Do the two photos show a match or a mismatch?

Match ○ Mismatch ○

Do the two photos show a match or a mismatch?

Match ○ Mismatch ○

Do the two photos show a match or a mismatch?

Match ○ Mismatch ○

Do the two photos show a match or a mismatch?

Match ○ Mismatch ○

Do the two photos show a match or a mismatch?

Match ○ Mismatch ○

B3.2. 50% Condition

Do the two photos show a match or a mismatch?

Match ○ Mismatch ○

Do the two photos show a match or a mismatch?

Match ○ Mismatch ○

Do the two photos show a match or a mismatch?

Match ○ Mismatch ○

Do the two photos show a match or a mismatch?

Match ○ Mismatch ○

Do the two photos show a match or a mismatch?

Match ○ Mismatch ○

Do the two photos show a match or a mismatch?

Match ○ Mismatch ○

Do the two photos show a match or a mismatch?

Match ○ Mismatch ○

Do the two photos show a match or a mismatch?

Match ○ Mismatch ○

Do the two photos show a match or a mismatch?

Match ○ Mismatch ○

Do the two photos show a match or a mismatch?

Match ○ Mismatch ○

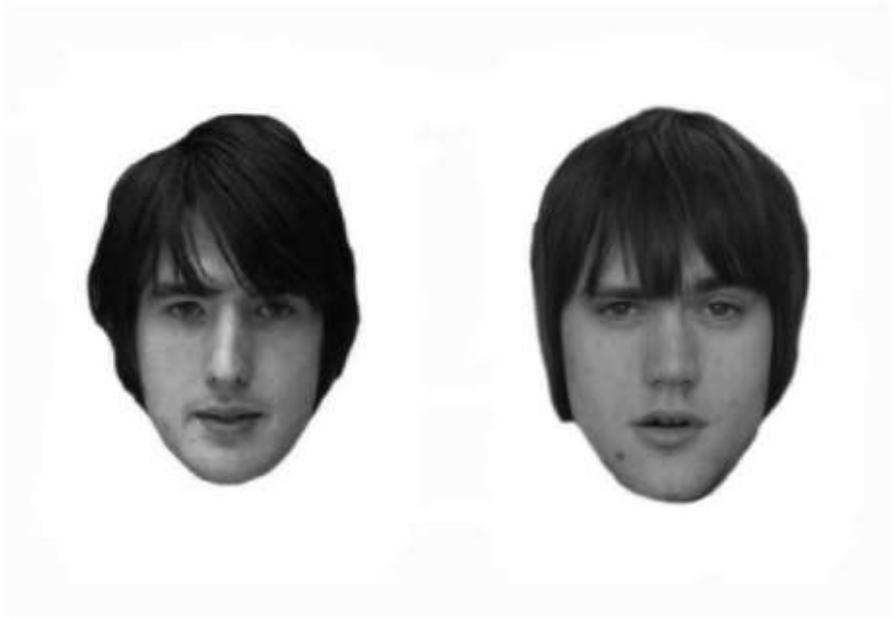
Do the two photos show a match or a mismatch?

Match ○ Mismatch ○

Do the two photos show a match or a mismatch?

Match ○ Mismatch ○

Do the two photos show a match or a mismatch?



Match ○ Mismatch ○

Do the two photos show a match or a mismatch?

Match ○ Mismatch ○

Do the two photos show a match or a mismatch?

Match ○ Mismatch ○

Do the two photos show a match or a mismatch?

Match ○ Mismatch ○

Do the two photos show a match or a mismatch?

Match ○ Mismatch ○

Do the two photos show a match or a mismatch?

Match ○ Mismatch ○

Do the two photos show a match or a mismatch?

Match ○ Mismatch ○

Do the two photos show a match or a mismatch?

Match ○ Mismatch ○

Do the two photos show a match or a mismatch?

Match ○ Mismatch ○

Do the two photos show a match or a mismatch?

Match ○ Mismatch ○

Do the two photos show a match or a mismatch?

Match ○ Mismatch ○

Do the two photos show a match or a mismatch?

Match ○ Mismatch ○

Do the two photos show a match or a mismatch?

Match ○ Mismatch ○

Do the two photos show a match or a mismatch?

Match ○ Mismatch ○

Do the two photos show a match or a mismatch?

Match

Mismatch

Do the two photos show a match or a mismatch?

Match

Mismatch

Do the two photos show a match or a mismatch?

Match ○ Mismatch ○

Do the two photos show a match or a mismatch?

Match ○ Mismatch ○

Do the two photos show a match or a mismatch?

Match ○ Mismatch ○

Do the two photos show a match or a mismatch?

Match ○ Mismatch ○

Do the two photos show a match or a mismatch?

Match ○ Mismatch ○

Do the two photos show a match or a mismatch?

Match ○ Mismatch ○

Do the two photos show a match or a mismatch?

Match ○ Mismatch ○

Do the two photos show a match or a mismatch?

Match Mismatch

Do the two photos show a match or a mismatch?

Match ○ Mismatch ○

Do the two photos show a match or a mismatch?

Match ○ Mismatch ○

Do the two photos show a match or a mismatch?

Match ○ Mismatch ○

Do the two photos show a match or a mismatch?

Match ○ Mismatch ○

B3.3. 90% Condition

Do the two photos show a match or a mismatch?

Match ○ Mismatch ○

Do the two photos show a match or a mismatch?

Match ○ Mismatch ○

Do the two photos show a match or a mismatch?

Match ○ Mismatch ○

Do the two photos show a match or a mismatch?

Match ○ Mismatch ○

Do the two photos show a match or a mismatch?

Match ○ Mismatch ○

Do the two photos show a match or a mismatch?

Match ○ Mismatch ○

Do the two photos show a match or a mismatch?

Match

Mismatch

Do the two photos show a match or a mismatch?

Match

Mismatch

Do the two photos show a match or a mismatch?

Match ○ Mismatch ○

Do the two photos show a match or a mismatch?

Match ○ Mismatch ○

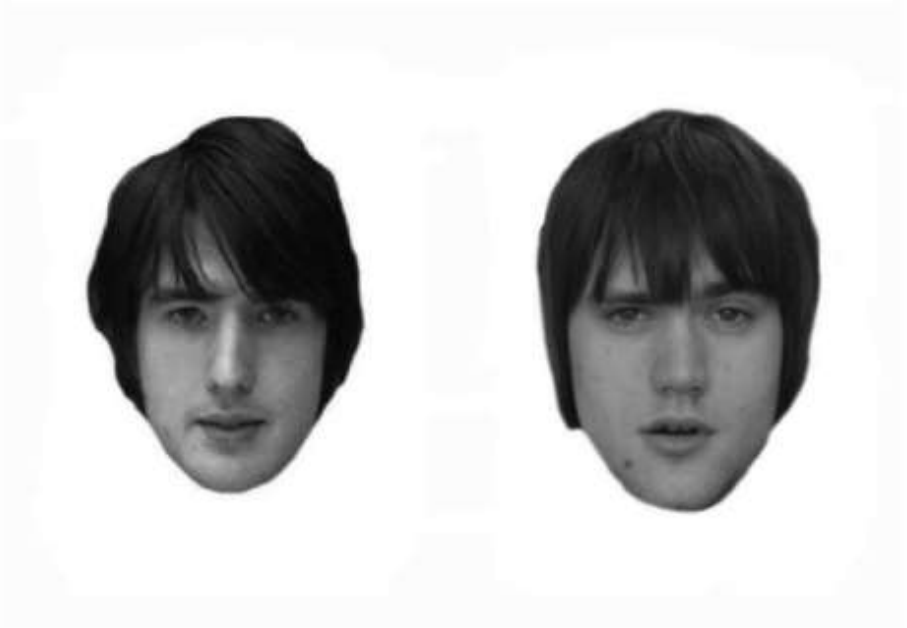
Do the two photos show a match or a mismatch?

Match ○ Mismatch ○

Do the two photos show a match or a mismatch?

Match ○ Mismatch ○

Do the two photos show a match or a mismatch?



Match ○ Mismatch ○

Do the two photos show a match or a mismatch?

Match ○ Mismatch ○

Do the two photos show a match or a mismatch?

Match ○ Mismatch ○

Do the two photos show a match or a mismatch?

Match ○ Mismatch ○

Do the two photos show a match or a mismatch?

Match ○ Mismatch ○

Do the two photos show a match or a mismatch?

Match ○ Mismatch ○

Do the two photos show a match or a mismatch?

Match ○ Mismatch ○

Do the two photos show a match or a mismatch?

Match ○ Mismatch ○

Do the two photos show a match or a mismatch?

Match ○ Mismatch ○

Do the two photos show a match or a mismatch?

Match ○ Mismatch ○

Do the two photos show a match or a mismatch?

Match ○ Mismatch ○

Do the two photos show a match or a mismatch?

Match ○ Mismatch ○

Do the two photos show a match or a mismatch?

Match ○ Mismatch ○

Do the two photos show a match or a mismatch?

Match ○ Mismatch ○

Do the two photos show a match or a mismatch?

Match ○ Mismatch ○

Do the two photos show a match or a mismatch?

Match ○ Mismatch ○

Do the two photos show a match or a mismatch?

Match ○ Mismatch ○

Do the two photos show a match or a mismatch?

Match ○ Mismatch ○

Do the two photos show a match or a mismatch?

Match ○ Mismatch ○

Do the two photos show a match or a mismatch?

Match ○ Mismatch ○

Do the two photos show a match or a mismatch?

Match ○ Mismatch ○

Do the two photos show a match or a mismatch?

Match ○ Mismatch ○

Do the two photos show a match or a mismatch?

Match ○ Mismatch ○

Do the two photos show a match or a mismatch?

Match ○ Mismatch ○

Do the two photos show a match or a mismatch?

Match ○ Mismatch ○

Do the two photos show a match or a mismatch?

Match ○ Mismatch ○

Do the two photos show a match or a mismatch?

Match ○ Mismatch ○

Do the two photos show a match or a mismatch?

Match ○ Mismatch ○

B4. Demographics & Debrief (same for all conditions)

Which mismatch condition were you told you were in?

- ○ 1%
- ○ 50%
- ○ 99%
- ○ I was not informed of my mismatch condition

What ethnicity/race do you most identify with?

- ○ African American/Black
- ○ Caucasian/White
- ○ East Asian
- ○ Arabic
- ○ Latinx
- ○ Asian
- ○ South Asian
- ○ Pacific Islander
- ○ First Nations/Inuit/Métis
- ○ Other
- ○ Prefer not to answer

How old are you? Please enter age below.

What gender do you most identify with?

- ○ Male
- ○ Female
- ○ Non-Binary
- ○ Transgender Male
- ○ Transgender Female
- ○ Gender non-conforming
- ○ Other
- ○ Prefer not to answer

Thank you

Thank you! Thank you for participating in our study about face matching decision making. Research into face matching is important because face matching is used to verify a person's identity at a number of locations, such as when buying a controlled substance, when interacting with law enforcement, or when entering/leaving the country. Face matching of unfamiliar faces is notoriously poor. The goal of this research is to find out more about how observers make face matching decisions so that we can find ways to improve unfamiliar face matching accuracy. Data from the study will be available in the beginning of 2022. If you have any questions about the study or face matching, please email the student lead at danielle.rumschik@uoit.ca

****CLICK THROUGH TO BE REDIRECTED AND RECIEVE YOUR CREDIT*****

Appendix C.

Results for Full Sample in Study 2 (Base Rates of Mismatch)

Significant results were found for trial ($F(1) = 3.273$, $p = 0.072$, $\eta_p^2 = 0.017$), trial x prevalence ($F(2) = 21.555$, $p < 0.001$, $\eta_p^2 = 0.183$), and the trial x prevalence x awareness interaction ($F(2) = 6.128$, $p = 0.003$, $\eta_p^2 = 0.06$). See tables below for details. No significant results were found for prevalence ($F(2) = 1.412$, $p = 0.246$, $\eta_p^2 = 0.014$), awareness ($F(1) = 1.290$, $p = 0.258$, $\eta_p^2 = 0.007$), prevalence x awareness ($F(2) = 0.543$, $p = 0.582$, $\eta_p^2 = 0.006$), or the trial x awareness interaction ($F(1) = 0.30$, $p = 0.862$, $\eta_p^2 = 0.00$).

Table C1. Means for Trial Type

			95% Confidence Interval	
Trial	Mean	SD	Lower Bound	Upper Bound
Match	.812	.014	.784	.839
Mismatch	.845	.012	.821	.868

Table C2. Means for Trial x Prevalence interaction

				95% Confidence Interval	
Prevalence	Trial	Mean	SD	Lower Bound	Upper Bound
10%	Match	.733	.024	.686	.781
	Mismatch	.916	.021	.875	.957
50%	Match	.797	.024	.749	.844
	Mismatch	.825	.021	.784	.866
90%	Match	.904	.024	.857	.952
	Mismatch	.793	.021	.751	.834

Table C3. Means for Trial x Prevalence x Awareness interaction

Prevalence	Awareness	Trial	Mean	SD	95% Confidence Interval

					Lower Bound	Upper Bound
10%	Aware	Match	.799	.035	.731	.867
		Mismatch	.898	.030	.840	.957
	Not Aware	Match	.668	.033	.602	.734
		Mismatch	.934	.029	.877	.991
50%	Aware	Match	.787	.035	.719	.856
		Mismatch	.837	.030	.779	.896
	Not Aware	Match	.806	.033	.740	.872
		Mismatch	.813	.029	.756	.870
90%	Aware	Match	.875	.035	.807	.943
		Mismatch	.834	.030	.775	.893
	Not Aware	Match	.934	.033	.868	1.000
		Mismatch	.751	.029	.694	.808

Appendix D.

Study Materials for Study 3 (Expressions)

D1. Pilot Study Materials

Instructions

In the next section you will be shown pairs of faces and asked how similar you think the faces look to each other. You will rate the similarity on a scale from 1 to 5 with 1 = not at all similar and 5 = extremely similar. Please take your time and try to answer as accurately as possible.

Similarity Ratings

How similar are these two faces?

1 - Not at all similar	2	3	4	5 - Extremely similar
○	○	○	○	○

How similar are these two faces?

1 - Not at all similar	2	3	4	5 - Extremely similar
○	○	○	○	○

How similar are these two faces?

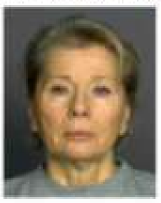
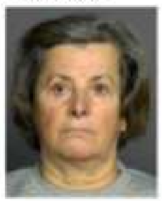
1 - Not at all similar	2	3	4	5 - Extremely similar
○	○	○	○	○

How similar are these two faces?

1 - Not at all similar	2	3	4	5 - Extremely similar
○	○	○	○	○

How similar are these two faces?

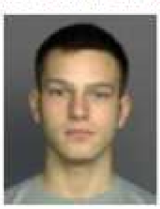
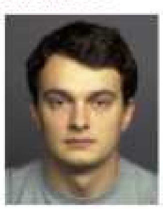
1 - Not at all similar	2	3	4	5 - Extremely similar
○	○	○	○	○

How similar are these two faces?

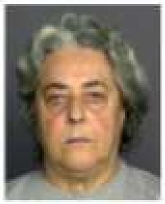
1 - Not at all similar	2	3	4	5 - Extremely similar
○	○	○	○	○

How similar are these two faces?

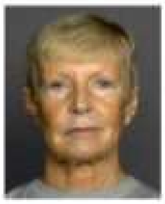
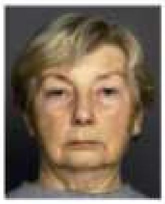
1 - Not at all similar	2	3	4	5 - Extremely similar
○	○	○	○	○

How similar are these two faces?

1 - Not at all similar	2	3	4	5 - Extremely similar
○	○	○	○	○

How similar are these two faces?

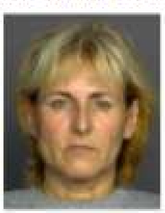
1 - Not at all similar	2	3	4	5 - Extremely similar
○	○	○	○	○

How similar are these two faces?

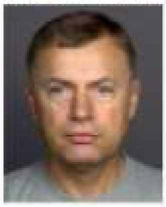
1 - Not at all similar	2	3	4	5 - Extremely similar
○	○	○	○	○

How similar are these two faces?

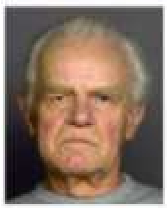
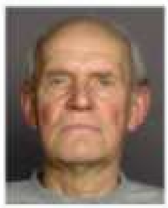
1 - Not at all similar	2	3	4	5 - Extremely similar
○	○	○	○	○

How similar are these two faces?

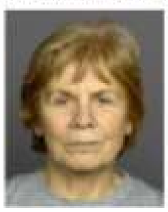
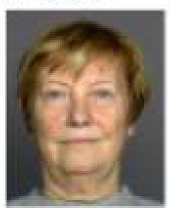
1 - Not at all similar	2	3	4	5 - Extremely similar

How similar are these two faces?

1 - Not at all similar	2	3	4	5 - Extremely similar

How similar are these two faces?

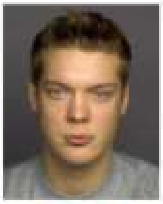
1 - Not at all similar	2	3	4	5 - Extremely similar

How similar are these two faces?

1 - Not at all similar	2	3	4	5 - Extremely similar

How similar are these two faces?

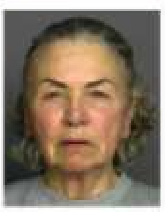
1 - Not at all similar	2	3	4	5 - Extremely similar
○	○	○	○	○

How similar are these two faces?

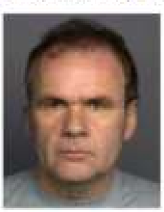
1 - Not at all similar	2	3	4	5 - Extremely similar
○	○	○	○	○

How similar are these two faces?

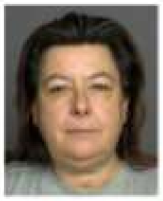
1 - Not at all similar	2	3	4	5 - Extremely similar
○	○	○	○	○

How similar are these two faces?

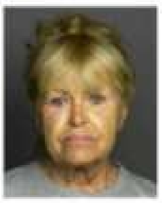
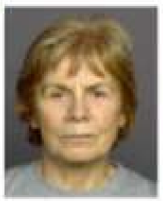
1 - Not at all similar	2	3	4	5 - Extremely similar
○	○	○	○	○

How similar are these two faces?

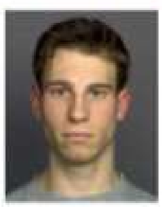
1 - Not at all similar	2	3	4	5 - Extremely similar

How similar are these two faces?

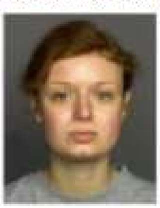
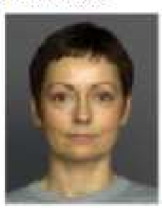
1 - Not at all similar	2	3	4	5 - Extremely similar

How similar are these two faces?

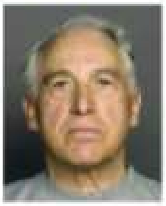
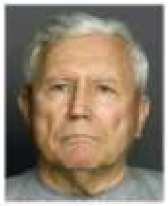
1 - Not at all similar	2	3	4	5 - Extremely similar

How similar are these two faces?

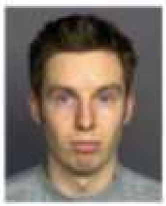
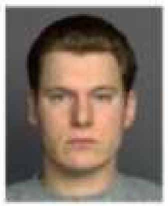
1 - Not at all similar	2	3	4	5 - Extremely similar

How similar are these two faces?

1 - Not at all similar	2	3	4	5 - Extremely similar
○	○	○	○	○

How similar are these two faces?

1 - Not at all similar	2	3	4	5 - Extremely similar
○	○	○	○	○

How similar are these two faces?

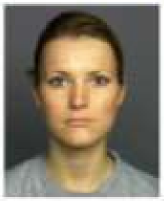
1 - Not at all similar	2	3	4	5 - Extremely similar
○	○	○	○	○

How similar are these two faces?

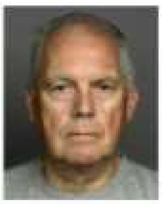
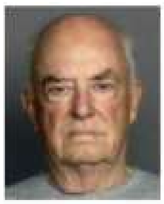
1 - Not at all similar	2	3	4	5 - Extremely similar
○	○	○	○	○

How similar are these two faces?

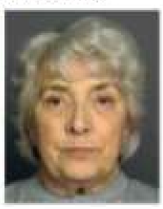
1 - Not at all similar	2	3	4	5 - Extremely similar
○	○	○	○	○

How similar are these two faces?

1 - Not at all similar	2	3	4	5 - Extremely similar
○	○	○	○	○

How similar are these two faces?

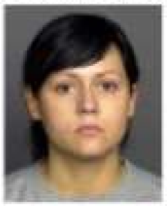
1 - Not at all similar	2	3	4	5 - Extremely similar
○	○	○	○	○

How similar are these two faces?

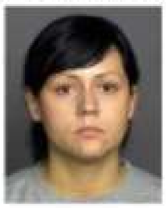
1 - Not at all similar	2	3	4	5 - Extremely similar
○	○	○	○	○

How similar are these two faces?



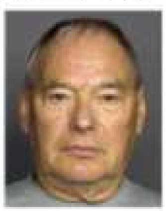
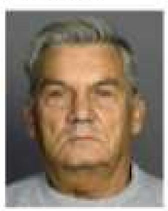

1 - Not at all similar	2	3	4	5 - Extremely similar
○	○	○	○	○

How similar are these two faces?




1 - Not at all similar	2	3	4	5 - Extremely similar
○	○	○	○	○

How similar are these two faces?

1 - Not at all similar	2	3	4	5 - Extremely similar
○	○	○	○	○

How similar are these two faces?


1 - Not at all similar	2	3	4	5 - Extremely similar
○	○	○	○	○

How similar are these two faces?

1 - Not at all similar	2	3	4	5 - Extremely similar
○	○	○	○	○

How similar are these two faces?

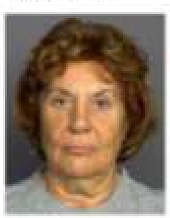
1 - Not at all similar	2	3	4	5 - Extremely similar
○	○	○	○	○

How similar are these two faces?

1 - Not at all similar	2	3	4	5 - Extremely similar
○	○	○	○	○

How similar are these two faces?

1 - Not at all similar	2	3	4	5 - Extremely similar
○	○	○	○	○

How similar are these two faces?

1 - Not at all similar	2	3	4	5 - Extremely similar
○	○	○	○	○

How similar are these two faces?

1 - Not at all similar	2	3	4	5 - Extremely similar
○	○	○	○	○

How similar are these two faces?

1 - Not at all similar	2	3	4	5 - Extremely similar
○	○	○	○	○

How similar are these two faces?

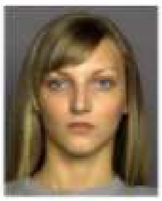
1 - Not at all similar	2	3	4	5 - Extremely similar
○	○	○	○	○

How similar are these two faces?

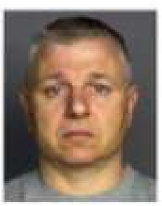
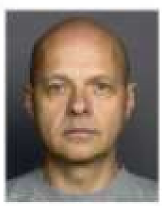
1 - Not at all similar	2	3	4	5 - Extremely similar
○	○	○	○	○

How similar are these two faces?

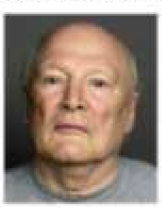
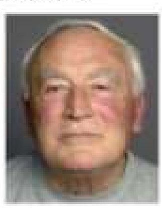
1 - Not at all similar	2	3	4	5 - Extremely similar
○	○	○	○	○

How similar are these two faces?

1 - Not at all similar	2	3	4	5 - Extremely similar
○	○	○	○	○

How similar are these two faces?

1 - Not at all similar	2	3	4	5 - Extremely similar
○	○	○	○	○

How similar are these two faces?

1 - Not at all similar	2	3	4	5 - Extremely similar
○	○	○	○	○

How similar are these two faces?

1 - Not at all similar	2	3	4	5 - Extremely similar
○	○	○	○	○

How similar are these two faces?

1 - Not at all similar	2	3	4	5 - Extremely similar
○	○	○	○	○

Thank you

Thank you! Thank you for participating in our study about face matching decision making. Research into face matching is important because face matching is used to verify a person's identity at a number of locations, such as when buying a controlled substance, when interacting with law enforcement, or when entering/leaving the country. Face matching of unfamiliar faces is notoriously poor. The goal of this research is to find out more about how observers make face matching decisions so that we can find ways to improve unfamiliar face matching accuracy. Data from the study will be available in the beginning of 2022. If you have any questions about the study or face matching, please email the student lead at danielle.rumschik@uoit.ca.

****CLICK THROUGH TO BE REDIRECTED AND RECIEVE YOUR CREDIT****

B2. Full Study Materials

Instructions

In the following section you will be asked to decide whether the presented photos show the same or different people. If you think the two photos show the same people, select the "Match" option. If you think the two photos show different people, select the "Mismatch" option. As you answer the questions, try to be as quick and as accurate as possible. Keep in mind that time may have passed between the two photos and that hairstyles, hair colors, and other aspects of the person's face, such as makeup, may have changed. Please do not zoom in while taking the survey.

Stimuli

Do the two photos show a match or a mismatch?


Match ○ Mismatch ○

What is 2+2?

- ○ 1
- ○ 4
- ○ 2
- ○ 5
- ○ 6

Do the two photos show a match or a mismatch?


Match ○ Mismatch ○

Do the two photos show a match or a mismatch?


Match ○ Mismatch ○

Do the two photos show a match or a mismatch?


Match ○ Mismatch ○

Do the two photos show a match or a mismatch?


Match ○ Mismatch ○

Do the two photos show a match or a mismatch?

Match ○ Mismatch ○

Do the two photos show a match or a mismatch?

Match ○ Mismatch ○

Do the two photos show a match or a mismatch?

Match ○ Mismatch ○

Do the two photos show a match or a mismatch?


Match ○ Mismatch ○

Do the two photos show a match or a mismatch?


Match ○ Mismatch ○

Do the two photos show a match or a mismatch?


Match ○ Mismatch ○

Do the two photos show a match or a mismatch?


Match ○ Mismatch ○

Do the two photos show a match or a mismatch?


Match ○ Mismatch ○

Do the two photos show a match or a mismatch?


Match ○ Mismatch ○

Do the two photos show a match or a mismatch?


Match ○ Mismatch ○

Do the two photos show a match or a mismatch?


Match ○ Mismatch ○

Do the two photos show a match or a mismatch?


Match ○ Mismatch ○

Do the two photos show a match or a mismatch?


Match ○ Mismatch ○

Do the two photos show a match or a mismatch?


Match ○ Mismatch ○

Do the two photos show a match or a mismatch?


Match ○ Mismatch ○

Do the two photos show a match or a mismatch?


Match ○ Mismatch ○

Do the two photos show a match or a mismatch?


Match ○ Mismatch ○

Do the two photos show a match or a mismatch?


Match ○ Mismatch ○

Do the two photos show a match or a mismatch?

Match

Mismatch

Demographics

What ethnicity/race do you most identify with?

- ○ African American/Black
- ○ Caucasian/White
- ○ East Asian
- ○ Arabic
- ○ Latinx
- ○ Asian
- ○ South Asian
- ○ Pacific Islander
- ○ First Nations/Inuit/Métis
- ○ Other
- ○ Prefer not to answer

How old are you? Please enter age below.

What gender do you most identify with?

- ○ Male
- ○ Female
- ○ Non-Binary
- ○ Transgender Male
- ○ Transgender Female
- ○ Gender non-conforming
- ○ Other
- ○ Prefer not to answer

Thank You

Thank you! Thank you for participating in our study about face matching decision making. Research into face matching is important because face matching is used to verify a person's identity at a number of locations, such as when buying a controlled substance, when interacting with law enforcement, or when entering/leaving the country. Face matching of unfamiliar faces is notoriously poor. The goal of this research is to find out more about how observers make face matching decisions so that we can find ways to improve unfamiliar face matching accuracy. Data from the study will be available in the beginning of 2022. If you have any questions about the study or face matching, please email the student lead at danielle.rumschik@uoit.ca.

****CLICK THROUGH TO BE REDIRECTED AND RECIEVE YOUR CREDIT*****