Facing Our Standards:

On the validity of standardised faces in face perception research

Friso Siemensma

6298559

f.e.f.c.siemensma@students.uu.nl

Applied Cognitive Psychology

Master thesis 27.5 ECTS

4-4-2023

First assessor: Dr. Ignace Hooge (i.hooge@uu.nl)

Second assessor: Dr. Roy Hessels (r.s.hessels@uu.nl)

Utrecht University

Word count: 9401

Abstract

Many studies in face perception research use standardised facial stimuli: digitally manipulated facial images used in eye tracking experiments studying facial viewing behaviour and emotion perception. With technological innovations in eye tracking equipment inviting us to study faces out in the wild, the present study investigated this seeming trend in research literature, identifying a predilection for stimulus control as an underlying line of reasoning for standardisation, alongside a potential problem if viewing behaviour showing visual preference for the eyes found in standardised stimuli, does not generalise to viewing behaviour showing visual preference for the eyes in non-standardised, "real" faces. As typical viewing behaviour in the literature shows the eyes have a strong attention maintaining capacity, a data analysis using a range of unusual non-standardised stimuli was conducted, estimating relative dwell time to the eyes. Results showed that while viewing behaviour differed between standardised and non-standardised stimuli, these differences were minor and did not substantially differ from results found in face perception literature. If additional studies confirm this generalisability, increasing the use of non-standardised stimuli could prove useful in in bridging the transition of face perception from the lab to the wild.

1. Introduction

Stepping into an full elevator on a busy day temporarily brings strangers into close proximity with one another. To ascertain whether it is safe to temporarily lock themselves in a little metal box with people they do not know, newcomers will furtively fixate their elevator companions' faces before stepping in, while those in the elevator do the same: can they trust these people? During the ride, some greetings and small talk may be exchanged until one by one each person exits the confines of the elevator.

During such a close proximity interaction, with multiple people standing face to face, it is exactly the face that provides the most informative cues for correctly assessing the situation, such as whether someone is young or old, what sex they belong to, what emotional state they are in and what their attitude is towards us, to name a few (Bruce & Young, 2012). By studying faces, people can inform their judgment as to whether their potential elevator companions are likely to be good company, or whether today it might be safer to take the stairs. With so much information packed in our facial features, it is no wonder that the face has been of much interest to social and behavioural sciences.

Faces have been studied for different purposes within different fields of research for years. At the start however, the cognitive elements of face perception were considered topics for cognitive psychology and neuropsychology, while the more social face perception elements were considered to fall under social psychology. Nowadays such categorisations have been fading, with researchers from either field taking interests in both the cognitive and social dimensions of face perception (Rhodes, Calder, Johnson, Haxby, 2012).[1] For example, the human ability to remember a great number of faces suggests that humans are capable of extracting and encoding the information that makes each perceived face unique. Because of this presumed capability, research in the field of engineering has put much effort into studying how people perceive faces, so that obtained knowledge may be applied to the development of computer algorithms for face recognition systems (O'Toole, 2012).

No matter what aspect of face perception researchers wish to study, face perception research has always primarily relied on photographs of faces as stimuli, with advancements in computer graphics in the last 40 years being accompanied by greater understanding of face perception, as

[1] Throughout the present study, "face perception" and "face perception research" are used as umbrella terms, used to refer to all elements of human facial cognition and all research employing facial stimuli or discussing the face and how faces are visually processed.

computer graphics have allowed researchers to vary and manipulate faces and facial patterns in ways that photographs alone simply do not allow for (Bruce & Young, 2012), as well as enabling computers to interface with eye tracking equipment (Rayner, 1998).

The ways in which faces are studied nowadays often employ eye tracking equipment to do so. Yet whereas eye tracking technology has changed quite a bit throughout the years, the present study contends that face stimuli have not changed in an comparable amount, and that much research studying face processing in essence still relies on photographs in the same way that face perception research started. In order to illustrate this, the place to start is the development of eye tracking.

1.1 An overview of eye tracking development

Studying how humans perceive faces begins with studying how humans perceive their environment in general. Human vision concerns peripheral vision, seeing without looking at anything in particular, and foveal vision, which concerns looking at particular things within the wider field of peripheral vision. To get an idea of someone's overt visual attention, three types of eye movements are particularly useful when using eye tracking methods: fixations, smooth pursuits and saccades.

Fixations are eye movements holding the eye's gaze aimed at a stationary object or detail in the environment. Smooth pursuits in contrast, are eye movements trying to hold the gaze on a moving object without interruption. Lastly, saccades are fast eye movements which consist of the eye moving from one fixation to another in one consecutive movement (Duchowski, 2017). All three types of eye movements employ the fovea, a small area at the centre of the eye where visual receptor cells have the highest density, resulting in the most high-resolution vision (Bahill & Stark, 1979). Not all mammals have fovea, but certain primates, including humans, do have it, allowing for high resolution colour vision (Bringmann et al., 2018), and through it we are able to employ fixations, smooth pursuits and saccades.

In order to track and measure these eye movements, researchers have deployed all kinds of eye trackers. Interest in the workings of human visual perception started quite early, with Rayner (1998) arguing that interest for eye movements in reading took off as early as 1879, and Dodge (1900) already discussing visual perception during eye movement at the start of the 20th century.

In 1935, Buswell published a book on eye movements when viewing pictures and geometrical patterns, using a camera recording corneal reflections as well as one for filming the few head

movements that were allowed. With his approach to eye tracking research, Buswell's methodology contributed greatly to the development of scan paths and heatmaps as we know them today, as well as discovering that participants show visual preference for stimuli areas with a higher density of information (Buswell, 1935; Wade, 2020).

Early implementations in the 1950s, most famously those of Yarbus (1967), often involved keeping participants' heads fixed with bite bars and placing mirror systems with suction caps and coiled scleral lenses placed directly on the eyes (Hayhoe & Ballard, 2005; Kowler, 2011; Tatler et al., 2010). Aside from his methods, Yarbus's research most notably confirmed one of Buswell's earlier observations, namely, that the instructions given to an observer could radically change the locations fixated (Buswell, 1935; Wade, 2020).

Although modern iterations of head-mounted eye trackers remain useful, it is fair to say that technological advancements have made modern eye tracking much less invasive and restrictive: the eye trackers most commonly used nowadays are non-invasive and often video-based, getting measurements by recording the reflections of the cornea and pupils with infrared cameras (Duchowski, 2017).

Coupled to these advancements is not only a decrease in invasive eye tracking methods, but also an increase in possibilities when designing eye tracking experiments, giving both the experimenter and the participant more freedom than before. Wearable eye trackers are a good example: Hayhoe and Ballard (2005) noted how portable eye trackers, with all the necessary equipment attached to the participant with little movement restrictions, allow for eye tracking during extended tasks in natural settings, in turn allowing for a greater variety of natural coordinated behaviours. With technology like portable eye trackers enabling a greater variety in experimental design, the possibilities on how to conduct modern eye tracking research are growing. However, when the ways in which we can research visual processing increase, it would be fair to assume an accompanying increase in the variety of stimuli presented for visual processing. Such variety however, in the use of facial stimuli specifically, has been lacking.

1.2 Face stimuli

In everyday interactions there are visual and non-visual differences between the faces observed, and faces presented as stimuli in face perception research: standardised face databases often showcase face stimuli with non-spontaneous, posed facial expressions, which coincidentally are frequently perceived as faked emotions (Dawel et al., 2017).

With image manipulation being self-evident since a couple of decades, researchers can present participants in eye tracking experiments with virtually any image, or tweak minute details to compare visual processing performance in seemingly identical pictures. Image manipulation to such extent comes in handy, for example, in the testing of advertisements, where eye tracking experiments allows for the testing of advertisement images with very subtle changes in design.

These possibilities for stimulus manipulation have also been recognised and capitalised upon intensively in face perception research: a great many face stimulus sets have been created over the years, such as the Pictures of Facial Affect (PoFA, 1976), the Karolinska Directed Emotional Faces (KDEF, 1998), the NimStim set of facial expressions (NimStim, 2009) and the Radboud Faces Database (RaFD, 2010).

While the previously described advancements in computer graphics provide an explanation for the heightened opportunity for altering stimuli, making it possible to manipulate virtually every aspect of stimulus imagery after collecting images or having pictures taken, this access to a greater degree of stimulus manipulation does not directly explain a trend of rigorous standardisation taking place through the creation of face stimulus databases in the last two decades. Actors are often asked to remove any distractors, or are excluded beforehand on the presence of those distractors, meaning face stimulus sets are often devoid of people shown with e.g. beards, glasses, facial blemishes and marks, and makeup (see **Figure 1**). After these requirements have been met, stimulus pictures, often of basic emotional expressions (e.g. fearful, angry, disgusted, cheerful) are taken, under specific lighting conditions, by instructing actors to pose their face in very specific ways to resemble the descriptions of these emotions seen on the face (Dawel et al., 2021).

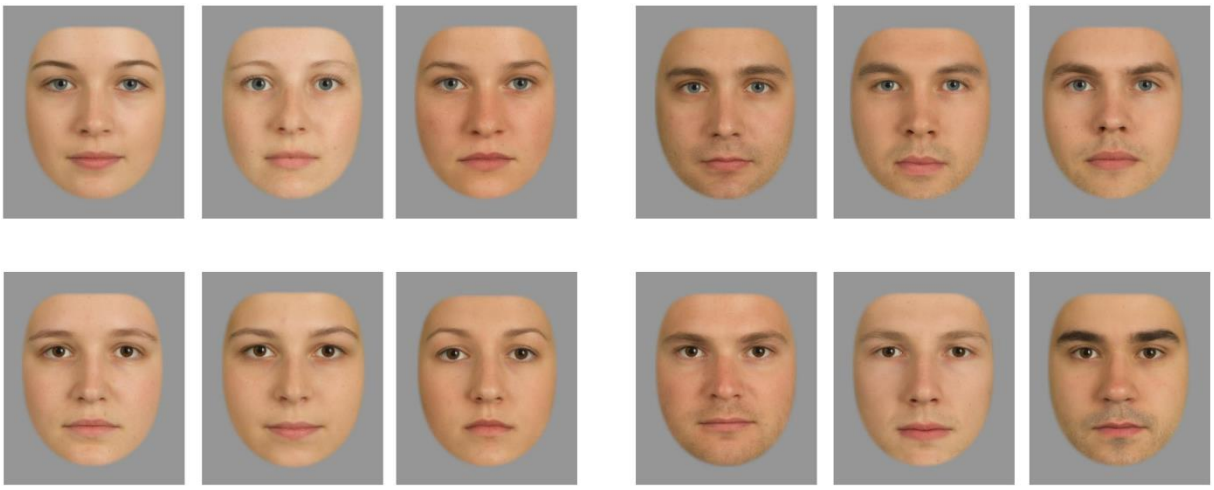

Figure 1 Examples of standardised face stimuli used in research. From "Using eye tracking to test for individual differences in attention to attractive faces" by C. Valuch et al., 2015, Frontiers in Psychology, 6:42. CC BY.

Now, this intensive standardisation of facial stimuli could suggest a desire for high quality data: in eye tracking research, this would mean data that has the least amount of spatial and temporal deviation between the actual gaze (subjective, as reported by the participant), and the gaze measured. This can be achieved by having a certain level of control over the research conditions, safeguarding data quality by keeping factors which may inadvertently influence the data to an acceptable minimum. For eye tracking specifically, inaccuracy and poor precision present two data quality issues which can render gathered data invalid (Holmqvist et al., 2012).

1.3 Structure of the present study

One way to try minimising systematic error and noise is through the use of a controlled lab environment. However, as the technological advancements in eye tracking research – and with it advancements in face perception research – have made it possible to do research beyond traditional lab constraints, the question arises whether standardised face stimuli may inadvertently limit the possibilities of face perception research.

If it is possible for a participant to comfortably go outside with a portable eye tracker registering their eye movements, it also becomes an inviting prospect to have those participants look at real faces outside, having face perception research go beyond the use of more artificial facial images it was forced to start out with due to technological constraints, and with it, going beyond the currently used standardised stimuli.

With technological innovations in eye tracking equipment inviting us to study faces out in the wild, the present study investigates the seeming trend of standardisation in research literature, in order to identify underlying line of reasoning for standardisation. Subsequently, a data analysis using a range of unusual non-standardised stimuli is conducted, to compare viewing behaviour between standardised and non-standardised facial stimuli through relative dwell time to the eyes.

2. Literature Study

With the context described above, the present study asked the following: what kind of standardised stimuli are used in face perception research, and what are the most salient considerations and questions in face perception research that motivate the use of standardised stimuli?

To answer these questions, a literature study was conducted, for which a selection of texts was compiled by searching internet databases with certain keywords. For this purpose, a relatively wide range of keywords was defined (see **Table 1**), based on known public stimulus sets and terminology observed in preliminary reading. Very broad and basic terms like *eye tracking* were also used, as standardised face stimuli could also be used as general eye tracking stimuli, outside of face perception research, and as literature not directly related to face perception research could still prove relevant through the use of, and reasoning for the use of, standardised face stimuli.

(ab)normal	cognition	facial	manipulated	PoFA	scan(ning)
affect	cognitive	fear(ful)	memory	processing	standardised
ambient	Dartmouth	fixation	movement	psychological	standardized
angry	database	gaze	MR2	psychology	static
artificial	distractor	generated	naturalistic	Qingdao	stimulus (set)
attention	dynamic	happy	neutral	Radboud	stimuli
attractive(ness)	ecological	image	NIMH-ChEFS	RADIATE	validation
avatar	emotion(al)	information	NimStim	RaFD	validity
Bogazici	expression	Karolinska	perception	realism	virtual
BP4D	eye (tracking)	KDEF	photo	realistic	
Chicago	face	manipulation	picture	recognition	

Table 1 Keywords and terms varied and combined in search engines.

After applying the keywords and terms above, the primary criterium for relevant literature was the following: texts gathered had to either mention or make use of a standardised face stimulus set. As identifying standardised stimuli concerned either their use or validation, two types of literature were chosen: regular research articles in which standardised stimuli were used for experimental design, and validation study articles in which stimulus sets were specifically discussed and validated. As identifying the use of standardised stimuli was the focus, the choice was made to compile a larger number of regular research articles using stimuli relative to a

smaller number of validation study articles. Additionally, as it was assumed that regular articles would understandably pay more attention to describing their own research than to describe the stimuli used, more regular research articles were used to compensate for an expected lack of information, compared to the validation studies, which specifically focused on the stimulus sets themselves.

The first phase consisted of identifying the stimulus sets used or mentioned in-text, assessing the degree of standardisation if applicable, and ascertaining whether the text authors had created the stimulus set. In identifying standardised stimuli in text and figures, the following passage from Dawel et al. (2021) proved useful for defining characteristics of standardised stimuli:

> Across psychology, there has been a longstanding tradition of using highly standardized face images. Face stimuli are typically shown under controlled lighting conditions, in frontal view, or a small number of other standard viewpoints. Some studies take this standardization a step further by editing out hair and identifying marks (e.g. moles). Additionally, emotional face images (e.g., happy, sad) are often generated primarily by asking models to pose expressions… (page 1).

This passage served as a baseline for categorising standardisation. Lighting as a category referred to lighting at the time of the photo or video shoot, as well as controlling lighting digitally afterwards. Facial angle was categorised as either frontal or alternative view. Editing the images, whether the removal of an specific feature or a complete oval crop around the face, was categorised as editing distractors. This also applied to model requirements before a shoot, e.g. researchers only wanting individuals lacking facial hair, piercings, glasses, etc. Additionally, whether stimuli were static or dynamic, coloured or grayscaled and posed or evoked was categorised as well. Expressions were "posed" when models were given very specific instructions on how their facial expression should appear, for example by using the Facial Action Coding System (FACS, Ekman & Friesen, 1978, revised and updated in 2002). Meanwhile, "evoked" meant cases in which the models were given instruction in acting out the emotion tied to the desired expression, or had the expression literally evoked through a joke, picture or other association. Finally, computer-generated and morphed stimuli were added as a category as well.

The second phase consisted of identifying reasoning for the use of standardised stimuli in-text. This included explicit reasoning, e.g. arguments made for or against standardisation measures, and implicit reasoning, e.g. assumptions deriving from certain statements and word choices.

2.1 Literature findings

In total, 50 articles were chosen which either provided images of the standardised stimuli their text mentioned, or which used standardised stimuli that could be viewed in other articles or viewed online, such as stimulus sets available for public use. This way the featured stimuli could be categorised in terms of standardisation based on written descriptions as well as image observations. These articles for the most part consisted of regular research articles (N = 35), and a number of validation study articles (N = 15), published between 2000 and 2021. Stimulus sets identified in these articles were created between 1976 and 2021 (see Table 2), with 23 being known stimulus sets available for public use, and 9 study-specific stimulus sets, created only for the research they appeared in.

Pictures Of Facial Affect (POFA, Ekman & Friesen, 1976)
Japanese And Caucasian Facial Expressions of Emotion (JACFEE, Matsumoto & Ekman, 1988)
Diagnostic Analysis of Non-verbal Accuracy-2, Adult Facial Expressions
(DANVA 2-AF, Nowicki & Duke, 1994)
Karolinska Directed Emotional Faces (KDEF, Lundqvist et al., 1998)
Chinese Faces (Wang & Markham, 1999)
The Montreal Set of Facial Displays of Emotion (MSFDE, Beaupré et al., 2000)
3D Facial Emotional Stimuli (Gur et al., 2002)
MIT-CBCL face recognition database (Weyrauch et al., 2004)
CAL/PAL Face Database (Minear & Park, 2004)
NimStim set of facial expressions (NimStim, Tottenham et al., 2009)
FACES Database (FACES, Ebner et al., 2010)
Radboud Faces Database (RaFD, Langner et al., 2010)
Amsterdam Dynamic Facial Expression Set (ADFES, van der Schalk et al., 2011)
NIMH Child Emotional Faces Picture Set (NIMH-ChEFS, Egger et al., 2011)
Umeå University Database of Facial Expressions (Samuelsson et al., 2012)
Dartmouth Database of Children's Faces (Dalrymple et al., 2013)
BP4D-Spontaneous (Zhang et al., 2014)
Chicago face database (CFD, Ma et al., 2015)
MR2 face database (MR2, Strohminger et al., 2016)
Warsaw Set of Emotional Facial Expression Pictures (WSEFEP, Olszanowski et al., 2015)
Bogazici face database (Saribay et al., 2018)
Racially Diverse Affective Expression face stimulus set (RADIATE, Conley et al., 2018)
Qingdao Preschooler Facial Expression Set (QPFE, Chen et al., 2021)

Table 2. Known stimulus sets identified.

Almost all articles featured frontal view stimuli exclusively, with only a handful of articles also employing alternative viewing angles. Methods of expression were relatively balanced, with posed and evoked expressions being documented in equal measure, as well as appearing simultaneously (e.g. cases in which both static and dynamic stimuli were used). While static stimuli were used predominantly, the categorised literature also featured some cases in which dynamic stimuli were used instead of or alongside static stimuli, with their use (or possible use in the future) argued for with a multitude of reasons: breaking with the apparent trend of only using static stimuli (Naples et al., 2014; Martin-Key et al., 2018; Bek et al., 2020), wishing to generalize results produced using static stimuli to dynamic ones as well (Martin-Key et al., 2018; Kaiser et al., 2019; Bek et al., 2020; Reisinger et al., 2020), increasing accuracy and neural activation in viewing participants when viewing dynamic stimuli as opposed to static ones (Trautmann et al., 2009; Van der Schalk et al., 2011; Bek et al., 2020), and lastly, providing 'ecological validity' (Van der Schalk et al., 2011; Trautmann et al., 2009; Martin-Key et al., 2018).

In many contemporary studies, ecological validity is brought up as a statement on the extent to which some aspect of experimental research resembles and generalises to the real-world (Holleman et al., 2020). Following Holleman et al. (2020), the present study considers ecological validity to be a rather vague term in its contemporary usage, requiring description of the context it supposedly refers to in order to be used constructively, if used at all.

Regardless of the meaning intended when mentioned, ecological validity often appeared as a prime concern among the literature employing static stimuli as well, with many of the research articles and validation studies considering a lack of it as a worry and a perceived increase of it as an asset (Goeleven et al., 2008; Schmid et al., 2011; Samuelsson et al., 2012; Lea et al., 2018). This interest in having stimuli be ecologically valid never seemed to be the only concern however, as it was often mentioned alongside with, and in contrast to, the importance of carefully controlled stimuli in research (Wieser et al., 2009; Langner et al., 2010; Naples et al., 2014; Strohminger et al., 2016). A good example is the considerations for the production of expressions in facial stimuli, mentioned in two validation studies: whether to have uniform posed expressions in the stimuli at the potential cost of ecological validity, posed through extensive instruction on which facial muscles to move, or to have more natural and authentic, but more varied emotional expressions by evoking them in the models (Tottenham et al., 2009; Samuelsson et al., 2012). While these are things to consider for every facial stimulus to be made

or used, stimulus sets containing posed expressions are noted to be standard in the field because of the heightened control they allow (Lewinski, 2015),[2] with this predominant measure of standardisation highlighting the priority of control over ecological validity in modern stimulus sets.

Reasoning found for the use of computer-generated stimuli emphasised this as well: while only a small number of the categorised literature featured computer-generated or morphed stimuli, both the texts that did not use them as well as the ones that did recognised that, while the ecological validity of computer-generated stimuli may be called into question (Goeleven et al., 2008; Crookes et al., 2015), this trade-off is generally accepted for increased control over stimuli standardisation (Gur et al., 2002; Wieser et al., 2009; Strohminger et al., 2016) to avoid potential confounds (Becker et al., 2011; Naples et al., 2014).

Around half of the categorised literature featured distractor edits, either beforehand, with models being required to not have facial hair, glasses, jewellery, makeup or piercings, or afterwards, with features digitally covered with an oval crop or edited out individually. Interestingly, while these distractor edits were often described in detail, these descriptions just as often lacked any reasoning for doing so (Goeleven et al., 2008; Hernandez et al., 2009; Sui & Lui, 2009; Haensel et al., 2012; Samuelsson et al., 2012; Pavlov et al., 2015; Olszanowski et al., 2015; Kaiser et al., 2019; Chen et al., 2021).

Some of the literature featured somewhat more implicit reasoning: removing jewellery deemed "excessive" or reducing hairstyles and makeup regarded as "eye-catching" (Ebner et al., 2008); cropping away hair and clothing or darkening "irrelevant" features or aspects around the face (Schmid et al., 2011; Macatee et al., 2016); to hide external features such as hair and ears (Wolf et al., 2014; Martin-Key et al., 2018), for example, for also deeming them "excessive" as non-facial features (Menks et al., 2021).

Now, the lack of clear reasoning for distractor edits could in part be explained through one of the tasks often present in experimental design in eye tracking: recognition tasks. Controlling for unique and distinguishing features that may confound recognition accuracy is highly

[2] While Lewinski argues that posed expressions are standard in facial stimuli, it should be noted that the ratio of posed to evoked expressions, as categorised in the literature overview of the present study, is more balanced: the explanation for this lies in the chosen definition for a posed or evoked expression. To illustrate, Lewinski considers the KDEF (1998) and WSEFEP (2015) stimuli as posed expressions, while this study considers them as evoked because of the freedom in expression still given to the models. As such, the present study uses a wider definition of evoked expressions.

desirable, with arguments and wording choice in a number of articles suggesting this as motivation. It is noted that editing included "prominent details" such as pimples, moles or gold teeth or the exclusion of bald headed models or models with braces (Ebner et al., 2010), emphasising stimulus control to minimise the possibility of individual faces being recognised on extra-facial cues such as hair or glasses (Dalrymple et al., 2013), removing accessories that would visually separate models from each other (Conley et al., 2018) or the research simply containing a recognition task (Crookes et al., 2015; Hunnikin et al., 2021).

Apart from specifically safeguarding the function of recognition tasks, distractor editing was also argued for with the mention of more general benefits, such as minimising variation in common databases to allow for more comparisons across studies (Ma et al., 2015), avoiding images that might look dated or fixed to a specific culture through facial features such as makeup, facial hair or hair styles (Strohminger et al., 2016) and isolating parameters of interest for more experimental control (Stephani et al., 2020).

With distractor editing, the balance of ecological validity and control over stimuli was emphasised again when categorised literature specifically mentioned not having edited out distractors, and subsequently argued how this might add to how natural or representative images in the database might be (Tottenham et al., 2009; Saribay et al., 2018), or that the risk of confounded results might have been increased while the stimuli themselves however might be closer to emotional faces in the real world (Bours et al., 2018).

2.2 Motivation for standardisation

Categorising the literature for standardisation measures found that standardised facial stimuli used in face perception research predominantly consisted of static frontal stimuli for which lighting conditions were almost always controlled during production or edited digitally afterwards. Posed and evoked expressions appeared in equal measure as well as alongside each other. Many stimuli also featured distractor edits through controlled exclusion or omittance of facial features such as (facial) hair, ears, pimples and moles, bald models, and extra-facial features such as glasses, jewellery, makeup or piercings, models wearing braces, and hairstyles. Outside of the facial area, clothing was often also controlled by either instructing models to wear certain types of clothing or by cropping the face as to exclude visibility of clothing.

Evaluating the explicit and implicit reasoning for doing so identified control over facial stimuli as the most salient consideration for standardising stimuli. The priority of control over stimuli before other concerns was most clearly accentuated through the concern for ecological validity,

mentioned in many of the texts which simultaneously considered increase or decrease in stimulus control, while arguments made for the use of computer-generated stimuli or editing out distractors provided further confirmation. As such, whether stimuli are ecologically valid enough presents one of the most salient questions asked when considering stimuli standardisation.

This consideration of control over stimuli through standardisation, conveys a very pragmatic mindset present in face perception research: many researchers are interested in identifying certain effects in viewing behaviour, and what is deemed necessary to elicit that effect is allowed to remain; any other variables are deemed either excessive, irrelevant, potentially confounding or distracting are cut out to isolate the desired effect. As such, researchers appear to be interested in identifying a standard, or general effect: an interest in the viewing behaviour elicited when viewing angry faces might best be identified through an oval crop of a posed angry expression without any hair, piercings or makeup, so that a standard angry face elicits only standard viewing behaviour when viewing angry faces. This pragmatic mindset influences how stimuli sets are developed and makes standardisation of stimuli the standard in face perception.

This identified trend of standardisation highlights two things that could prove problematic for face perception research in the long term. The first concerns the lack of non-standardised stimulus sets: while the preference for standardisation prioritises standard and general effects in viewing behaviour by controlling for unwanted details, there are many possible applications of face perception where it is exactly those facial details that matter. For example, law enforcement and forensics, where facial details such as scars, moles and other features can add to the training of automated facial recognition software (Leone, 2020), or healthcare aid applications where detailed facial information is essential for monitoring and diagnosing patients (Leo et al., 2020), improving access to healthcare in poor or remote communities.

The second potential issue concerns generalisation: with standardised facial stimuli as the norm in many fields in research, the viewing behaviour observed in many studies was produced with facial stimuli featuring a lack of extra-facial features (makeup, facial hair, piercings, glasses, etc.) as well as a lack of variety of facial expressions, through the preference for using posed expressions to produce emotional faces. As such, the viewing behaviour found in many studies was found using relatively uniform facial stimuli compared to the diversity of faces found in everyday life. The subsequent question for this issue, is whether the viewing behaviour found

using standardised stimuli is generalisable to non-standardised facial stimuli, namely, real faces, as well.

2.3 Testing for generalisability

The second part of this study investigates the potential issue of generalisation mentioned above through a data analysis of a 2018 study: an eye tracking experiment named "Face It" was conducted during the 2018 Betweter Festival in Utrecht. Like many other eye tracking experiments, participants were given a free viewing task of a random selection of stimuli. However, unlike most face stimulus sets commonly used for eye tracking research, the common denominator in the stimuli was that they mostly contained unusual or rather extreme faces, alongside presumably distracting extra-facial features in the form of glasses, piercings, makeup and the like (see **Figure 2**).

Figure 2 Examples of stimuli from the 2018 Face It experiment.

In face perception literature, a specific viewing pattern has often been observed in the viewing behaviour of participants presented with facial stimuli. When observing the face, fixations often cluster around the eyes, nose and mouth of facial stimuli in a cyclic manner (Tatler et al, 2010; Yarbus, 1967). With this characteristic triangular pattern being widely acknowledged within the research community, the data analysis of the present study asks where people look when viewing non-standardised faces, and to what extent this differs from standardised faces. To answer this question, areas of interest (AOIs) will be determined for the eyes, nose and mouth of the stimuli. For all AOIs, attention-maintaining capacity will be measured in relative dwell time, to observe whether the viewing behaviour differs between standardised and non-standardised faces.

Specifically, a focus has been placed on detecting differences in the attention-maintaining capacities of the eyes relative to the nose, mouth and the rest of the face. The eyes have been chosen because of the strong visual preference people's viewing behaviour expresses for the eyes when observes faces, with the eyes often being the first point of fixation (Birmingham et

al., 2009). With this widely observed urge for people to pay attention the eyes, the argument here is that any disruptions in viewing behaviour due to extreme face stimuli or disruptors within the stimuli will be most telling relative to the viewing behaviour in regard to the eyes. As such, the following null- and alternative hypotheses are posited:

H^0; The attention-maintaining capacities of the eyes measured in relative dwell time differ greatly between viewing behaviour observed in standardised faces relative to non-standardised faces.

H^1: The attention-maintaining capacities of the eyes measured in relative dwell time do not differ greatly between viewing behaviour observed in standardised faces relative to non-standardised faces.

3. Methods

3.1 Face It 2018

The Face It experiment was conducted at the "Betweter Festival" in Utrecht, a science and art festival hosted in collaboration with Utrecht University. 86 participants (47 women, 36 men) took part in the experiment, with an age range of 18 to 70 (M = 32.56). All participants were visitors of the festival who joined after giving written informed consent. This was after reading a briefing letter and viewing a print showing a sample with some of the more extreme stimuli, to be sure participants were aware of the possibility of seeing potentially startling images.

Participants were seated in an eye tracking booth with a chin-rest attached at the front of the desk, restricting viewer movement 65 cm in front of a Tobii TX300 eye tracker (Tobii Technology, Stockholm, Sweden, 300 Hz), mounted underneath a 23 inch screen (1920x1080 resolution, 60 Hz). Stimuli on the screen were running in PsychToolBox (version 3.0.11, Brainard 1997) in Matlab (version 2012b), alongside Tobii Pro Analytics SDK (version 3.0), all running on a MacBook Pro (2.8 GHz i7 processor, OS X 10.9) (Wijman, 2020).

For the stimuli presented, 280 images were gathered from the internet and compiled to form 35 categories of 8 images per category. Standardisation was only applied in the form of cropping all images to the same height and adding in grey areas to the sides. Facial size in the images varies, with pictures being taking from different distances, and varying in the amount of body included alongside the face.

The experiment consisted of both a free-viewing task and a recall task. The present study concerns only the former. For the free-viewing task, participants were presented a randomised selection of 24 images with a viewing duration of 3 seconds per image with a 1 second interval. Prior to each image, a fixation point was shown with random placement (Arizpe et al., 2012).

3.2 AOI method

For the present data analysis of the 2018 experiment data, the AOI method of choice was the limited-radius Voronoi tessellation (LRVT) method, due to its relative robustness to noise in face stimuli, as well as the relative ease which with the radii can be manipulated (see Hessels et al., (2016) for more details).

AOI cell centres were selected manually by mouse for the main features (eyes, nose, mouth), through Matlab, with the stimulus area outside of these AOIs being labelled the "none" AOI.

While most standardised face stimuli are produced in a controlled frontal perspective, the Face It stimuli also featured slightly turned faces. In order to account for this variation, placement coordinates for the eyes and nose were used to estimate the position of the nose between the distance from eye to eye. As one could expect the AOI centre of the nose to be in the middle of that distance on a frontal angle and to move away from the middle when turning, only stimuli of which the nose fell within 1 standard error of the mean distance of the eyes were included. Additionally, as the present study focused on attention maintaining capacity of the eyes, stimuli with only one eye such as cyclopes were left out. As such, 234 of the 280 stimuli were used for data analysis. For this selection, AOI radii could then be computed through Matlab, defining main feature AOIs for the eyes, nose and mouth, and non-AOIs as the space outside of those.

3.3 Categorisation & data quality

Apart from the NimStim, there were arguably no other stimuli included in the Face It set standardised to an extent comparable to standardisation commonly observed in the literature. In comparison, all other Face It stimuli would be viewed as unusual stimuli in face perception research. As such, in order to compare viewing behaviour from standardised facial stimuli with non-standardised ones using the Face It stimuli, one category of the original 35 was left out, while the rest were formed two sets for the present study: Normal and Abnormal, each with two subsets; Common and Atypical (see **Figure 3**).

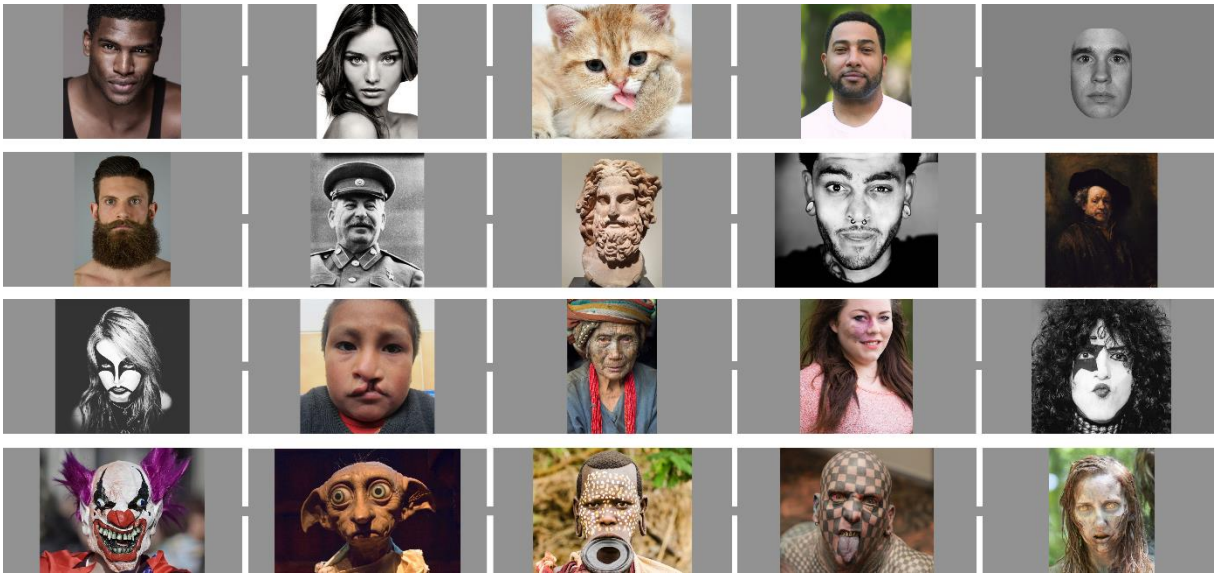

Figure 3 Examples from stimuli within the four categories from top to bottom: Normal-Common (N-C), Normal-Atypical (N-A), Abnormal-Common (A-C) and Abnormal-Atypical (A-A).

With every participant from the 2018 experiment viewing multiple stimuli, applying the aforementioned selection of only counting stimuli data with a sufficiently frontal angle resulted

in data from 86 participants totalling at 1690 lines of eye tracking data from included trials providing relative dwell time. Data loss was accounted for by accepting a maximum data loss of 20%, resulting in 1466 remaining lines, for which relative dwell time was recalculated to incorporate data loss. As data analysis after dividing the stimuli into four main categories required all participants viewing stimuli of all four categories, data from five participants having viewed stimuli from less than four categories was removed, leaving 1445 lines of data based on the viewing behaviour of 81 participants.

Relative dwell time data from the Face It stimuli was divided into four categories based on a subjective rating of (ab)normality and occurrence of the stimuli: Normal-Common (N-C), Normal-Atypical (N-A), Abnormal-Common (A-C) and Abnormal-Atypical (A-A). Stimuli categorisation was based on perceived (ab)normality and occurrence in daily life in a West-European provincial capital, encountered both in real life as well as via media in the form of news reports, online applications and films. A such, the subjective nature of this categorisation should be emphasised. Each of the four categories retained subcategories from the original 2018 Face It set, with multiple stimuli varying in image while having a common subject, such as different images from the 2009 NimStim set, different people with missing teeth, various attractive male and female models with makeup and a number of disturbing images depicting clowns, to name a few.

Categorising (ab)normality and occurrence concerned every stimulus as a whole as well as specific distractors it contained: Normal-Common (N-C) contained pictures of people featuring makeup, glasses, missing teeth, elderly faces and cats, as such variation could be considered both relatively normal as well as often occurring; glasses and makeup are used by both young and old, most families have elderly members, and many households have a cat seen in- and outside, while gap-toothed mouths are a common and understandable sight among both children as well as elderly.

In addition, the N-C category also included various stimuli from the NimStim database, with the following reasoning: being the only standardised stimuli in the Face It set, the NimStim stimuli represented standardised stimuli as they are commonly used in research, making the viewing behaviour from these specific stimuli salient for detecting potential differences in viewing behaviour between standardised and non-standardised stimuli. Placing the NimStim faces in the N-C category was done to accentuate findings, as any resulting differences between standardised and non-standardised stimuli could then be compared on a gradual scale, within the N-C category itself as well as specific stimuli from other categories.

Following N-C, category Normal-Atypical (N-A) featured bearded, large-nosed, pierced and infamous individuals alongside unusual haircuts and paintings and statues of people, as all these could be considered to occur less while still being relatively normal, while images of people, for example featured in advertisements, as digital avatars online or physically as statues or paintings, also naturally appear less often than we see real people whilst remaining a normal sight.

In contrast, categories Abnormal-Common and Abnormal-Atypical (A-C and A-A respectively) on the one hand featured burkas covering the face, intimidating black-and-white makeup for musical concerts, individuals with cleft lips, wine stains and facial injuries as faces which might be considered relatively abnormal without being all that rare, while masks, extensive facial tattooing, lip plates and horror- and fantasy related faces on the other hand would probably be considered as both abnormal as well as seldom occurring.

As each participant had viewed multiple stimuli belonging to each of the four main categories, a repeated measures one-way Analysis of Variance (ANOVA) was chosen for the first part of the data analysis, to establish whether the attention-maintaining capacity of the eyes differed in a statistically significant way between categories. In accordance with ANOVA requirements, normality of the data was estimated through Q-Q plots and checking skewness and kurtosis, with all values landing between 2 and -2. Independence was also met as all participants were randomly recruited at the 2018 Betweter festival in Utrecht, although this likely also meant most visitors were college-educated and living within the urban agglomeration. Performing Mauchly's Test of Sphericity also met the assumption of sphericity, validating the ANOVA results.

In order to compare viewing behaviour of standardised and non-standardised stimuli, the second part of the analysis focused on comparing average relative dwell time to the eyes for all original Face It subcategories within each of the four present categories, accompanied by scatterplots depicting viewing behaviour of a number of subcategories (**see Figure 4**). Gaze behaviour to the NimStim stimuli was used as a primary means of comparison, as they represented standardised stimuli as they are commonly used in face perception research, and could help spot noteworthy differences and similarities.

4. Results

Averages of relative dwell time on the eyes relative to the combined nose, mouth and none AOIs reported the highest attention maintaining capacity of all four categories (see Figure for N-C (M = 0.56, SD = 0.14), followed by A-C (M = 0.51, SD = 0.14), N-A (M = 0.45, SD = 0.16) and A-A (M = 0.44, SD = 0.15). Furthermore, a significant difference between the four categories was shown: $F(3, 240) = 30.63$, $p < .001$.

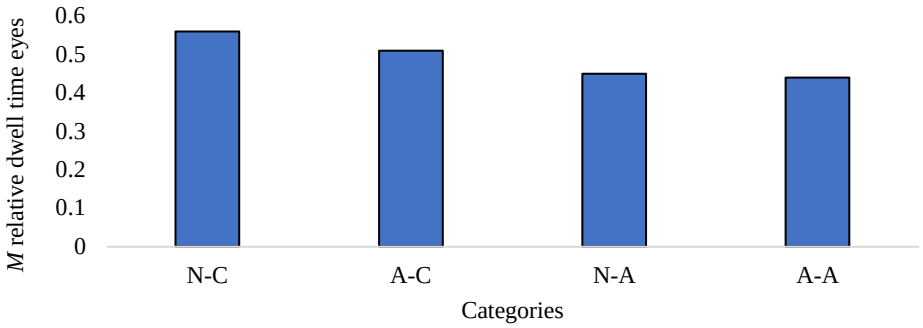

Figure 4 Average relative dwell time of the eyes per category in descending order.

Post-hoc analysis was conducted with Bonferroni-correction to see which categories specifically differed in a significant manner. Significant differences were reported for all categories compared to one another, except for categories N-A and A-A (p = 1.00).

4.1 Within categories

Looking within the four categories, starting with category Normal-Common, the NimStim subcategory had the strongest attention maintaining capacity of the eyes relative to the rest of the face in the N-C category (M = 0.61), followed closely stimuli featuring elderly people (M = 0.59) and pictures of people wearing glasses (M = 0.57). The rest of the stimuli subcategories descended gradually in terms of relative dwell time on the eyes, with the three subcategories with the lowest attention maintaining capacity of the eyes depicting a selection of pictures of everyday individuals dubbed as subcategory “Normalo” (M = 0.54), attractive male models (M= 0.54), with the only sudden in-category decrease to the eyes found in pictures of adults and children with missing teeth (M = 0.44).

Highest in category Normal-Atypical were stimuli featuring people with ninja masks (M = 0.59). The highest mean relative dwell time to the eyes after these were found in stimuli of people with large noses (M = 0.48) and individuals with nose piercings (M = 0.47), with a slow descent for the remaining subcategories, ending with people with mullet haircuts (M = 0.39),

people wearing blindfolds (*M* = 0.38) and stimuli of people eating scoring the lowest average relative dwell time (*M* = 0.37) within category N-A.

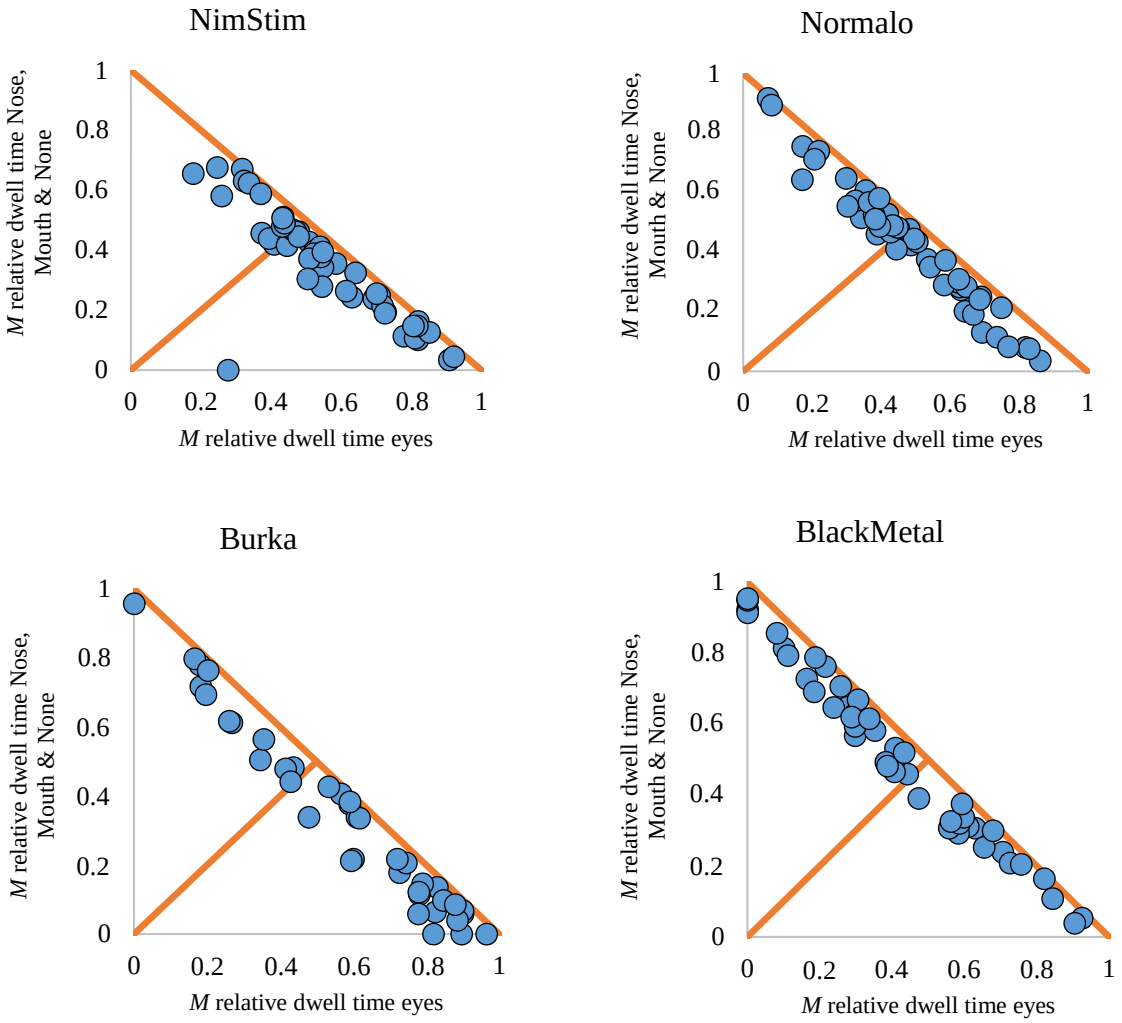

Figure 5 Relative dwell time of the eyes compared to the combined relative dwell time of the nose, mouth and none-AOIs per subcategory. Every blue point represents the relative dwell time for each participant presented with stimuli from a subcategory. Percentages of participants viewing the eyes relatively longer than the rest of the face were 70.59% for subcategory "NimStim" and 50.98% for "Normalo" (both belonging to category N-C), 66.67% for "Burka" and 39.58% for "BlackMetal" (both belonging to category A-C).

Within category Abnormal-Common, the three subcategories with the highest average relative dwell time were stimuli featuring burkas (*M* = 0.64), individuals with port-wine stain skin discolorations (*M* = 0.57) and people wearing face paint for rock music concerts (M = 0.54). The three lowest scoring subcategories within category A-C were formed by stimuli of people with cleft lip and cleft palate birth defects (*M* = 0.47), people with black and white face paint for black metal music concerts (*M* = 0.45) and people of different ethnicities with elaborate facial decorations (*M* = 0.43).

Mean relative dwell time to the eyes in the last category, Abnormal-Atypical, was highest for stimuli belonging to the horror genre (*M* = 0.50), disturbing pictures of clowns (*M* = 0.48) and various unsavoury looking film antagonists (*M* = 0.48). The lowest scoring stimuli were pictures depicting heavily tattooed faces *(M* = 0.42), masked individuals (*M* = 0.37) and people wearing lip plates (*M* = 0.36).

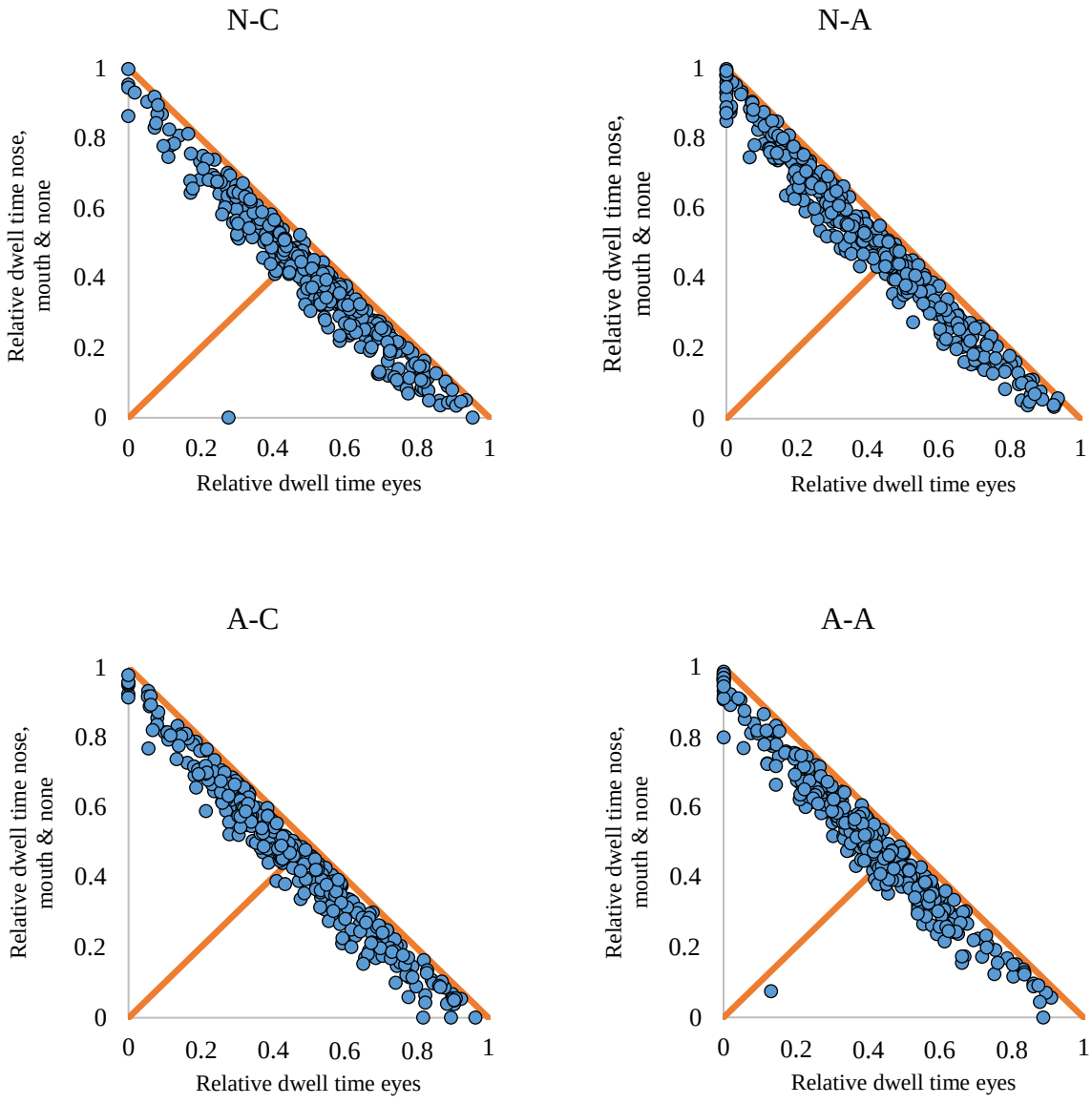

Figure 6 Relative dwell time of the eyes compared to combined relative dwell time of the nose, mouth and none-AOIs per category. Every blue point represents the relative dwell time for each participant presented with stimuli from a category. Percentages of participants viewing the eyes relatively longer than the rest of the face were as follows: N-C (60.84%), N-A (41.62%), A-C (48.73%) and lastly A-A (40.19%).

5. Discussion

Descriptive statistics accompanying the ANOVA imparted a number of insights. First, the most important ones will be summed up briefly, to be explained in more detail alongside the other findings below.

5.1 Most important findings

Arranging the four categories in descending order in terms of average relative dwell time to the eyes per category suggested the attention maintaining capacities of the eyes decrease as people view faces that can be considered less normal as well as less common within their sphere of perceived (ab)normality.

ANOVA results reported significant differences between the four main categories in terms of mean relative dwell time to the eyes. However, with the mean relative dwell time to the eyes of the NimStim subcategory being the representation for the viewing behaviour of standardised stimuli in this study, subcategory differences across all four main categories were deemed more relevant to ascertain whether viewing behaviour between standardised and non-standardised facial stimuli differed substantially. As expected, the NimStim subcategory featured a high average relative dwell time to the eyes, showing a preference for the eyes, and almost all other subcategories fell below this score. However, average relative dwell time across all subcategories showed a preference for the eyes, and in this aspect none of the subcategories deviated from the NimStim findings, or findings in literature in general, in a substantial way.

5.2 Overview of all results

Through average relative dwell time per category: with Normal-Common featuring the highest mean, followed by Abnormal-Common, Normal-Atypical and Abnormal-Atypical, there is a subtle decrease, suggesting that relative dwell time to the eyes decreases as participants are presented with facial stimuli which can be perceived as less normal as well as less common (see **Figure 4**). Interestingly, this observed sequence did not follow (ab)normality in the way the categories were created: the assumption when dividing the stimuli into four categories was that perceived normality of a face would precede the perceived occurrence of that face. Instead, with average relative dwell time higher for Abnormal-Common ($M = 0.51$) than for Normal-Atypical ($M = 0.45$), the observed pattern could suggest that while perceived normality of a face matters in the relative dwell time on the eyes in facial stimuli, it does not necessarily precede perceived occurrence in retaining higher relative dwell time to the eyes.

As reported, the ANOVA observed significant differences between all four categories, except for categories Normal-Atypical and Abnormal-Atypical. This could suggest a point of saturation, where increasing the rarity and abnormality of such faces as these two categories contained would no longer result in an accompanying decrease in relative dwell time to the eyes. Further research could investigate this in more detail.

Results between categories were quite interesting. Within category N-C the NimStim stimuli, as the only traditionally standardised stimuli in the set, performed best in terms of attention maintaining capacity of the eyes, while the other subcategories produced lower yet similar results in terms of relative average dwell time, with a gradual descent. The subcategory featuring missing teeth was the only exception, by showing an abrupt decrease from the previous results. With their relative average dwell time performance being the lowest in the category, the missing teeth displayed through broad smiles seem the likely explanation, as the relative average dwell time to the mouth AOI was highest of all subcategories across the main categories (M = 0.28).

For category N-A, faces covered with ninja masks featured the highest scores for the eyes, much higher than the rest. One possible explanation for this could be the fact that for most stimuli the nose and mouth were covered, which mean relative dwell time to the mouth seem to confirm (M = 0.04). The stimuli featuring large noses and nose piercings, while retaining relatively high dwell time to the eyes, also featured the highest (and same) average dwell time to the nose of all categories (M = 0.27). For people with mullet haircuts, belonging to the lowest performing N-A subcategories, it was interesting to observe that mean relative dwell time to the none AOI, meaning the area outside of the main facial features, was the highest across all four categories (M = 0.30), suggesting that unusual haircuts may have a strong attention maintaining capacity. Lastly, the lowest performing subcategory in terms of the eyes, featuring people eating, probably did so because of the attention maintaining capacity of the mouth AOI (M = 0.28) proved strong competition, suggesting that eating proves a strong potential distractor relative to the attention maintaining capacity of the eyes.

Category A-C featured the subcategory with the highest attention maintaining capacity across all categories with stimuli depicting burkas. With very low mean relative dwell time for both the nose as well as the mouth AOIs but a relatively high amount for the none AOI (M = 0.24), it seems as though most participants tried their best to catch any glimpse of facial or bodily details of any kind, and then focused their viewing efforts on the eye slit of the robes as they failed to observe any other features. Category A-C featured two subcategories with face painted

stimuli, entitled “RockFace” and “BlackMetal”. While it could be argued that the two were rather similar, they differed in terms of attention maintaining capacity, suggesting there may have been something causing a consistent difference in the attention maintaining capacity of the eyes. This could be the perhaps more threatening visages of the latter subcategory, but this would require further investigation.

Subcategory “Zombie” produced the highest average relative dwell time for the none AOI across all four categories (M = 0.27) while retaining a moderately high average relative dwell time for the eyes (M = 0.46). This could in some way be due to the unsettling nature of the stimuli, although it was not apparent in what way. Average relative dwell time for the none AOI was also high for stimuli showing lip plates (M = 0.24), the lowest performing subcategory in A-C. As the AOIs created with the LRVT method retained equal size and distance based on the main facial features while the lip plates’ size went beyond the mouth AOI, it seems likely that most of the average relative dwell time spent in the none AOI maintained attention on the lip plates.

While the ANOVA results reported significant differences between the four categories, it also mattered whether those results are relevant, both in terms of how the data looks as well as how it compares in context of other studies reporting relative dwell time to the eyes. Average relative dwell time percentages as well visualisations of the categories in scatterplots (see **Figure 6**) did indeed show differences, yet the question remained whether those differences necessarily suggest radically different capacities of maintaining attention to the eyes compared to standardised facial stimuli. Looking at the results from the subcategories, it seems clear that standardised facial stimuli such as the NimStim feature a strong attention maintaining capacity for the eyes. While most subcategories fell below this capacity in terms of average relative dwell time for maintaining attention to the eyes, almost all retained a strong preference for the eyes, while present differences rarely appeared major, with only a limited number of stimuli showcasing strong average relative dwell time for other specific AOIs beside the eyes. In the context of other research literature, neither the relative dwell time averages of the four categories nor of its subcategories appeared to deviate from results in face perception literature, while depicting a large number of highly unusual stimuli: a study investigating own-race bias reported an average dwell time percentage to the eyes of 50.7% for all their participants (Wu et al., 2012), while a study researching social phobias reported the average relative dwell time for the eyes at 60.3% for their control group (Moukheiber et al., 2010). Lastly, a study researching

differences in viewing behaviour between men and women reported dwell time percentages to the eyes of 26.39% for the men and 37.04% for the women (Hall et al., 2010).

5.3 Implications

Considering the relatively limited degree of variation found in the results, alongside variation found in face perception literature, the potential issue for generalisation identified in the literature study does not seem to be present: while relative dwell time to the eyes does seem to decrease as perceived abnormality and occurrence of observed non-standardised faces increases, this decrease appears gradual, while remaining relatively close to standardised ones. Or, put differently, the findings suggest that human facial processing might be more able to account for variation in faces and (extra) facial features than might have been expected, namely expectations that increased variation in faces and (extra-) facial features might disrupt typical facial viewing patterns.

If that is the case, researchers could be saved a lot of time and effort in regard to producing and standardising stimuli: facial models could be asked to meet less requirements, potentially speeding up the recruitment and photo production process, while more inclusivity due to less stringent requirements might allow for more representative stimuli samples, depending on the kind of representation desired. Post photo production, conceding less standardisation through photo selection and editing could increase the number of usable stimuli while potentially alleviating concerns for ecological, for example through increased variety, provided concerns are clearly defined per case (Holleman et al., 2020).

5.4 Limitations

Considering the data analysis focused primarily on relative dwell time to the eyes, future studies researching non-standardised stimuli could do so by focusing on alternative gaze patterns, for example when further investigating possible correlation between often observed facial viewing patterns and the perceived (ab)normality and occurrence of facial stimuli. Doing so would also allow for more comparisons with results found in face perception research.

It should also be emphasised that the four stimulus categories, based on perceived (ab)normality and occurrence, were categorised through a subjective reasoning process, meaning biases influenced judgement of classification. Future studies categorising stimuli or creating new stimulus sets might alleviate such influences by instead categorising via a more objective approach.

Considering the participant recruitment of the original 2018 Face It experiment at the Betweter Festival in Utrecht, it should also be taken into account that although onsite recruitment was random, visitors to the festival itself were likely to be college-educated and living in the urban agglomeration. Future studies could investigate whether the results of the present study could be generalised through the use of different population samples as well, either within different cities or different countries altogether.

The question of generalisability across different populations could also be of interest in regard to the Face It stimuli, as these consisted of images gathered from the internet. Using the internet as an international facial image database, new studies could also easily produce new variations of the original Face It stimulus set as well as expand on existing subcategories of interest. For example, the subcategory of female models wearing makeup could be expanded with images of models wearing makeup as worn across different parts of the world to compare viewing behaviour when observing faces wearing makeup.

6. Conclusion

The present study investigated the use of standardised stimuli in face perception research by means of a literature study to find a line of reasoning behind this trend. The literature study identified control over facial stimuli as the principal motivation for standardising stimuli, where the reasoning for the degree of control is often closely linked to reasoning for securing some degree of ecological validity. It also emphasised how this trend could pose an issue for generalisation if it turned out that viewing behaviour observed in standardised stimuli would not generalise to viewing behaviour in "real", non-standardised faces. A data analysis was conducted, using a range of highly unusual stimuli from the 2018 Face It experiment, and found that relative dwell time to the eyes did not differ greatly when observed viewing behaviour in standardised and non-standardised stimuli, suggesting that viewing behaviour from standardised facial stimuli may be generalisable to non-standardised faces. If other studies confirm the observed gaze behaviour found here, use of non- and less standardised stimuli could help in more efficient stimulus set production, as well as provide a way of bridging the knowledge gap concerning facial attention in natural settings (Varela et al., 2023).

References

Aguillon-Hernandez, N., Roché, L., Bonnet-Brilhault, F., Roux, S., Barthelemy, C., & Martineau, J. (2016). Eye Movement Monitoring and Maturation of Human Face Exploration. *Medical Principles and Practice*, *25*(6), 548–554. doi: 10.1159/000447971

Arizpe, J., Kravitz, D. J., Yovel, G., & Baker, C. I. (2012). Start Position Strongly Influences Fixation Patterns during Face Processing: Difficulties with Eye Movements as a Measure of Information Use. *PLOS One*, *7*(2), e31106. doi: 10.1371/journal.pone.0031106

Bahill, A. T., & Stark, L. (1979). The Trajectories of Saccadic Eye Movements. *Scientific American*, *240*(1), 108–117. doi: 10.1038/scientificamerican0179-108

Becker, D. V., Anderson, U. S., Mortensen, C. R., Neufeld, S. L., & Neel, R. (2011). The face in the crowd effect unconfounded: Happy faces, not angry faces, are more efficiently detected in single- and multiple-target visual search tasks. *Journal of Experimental Psychology: General*, *140*(4), 637–659. doi: 10.1037/a0024060

Bek, J., Poliakoff, E., & Lander, K. (2020). Measuring emotion recognition by people with Parkinson's disease using eye-tracking with dynamic facial expressions. *Journal of Neuroscience Methods*, *331*, 108524. doi: 10.1016/j.jneumeth.2019.108524

Billeci, L., Muratori, P., Calderoni, S., Chericoni, N., Levantini, V., Milone, A., Nocentini, A., Papini, M., Ruglioni, L., & Dadds, M. (2019). Emotional processing deficits in Italian children with Disruptive Behavior Disorder: The role of callous unemotional traits. *Behaviour Research and Therapy*, *113*, 32–38. doi: 10.1016/j.brat.2018.12.011

Birmingham, E., Bischof, W. F., & Kingstone, A. (2009). Saliency does not account for fixations to eyes within social scenes. *Vision Research*, *49*(24), 2992–3000. doi: 10.1016/j.visres.2009.09.014

Bodenschatz, C. M., Kersting, A., & Suslow, T. (2019). Effects of Briefly Presented Masked Emotional Facial Expressions on Gaze Behavior: An Eye-Tracking Study. *Psychological Reports*, *122*(4), 1432–1448. doi: 10.1177/0033294118789041

Bours, C. C. A. H., Bakker-Huvenaars, M. J., Tramper, J., Bielczyk, N., Scheepers, F., Nijhof, K. S., Baanders, A. N., Lambregts-Rommelse, N. N. J., Medendorp, P., Glennon, J. C., & Buitelaar, J. K. (2018). Emotional face recognition in male adolescents with autism spectrum disorder or disruptive behavior disorder: an eye-tracking study.

European Child &Amp; Adolescent Psychiatry, *27*(9), 1143–1157. doi: 10.1007/s00787-018-1174-4

Bringmann, A., Syrbe, S., Görner, K., Kacza, J., Francke, M., Wiedemann, P., & Reichenbach, A. (2018). The primate fovea: Structure, function and development. *Progress in Retinal and Eye Research*, *66*, 49–84. doi: 10.1016/j.preteyeres.2018.03.006

Bruce, V., Young, A. (2012). *Face perception* (1st ed.). Psychology Press. doi: 10.4324/9780203721254

Calder, A. J., Keane, J., Manes, F., Antoun, N., & Young, A. W. (2000). Impaired recognition and experience of disgust following brain injury. *Nature Neuroscience*, *3*(11), 1077–1078. doi: 10.1038/80586

Capriola-Hall, N. N., Ollendick, T. H., & White, S. W. (2021). Attention Deployment to the Eye Region of Emotional Faces among Adolescents with and without Social Anxiety Disorder. *Cognitive therapy and research*, *45*(3), 456–467. doi: 10.1007/s10608-020-10169-2

Chen, J., Zhang, Y., & Zhao, G. (2021). The Qingdao Preschooler Facial Expression Set: Acquisition and Validation of Chinese Children's Facial Emotion Stimuli. *Frontiers in Psychology*, *11*. doi: 10.3389/fpsyg.2020.554821

Conley, M. I., Dellarco, D. V., Rubien-Thomas, E., Cohen, A. O., Cervera, A., Tottenham, N., & Casey, B. (2018). The racially diverse affective expression (RADIATE) face stimulus set. *Psychiatry Research*, *270*, 1059–1067. doi: 10.1016/j.psychres.2018.04.066

Crookes, K., Ewing, L., Gildenhuys, J. D., Kloth, N., Hayward, W. G., Oxner, M., Pond, S., & Rhodes, G. (2015). How Well Do Computer-Generated Faces Tap Face Expertise? *PLOS One*, *10*(11). doi: 10.1371/journal.pone.0141353

Dalrymple, K. A., Gomez, J., & Duchaine, B. (2013). The Dartmouth Database of Children's Faces: Acquisition and Validation of a New Face Stimulus Set. *PLOS One*, *8*(11). doi: 10.1371/journal.pone.0079131

Dawel, A., Miller, E.J., Horseburgh, A., & Ford, P. (2021). A systematic survey of face stimuli used in psychological research 2000-2020. *Behaviour Research Methods* (2021). doi: 10.3758/s13428-021-01705-3

Dawel, A., Wright, L., Irons, J., Dumbleton, R., Palermo, R., O'Kearney, R., & McKone, E. (2017). Perceived emotion genuineness: normative ratings for popular facial expression stimuli and the development of perceived-as-genuine and perceived-as-

fake sets. *Behavior Research Methods*, *49*(4), 1539–1562. doi: 10.3758/s13428-016-0813-2

Dodge, R. (1900). Visual perception during eye movement. *Psychological Review*, *7*(5), 454–465. doi: 10.1037/h0067215

Duchowski, A. T. (2017). *Eye Tracking Methodology: Theory and Practice* (3rd ed.). Springer Publishing. doi: 10.1007/978-3-319-57883-5

Eberhardt, L. V., Huckauf, A., & Kliegl, K. M. (2016). Effects of Neutral and Fearful Mood on Duration Estimation of Neutral and Fearful Face Stimuli. *Timing & Amp; Time Perception*, *4*(1), 30–47. doi: 10.1163/22134468-00002060

Ebner, N. C. (2008). Age of face matters: Age-group differences in ratings of young and old faces. *Behavior Research Methods*, *40*(1), 130–136. doi: 10.3758/brm.40.1.130

Ebner, N. C., Riediger, M., & Lindenberger, U. (2010). FACES—A database of facial expressions in young, middle-aged, and older women and men: Development and validation. *Behavior Research Methods*, *42*(1), 351–362. doi: 10.3758/brm.42.1.351

Egger, H. L., Pine, D. S., Nelson, E., Leibenluft, E., Ernst, M., Towbin, K. E., & Angold, A. (2011). The NIMH Child Emotional Faces Picture Set (NIMH-ChEFS): a new set of children's facial emotion stimuli. *International Journal of Methods in Psychiatric Research*, *20*(3), 145–156. doi: 10.1002/mpr.343

Ellingsen, E. F., Drevsjø, S., Volden, F., & Watten, R. G. (2019). Extraversion and focus of attention on facial emotions: an experimental eye-tracking study. *Current Issues in Personality Psychology*, *7*(2), 91–97. doi: 10.5114/cipp.2019.85413

Goeleven, E., De Raedt, R., Leyman, L., & Verschuere, B. (2008). The Karolinska Directed Emotional Faces: A validation study. *Cognition & Amp; Emotion*, *22*(6), 1094–1118. doi: 10.1080/02699930701626582

Gur, R. C., Sara, R., Hagendoorn, M., Marom, O., Hughett, P., Macy, L., Turner, T., Bajcsy, R., Posner, A., & Gur, R. E. (2002). A method for obtaining 3-dimensional facial expressions and its standardization for use in neurocognitive studies. *Journal of Neuroscience Methods*, *115*(2), 137–143. doi: 10.1016/s0165-0270(02)00006-7

G. T. (1935). *How people look at pictures: a study of the psychology and perception in art.* Univ. Chicago Press.

Haensel, J. X., Ishikawa, M., Itakura, S., Smith, T. J., & Senju, A. (2020). Cultural influences on face scanning are consistent across infancy and adulthood. *Infant Behavior and Development*, *61*, 101503. doi: 10.1016/j.infbeh.2020.101503

Hall, J., Hutton, S. B., & Morgan, M. J. (2010). Sex differences in scanning faces: Does attention to the eyes explain female superiority in facial expression recognition? *Cognition & Emotion*, *24*(4), 629–637. doi: 10.1080/02699930902906882

Hayhoe, M., & Ballard, D. (2005). Eye movements in natural behavior. *Trends in Cognitive Sciences*, *9*(4), 188–194. doi: 10.1016/j.tics.2005.02.009

Hernandez, N., Metzger, A., Magné, R., Bonnet-Brilhault, F., Roux, S., Barthelemy, C., & Martineau, J. (2009). Exploration of core features of a human face by healthy and autistic adults analyzed by visual scanning. *Neuropsychologia*, *47*(4), 1004–1012. doi: 10.1016/j.neuropsychologia.2008.10.023

Hessels, R. S., Kemner, C., van den Boomen, C., & Hooge, I. T. C. (2015). The area-of-interest problem in eyetracking research: A noise-robust solution for face and sparse stimuli. *Behavior Research Methods*, *48*(4), 1694–1712. doi: 10.3758/s13428-015-0676-y

Holas, P., Krejtz, I., Cypryanska, M., & Nezlek, J. B. (2014). Orienting and maintenance of attention to threatening facial expressions in anxiety – An eye movement study. *Psychiatry Research*, *220*(1–2), 362–369. doi: 10.1016/j.psychres.2014.06.005

Holleman, G. A., Hooge, I. T. C., Kemner, C., & Hessels, R. S. (2020). The 'Real-World Approach' and Its Problems: A Critique of the Term Ecological Validity. *Frontiers in Psychology*, *11*. doi: 10.3389/fpsyg.2020.00721

Holmqvist, K., Nyström, M., & Mulvey, F. (2012). Eye tracker data quality: What it is and how to measure it. *Proceedings of the Symposium on Eye Tracking Research and Applications*. doi: 10.1145/2168556.2168563

Hunnikin, L. M., Wells, A. E., Ash, D. P., & van Goozen, S. H. M. (2021). Can facial emotion recognition be rapidly improved in children with disruptive behavior? A targeted and preventative early intervention study. *Development and Psychopathology*, *34*(1), 85–93. doi: 10.1017/s0954579420001091

Kaiser, D., Jacob, G. A., van Zutphen, L., Siep, N., Sprenger, A., Tuschen-Caffier, B., Senft, A., Arntz, A., & Domes, G. (2019). Biased Attention to Facial Expressions of Ambiguous Emotions in Borderline Personality Disorder: An Eye-Tracking Study. *Journal of Personality Disorders*, *33*(5), 671-S8. doi: 10.1521/pedi_2019_33_363

Kowler, E. (2011). Eye movements: The past 25 years. *Vision Research*, *51*(13), 1457–1483. doi: 10.1016/j.visres.2010.12.014

Langner, O., Dotsch, R., Bijlstra, G., Wigboldus, D. H. J., Hawk, S. T., & van Knippenberg, A. (2010). Presentation and validation of the Radboud Faces Database. *Cognition &Amp; Emotion*, *24*(8), 1377–1388. doi: 10.1080/02699930903485076

Lea, R. G., Qualter, P., Davis, S. K., Pérez-González, J. C., & Bangee, M. (2018). Trait emotional intelligence and attentional bias for positive emotion: An eye tracking study. *Personality and Individual Differences*, *128*, 88–93. doi: 10.1016/j.paid.2018.02.017

Leo, M., Carcagnì, P., Mazzeo, P. L., Spagnolo, P., Cazzato, D., & Distante, C. (2020). Analysis of Facial Information for Healthcare Applications: A Survey on Computer Vision-Based Approaches. *Information*, *11*(3), 128. doi: 10.3390/info11030128

Leone, M. (2020). From Fingers to Faces: Visual Semiotics and Digital Forensics. *International Journal for the Semiotics of Law - Revue Internationale De Sémiotique Juridique*, *34*(2), 579–599. doi: 10.1007/s11196-020-09766-x

Lewinski, P. (2015). Automated facial coding software outperforms people in recognizing neutral faces as neutral from standardized datasets. *Frontiers in Psychology*, *6*. doi: 10.3389/fpsyg.2015.01386

Ma, D. S., Correll, J., & Wittenbrink, B. (2015). The Chicago face database: A free stimulus set of faces and norming data. *Behavior Research Methods*, *47*(4), 1122–1135. doi: 10.3758/s13428-014-0532-5

Macatee, R. J., Albanese, B. J., Schmidt, N. B., & Cougle, J. R. (2017). The moderating influence of heart rate variability on stressor-elicited change in pupillary and attentional indices of emotional processing: An eye-Tracking study. *Biological psychology*, *123*, 83–93. doi: 10.1016/j.biopsycho.2016.11.013

Madera, J. M., & Hebl, M. R. (2012). Discrimination against facially stigmatized applicants in interviews: An eye-tracking and face-to-face investigation. *Journal of Applied Psychology*, *97*(2), 317–330. doi: 10.1037/a0025799

Menks, W. M., Fehlbaum, L. V., Borbás, R., Sterzer, P., Stadler, C., & Raschle, N. M. (2021). Eye gaze patterns and functional brain responses during emotional face processing in adolescents with conduct disorder. *NeuroImage: Clinical*, *29*, 102519. doi: 10.1016/j.nicl.2020.102519

Moukheiber, A., Rautureau, G. J. P., Perez-Diaz, F., Soussignan, R., Dubal, S., Jouvent, R., & Pelissolo, A. (2010). Gaze avoidance in social phobia: Objective measure and correlates. *Behaviour Research and Therapy*, *48*(2), 147–151. doi: 10.1016/j.brat.2009.09.012

Naples, A., Nguyen-Phuc, A., Coffman, M., Kresse, A., Faja, S., Bernier, R., & McPartland, J. C. (2015). A computer-generated animated face stimulus set for psychophysiological research. *Behavior Research Methods*, *47*(2), 562–570. doi: 10.3758/s13428-014-0491-x

Olszanowski, M., Pochwatko, G., Kuklinski, K., Scibor-Rylski, M., Lewinski, P., & Ohme, R. K. (2015). Warsaw set of emotional facial expression pictures: a validation study of facial display photographs. *Frontiers in Psychology*, *5*. doi: 10.3389/fpsyg.2014.01516

O'Toole, A. (2012). Cognitive and computational approaches to face recognition. Rhodes, G., Calder, A., Johnson, M., & Haxby, J.V. (Eds.), *The Oxford handbook of face perception* (pp. 43-56). Oxford University Press. doi: 10.1093/oxfordhb/9780188559053.001.0001

Pavlov, S. V., Korenyok, V. V., Reva, N. V., Tumyalis, A. V., Loktev, K. V., & Aftanas, L. I. (2015). Effects of long-term meditation practice on attentional biases towards emotional faces: An eye-tracking study. *Cognition and Emotion*, *29*(5), 807–815. doi: 10.1080/02699931.2014.945903

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, *124*(3), 372–422. doi: 10.1037/0033-2909.124.3.372

Reisinger, D. L., Shaffer, R. C., Horn, P. S., Hong, M. P., Pedapati, E. V., Dominick, K. C., & Erickson, C. A. (2020). Atypical Social Attention and Emotional Face Processing in Autism Spectrum Disorder: Insights From Face Scanning and Pupillometry. *Frontiers in Integrative Neuroscience*, *13*. doi: 10.3389/fnint.2019.00076

Rhodes, G., Calder, A., Johnson, M., & Haxby, J.V. (2012). Preface. Rhodes, G., Calder, A., Johnson, M., & Haxby, J.V. (Eds.), *The Oxford handbook of face perception* (pp. 6-9). Oxford University Press. doi: 10.1093/oxfordhb/9780188559053.001.0001

Samuelsson, H., Jarnvik, K., Henningsson, H., Andersson, J., & Carlbring, P. (2012). The Umeå University Database of Facial Expressions: A Validation Study. *Journal of Medical Internet Research*, *14*(5), e136. doi: 10.2196/jmir.2196

Saribay, S. A., Biten, A. F., Meral, E. O., Aldan, P., Třebický, V., & Kleisner, K. (2018). The Bogazici face database: Standardized photographs of Turkish faces with supporting materials. *Plos ONE*, *13*(2), e0192018. doi: 10.1371/journal.pone.0192018

Schmid, P. C., Schmid Mast, M., Bombari, D., Mast, F. W., & Lobmaier, J. S. (2011). How Mood States Affect Information Processing During Facial Emotion Recognition: An Eye Tracking Study. *Swiss Journal of Psychology*, *70*(4), 223–231. doi: 10.1024/1421-0185/a000060

Stanley, J. T., Zhang, X., Fung, H. H., & Isaacowitz, D. M. (2013). Cultural differences in gaze and emotion recognition: Americans contrast more than Chinese. *Emotion*, *13*(1), 36–46. doi: 10.1037/a0029209

Stephani, T., Kirk Driller, K., Dimigen, O., & Sommer, W. (2020). Eye contact in active and passive viewing: Event-related brain potential evidence from a combined eye tracking and EEG study. *Neuropsychologia*, *143*, 107478. doi: 10.1016/j.neuropsychologia.2020.107478

Strnádelová, B., Halamová, J., & Kanovský, M. (2019). Eye-tracking of Facial Emotions in Relation to Self-criticism and Self-reassurance. *Applied Artificial Intelligence*, *33*(10), 839–862. doi: 10.1080/08839514.2019.1646004

Strohminger, N., Gray, K., Chituc, V., Heffner, J., Schein, C., & Heagins, T. B. (2016). The MR2: A multi-racial, mega-resolution database of facial stimuli. *Behavior Research Methods*, *48*(3), 1197–1204. doi: 10.3758/s13428-015-0641-9

Sui, J., & Liu, C. H. (2009). Can beauty be ignored? Effects of facial attractiveness on covert attention. *Psychonomic Bulletin &Amp; Review*, *16*(2), 276–281. doi: 10.3758/pbr.16.2.276

Tatler, B. W., Wade, N. J., Kwan, H., Findlay, J. M., & Velichkovsky, B. M. (2010). Yarbus, Eye Movements, and Vision. *I-Perception*, *1*(1), 7–27. doi: 10.1068/i0382

Tottenham, N., Tanaka, J. W., Leon, A. C., McCarry, T., Nurse, M., Hare, T. A., Marcus, D. J., Westerlund, A., Casey, B., & Nelson, C. (2009). The NimStim set of facial expressions: Judgments from untrained research participants. *Psychiatry Research*, *168*(3), 242–249. doi: 10.1016/j.psychres.2008.05.006

Trautmann, S. A., Fehr, T., & Herrmann, M. (2009). Emotions in motion: Dynamic compared to static facial expressions of disgust and happiness reveal more widespread emotion-specific activations. *Brain Research*, *1284*, 100–115. doi: 10.1016/j.brainres.2009.05.075

Valuch, C., Pflüger, L. S., Wallner, B., Laeng, B., & Ansorge, U. (2015). Using eye tracking to test for individual differences in attention to attractive faces. *Frontiers in Psychology*, *6*. doi: 10.3389/fpsyg.2015.00042

van der Schalk, J., Hawk, S. T., Fischer, A. H., & Doosje, B. (2011). Moving faces, looking places: Validation of the Amsterdam Dynamic Facial Expression Set (ADFES). *Emotion*, *11*(4), 907–920. doi: 10.1037/a0023853

Varela, V.P.L., Towler, A., Kemp, R.I. *et al.* Looking at faces in the wild. *Sci Rep* **13**, 783 (2023). doi: 10.1038/s41598-022-25268-1

Wade, N. J. (2020). Looking at Buswell's pictures. *Journal of Eye Movement Research*, *13*(2). doi: 10.16910/jemr.13.2.4

Wieser, M. J., Pauli, P., Alpers, G. W., & Mühlberger, A. (2009). Is eye to eye contact really threatening and avoided in social anxiety?—An eye-tracking and psychophysiology study. *Journal of Anxiety Disorders*, *23*(1), 93–103. doi: 10.1016/j.janxdis.2008.04.004

Wijmans, M. (2020). Extreme faces: An exploratory eye-tracking study into different types of facial stimuli.

Wolf, R. C., Philippi, C. L., Motzkin, J. C., Baskaya, M. K., & Koenigs, M. (2014). Ventromedial prefrontal cortex mediates visual attention during facial emotion recognition. *Brain*, *137*(6), 1772–1780. doi: 10.1093/brain/awu063

Wu, E. X. W., Laeng, B., & Magnussen, S. (2012). Through the eyes of the own-race bias: Eye-tracking and pupillometry during face recognition. *Social Neuroscience*, *7*(2), 202–216. doi: 10.1080/17470919.2011.596946

Yarbus, A. L. (1967). *Eye Movements and Vision* (1st ed.). Springer.

Zhang, X., Yin, L., Cohn, J. F., Canavan, S., Reale, M., Horowitz, A., Liu, P., & Girard, J. M. (2014). BP4D-Spontaneous: a high-resolution spontaneous 3D dynamic facial expression database. *Image and Vision Computing*, *32*(10), 692–706. doi: 10.1016/j.imavis.2014.06.002