THE UNIVERSITY OF HULL

Rapid Detection of Human Facial Attractiveness in Groups

being a Thesis submitted for the Degree of

Doctor of Philosophy

in the University of Hull

by

Richard J. Carvey

MSc Foundations of Clinical Neuropsychology, University of Wales, Bangor

Supervisors: Professor Chang Hong Liu, Dr Tjeerd Jellema, and Dr Kazuyo Nakabayashi

May 2017

Abstract

In a world full of great visual repetition, humans have evolved to simplify visual processing, taking redundant information and compressing it into a simpler form (Alvarez, 2011). This compressed form is an ensemble representation, an abstract singular entity that conveys the relevant information about its constituents.

Haberman & Whitney (2009) demonstrated that even with stimuli as complex as human faces, and specifically their emotional expressions, such a representation can be generated, and the mean expression of a group can be accurately identified from brief presentations. Other research has shown that the attractiveness of faces can be rapidly assessed from very brief exposures (Olson & Marshuetz, 2005; Willis & Todorov, 2006), but this has not considered more than a single face in a presentation. Those that have, only considered estimates of frequency of attractiveness comparing between brief exposures and longer presentation times, not taking into account how accurate these estimates were.

The aim of this thesis was to explore the accuracy with which participants could judge the attractiveness of a group of faces, either as a two-alternative-forced-choice task judging which of two groups contained more attractive faces, whether a single group contained more attractive or more unattractive faces, and estimating the number of attractive faces in a group. The results showed that the judgements of attractiveness were accurate from brief exposures, but this judgement was modulated partially by the task at hand. This modulation was further explored by comparing various ratings of attractiveness of the groups, and suggested that the ensemble representation might be formed by some combination of statistical and visual averaging. Finally, the use of eye-tracking technology showed no bias in visual attention towards more attractive faces, and that fixation duration patterns were, to some extent, also modulated by the task.

Acknowledgements

I would like to thank Professor Chang Hong Liu for his input, guidance, and support during our time working together. Even after moving to the south coast, he was still able and willing to provide input and feedback for my work. His trust in my abilities, and his willingness to vouch for me for the second round of funding that meant I could actually undertake my PhD studies, have, I hope, been vindicated. I would also like to thank Dr Tjeerd Jellema, who stepped up to the role of primary supervisor and, despite this not being his area of expertise, was still able to provide useful feedback and support.

I also owe a debt of thanks to Elisabeth Cratchley at Selby College, who has provided moral support to help me push through this work, and has given flexibility where possible to allow me time to work on it - this could have been a lot more difficult! Some other staff at the college also deserve a mention; thank you to Keri O'Shea, Lucy Stanworth, Emily Davies, and Shelley Bratton.

I must thank my family for their belief in my ability to achieve this, and their tolerance of me while I attempted it. It is easy to forget that life goes on, for yourself and others, when inside this bubble, but there have been moments that have strongly reminded me of that. Harry, in particular, was a source of hope and a welcome distraction at times.

Finally, there is Katherine, without whom I genuinely do not know where I would be today, the shape this work would have taken, or indeed how I would have even begun my research. This has not been an easy journey, but we have travelled it together and got further than either of us would have alone. This thesis represents an overdue conclusion to the first chapter of our story. May we continue as we began - writing together.

Table of Contents

1. Chapter 1: Literature Review

1.1. Background and context

For a young man out on the town, perhaps in a bar, looking for a mate (or something more fleeting) is a time-contingent task. The longer it takes him to locate and approach a suitable target, the more chance that another suitor might step in, that she might get bored and go elsewhere, or that she might become inebriated to excess (although in some situations this could prove more help than hindrance). As such, it is key that he quickly appraise the situation and find a target worth his time and effort.

A worthwhile target needs (in most cases) to be not obviously romantically linked to somebody else, attractive to the young man, and realistically within his reach - if she is pointedly more attractive than him, he may stand little chance, irrespective of charm. So he scans the room. He looks for groups of women, ideally with no men, looking for an attractive woman with whom he stands a chance. But it is not only the attractiveness of the target woman that is of interest to him; he also needs to consider the women she is with.

Firstly, he needs to ensure that he has picked the best woman in the group to whom to devote his attention, that she is the most appealing to him. But part of his success may hinge on the woman's perceptions of herself; if she is highly attractive, she will likely be aware of this, but if her friends are also highly attractive, she may not consider herself quite so attractive, in comparison to them. On the other side of this coin, if her friends are mostly unattractive (or at least, less attractive than her) then she may have a heightened concept of her own attractiveness. As such, the young man must consider her social group as a whole, and her place amongst it.

It stands to reason that the woman closest to the average attractiveness for her group will have the most balanced self-opinion out of the women in that group - seeing

those more attractive than her, and understanding any imperfections she might have, while also seeing those who are less attractive than her, and being able to appreciate her own features - although this may not be an accurate opinion, depending upon the general attractiveness of the group. As such, it is also beneficial to the young man that he should be able to quickly and accurately assess this average, so that he can use it as a benchmark against which to compare each of the women in the group, and thus select the optimum target for his attention.

It is to this end that the research described here has been conducted, with a view to understanding some of the key concepts of the young man's process. How quickly, easily, and accurately can the attractiveness of a group be ascertained? Are there limits to this ability in terms of minimum exposure time, or maximum size of the group, and are these inversely related? Are the highly attractive (or highly unattractive) faces more selectively attended, and are these faces also more accurately remembered? If an average is extracted, is it a visual averaging, a statistical averaging, or something more abstract and gist-like?

1.2. Studies of attractiveness

For the aforementioned young man, the attractiveness of his potential mate is important. Assuming that he is seeking to mate, and to ensure the quality and survival of his offspring, attractiveness can serve as a cue to the fitness of potential mates. Humans, like many other animals, seek out mates, at least in part, who are able to ensure the health and survival of their offspring, be that through the provision of nutrition, resistance to parasites and disease, defence against predators and environmental dangers, or similar (E.g. Thornhill & Gangestad, 1993). As such, it is suggested that judgements of attractiveness are reflective of mate preferences that have been shaped

through evolution in response to selection pressures caused by such things as parasites (Hamilton & Zuk, 1982; Thornhill & Gangestad, 1993; grammer & Thornhill, 1994), and that evolution should have steered humans to observe physical traits that vary with mate value and be drawn to such traits that reflect high mate value (Thornhill & Gangestad, 1999).

Research has considered what features we deem to be beautiful, and how we respond to beautiful faces. Tying in to the Parasite Theory (Hamilton & Zuk, 1982), ties have been established between facial adiposity and judgements of health and of attractiveness (Coetzee, Re, Perrett, Tiddeman, & Xiao, 2011), and between attractiveness and homogeneity of skin texture, which is indicative of health and fertility (Fink, Grammer, & Thornhill, 2001). There is also a great deal of research into the impact of averageness (E.g. Langlois & Roggman, 1990; Langlois, Roggman, & Musselman, 1994; Rhodes & Tremewan, 1996; Rhodes, Sumich, & Byatt, 1999; Peskin & Newell, 2004) and symmetry (E.g. Rhodes, Proffitt, Grady, & Sumich, 1998; Rhodes, Sumich, & Byatt, 1999; Perrett, Burt, Penton-Voak, Lee, Rowland, & Edwards, 1999; Mealey, Bridgstock, & Townsend, 1999; Cardenas & Harris, 2007) on judgements of attractiveness.

It has been found that composite images of multiple faces are considered more attractive than the individual faces contributing to the composite (E.g. Langlois & Roggman, 1990, among many others), and it is suggested that what makes these images more attractive is that they are closer to a population average, and this averageness is indicative of having the genetic qualities needed to succeed within a population. A general preference for averageness was also found for dogs, wristwatches, and birds, suggesting an attraction to the prototypical (Halberstadt & Rhodes, 2000). This was supported by findings that ratings of distinctiveness (the inverse to averageness) were

negatively correlated with ratings of attractiveness (Rhodes & Tremewan, 1996), which were in turn positively correlated with ratings of familiarity (Peskin & Newell, 2004).

Counter to this are suggestions that the process of producing composite images removes the varying asymmetries among the sample faces and generates a symmetry, which is suggestive of successful development and an ability to resist diseases and other environmental disruptions to development (Thornhill & Gangestad, 1999; Fink & Penton-Voak, 2002). Rhodes, et al. (1998) found that increasing the symmetry of an individual face increased the ratings of attractiveness, and decreasing the symmetry similarly reduced attractiveness ratings. This was supported by Perrett, et al. (1999), who altered the symmetry of a face, while retaining original skin textures (the smoothing of skin textures, and removal of blemishes being one of the arguments against the effect of averageness found by Langlois & Roggman), and found that increasing symmetry resulted in higher ratings of attractiveness. Further, Mealey, et al. (1999) found that when comparing monozygotic twins, who are identical genetically, but not developmentally, the more symmetrical of the two was consistently rated as the more attractive, and the degree of difference in symmetry between the twins was directly related the degree of difference between their attractiveness ratings.

However, Fink & Penton-Voak (1999) have suggested that symmetry might not have a direct impact on attractiveness, and may simply covary with features that do impact on attractiveness. This is supported by several studies that have found independent impacts of averageness and symmetry on attractiveness (E.g. Rhodes, et al., 1999), in particular, Valentine, Darling, & Donnelly (2004). Valentine, et al. found that when faces were morphed towards an average (and thus symmetry), ratings of attractiveness increased for a full-face view, but this effect was also true, although less pronounced, when the faces were viewed in profile, where symmetry would be

undetectable. Similarly, when the faces were morphed away from the average (and thus towards asymmetry) ratings of attractiveness dropped. That averageness impacted upon attractiveness ratings even when symmetry was not discernible, but this effect increased when it was, suggests that both averageness and symmetry contribute independently to ratings of attractiveness, even if they do co-vary somewhat.

But what impact does beauty have on observers, other than swaying judgements of mating potential? At a biological level, seeing beautiful faces triggers responses associated with reward. Aharon, Etcoff, Ariely, Chabris, O'Connor, & Breiter (2001) found that even passive viewing of beautiful faces activated reward-related brain circuitry, and similar findings were reported by O'Doherty, Winstow, Critchley, Perrett, Burt, & Dolan (2003). Further, Schacht, Werheid, & Sommer (2008) found that both highly attractive and highly unattractive faces (as compared with faces of middling attractiveness) elicited a rapid amplified response in ERP signals when being rated for attractiveness, and this effect was still present, although lesser, when judging the gender of faces. This suggests that some neurological response to attractive faces is fast, and somewhat automatic, although this is obviously modulated by specific attention to attractiveness.

It is understandable, then, that beautiful faces capture participants' attention (Maner, Kenrick, Becker, Delton, Hofer, Wilbur, & Neuberg, 2003; Maner, Gailliot, & DeWall, 2007; DeWall & Maner, 2008; Sui & Liu, 2009), and that observers will actively expend effort in order to view beautiful faces for longer (Levy, Ariely, Mazar, Chi, Lukas, & Elman, 2007). There is also evidence that highly attractive faces are more memorable than medium attractive ones, with pointedly higher recognition, even 35 days after exposure (Shepherd & Ellis, 1973), and this ties in with findings that ratings of attractiveness and familiarity are correlated (Peskin & Newell, 2004).

However much observers might care to direct their attention to beautiful faces, there is evidence, alongside the neurological mentioned earlier, that assessment of attractiveness can be reliably performed from very brief exposures. Willis & Todorov (2006) found that when participants rated faces for attractiveness, among other things, the ratings made with no viewing time constraints correlated highly with those made after only 100ms exposures, and Olson & Marshuetz (2005) found that beauty of faces could be reliably perceived from exposures of as little as 13ms. While the rating of individual faces from limited exposure appears to be accurate, Maner, et al. (2003) found that when presented in a group, and with restricted viewing conditions, participants estimated there to be a higher frequency of attractive faces than when viewing the group with much fewer restrictions.

These results demonstrate that attractiveness can be assessed from very brief exposures, but do little to address how the attractiveness of a group is summarised and assessed. This thesis aims to explore this in more detail, considering the ways in which visual information might be summarised and represented to be used for such judgements.

1.3. How do we deal with large amounts of visual information, spatial vs. temporal?

The work discussed so far deals predominantly with the perceptions of individual faces; their structure, attractiveness, emotional expressions, and inferred personality from these. However, while we often see faces individually, we also frequently encounter people (and thus their faces) in groups. Such groups present an interesting challenge for the human visual system, owing to the sheer quantity of visual

information present therein. This amount of information, received concurrently, exceeds the capacity of the human visual system.

One of the more pertinent reasons for this limitation is the constraints of the visual short-term memory (VSTM). There is little use in processing any such information, if the results of that processing are unavailable for later use. The VSTM has very limited capacity for anything beyond the most basic of visual stimuli, which still faces tight constraints. Indeed, Alvarez and Cavanagh (2004) found a variance in the capacity of the VSTM dependent on the complexity of the visual information; ranging from 1.6 items for a shaded cube to 4.4 items for colour.

There are several ways that the human brain can potentially overcome such shortcomings, in order to still be able to process the information received from the environment. These include the selected direction of attention, and the compression and averaging of processed information.

Selective attention suggests that incoming information (across multiple modalities) is sampled in a serial fashion, and filtered for relevant information, thereby directing attention (Broadbent, 1958; Treisman, 1964; Treisman & Gelade, 1980). Broadbent (1958) suggested that any meaningful or semantic processing was performed only after items had been filtered to be attended, with the primary filtering being based solely on physical characteristics. However, the ‘cocktail party effect’ (Cherry, 1953) would call this interpretation into question. In this phenomenon, unattended information - such as an unrelated conversation in the background in a crowded room - can be still be used to direct attention, despite being unattended - for instance, a salient piece of information, such as one’s own name, can drive attention to the erstwhile dismissed conversation.

Treisman (1964) suggested an alternative solution, that accounts for the cocktail party effect; unattended information is not filtered out, but merely attenuated. Thus, all information is processed, but meaning is only attributed to the attended information. This accounts for the cocktail party effect, because the information from the unattended conversation would still be processed, and then attention can be directed if needed.

Some research has suggested that the human visual system compresses and compiles visual information into a summary representation, filtering out redundancies and duplications (Ariely, 2001). Despite the world having many varied and nuanced facets, a great many attributes of it are actually very internally stable, and thus predictable, which in turn makes such filtering viable, without any great risk of overlooking or dismissing any vital information.

To explore this summary representation, Ariely (2001) showed participants sets of circles varying in size, and followed these sets with one of two questions about the stimuli. Participants had to report either the size of a specific circle from the set (chosen randomly), or the average size of the circles in the set. While participants were incredibly poor at reporting the size of a specific circle (member discrimination), they were actually very competent at reporting the average of the set (mean discrimination). The results suggested that participants were creating an average of the set, and using this as a singular representation, while discarding any information about the individual set members.

Chong and Treisman (2003; 2005) demonstrated a further robustness of mean discrimination with a similar experiment. They used sets of circles, either of heterogeneous size, or homogeneous size, with varying display durations (50 - 1000ms) and delay following stimulus display (up to 2s), and found that averaging was automatic and not driven by the intent of the participants, but did require attention be directed

towards processed items. Results showed that even with these manipulations in place, participants could judge an average of the set, but could not discriminate its members. It would appear that the limits of the human visual system (and the VSTM in particular) are countered by the creation of a single, simplified representation of (similar) visual information.

1.4. Summaries of visual information

In terms of creating an average representation of visual information, there are two principal strands of investigation; the Prototype Effect, concerning averaging over time, and Ensemble Representation, concerning averaging across space.

1.4.1. Prototype effect

The Prototype Effect works around the idea of creating a representation (or prototype) of multiple visual stimuli that are experienced individually over time that reflects the central value of the series. Within the research, this often leads to responses reflecting this central value, even without prior experience of that particular value (Cabeza, Bruce, Kato, & Oda, 1999). Thus, a participant seeing stimuli sufficiently close to this central value will respond as though having previously encountered it. Posner & Keele (1968) suggest that the prototype is a summary representation of 'central tendency'.

Posner & Keele (1968) explored the formation of prototypes in sets of dots. They generated prototype stimuli using 25 dots, and then variants of these stimuli, in which 10 dots appeared in different locations. Participants were trained to classify sets of dots from subsets of the variants, and were told that these training sets belonged to a single

category. They were then asked to indicate whether test items belonged to this same category. The test stimuli were either the variants on which participants trained, new variants, or the prototype sets. The results showed that responses and response times were very similar for training items and the prototypes, but both measures were pointedly poorer for the new variant stimuli. This suggests that participants were forming a representation of the items from training that resembled the prototype sets, hence the similar performance between the previously experienced training items and the previously unseen prototypes thereof, while the previously unseen new variants were less accurately, and less speedily classified.

This was expanded on by Cabeza, Bruce, Kato, and Oda (1999), who expanded this idea by using face stimuli. They presented participants with a series of faces, and then asked them to identify whether a test face had been present in the series. The faces in the series were manipulated with varying distance between the eyes and nose. When the target face had the mean properties of the series, participants tended to incorrectly respond that the face had been present. In some trials, the variation across stimuli was large, while in others it was smaller. This tendency to identify the mean as being present in the series only occurred in the lower-variation trials. This led to the conclusion that the averaging process likely only occurs in stimuli with low levels of variation.

Taken together, these findings suggest that such averaging of visual information over time (serial) only occurs in sets of items that are low in variation (when the stimuli are manipulated to be so, in the instance of high-level stimuli such as faces). However, it is possible that there are differences in these restrictions when comparing serial averaging (as seen above) and parallel averaging (when multiple items are presented simultaneously and a summary created over that plane), with perhaps greater diversity in one average as compared with the other.

1.4.2. Ensemble Representation

Where the Prototype effect relates to the formation of a summary representation of stimuli experience over time, Ensemble Representation relates to a similar statistic created for multiple stimuli presented simultaneously (Alvarez & Oliva, 2008). Like the Prototype Effect, ensemble representation has been studied in face research, by Haberman & Whitney (2007; 2009). They focused their research on two primary aspects; accurate summary of the group (mean discrimination), and the discrimination of individual group members (or lack thereof).

1.4.2.1 Discrimination of mean and members in groups of faces

Haberman and Whitney used groups of faces showing emotional expression to explore whether face stimuli allowed for good mean discrimination when presented simultaneously. The sets of faces contained expressions ranging from either happy to sad, or neutral to disgusted. Participants were shown these groups, followed by an individual test face, and then asked to indicate whether the test face was happier or sadder (or more neutral or disgusted) than the group (see *Figure* 1).

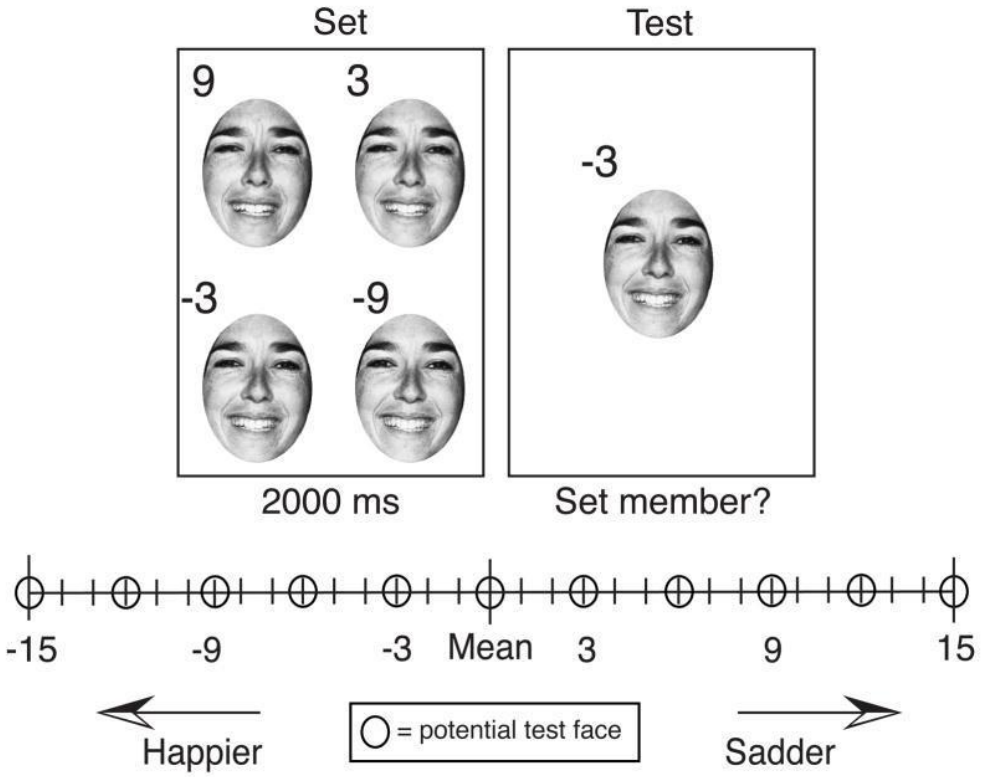


Figure 1: Design of Haberman & Whitney's (2009) task. Values reference the distance of each face from the mean of the set, on a scale of 1 (completely disgusted) to 50 (completely happy).

Haberman & Whitney (2009) created their stimuli by forming a continuum of 50 levels of expression for a single individual. This was achieved by morphing together two images of the same face showing the extreme of two different emotions. By varying the proportion of each extreme expression included in the image, the intensity of the expression could be manipulated. Strategic points along this continuum were then identified, and used to create face images, which were compiled into sets of faces of the same identity but varying levels of expression. The test face was also plucked from this continuum, from a point only slightly away from the average of the set.

The task required participants to identify whether the set was more happy/sad than the test face, or vice versa. The results showed that participants achieved a 75% correct threshold of as few as four steps on the continuum (approximately 8% difference, given

the 50 increments of the continuum), suggesting that participants were fairly accurately forming an average representation of the set, paralleling Ariely's (2001) findings with low-level information.

Haberman and Whitney also investigated participants' capacity for member discrimination. Participants were presented with sets of faces as in the previous experiment, followed by a test face. The task was to identify whether the test face had been present in the set - the individual faces being differentiated based on the subtleties of their expressions, rather than the identity of the face, which was the same within a given trial. The number of faces in the set varied from one to four, and a predicted accuracy was calculated based on a participant only being able to extract a single face from the set. Participants were very poor at the task, performing at about the same level as the predicted accuracy. These two experiments taken together suggest that participants are creating an accurate representation of the mean of the sets in some form of ensemble, while discarding individual member information.

1.5. Summary of Work Presented

The aim of the work in this thesis was to explore the methods used when summarising the attractiveness of a group of faces, and to assess the efficiency of this process. Haberman & Whitney (2007; 2009) have already demonstrated that the mean expression on a group of faces can be summarised into an ensemble representation and acted upon accurately, but there was little evidence regarding whether the same could be said for attractiveness.

Chapter 2 served to establish whether participants could make reasonably accurate comparisons between briefly presented groups of faces. Participants were asked to determine which of the two groups contained the most attractive faces, initially with a

group size of four faces, and the results showed that performance on the task was better than chance levels, even in the most difficult of conditions. This result was taken as a general confirmation of participants' ability to make a reasonably accurate general assessment of small groups of faces from brief presentations. The next experiment used the same task, but expanded the display array to include nine faces in each group, thereby expanding the potential difficulty of the task, but also allowing finer scale manipulation of the difficulty. Again, results suggested participants were able to perform this task, and only in the most difficult of conditions did performance drop to chance levels. This suggested that the number of faces in the group did not significantly impact the capacity to summarise elements of the group, in turn suggesting that the groups were being processed in a parallel fashion, rather than serial.

In Chapter 3 the display duration of the faces was manipulated to investigate the impact of the brief presentation on capacity for summarising the groups. Initially, the same 500ms display duration was used in conjunction with an unrestricted display duration. In order to accommodate the unrestricted condition, the task was altered to use only one group and to ask participants to judge whether there were more attractive or more unattractive faces in the group. The results demonstrated that while there was no difference in overall accuracy between the two display durations, the two did differ across the varying conditions of the task. When the majority of the group was unattractive, participants generally performed better when the display duration was unrestricted, and when there were five attractive faces and four unattractive faces (the joint most difficult condition in the task) accuracy was significantly below chance. This, combined with overall better performance when the majority of the group was unattractive, suggests that participants are more likely to indicate that a group contains more unattractive faces, even when given as much time as they like to inspect the group.

This same pattern of generally perceiving a group as having fewer attractive faces continued into the next experiment, in which a similar task, but using two additional restricted display durations (one longer and one shorter than the restricted condition of the previous experiment), was combined with eye-tracking technology, and a further task in which participants needed to estimate the number of attractive faces in the group (10AFC). The first of these tasks followed the same pattern as the previous experiment, while the second, 10AFC task, reinforced the point, with participants underestimating the number of attractive faces in all but the three conditions with zero, one, and two attractive faces (out of nine), when there was an overestimation. These results again suggesting a general perception of fewer attractive faces than actually present in the group, but especially so when estimating the exact number.

The eye-tracking data showed no impact of the attractiveness of the faces on the order in which they were fixated, and in fact revealed a more systematic approach to moving the eyes around the group, especially in trials with accurate responses. There were also very few differences in the duration of fixations, and where these were present, they were only to the order of around 5-10ms difference. There appeared to be no pop-out effect for attractive faces, and, certainly when actively seeking to make judgements about the faces, neither attractive nor unattractive faces appeared to draw or hold visual attention any more than the other. However, performance on trials did appear to be impacted by faces in the group that were not fixated during the trial, suggesting some non-foveal information was contributing to the representation of the group.

The final experiment of Chapter 3 to use groups of nine faces was intended to clarify some of the findings of the 10AFC task in the previous experiment. The results had suggested that accuracy was at its highest (or rather, the degree of error was least)

when there were three attractive faces in the group. It was unclear whether this tendency to respond with a three more frequently than any other number was due to an actual perception of that many attractive faces, or just a general tendency to respond more in that end of the scale. There was also a possibility of the instructions being ambiguous, with participants potentially misconstruing "attractive faces" to mean "highly attractive faces", rather than its intended meaning of any face scoring a five or higher on a 10-point scale of attractiveness. The instructions were clarified, and another task was added in which participants also estimated the number of unattractive faces (with similar clarification of instructions). By comparing the estimations of the number of attractive and unattractive faces, it was possible to tease out whether the response pattern was indicative of a response bias, or a perceptual bias. The results suggested that overall, the bias was perceptual, and that whether estimating the number of attractive or unattractive faces, the general perception of the group was the same. Results also indicated no impact of the clarified instructions, suggesting participants in the previous experiment had understood the intention of the instructions, despite lacking some explicit explanation.

The final experiment of Chapter 3 expanded the display size of the groups of faces to 16. This was to eliminate the previously observed confound that participants began each trial fixating on a central location, which would contain a face upon stimulus onset, potentially providing an immediate impact on the judgement of the group. By increasing the display size, the central face was removed, and trials began with a fixation on empty space. This experiment served as a test to ensure participants were still able to make rapid, reasonably accurate judgements about the attractiveness of a group with this further increased group size. The results suggested no detrimental impact on performance, and so future experiments used the 16-face group instead of the nine-face one.

Having found a lack of evidence for the pop-out effect of attractive faces, and suggestions that non-fixated faces were being processed with the group, Chapter 4 set out to explore whether participants would be able to identify the most or least attractive face in a group, both under restricted display durations (and thus find the target group member, and remember its location) and unrestricted ones (thus only requiring that the participant find the target face, having the opportunity to freely compare each group member to each other group member, if desired). The results suggested that performance at this task was moderate, but was pointedly better when given unrestricted viewing time of the groups. Apart from in conditions where participants were selecting the most attractive face from a group of faces all rated as unattractive (or the inverse), performance did not differ between the two tasks, suggesting participants were equally moderately good at selecting the most and the least attractive face from a group. This further questions any suggestion of a pop-out effect, and continues a theme of the attractiveness of the faces being modulated seemingly by simply being in a group context.

The final experimental chapter was intended to explore the nature of the representation of the attractiveness of a group. The task compared various different methods of rating the attractiveness of a group. Participants provided singular ratings for the overall attractiveness of a group, ratings of attractiveness for each face in the group while presented in the group (which was then averaged to form a singular value for the group), a rating of attractiveness for a morphed image that was a visual conglomeration of all sixteen faces in the group, and a rating of attractiveness of each stimulus face when presented individually (and the values for the sixteen members of the group were averaged to form another singular value for the group). By comparing these different measures, it was possible to establish several things: 1) how the group context changes perceptions of attractiveness of faces, 2) how closely the singular rating

of the group resembles the two averages of the individual members, and thus how likely it is that either of these methods reflects the approach taken to judge the attractiveness of a group, and 3) whether the morphed image of the group is a reasonable analogue to the summary representation of a group.

As expected (Perrett, Burt, Penton-Voak, Lee, Rowland, & Edwards, 1999; Valentine, Darling, & Donnelly, 2004), the morphed image of the group was rated as considerably more attractive than any of the other measures, and was, in fact, the least representative of the singular rating of the group. The closest approximation, though still substantially lower overall, was the average of the group members when rated in the group context, which was itself substantially higher than the average rating of all group members when rated individually, which was counter to expectations based on previous results that suggested the group context reduced the perceived attractiveness of faces. However, the previous results related the estimates of majority or number of attractive faces in the group, rather than an overall rating of attractiveness. While these two values must be somewhat linked (a group with a larger number of attractive faces should reasonably have a higher overall rating of attractiveness than a group with few attractive faces), they are not directly analogous, which could explain this unexpected finding.

From these results, it was concluded that the perception of attractiveness in groups is possible from to reasonable levels of accuracy, even from brief presentations, but this capacity appears to be modulated by the task at hand. Further, previously found patterns of attractive faces drawing more and longer fixations was not present when the attractiveness of the face was task relevant, further suggesting that perception is modulated by task. Finally, the method of ensemble representation is hypothesised to be

a combination of visual and statistical averaging, with neither alone being an accurate reflection of the overall judgement of the attractiveness of a group.

2. Chapter 2: Assessment of attractiveness - comparing two groups

2.1. Experiment 1: Can participants select the more attractive of two briefly presented groups?

2.1.1. Introduction

With previous studies only considering rapid appraisal of attractiveness of singular faces (Olson & Marshuetz, 2005) or not considering the accuracy of the response to groups (Maner, Kenrick, Becker, Delton, Hofer, Wilbur, & Neuberg, 2003), Experiment 1 was intended to establish a baseline for performance on the assessment of attractiveness of groups from brief presentations. Participants were simply tasked with determining which of two briefly-presented sets of faces contained the greater number of attractive faces. The aim of this was to establish if such a task was possible, whether performance would be better than at chance level, and the trends in performance across conditions that were arguably more difficult. It was hypothesised that conditions with smaller differences between the two groups would be more difficult, and thus show lower levels of performance. While a comparison of performance with chance levels was part of this experiment, it was unclear from existing literature whether the more difficult conditions might prove beyond participants' abilities, and thus result in performance at no better than chance.

The 250ms experiment was ended after 10 participants, because the results were showing the same trends as the 500ms experiment (as detailed later). As such, later experiments focused on increasing the number of stimuli to manipulate task difficulty.

2.1.2. Method

2.1.2.1. Participants

Ten undergraduates from the University of Hull (eight female) participated in the 250ms experiment. Their age ranged from 18 to 23 years (M = 19.8, SD = 1.40), and all had normal or corrected-to-normal vision.

Twenty undergraduates from the University of Hull (16 female) participated in the 500ms experiment. Their age ranged from 18 to 27 years (M = 19.9 years, SD = 2.38), and all had normal or corrected-to-normal vision.

2.1.2.2. Stimuli

The images used in this study were taken from the University of St Andrews. All of the images in this database were already masked and aligned such that the pupils were in a constant location across the images, with neutral expressions and no glasses or other items occluding the face. Although originally in colour, the images were converted to greyscale for this experiment, and processed for mean luminance and contrast. Only images of female Caucasian faces were used.

Nineteen observers (aged 18-29 years; 12 female) had already judged the images for attractiveness on a 1-7 scale, and these ratings were used to select the 32 faces with the most unattractive average rating, and the 32 faces with the most attractive average rating. Taking faces from the most extreme ends of the scale allowed the greatest possible difference between the attractive and the unattractive faces.

In each trial, two sets of four images were presented to the participant, and these were selected at random from the attractive and unattractive groups, with the number of each determined by the condition of the trial. There were five different possible

combinations, ranging from 0 attractive/four unattractive faces, through to four attractive/zero unattractive faces. A given face image could not appear more than once in each trial. Each image was placed with its centre at 7.20° diagonally from the centre of the screen (see *Figure 2*), and sized at approximately 4.75-5.00° horizontally and 6.00-7.00° vertically (at a viewing distance of 57cm).

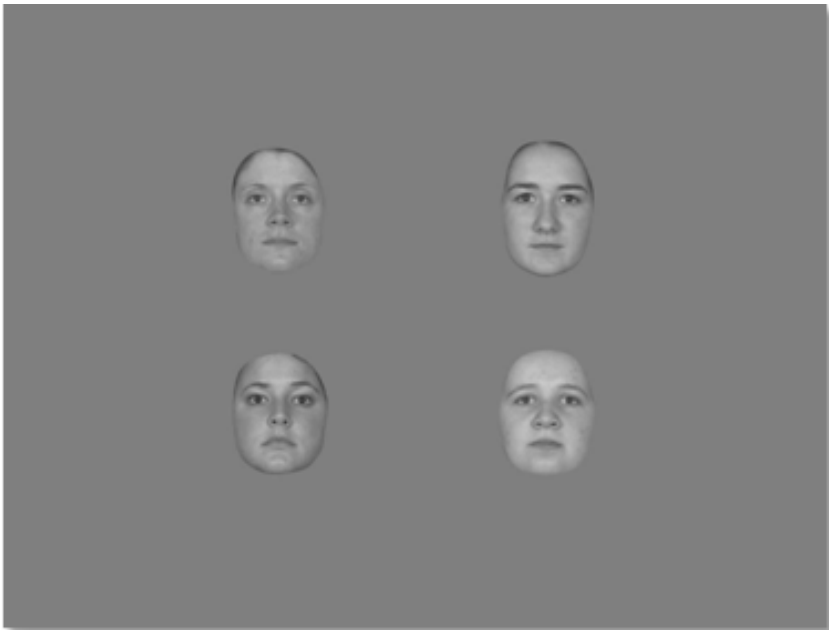


Figure 2: An example of the stimuli and their layout as used in Experiment 1.

2.1.2.3. Design and Procedure

The task in this experiment was to discern which of two groups of faces contained the greater number of attractive faces. The difficulty of the task was affected by the degree of difference between the number of attractive faces in the groups. There were four levels of difference (25%, 50%, 75%, and 100%), ranging from one group having one more attractive face than the other, through to one group having four unattractive faces, with the other group having four attractive faces. Each participant completed 20 trials of each level, and the trial procedure was identical between the timing conditions, excepting the display durations.

The participants were given a verbal explanation of the task, which was accompanied by a written explanation on screen. Each trial consisted of a fixation cross appearing in the centre of the screen for 250ms (or 500ms), followed by the first of the two groups of faces for 250ms (or 500ms), then the screen went blank for 250ms (or 500ms), followed by the second group for 250ms (or 500ms), after which there was a prompt for a response, which remained indefinitely until response. There were several breaks programmed into the experimental session, although the session was only brief.

At the end of each experimental session, the participant was asked to rate each of the faces in the experiment for attractiveness. This was to create a database of ratings of these faces when unclouded by surrounding faces. Each face was presented centrally on the screen, in a random order, and participants were asked to rate the face on a scale of 1 (highly unattractive) to 10 (highly attractive). The rating was made using the number keys across the top of the keyboard (with 0 being used to indicate 10), and each face remained on-screen until the rating was given, at which point, another face appeared, until all of the stimuli had been rated this way.

2.1.3. Results

2.1.3.1. 250ms:

In this task, and in all others that use the attractiveness ratings of the stimuli as provided by the original raters, the term "accuracy" is used as a general analog of agreement with the original ratings. Obviously, ratings of attractiveness are all subjective, and so the consensus of the original ratings is used as a baseline to determine the attractiveness of a group. An "incorrect" response does not necessarily indicate that the participant failed to ascertain the attractiveness of the group(s) (although this could,

of course, be the case), just that their response does not fall in line with this established baseline.

A repeated measures ANOVA on accuracy results showed a main effect of difference ($F(3,27) = 9.33$, $p < .001$, $\eta^2 = .51$), with accuracy increasing from 25% difference ($M = 58$, $SD = 10.6$) to 100% difference ($M = 79$, $SD = 15.78$), as illustrated in *Figure* 3. Planned comparisons showed that 25 and 50% were not significantly different ($p > .40$) , and that 75 and 100% were not significantly different ($p > .89$), but there was a significant difference between 50 and 75% ($F(1,9) = 17.19$, $p < .005$, $\eta^2 = .66$). One-samples t-tests showed performance to be above chance (50% accuracy) in all conditions (25%, $t(9) = 2.39$, $p < .05$; 50%, $t(9) = 4.29$, $p < .005$; 75%, $t(9) = 8.54$, $p < .001$; 100%, $t(9) = 5.81$, $p < .001$). A repeated measures ANOVA on reaction time data showed no effect of difference ($p > .40$), see *Figure* 4.

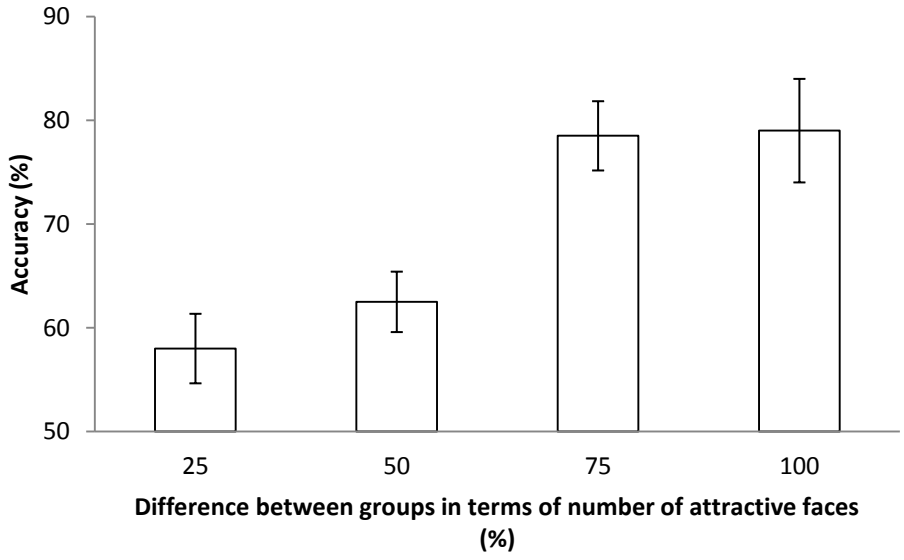


Figure 3: Accuracy at each level of difference between groups. Error bars indicate standard error.

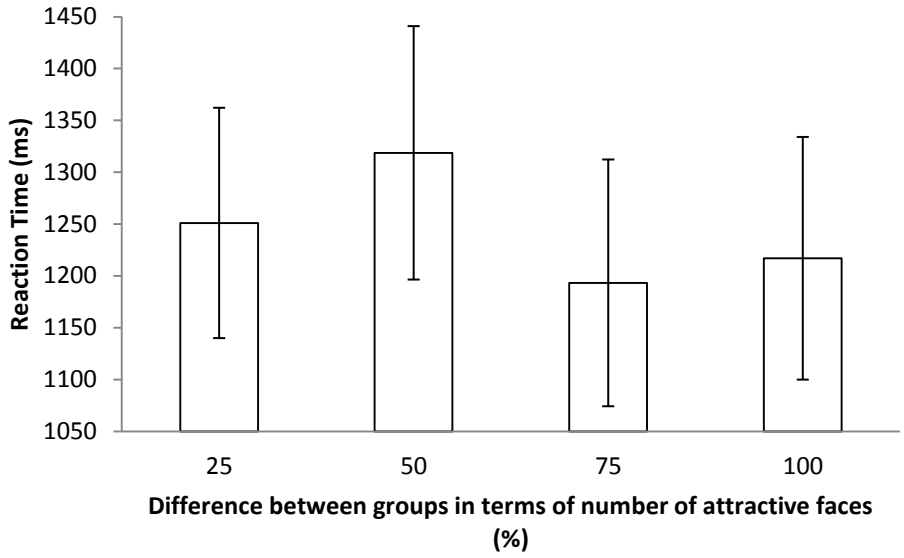


Figure 4: Mean reaction time at each level of difference between groups. Error bars indicate standard error.

2.1.3.2. 500ms:

A repeated measures ANOVA on accuracy results showed a main effect of difference ($F(3,57) = 51.69$, $p < .001$, $\eta^2 = .73$), with accuracy increasing from 25% difference ($M = 60.07$, $SD = 8.57$) to 100% difference ($M = 88.03$, $SD = 10.78$), as illustrated in *Figure* 5. Planned comparisons showed significant increases in accuracy between 25 and 50% ($F(1,19) = 20.50$, $p < .001$, $\eta^2 = .52$), and 50 and 75% ($F(1,19) = 20.55$, $p < .001$, $\eta^2 = .52$), but not between 75 and 100% ($p > .12$). One-sample t-tests showed performance to be above chance in all conditions (25% $t = 5.26$; 50% $t = 7.25$; 75% $t = 13.80$; 100% $t = 15.78$, all ps $< .001$, all dfs $= 19$).

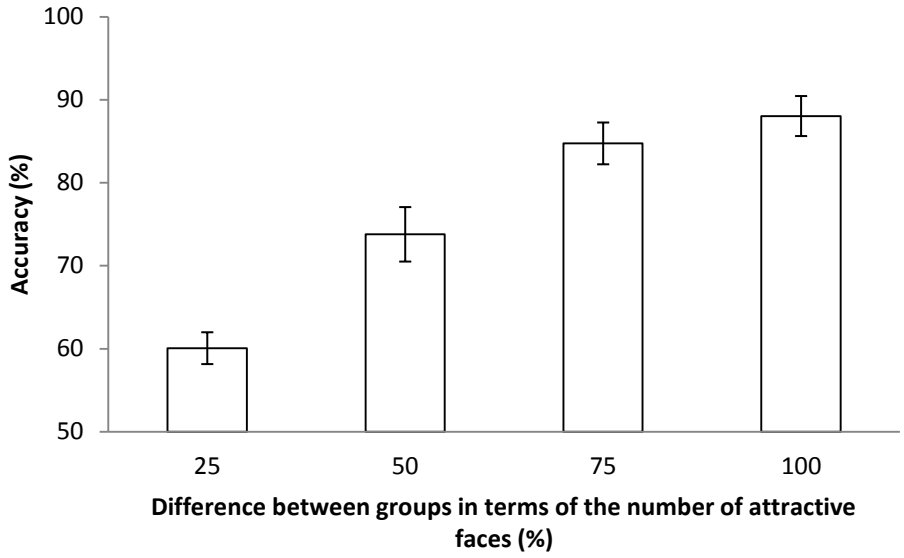


Figure 5: Accuracy at each level of difference between groups. Error bars indicate standard error.

A repeated measures ANOVA on reaction time data also showed a main effect of difference ($F(3,57) = 10.41$, $p < .001$, $\eta^2 = .35$), with a decrease in RT from 25% difference (M = 921ms, SD = 297ms) to 100% difference (M = 780ms, SD = 319ms), as illustrated in *Figure* 6. Planned comparisons showed that 25 and 50% were not significantly different ($p > .73$) , and that 75 and 100% were not significantly different ($p > .43$), but there was a significant difference between 50 and 75% ($F(1,19) = 15.59$, $p < .005$, $\eta^2 = .45$).

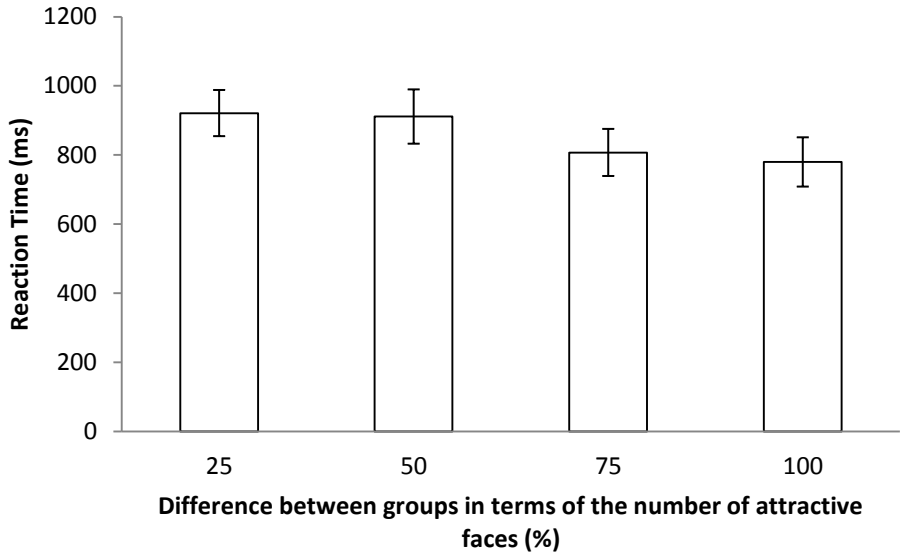


Figure 6: Mean reaction time at each level of difference between groups. Error bars indicate standard error.

2.1.4. Discussion

This experiment aimed to explore whether the number of attractive faces in two groups could be accurately assessed and compared with short display durations. The results suggest that even with only 250ms displays per group, and with a difference between the groups of only one attractive face, participants still performed the task better than chance levels (50%).

Accuracy in both timing conditions increased as the difference between the two groups increased, supporting the idea that the larger the difference, the easier the task. However, this appears to only be true to a point; in both timings, performance did not improve between 75% and 100% differences. This suggests that although the task was clearly slightly easier with longer exposure to the groups, some of the subjectivity inherent in judgements of attractiveness may be leading to some general "error" when performing the task, hence performance peaking at 75%.

If the performance plateau observed is partly due to certain individual face stimuli being attractive or unattractive to some participants, when the general consensus is otherwise, then larger stimuli sets, with the difference between difficulty levels incrementing by two faces, instead of one, should offset this. In a group of four faces, if one of the faces is deemed to be unattractive by the participant when all other observers have classified it as attractive, then the perceived composition of that group could change wildly. Whereas, with more faces in the group, such an anomaly will have less dramatic an impact on the composition.

2.2. Experiment 2: Is performance degraded by a larger group size?

2.2.1. Introduction

With Experiment 1 establishing a baseline capacity for making observations about the attractiveness of members of a group from brief exposure, Experiment 2 was intended to build upon this. By increasing the number of faces in each group, the task potentially became more difficult, with more visual information to process.

If the members of the group are all processed simultaneously, then there should be limited impact on performance of the task, because the same amount of information should be available to inform the response. Of course, with a larger group, there is also more information to store for the sake of comparisons, but levels of performance similar to the smaller groups in Experiment 1 would suggest that any limitations in the task are a result of the variability of the assessment of attractiveness, rather than reaching a peak of cognitive capacity. This in turn would allow for finer control in the variability of groups for further experiments, where using a larger number of faces in a group means that a change to a single member of the group has a proportionally smaller effect.

Based on the results of Experiment 1, in tandem with the findings of Haberman & Whitney (2007; 2009), it was hypothesised that there would be some signs of parallel processing, I.e. that the increase in group size would not show a significant degradation in performance.

2.2.2. Method

2.2.2.1. Participants

Twenty one undergraduate students from the University of Hull (18 female) participated in this experiment. Their age ranged from 18 to 40 (M = 21.14, SD = 5.04) years, and they all had normal or corrected-to-normal vision.

2.2.2.2. Stimuli

This experiment used the same images as the one previously described in this chapter. However, the images were presented in groups of nine for this task. The faces were displayed equidistantly in a 3x3 grid, which subtended approximately 27°x27° of visual angle, at a viewing distance of 57cm (see *Figure 7*).

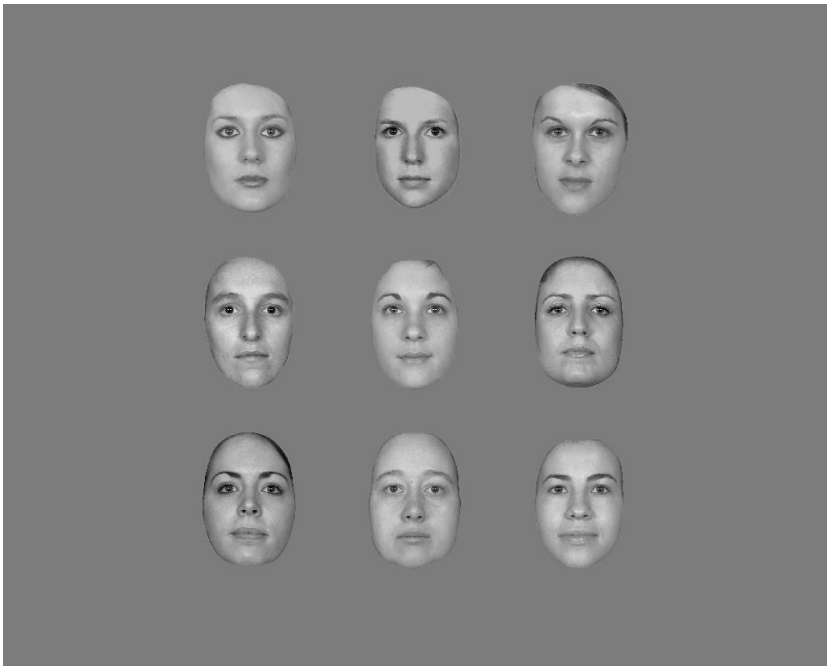


Figure 7: An example of the stimuli and their layout as used in Experiment 2.

2.2.2.3. Design and Procedure

This experiment followed the same trial and session procedure as the two sets of four faces, but the difficulty of the task, as modulated by the difference in the number of attractive faces in the two groups, varied slightly. There were five levels of difference

(11%, 33%, 56%, 78%, and 100%), ranging from one group having one more attractive face than the other, through three, five, and seven, up to one group being solely attractive faces, and the other being solely unattractive.

2.2.3. Results

A repeated measures ANOVA on accuracy results showed a main effect of difference ($F(4,80) = 41.69$, $p < .001$, $\eta^2 = .68$), with an increase from 11% difference ($M = 52.03$, $SD = 10.77$) to 100% difference ($M = 89.97$, $SD = 8.22$), as illustrated in *Figure* 8. Planned comparisons showed an increase between 11% and 33% ($F(1,20) = 13.12$, $p < .005$, $\eta^2 = .40$), between 33% and 56% difference ($F(1,20) = 7.69$, $p < .05$, $\eta^2 = .28$), and between 78% and 100% ($F(1,20) = 28.31$, $p < .001$, $\eta^2 = .59$) but no difference between 56% and 78% ($p > .39$). One-sample t-tests showed performance at 11% difference to be no better than chance ($p > .39$), but at all other differences it was (33%, $t = 5.56$; 56%, $t = 12.62$; 78%, $t = 15.00$; 100%, $t = 22.29$, all ps <.001, all dfs = 20).

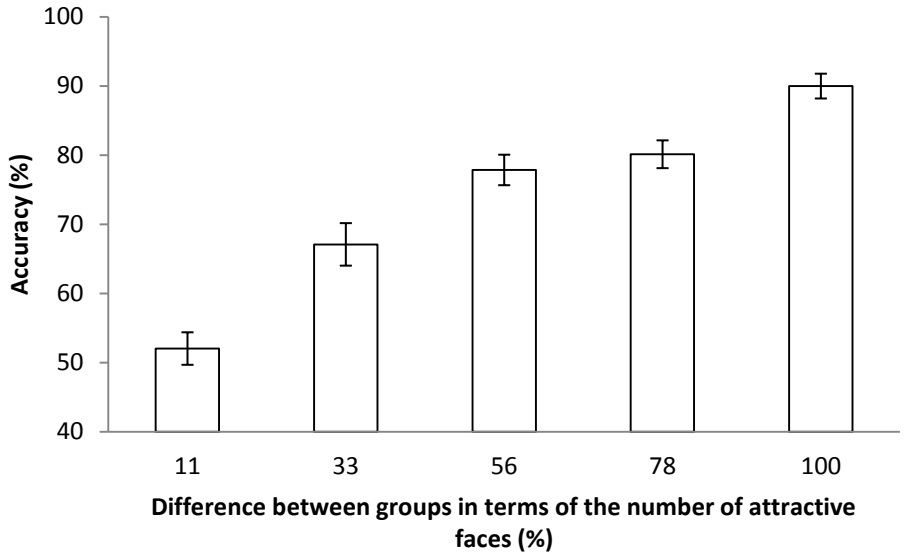


Figure 8: Accuracy at each level of difference between groups. Error bars indicate standard error.

A repeated measures ANOVA on reaction time data showed a main effect of difference ($F(4,80) = 14.21$, $p < .001$, $\eta^2 = .42$), with a decrease from 11% difference (M = 980ms, SD = 497ms) to 100% difference (M = 793ms, SD = 447ms), as illustrated in *Figure* 9. Planned comparisons showed a significant difference between 11% and 33% difference ($F(1,20) = 14.89$, $p < .005$, $\eta^2 = .43$), and between 78% and 100% difference ($F(1,20) = 10.17$, $p < .01$, $\eta^2 = .34$), but no difference between 33% and 56% difference ($p > .77$) or between 56% and 78% difference ($p > .14$).

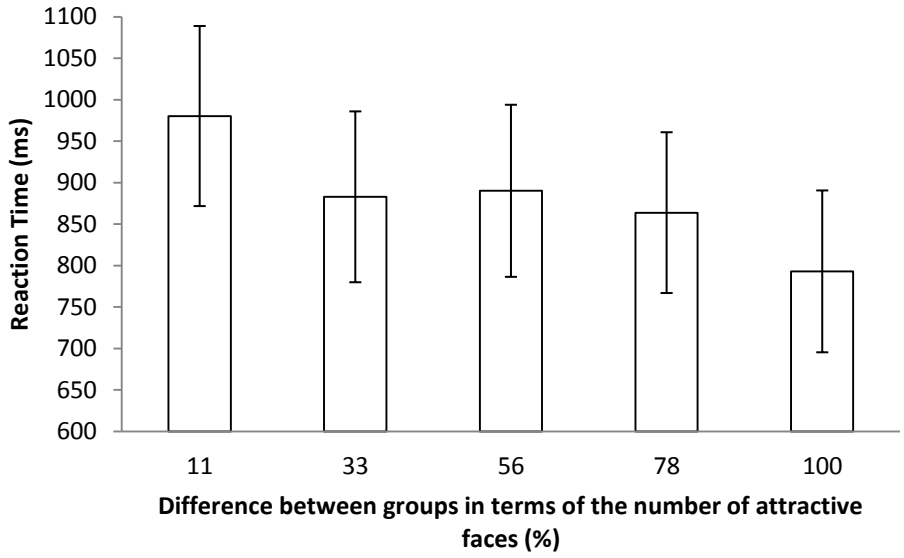


Figure 9: Mean reaction time at each level of difference between groups. Error bars indicate standard error.

2.2.4. Discussion

These results suggest that participants were still capable of making informed decisions about the attractiveness of two different groups, even with a group size more than double that of Experiment 1. Although few of the difficulty levels are directly comparable between the two experiments, owing to the different percentage changes between the groups in each, performance at 100% still caps out at around 90% accuracy in Experiment 2, as it did in the same exposure length in Experiment 1. The same trend in performance can be seen between the two, with only 11% in Experiment 2 sitting at chance level.

It is a reasonable conclusion to draw that the larger group does not significantly increase the difficulty of the task such that it can no longer be performed. This implies a certain level of parallel processing of the stimuli, and that, whatever form the summary

of the groups might be taking, it is one that can be held in memory sufficiently long enough to allow the processing of another group, and a comparison between these.

However, it is unclear from these results whether the groups are being processed and stored as a singular representation, or average, or whether a certain amount of meta data about the groups is being represented. Are participants representing, and thus able to recall, the composition of the group (I.e. the number of faces of differing levels of attractiveness), or simply the general attractiveness of the group (I.e. a singular representation, perhaps that the group has an average attractiveness of a certain level or value). Further experiments will seek to expand on this idea.

3. Chapter 3: Determining the majority of a single group

3.1. Experiment 3: Are brief presentations detrimental to judgements of a single group?

3.1.1. Introduction

With Experiments 1 and 2 illustrating that participants were able to make assessments regarding the relative attractiveness of two different groups from brief exposures at reasonable accuracy levels, there was still a question of whether these display constraints were still having a negative impact on the task. While the larger groups in Experiment 2 did not seem to suffer too greatly from the extra information being presented, performance was not entirely accurate in either experiment.

Some of the errors observed previously might stem from the time constraints of the tasks, but they might also be couched in the subjective differences in perceptions of attractiveness, and indeed in the slightly unorthodox and unfamiliar method of presentation. Experiment 3 was designed to address the possible impact of time constraints on previous tasks. In this experiment, participants would perform the same task under two different conditions; in one condition the stimuli would be presented for 500ms, as before, but in the other, participants would face no time constraints at all.

In order to facilitate the lack of time constraints, the task was altered to use only a single group of faces, thereby allowing dedicated study of the stimuli in the self-paced condition, without then imposing a longer period for which to store the details of the first group while studying the second. The result of this was a change of task to identify whether the group contained a greater number of attractive faces, or of unattractive faces. This also served the purpose of removing any possible issues with memory that may have occurred in Experiments 1 and 2, with participants now not needing to retain information about two groups for comparison.

It was hypothesised that, where comparable, performance in this task would improve compared with that of Experiment 2, given the reduced cognitive load of the task. Further, it was expected that the self-paced condition would result in much higher levels of accuracy than the timed condition, given the opportunity for participants to fully explore the stimuli.

3.1.2. Method

3.1.2.1. Participants

Twenty undergraduate students from the University of Hull (14 male) participated in this experiment. Their age ranged from 18 to 22 (M = 20.15, SD = 1.01) years, and they all had normal or corrected-to-normal vision.

3.1.2.2. Stimuli

This experiment used the same images as previously described in other experiments. The images were presented in the same way as Experiment 2; in groups of nine, spaced equidistantly in a 3x3 grid, which subtended approximately 27°x27° of visual angle, at a viewing distance of 57cm.

3.1.2.3. Design and Procedure

This experiment was broken into two test sessions, counterbalanced for order. Both sessions involved the same task, but differed in the display duration of the stimuli.

Participants were asked to judge whether a group of nine faces contained a greater number of attractive faces or a greater number of unattractive faces. This instruction was provided verbally, and in writing at the start of the experimental procedure.

In each trial, a fixation cross appeared in the centre of the screen for 500ms, followed by a group of nine faces. In the timed condition, this group was presented for 500ms, and was then replaced with a response prompt, asking for a keypress to indicate the response. In the self-paced condition, the group of faces remained on screen until the participant made a response; there was an instruction to respond (with the relevant keys listed) beneath the group of faces, which remained on screen while the faces were present. In either condition, following the response, a blank screen was presented for 500ms, before the next trial began. As before, breaks were included in the experimental procedure, despite its brevity.

The groups of faces were generated randomly in each trial, with the restriction that no face could appear more than once in a single trial. The number of attractive faces in each group ranged from zero to nine, and this was selected randomly for each trial, such that a total of 10 trials of each number of attractive faces appeared in each session.

The difficulty of the task was modulated by the relative proportions of the group, with nine attractive faces and zero unattractive faces (or nine unattractive and zero attractive) being the easiest condition, while five attractive and four unattractive faces (or vice versa) was the most difficult. As such, performance would be expected to be better at the very low and very high numbers of attractive faces, compared with trials with four or five attractive faces.

3.1.3. Results

A 2 (timing) x 2 (majority) x 5 (proportion) repeated measures ANOVA on accuracy data showed a main effect of majority ($F(1,19) = 6.43$, $p < .05$, $\eta^2 = .25$), with a higher accuracy in majority unattractive ($M = 75.2$, $SD = 14.35$) than in majority attractive ($M = 62.35$, $SD = 13.94$). There was also a main effect of proportion ($F(4,76)$

= 32.00, $p < .001$, $\eta^2 = .63$), illustrated in *Figure* 10, ranging from the lowest accuracy in 5/4 proportion ($M = 53.25$, $SD = 11.39$), through to the highest accuracy in 9/0 proportion ($M = 81.00$, $SD = 12.55$).

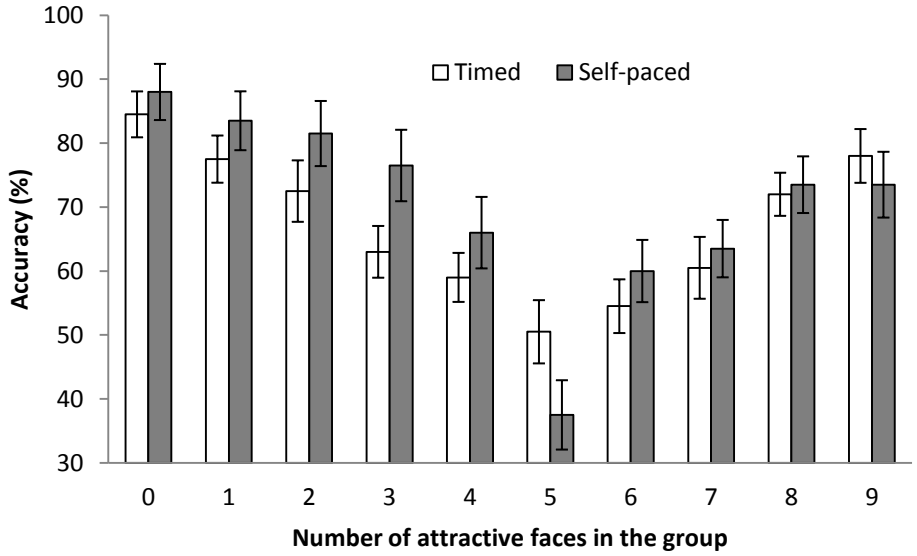


Figure 10: Accuracy at each level of number of attractive faces in the group, separated by timing. Error bars indicate standard error.

The ANOVA also found two significant interactions (but no three-way interaction): timing x majority ($F(1,19) = 6.79$, $p < .05$, $\eta^2 = .26$); and timing x proportion ($F(4,76) = 2.63$, $p < .05$, $\eta^2 = .12$). Simple effects analyses showed that in the timed condition, there was no effect of majority ($p > .07$), but in the self-paced condition there was ($F(1,19) = 7.92$, $p < .05$, $\eta^2 = .92$), with higher accuracy in majority unattractive ($M = 79.10$, $SD = 18.83$) than in majority attractive ($M = 61.60$, $SD = 15.51$). The timing x proportion interaction shows that in the timed condition, the drop in accuracy from a 9/0 proportion to a 5/4 proportion is quite linear, whereas in the self-paced condition, the decline in performance is less pronounced until the 5/4 proportion, when there is a clear drop in performance.

One-sample t-tests showed that most conditions had accuracy higher than 50% chance ($p < .05$), excepting six attractive faces in the timed condition ($t(19) = 1.07$, $p > .29$), five attractive faces in the timed condition ($t(19) = 0.10$, $p > .92$), and six attractive faces in the self-paced condition ($t(19) = 2.06$, $p > .05$), where performance was not significantly different from chance. Further, with five attractive faces in the self-paced condition, performance actually dropped significantly below chance ($M = 37.5$, $SD = 24.25$, $t(19) = -2.31$, $p < .05$).

A 2 (timing) x 2 (majority) x 5 (proportion) repeated measures ANOVA on reaction time data showed main effects of timing ($F(1,19) = 45.37$, $p < .001$, $\eta^2 = .71$), with self-paced understandably having higher RT ($M = 4093$, $SD = 2320$) than timed ($M = 700$, $SD = 274$), majority ($F(1,19) = 11.45$, $p < .005$, $\eta^2 = .38$), with majority attractive having higher RT ($M = 2616$, $SD = 1366$) than majority unattractive ($M = 2178$, $SD = 1105$), and proportion ($F(4,76) = 3.73$, $p < .01$, $\eta^2 = .16$), with a general increase in RT as proportion tended towards 5/4.

There was also a significant interaction between timing and majority ($F(1,19) = 10.19$, $p < .01$, $\eta^2 = .35$), with a smaller increase in RT from majority unattractive to majority attractive in the timed condition ($F(1,19) = 5.11$, $p < .05$, $\eta^2 = .21$) than in the self-paced condition ($F(1,19) = 10.94$, $p < .005$, $\eta^2 = .37$). Further, there was a marginally significant interaction between timing and proportion ($F(4,76) = 2.46$, $p = .053$, $\eta^2 = .12$), with no significant effect of proportion in the timed condition ($p > .51$), but a significant effect in the self-paced condition ($F(4,76) = 3.25$, $p < .05$, $\eta^2 = .15$).

These two interactions were further qualified by a marginally significant three-way interaction between timing, majority, and proportion ($F(4,76) = 2.45$, $p = .053$, $\eta^2 = .11$). Simple effects analysis of this interaction revealed that there was no interaction between majority and proportion in the timed condition ($p > .62$), but there was a

marginally significant one in the self-paced condition ($F(4,76) = 2.30$, $p = .066$, $\eta^2 = .11$). This can be further explained by a lack of main effect of proportion when the majority of the group was attractive ($p > .96$), with RTs remaining around 4500ms, but a main effect of proportion when the majority of the group was unattractive ($F(4,76) = 4.92$, $p < .005$, $\eta^2 = .21$), with a general increase in RT as the proportion moved towards 5/4. These effects are all illustrated in *Figure* 11.

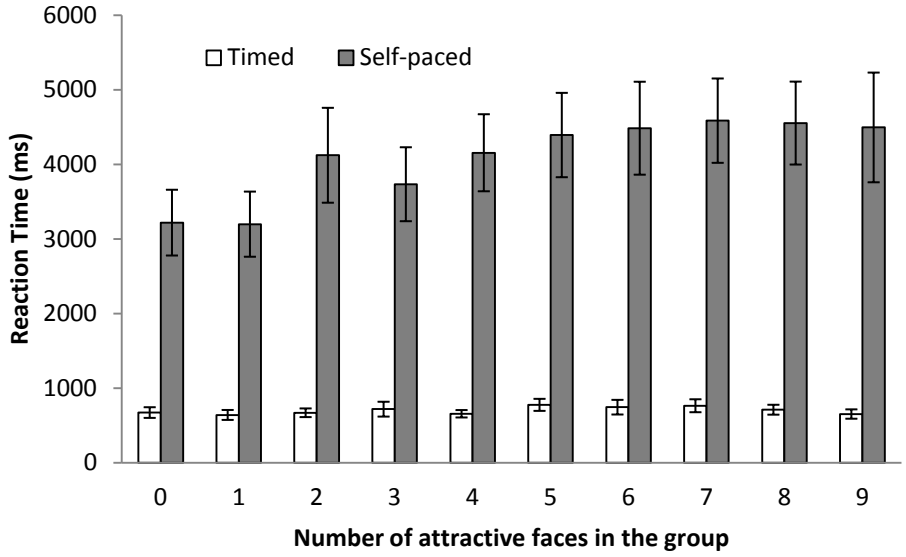


Figure 11: Mean reaction time at each level of number of attractive faces in the group, separated by timing. Error bars indicate standard error.

3.1.4. Discussion

This experiment was designed to identify whether an unconstrained viewing time would affect the judgement of attractiveness of a group of faces, while also investigating the overall perception of the attractiveness of a group across differing configurations of the groups. The results suggested that increased viewing time did not provide a general improvement to the performance of the task, but that it did, to some

extent, lessen the impact of the increasing difficulty as the groups became more evenly balanced.

When given unrestricted time to study the groups, participants' performance showed less of a drop in accuracy as the composition of the groups became less skewed towards attractive (or unattractive) faces, compared with the timed condition. The only point when this reduced impact was not evident was when the group contained five attractive faces (and thus, four unattractive faces), in which conditions, participants seemed more likely to identify the group as containing a majority of unattractive faces. This effect likely ties in with the interaction between majority and timing.

Performance suggests that when given time to study the groups of faces, there is a significant impact of the majority of those faces being unattractive, with accuracy increasing. This might be because participants linger longer on the unattractive faces, thereby lending them more weight in whatever representation of the group the participant generates, or perhaps there is something more salient in the unattractive faces, such as the threat of disease (Thornhill & Gangestad, 1993; Thornhill & Gangestad, 1999) that gives them more impact. It could also be that it is the number of unattractive faces that is being overestimated, rather than a general level of unattractiveness. While this impact could (and, perhaps, should) be lesser when attractive faces make up the majority of the group, the slight majority of five over four might not be enough to override this effect, leading to the effect of responses being skewed towards the group being unattractive, as observed in the five-attractive face groups in the self-paced condition.

Of course, reaction times in the self-paced condition can be taken as a strong indicator of how much time participants committed to studying the groups, and the general reduction in RT as unattractive faces became more prevalent in the group would

suggest that, actually, unattractive faces were not holding attention as much as attractive faces might have been. This, in turn, feeds the earlier suggestion that there is something inherent in unattractive faces that means they have a greater impact on the assessment of the attractiveness of a group.

Both the possibility of differing amounts of visual attention being directed to attractive versus unattractive faces, and of over- or underestimation of the number of faces of either type, will be addressed in later experiments.

3.2. Experiment 4: How do judgements of majority differ from estimates of frequency, and do attractive faces draw visual attention during these tasks?

3.2.1. Introduction

The results of Experiment 3 hinted at unattractive faces having some stronger influence or greater import in the assessment of the attractiveness of a group. As such, Experiment 4 set out to understand both facets of whether unattractive faces draw more visual attention than do attractive faces, and whether the absolute numbers of faces of either type are being incorrectly estimated.

Previous research has suggested that attractive faces capture attention (Maner, Kenrick, Becker, Delton, Hofer, Wilbur, & Neuberg, 2003; Maner, Gailliot, & DeWall, 2007; DeWall & Maner, 2008; Sui & Liu, 2009), and that attractive faces garner longer fixations (Leder, Tinio, Fuchs, & Bohrn, 2010). There has also been some question over whether faces "pop-out" from among other stimuli (Treisman & Gelade, 1980; Nothdurft, 1993; Brown, Huey, & Findlay, 1997; Santhi & Reeves, 2004; Hershler & Hochstein, 2005). Faces do at least appear to pull visual attention, even when it is counterproductive to a task (Cerf, Frady, & Koch, 2009; Crouzet, Kirchner, & Thorpe, 2010), but there is little information regarding whether attractive faces elicit a similar pop-out effect when presented among other faces.

The question of visual attention potentially being allocated more towards unattractive faces can be addressed through the use of eye-tracking methods. Such methods can follow the gaze of participants, logging exactly where on the screen they are looking at a given moment. This information can then be used to determine which faces drew visual attention, for how long faces (unattractive or otherwise) were fixated upon, and the order in which faces were fixated.

The inclusion of a different task sought to address the issue of possible inaccurate assessment of the number of attractive or unattractive faces in the groups. It is unclear from the previous results whether participants are considering the number of attractive faces in the group when making their judgements, or are simply using a gist value. By asking participants to estimate the number of attractive faces in the group, and by building these groups based on participants' own ratings of the attractiveness of the faces, it is possible to determine whether participants are indeed making a count of the number of attractive faces, and whether the group setting somehow influences the perceived number of attractive faces in the group, and thus overall perceived attractiveness of the group.

Based on suggestions that attractive faces hold visual attention, it was hypothesised that more attractive faces would receive longer fixations, especially in the self-paced conditions. However, it is unclear from previous works whether visual attention would actually be more drawn to these faces, rather than simply held for longer once there. It was also hypothesised that when using participants' own ratings of the attractiveness of the stimuli, there would be a general increase in accuracy during the 10AFC, and in particular when comparing directly with the 2AFC.

3.2.2. Method

3.2.2.1. Participants

Twenty one undergraduates from the University of Hull (16 females) participated in this experiment. Their age ranged from 18 to 35 years (M = 21.62 years, SD = 5.55), and all had normal or corrected-to-normal vision. The 10AFC task was performed in a separate session, after the 2AFC task, and the same participants took part in both, excepting one participant who did not return for the 10AFC. The separation of the tasks

was to allow collection of the rating data to then generate the groups accurately for each participant in the 10AFC.

3.2.2.2. Stimuli

The faces used for this experiment were the same as in previous experiments, with the layout being the same as in Experiment 3.

3.2.2.3. Design and Procedure

In the 2AFC experiment, the task was primarily identical in set-up to Experiment 3, with the difficulty of the task varying by the number of attractive faces in the group (and thus varying the proportion of attractive:unattractive faces), and the order of trials and the faces in the group and their location all being randomised in the same way. However, this experiment was also expanded, to include an additional two time-restricted conditions, of 250ms and 1000ms displays, alongside the original 500ms display and the self-paced conditions. The inclusion of these additional conditions was intended to explore the effect of varying display time constraints on participants' eye movements. The order in which these timing conditions were conducted was counterbalanced across participants.

All participants took part in the 2AFC experiment session first, with the task being the same as in Experiment 3 - to indicate whether the group contained more attractive or more unattractive faces - but with the addition of the two new timing conditions. After completing this task, participants performed a rating task, in which they were presented with each of the faces used in the experiment, one at a time, in a randomised order. Each face was presented once, centrally on the screen, and participants were asked to rate the attractiveness of the face, on a 1-10 scale, with 1 being the most unattractive, and 10 being the most attractive. Each face remained on-screen until the participant responded,

using the number keys across the top of the keyboard (with the 0 key functioning as a rating of 10), at which point another face was displayed. These results were used to inform the experimental procedure for the 10AFC.

In the 10AFC, participants were asked to estimate the number of attractive faces present in a group, responding using the number keys on the keyboard. The task was performed in the same four display durations as the 2AFC, and again, the order of these was counterbalanced between participants. However, the groups of faces were constructed based on each participant's own ratings of their attractiveness. Any faces rated as 5 or below were considered to be unattractive to that participant, and any rated 6 or above were considered to be attractive to the participant. As such, the number of attractive faces in a given group was determined by these ratings, with each combination of zero to nine attractive faces appearing ten times during each display duration. It was therefore possible to know, subjectively, how many faces that a participant deemed to be attractive were actually present in a given group.

In both experiments, participants undertook a small selection of practice trials of the 500ms and self-paced conditions before the experiment began, and each block of a given display duration contained some preset breaks for participants, with a break between timing conditions as well.

During all experimental trials (excluding the rating task), participants' eye movements were recording using an Eyelink eye tracker. Participants were sat, placing their head on a chin rest, which was fixed at 57cm from the display screen, and were instructed that following calibration, they should remain in the chin rest until prompted otherwise. The eye tracker was calibrated at the beginning of each block of the experiment, and after each break, whether or not the participant removed their head from the chin rest.

The camera was positioned and focused to ensure a clear image of the eye, before beginning the calibration of the software. During calibration, participants were instructed to follow a small circle on the screen, moving their eyes while keeping their head still in the rest. The circle would move through nine positions on the screen in a random order. After this, the process would repeat in order to validate the calibration, all of which was also monitored for quality by the present experimenter. Following a successful calibration and validation, participants were instructed to begin the trials with a key press when ready.

3.2.3. Results

3.2.3.1. Accuracy and Reaction Times

3.2.3.1.1. 2AFC

A 4 (timing) x 2 (majority) x 5 (proportion) repeated measures ANOVA on accuracy data showed main effects of each factor, but no higher-order effects (all *p*s > .10). The effect of timing ($F(3,60) = 4.09$, $p < .05$, $\eta^2 = .17$) reflected a general increase in accuracy as display time increased, and planned contrasts showed that 250ms and 500ms were no different ($p > .91$), nor that 1000ms and self-paced were different ($p > .34$), but that there was a difference between 500ms and 1000ms ($F(1,20) = 5.71$, $p < .05$, $\eta^2 = .22$).

The main effect of majority ($F(1,20) = 13.42$, $p < .005$, $\eta^2 = .40$) showed that accuracy was higher when the majority of the group was unattractive ($M = 78.57$, $SD = 7.16$) than when it was attractive ($M = 62.50$, $SD = 14.73$). The effect of proportion ($F(2.63,52.66) = 119.26$, $p < .001$, $\eta^2 = .86$, Greenhouse-Geisser corrected) showed a

consistent increase in accuracy from the 5/4 proportion (*M* = 52.38, *SD* = 4.84) to the 9/0 proportion (*M* = 84.94, *SD* = 8.83), as illustrated in *Figure* 12.

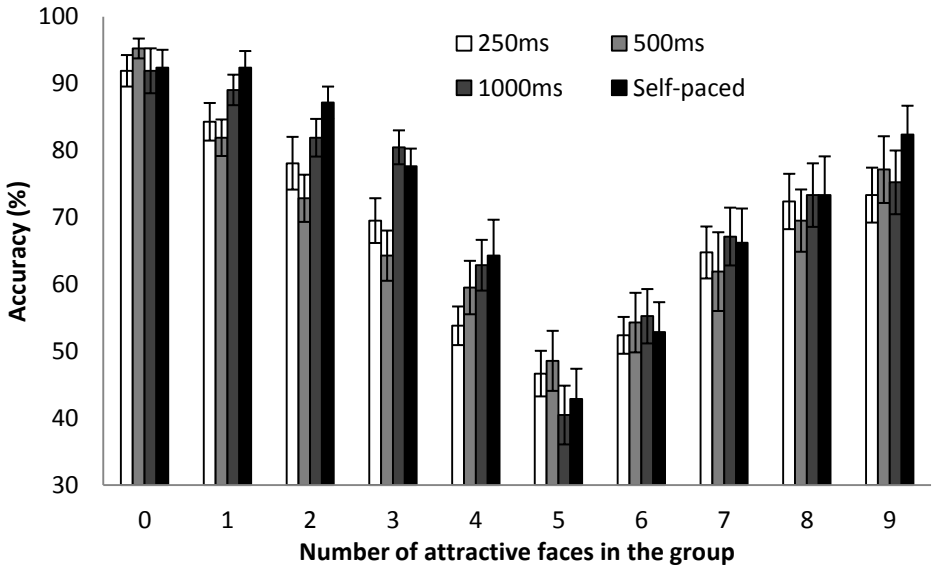


Figure 12: Accuracy at each level of number of attractive faces in the group, separated by timing. Error bars indicate standard error.

Reaction time data was analysed in the same manner, and showed only significant main effects of timing ($F(1.07,21.42) = 41.95$, $p < .001$, $\eta^2 = .68$, Greenhouse-Geisser corrected) and proportion ($F(2.61,52.12) = 7.29$, $p < .005$, $\eta^2 = .27$, Greenhouse-Geisser corrected), with no other effects or interactions. Planned contrasts showed no significant differences between 250ms and 500ms ($p = .234$), or between 500ms and 1000ms ($p = .625$), but there was a significant difference between 1000ms and self-paced ($F(1,20) = 45.64$, $p < .001$, $\eta^2 = .70$), with RT in the self-paced condition being understandably much greater (*M* = 2907, *SD* = 1561) than in 1000ms (*M* = 703, *SD* = 163). The effect of proportion showed a general increase in RT from the 9/0 proportion (*M* = 1203, *SD* = 460) to the 5/4 proportion (*M* = 1385, *SD* = 569). Reaction time data is displayed in *Figure* 13.

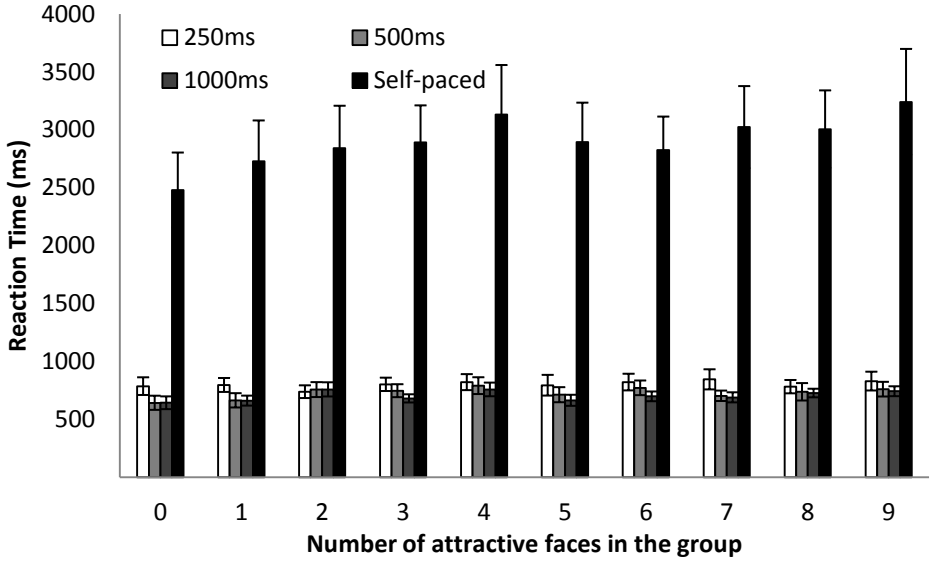


Figure 13: Mean reaction time at each level of number of attractive faces in the group, separated by timing. Error bars indicate standard error.

3.2.3.1.2. 10AFC

In the 10AFC, performance on the task is assessed by the degree by which participants' responses deviate from the number of faces in the group that they had previously identified as being attractive. As such, a positive value indicates an overestimation of this number, while a negative value indicates an underestimation. Unlike in most other tasks described and analysed in this thesis, this error does reflect a deviation away from the participants' own ratings, rather than the previously established baseline derived from earlier ratings of attractiveness. *Figure* 14 illustrates this degree of error across conditions, and for clarity, *Figure* 15 illustrates this same data but as the perceived number of attractive faces in the group.

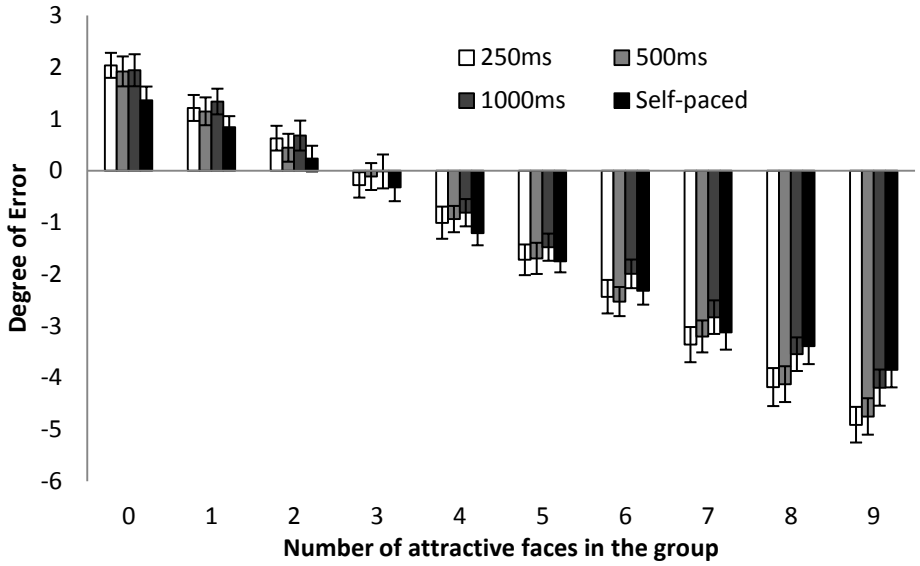


Figure 14: Degree of error in responses at each level of number of attractive faces in the group, separated by timing. Error bars indicate standard error.

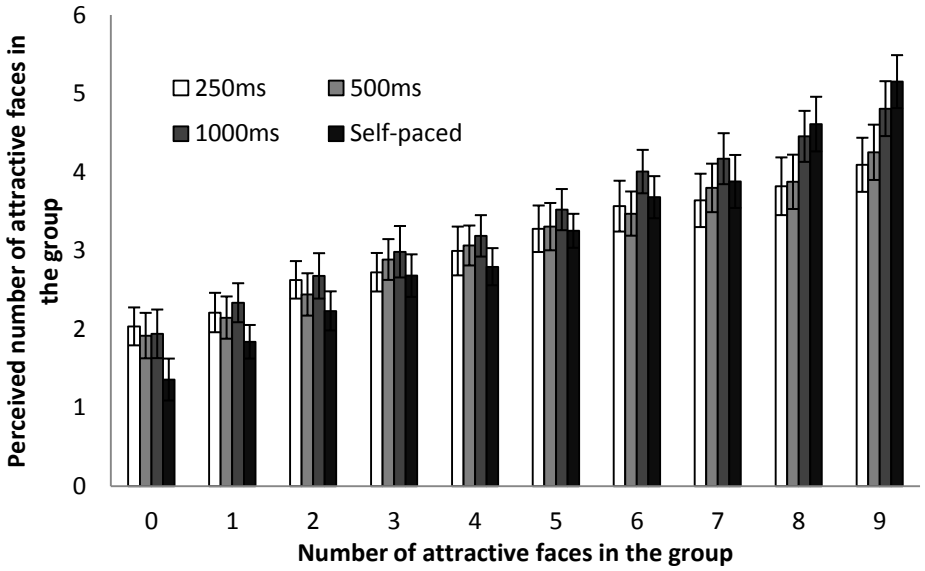


Figure 15: Perceived number of attractive faces at each level of number of attractive faces in the group, separated by timing. Error bars indicate standard error.

This data was analysed using a 4 (timing) x 10 (number of attractive faces) repeated measures ANOVA. This showed a main effect of number of attractive faces ($F(1.58,28.35) = 410.46$, $p < .001$, $\eta^2 = .96$, Greenhouse-Geisser corrected), with overestimations when the number of attractive faces is low, and underestimations when it is high. There was no main effect of timing ($p > .36$), but it did interact with number of attractive faces ($F(27,486) = 4.47$, $p < .001$, $\eta^2 = .20$). This interaction is clarified by simple effects analyses, which demonstrate a significant effect of timing only for zero ($F(3,54) = 5.12$, $p < .005$, $\eta^2 = .221$), eight ($F(3,54) = 3.60$, $p < .05$, $\eta^2 = .17$), and nine ($F(3,54) = 6.10$, $p < .005$, $\eta^2 = .25$) attractive faces (all other ps > .11). In each of these cases, there is a general decrease in error as display duration increases.

The general trend of this data shows that when there are fewer than three attractive faces in the group, participants generally overestimate the number, whereas this becomes consistently an underestimate at four or more attractive faces. One-sample t-tests confirm that with three attractive faces in the group, errors were not significantly different from 0 in any of the display durations (all ps > .25), and with two attractive faces responses were not significantly different from 0 in the 500ms display condition ($p > .11$) or the self-paced condition ($p > .36$), so in each of these situations, performance was essentially accurate. In all others, responses differed significantly from 0, following the trend mentioned above.

However, the relative accuracy at two and three attractive faces may also be due to a higher propensity to respond with a two or three. Because of this, it is difficult to derive whether the pattern of performance seen here is due to a tendency to specifically estimate the number of attractive faces more frequently at these numbers, or a more general tendency to respond in this range more often when performing a 10AFC of any sort.

Reaction times in this task were analysed using a 4 (timing) x 2 (majority) x 5 (proportion) repeated measures ANOVA. This showed main effects of timing ($F(1.04,18.78) = 61.35$, $p < .001$, $\eta^2 = .77$, Greenhouse-Geisser corrected), and majority ($F(1,18) = 15.16$, $p < .005$, $\eta^2 = .46$). Planned contrasts showed that 250ms, 500ms, and 1000ms all had similar RTs (all *p*s > .5), while self-paced had a significantly higher RT ($M = 4845$, $SD = 2237$) than 1000ms ($M = 1137$, $SD = 416$) ($F(1,18) = 60.37$, $p < .001$, $\eta^2 = .77$). RTs were slower when the majority of the group was attractive ($M = 2177$, $SD = 845$) than when it was unattractive ($M = 1974$, $SD = 736$).

There were also interactions between timing and majority ($F(1.14,20.57) = 12.79$, $p < .005$, $\eta^2 = .42$, Greenhouse-Geisser corrected), majority and proportion ($F(4,72) = 10.62$, $p < .001$, $\eta^2 = .37$), and majority, proportion, and timing ($F(3.87,69.68) = 6.26$, $p < .001$, $\eta^2 = .26$, Greenhouse-Geisser corrected). Further exploration shows that the interaction between proportion and majority is not present in any of the timed display conditions (all *p*s > .07), but that it is present in the self-paced condition ($F(4,72) = 8.62$, $p < .001$, $\eta^2 = .32$).

In the self-paced condition, when unattractive faces more strongly outnumber the attractive ones, reaction times drop ($F(2.34,42.05) = 3.57$, $p < .05$, $\eta^2 = .17$, Greenhouse-Geisser corrected), and the same is true when the majority is attractive; a higher proportion of attractive faces increases reaction time ($F(4,72) = 5.05$, $p < .005$, $\eta^2 = .22$). When given free time to inspect the group, while deciding the number of attractive faces in the group, participants took pointedly longer to make their response when there were more attractive faces, but this was not true in the restricted viewing conditions. This may reflect a genuine impact on the decision making process, or may simply be a result of attractive faces holding visual attention, with participants looking for longer at what is arguably more pleasant stimuli. RT data is displayed in *Figure* 16.

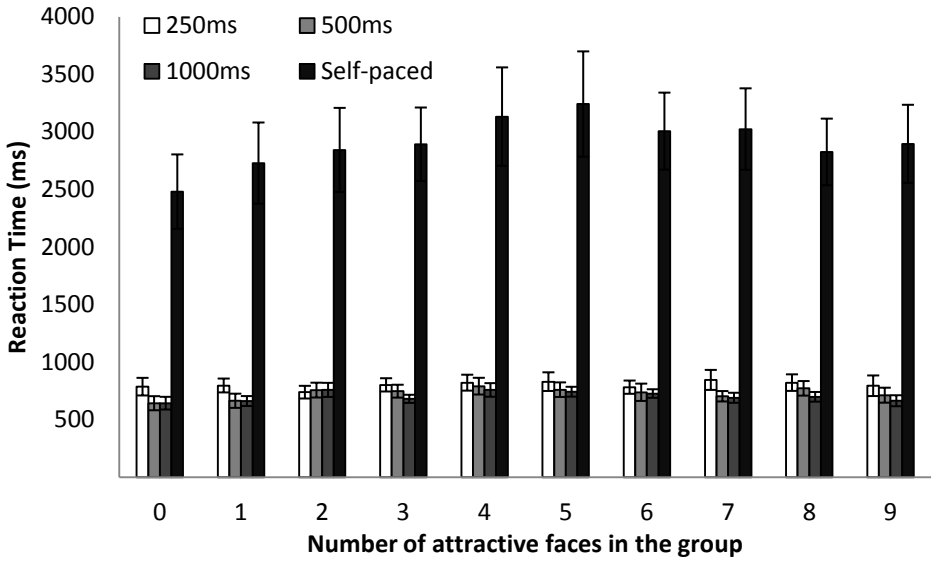


Figure 16: Mean reaction time at each level of number of attractive faces in the group, separated by timing. Error bars indicate standard error.

3.2.3.1.3. 2AFC and 10AFC comparison

In order to make a more direct comparison between performance in the 2AFC and that in the 10AFC, the data from the 10AFC was adjusted. To do this, all responses that indicated five or more attractive faces in the group were considered to be "more attractive" in a 2AFC, while all responses of four or below were considered "more unattractive". These responses were then coded as being simply accurate or inaccurate, based on the composition of the group.

Initially, a 4 (timing) x 2 (majority) x 5 (proportion) repeated measures ANOVA was conducted on just this data, and found several main effects and interactions. Wherever the assumption of sphericity is breached, *df*s and *p*s are corrected with the Greenhouse-Geisser correction.

There was a main effect of timing ($F(2,35.92) = 5.67$, $p < .01$, $\eta^2 = .24$), in the same pattern as seen in the 2AFC. Planned contrasts revealed the same patterns of

differences, with only a significant difference between 500ms and 1000ms ($F(1,18) = 5.01$, $p < .05$, $\eta^2 = .22$). There was also a main effect of majority ($F(1,18) = 48.15$, $p < .001$, $\eta^2 = .73$), with higher accuracy when the majority of the group was unattractive ($M = 87.05$, $SD = 11.27$) than when it was attractive ($M = 36.32$, $SD = 21.79$). Further, there was a main effect of proportion ($F(1.53,27.54) = 40.32$, $p < .001$, $\eta^2 = .69$), which also showed a similar pattern to that seen in the 2AFC, increasing from the 5/4 proportion ($M = 53.03$, $SD = 3.03$) to the 9/0 proportion ($M = 70.07$, $SD = 10.59$).

There were two-way interactions between timing and proportion ($F(6.53,117.45) = 2.37$, $p < .01$, $\eta^2 = .12$), majority and proportion ($F(4,72) = 4.96$, $p < .005$, $\eta^2 = .22$), and a three-way interaction between all three factors ($F(5.76,103.59) =$, $p < .05$, $\eta^2 = .14$). Further exploration of this three-way interaction shows that majority and proportion do not interact in any of the timed display conditions (all ps > .09), but that they do in the self-paced condition ($F(4,72) = 8.20$, $p < .001$, $\eta^2 = .31$), enough to show an overall interaction that is clearly not present in the other conditions. Simple effects on this interaction show that while there is an effect of proportion both when the majority is unattractive ($F(2.83,50.91) = 4.77$, $p < .005$, $\eta^2 = .21$) and when it is attractive ($F(2.20,39.56) = 19.59$, $p < .001$, $\eta^2 = .52$), it is much larger in the latter. This interaction is illustrated in *Figure* 17.

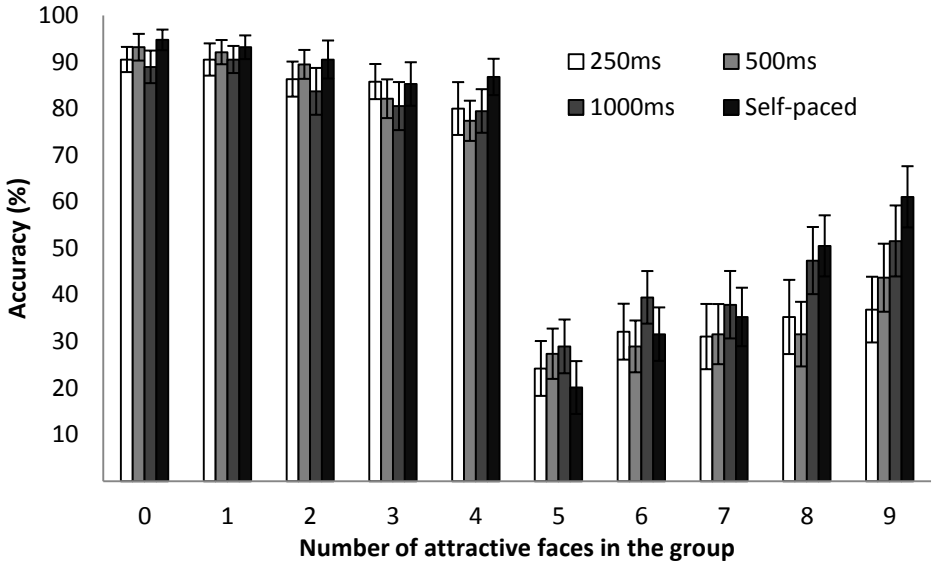


Figure 17: Accuracy at each level of number of attractive faces in the group, separated by timing. Error bars indicate standard error.

This data was then included in a 4 (timing) x 2 (majority) x 5 (proportion) x 2 (task) repeated measures ANOVA alongside the data from the 2AFC. Only effects involving the task variable are considered here, because all other effects are addressed in the analyses of each task. There was a main effect of task ($F(1,18) = 44.96$, $p < .001$, $\eta^2 = .71$), showing higher accuracy in the 2AFC ($M = 70.24$, $SD = 5.87$) than the 10AFC ($M = 61.69$, $SD = 6.86$). Task also had two-way interactions with majority ($F(1,18) = 31.03$, $p < .001$, $\eta^2 = .63$), and proportion ($F(4,72) = 2.00$, $p < .05$, $\eta^2 = .10$), and a three-way interaction with majority and proportion ($F(4,72) = 3.48$, $p < .05$, $\eta^2 = .16$). There was no four-way interaction.

For the interaction between majority and proportion, the effect of majority was much greater in the 10AFC ($F = 48.15$) than in the 2AFC ($F = 13.71$), whereas the effect of proportion was lesser in the 10AFC ($F = 40.32$) than in the 2AFC ($F = 97.77$). The interaction between majority and proportion was not significant in the 2AFC,

whereas it was in the 10AFC (F = 4.96), which, as seen previously, was only truly present in the self-paced condition. For clarity, *Figure* 18 shows the changing interaction between majority and proportion across the two tasks.

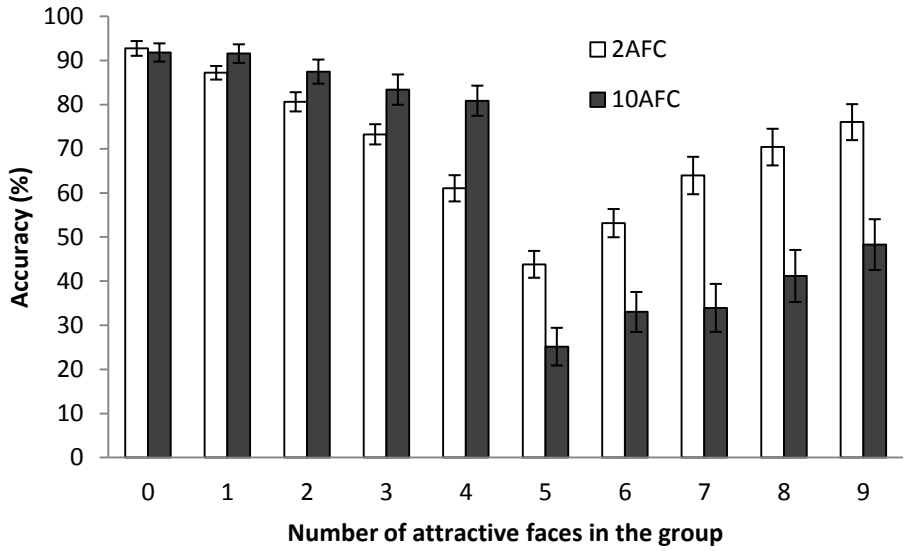


Figure 18: Accuracy at each level of number of attractive faces in the group, separated by task. Error bars indicate standard error.

3.2.3.2. Eye-tracking

The location and duration of each fixation made during both tasks was logged, and this was cross-referenced with the rating data collected between the two tasks. As such, for each fixation that is directed to a face, the participant's rating of the attractiveness of that face is known, as is the attractiveness ratings of each of the other faces in the group. Because in many cases an attractive face, or an unattractive face, will not be the only one of its kind in a group, it is unwise to consider simply whether a fixation is directed to an attractive or an unattractive face, so instead, consideration is based on whether the fixated face is more or less attractive than the mean attractiveness rating of the group. For example, in a group with nine attractive faces, any fixation to a

face will be to an attractive one, but this may actually be to one of the less attractive of these faces.

In the 2AFC, only 2% of all fixations were directed towards faces that were exactly equal in attractiveness to the mean of the group, and in the 10AFC this value was at 2.5%. As such, these fixations are not considered in the following analyses.

Because the number of faces above (or below) the average attractiveness of the group changes between trials, the number of fixations to faces of a certain type is somewhat uninformative. Instead, mean fixation duration will be used as a metric of visual attention. The pattern of fixation locations will also be explored, to understand whether certain types of faces drew visual attention more than others, and whether this affected performance on the respective tasks.

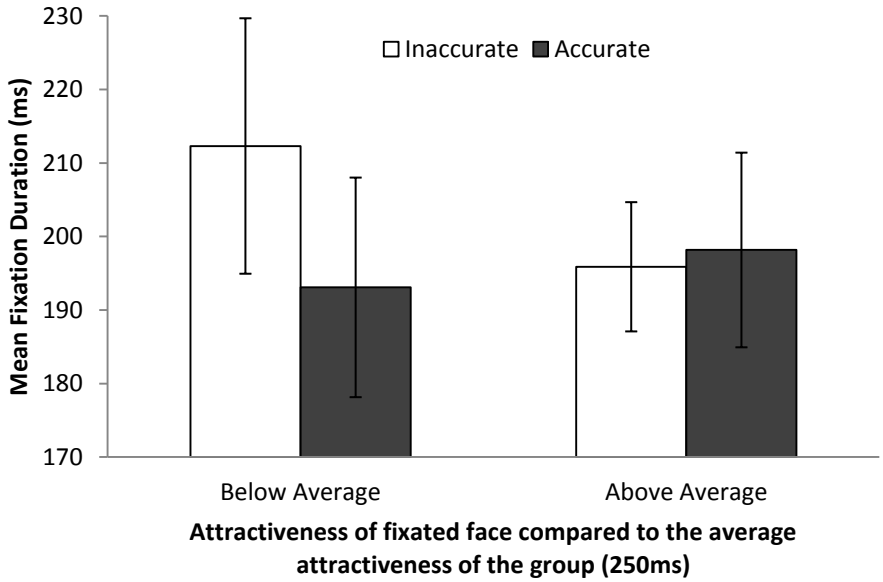
3.2.3.2.1. 2AFC fixation Duration

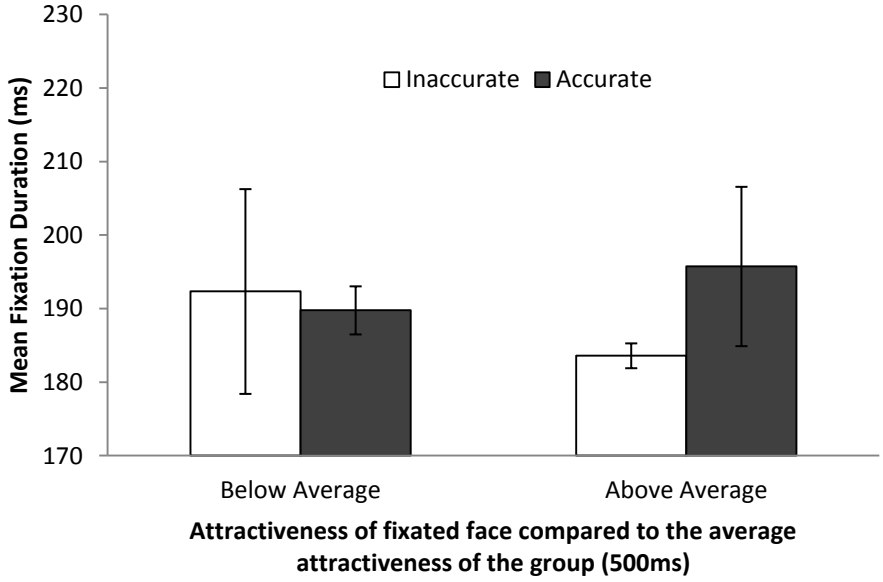
Fixation duration data (in ms) was analysed using a 4 (timing) x 2 (majority) x 5 (proportion) x 2 (accuracy) x 2 (above/below average) repeated measures ANOVA. Accuracy is included as a factor because performance on the task is likely to be directly related to which faces (or types of faces) were fixated.

This ANOVA found only two significant effects: a main effect of majority ($F(1,2) = 35.36$, $p > .05$, $\eta^2 = .95$); and an interaction between timing, accuracy, and above/below average ($F(3,6) = 8.79$, $p > .05$, $\eta^2 = .82$). The effect of majority simply shows that fixations were overall longer when the majority of the group was unattractive ($M = 206$, $SD = 20$) than when it was attractive ($M = 196$, $SD = 13$).

Further exploration of the interaction showed there to be no significant interaction between accuracy and above/below average in the 500ms, 1000ms, or self-paced

conditions (all *p*s > .05), but there was one present in the 250ms condition ($F(1,2) = 176.67$, $p < .01$, $\eta^2 = .99$). this is illustrated in *Figure* 19. Simple effects showed that in the 250ms condition, there was no difference between the duration of fixations to faces above and below the average of the group in trials that were responded to accurately ($p > .6$), whereas there was a marginally significant difference in inaccurate trials ($F(1,2) = 16.86$, $p = .054$, $\eta^2 = .89$), with faces below the average receiving slightly longer fixations ($M = 212$, $SD = 76$) than those above average ($M = 196$, $SD = 38$).





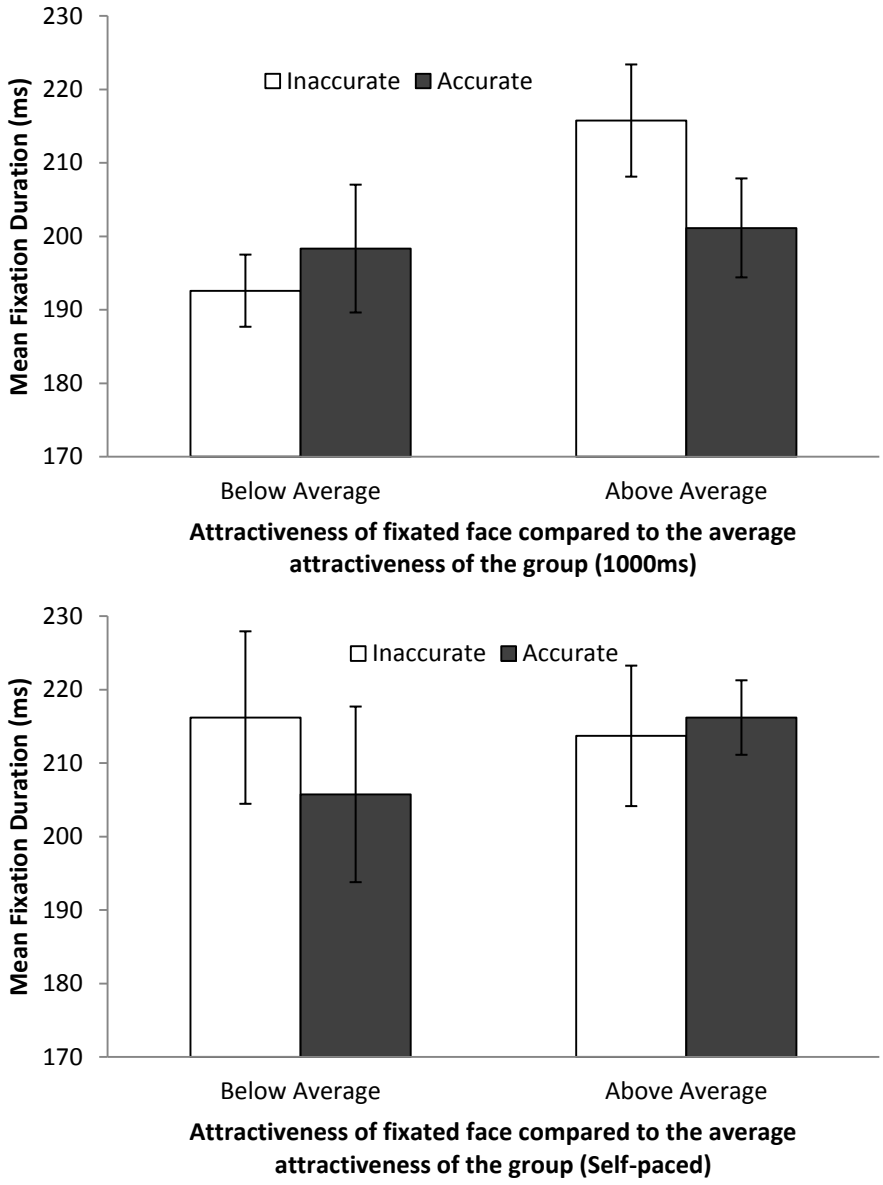


Figure 19: Mean fixation duration to faces above and below the average attractiveness of the group in trials with an accurate response and trials with an inaccurate response, separated by timing. Error bars indicate standard error.

3.2.3.2.2. 10AFC fixation duration

Performance on the trials is not included in the analyses of the 10AFC data, because using all ten levels of degree of error would result in a dataset with large portions simply missing because participants did not make an error of the given size in a

trial of the given condition, and this would make statistical tests impossible. As such, a 4 (timing) x 2 (majority) x 5 (proportion) x 2 (above/below average) repeated measures ANOVA was used to analyse fixation durations in the 10AFC.

This ANOVA found main effects of timing ($F(1.99,33.90) = 20.12$, $p < .001$, $\eta^2 = .54$, Greenhouse-Geisser corrected) and majority ($F(1,17) = 6.81$, $p < .05$, $\eta^2 = .29$), with slightly longer fixations when the majority of the group was attractive ($M = 201$, $SD = 18$) than when it was unattractive ($M = 197$, $SD = 17$), and a general increase in fixation duration as display duration increased, but planned contrasts showed that fixation durations were significantly longer in 250ms than in 500ms ($F(1,17) = 7.18$, $p < .02$, $\eta^2 = .30$), no difference between 500ms and 1000ms ($p > .79$), while durations in the self-paced condition were substantially larger than 1000ms ($F(1,17) = 40.50$, $p < .001$, $\eta^2 = .70$), as illustrated in *Figure* 21. There were also three two-way interactions: between majority and proportion ($F(4,68) = 3.02$, $p < .05$, $\eta^2 = .15$); between majority and above/below average ($F(1,17) = 6.44$, $p < .05$, $\eta^2 = .27$); and between timing and above/below average ($F(3,51) = 2.83$, $p < .05$, $\eta^2 = .14$).

The interaction between majority and above/below average stems from there being no difference between duration of fixations to faces above the average of the group when the majority is unattractive ($M = 199$, $SD = 17$) and when it is attractive ($M = 200$, $SD = 18$) ($p > .64$), but those faces below the average of the group garnered longer fixations when the majority was attractive ($M = 202$, $SD = 19$) than when it was unattractive ($M = 195$, $SD = 19$) ($F(1,17) = 12.98$, $p < .005$, $\eta^2 = .43$). This interaction is shown in *Figure* 20.

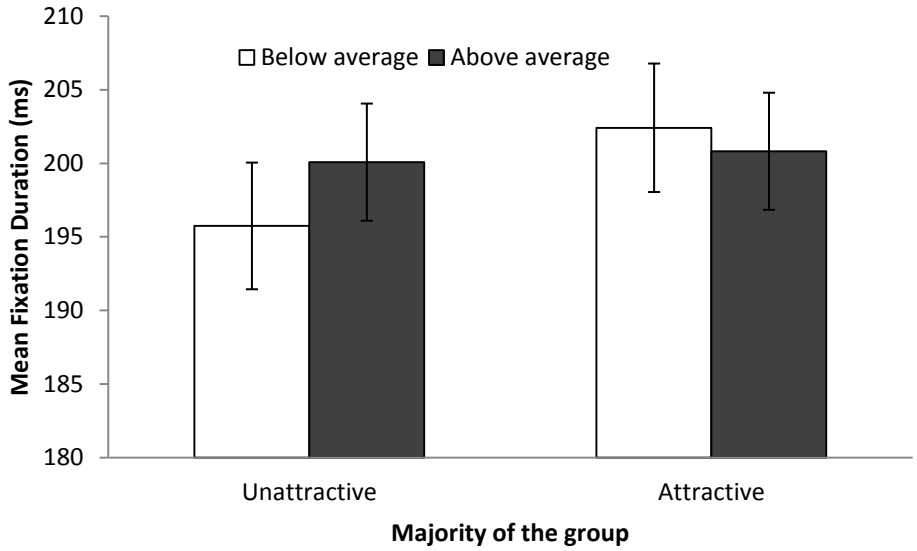


Figure 20: Mean fixation duration to faces above and below the average attractiveness of the group in trials when the majority of the group was either attractive or unattractive.

The interaction between timing and above/below average is explained simply enough; there is no difference between duration of fixations to faces above or below the average of the group in any of the timed conditions (all *p*s > .33), but there is in the self-paced condition ($F(1,17) = 12.31$, $p < .005$, $\eta^2 = .42$). In the self-paced condition, fixations to faces above the average ($M = 230$, $SD = 27$) are consistently longer ($F(1,17) = 12.31$, $p < .005$, $\eta^2 = .42$) than to those below the average attractiveness of the group ($M = 224$, $SD = 18$). This is illustrated in *Figure* 21.

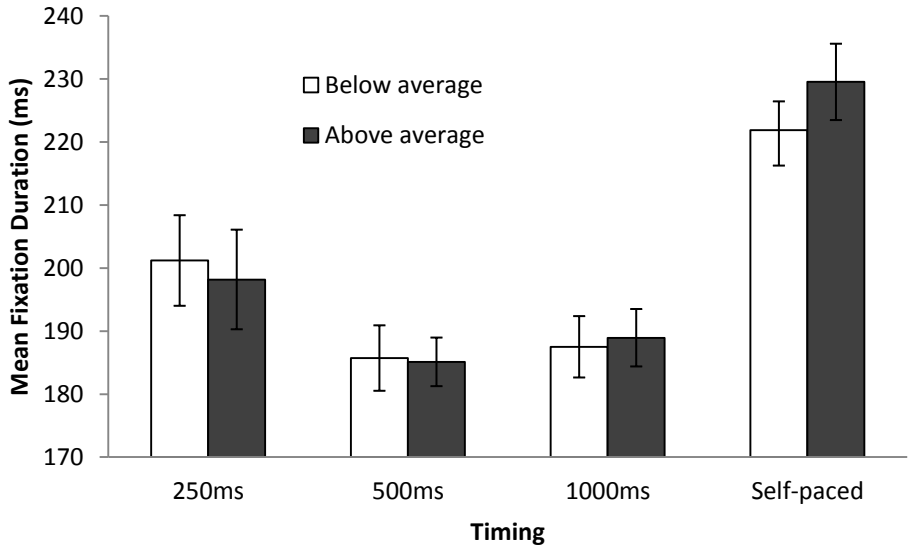


Figure 21: Mean fixation duration to faces above and below the average attractiveness of the group across timing conditions. Error bars indicate standard error.

Finally, the majority x proportion interaction is also relatively easy to understand. Simple effects showed no significant differences between fixation duration between one and eight attractive faces, three and six attractive faces, or five and four attractive faces (all *p*s > .18), with all six of these conditions having mean fixation durations around 198-200ms. However, there were significant differences between two and seven attractive faces ($F(1,17) = 9.00$, $p < .01$, $\eta^2 = .35$), with longer fixations when there were seven ($M = 205$, $SD = 19$) than when there were two ($M = 196$, $SD = 17$), and between zero and nine attractive faces ($F(1,17) = 12.21$, $p < .005$, $\eta^2 = .42$), with longer fixations to nine ($M = 203$, $SD = 20$) than to zero attractive faces ($M = 196$, $SD = 21$). This is illustrated in *Figure* 22.

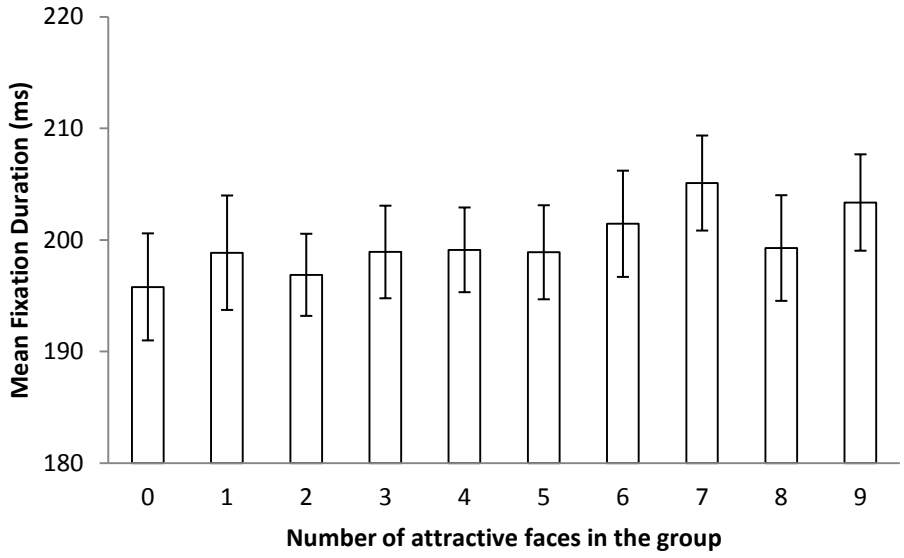


Figure 22: Mean fixation duration at each level of number of attractive faces in the group. Error bars indicate standard error.

3.2.3.2.3. 2AFC visual patterns

If more attractive and more unattractive faces capture visual attention differently, then the frequency of fixations to a given location over time should change dependent on the nature of the face displayed in that location, and of those displayed in the rest of the group. The frequency of fixations over time in one condition can be compared with those in another condition using a Kolmogorov-Smirnov (K-S) test. This test compares the shape and size of two distributions. In this instance, the x-axis of the distribution represents relative time in a trial as a function of number of fixations, while the y-axis is simply the frequency of recorded fixations to that location.

All data was split by timing condition, because they each allow the opportunity for different numbers of fixations in a trial. Then, with each location being analysed separately, data from trials with each number of attractive faces was compared with trials with each other number of attractive faces, and data from trials where the face in

that location was above the average attractiveness of the group was compared with those where it was below it, and trials that were responded to accurately were compared with trials that were responded to inaccurately.

With the K-S test, a non-significant result indicates that the two distributions being compared do not differ from one another. Of all of the comparisons listed, none were significant (all *p*s > .05). The number of attractive faces in the group, whether the face in the location was more or less attractive than the average of the group, and whether the participants responded correctly to the trial or not, did not in any way affect the likelihood, over time, of participants fixating on a given location. As such, it would seem that gaze patterns were dictated purely by spatial information (the grid layout itself), rather than by the stimuli presented in the grid. Although clearly the stimuli did have some effects on how long the eyes lingered after fixation.

Given that gaze patterns seem to be dictated by the spatial properties of the grid, and not its contents, it is sensible to consider which patterns recur most frequently. To explore this, the sequence of fixations in each trial was simplified to omit any fixations directed at space not occupied by a face, and to disregard repeated fixations to a given face if they were not separated by a fixation to another location. Each location in the grid is assigned a number from 1-9, in the same manner as a telephone number pad, and a sequence of fixations is represented by the number of the fixated location, with each subsequent fixation separated by a dash (E.g. 5-2-6 would indicate the first fixation to the central face, the second to the top middle face, and the third to the middle right face). For each timing condition, the most frequently observed patterns in accurate trials, and inaccurate trials, are listed in *Table* 1.

Table 1: The most frequently observed gaze patterns (and their frequency) in accurate and inaccurate trials in each timing condition

	250ms		**500ms**		**1000ms**		**Self-paced**	
	Seq.	(%)	Seq.	(%)	Seq.	(%)	Seq.	(%)
Inaccurate	5	57.6	5	21.5	5-4-1-2	2.5	2-1	1.6
	2	10.6	5-4	13.8	5-2	1.7	5	1.4
	5-6	5.4	5-2	12.8	5-6-3-2	1.7	2	1.2
	5-4	5.1	5-6	9.7	5-4-5	1.5	5-2	1.0
	5-2	3.4	2-5	4.9	2-1-5	1.4	5-2-1	1.0
Accurate	5	58.3	5	21.5	5-4-1-2	3.2	1-2-3-4-5-6-7-8-9	1.6
	2	9.3	5-4	16.0	5-2-5-8	1.7	2-1	1.2
	5-2	5.9	5-2	14.0	5-2-3	1.5	5-4	1.0
	5-4	5.8	5-6	11.4	5-6-9-8	1.4	5-4-5	0.9
	5-6	5.2	5-8	4.7	5-6-3-2	1.3	5	0.9

As can be seen from *Table* 1, there is little variation in patterns between the 250ms and 500ms conditions, and almost no variation between their respective accurate and inaccurate trials. These largely involve a central fixation followed by one other fixation in a cardinal direction. However, in the 1000ms condition, patterns suggest that following the initial central fixation, further fixations are being made in a circular pattern around the outer faces of the group, more so for accurate trials, with inaccurate trials demonstrating the same patterns observed in 250ms and 500ms trials, but with an additional subsequent central fixation.

Interestingly, despite the relatively low frequency of each sequence in the self-paced condition, and the less structured nature of most of the commonly observed patterns, the most frequent pattern in accurate trials follows a left-to-right, top-to-

bottom reading pattern. This demonstrates that when given the opportunity to do so, participants will systematically inspect each image in the group to make a decision.

The similarity between patterns observed in accurate trials and those in inaccurate trials, especially in the shorter display durations, leads to the question of whether performance at the task is largely dependent on whether the attractiveness of the faces fixated in these patterns happens to also reflect the attractiveness of the group as a whole. Are participants only basing their judgement on the fixated faces, or are they able to draw information from other faces in the group to allow a more informed response?

To investigate this, each trial was labeled as either a "representative" trial, or a "falsely representative" one. Representative trials were simply those trials in which the majority of fixated faces were the same attractiveness (more attractive/less attractive) as the majority of all faces in the group (E.g. the participant fixated on three attractive faces and one unattractive face, and the group contained seven attractive faces, thus both the fixated faces and the group as a whole contained a majority of attractive faces). Falsely representative cases were trials in which the majority of fixated faces were the same attractiveness as the minority of the group (E.g. the participant fixated on three attractive faces and one unattractive face, and the group only contained three attractive faces, thus the fixated faces had a majority of attractive faces, while the group actually had a majority of unattractive faces). There were also trials in which an equal number of attractive and unattractive faces were fixated, and thus are neither representative nor falsely representative. These “unrepresentative” trials are considered separately.

The effect of representativeness of fixations on the accuracy of a trial was examined using a 4 (timing) x 2 (majority) x 2 (representativeness) repeated measures

ANOVA. Proportion was not included in this analysis because many of the sub-levels of conditions did not contain any data, with no trials matching those specific criteria.

This ANOVA found only one significant effect; of representativeness ($F(1,11) = 90.79$, $p < .001$, $\eta^2 = .89$). Representative trials had 79.25% accuracy ($SD = 7.96$), compared to 44.29% accuracy ($SD = 6.22$) in falsely representative trials. This suggests that performance in a trial is, unsurprisingly, somewhat dependent on the faces fixated in that trial, but also that faces outside of fixation must play a role in task performance - accuracy in falsely representative trials would be much lower than approximately chance if the response was based solely on the fixated faces (in fact, one might reasonably expect it to be around 20%, given the error observed in representative trials).

A 4 (timing) x 2 (majority) repeated measures ANOVA on the unrepresentative trials showed no significant change in accuracy on the task across different conditions (all ps > .65). This data further supports the suggestion that performance is not solely reliant on the fixated faces, with an accuracy of 62.5% ($SD = 6.34$), when this would be expected at chance level (50%) if the decision was based solely on the observed faces.

From these results, it is clear that when fixating on faces representative of the group, performance is improved, but that information taken from outside of fixations must also be informing performance. As the majority of the group becomes more pronounced (I.e. towards zero or nine attractive faces), there is a potential for an increase in the number of faces that are representative of the group, but have not been fixated (indeed, with zero or nine attractive faces, each face in the group should be representative of that group – depending upon each participant's own ratings of the faces). As such, if non-fixated faces contribute to the assessment of the group, then performance should increase when the representativeness of these faces increases.

In order to test this, accuracy in only representative trials was assessed. A 4 (timing) x 2 (majority) x 5 (proportion) repeated measures ANOVA showed an effect of proportion ($F(1.79,19.63) = 8.70$, $p < .005$, $\eta^2 = .44$) and a just-significant effect of majority ($F(1,11) = 4.84$, $p = .05$, $\eta^2 = .31$), with no other significant effects or interactions. The effect of majority shows that accuracy is higher ($M = 83.20$, $SD = 7.77$) when unattractive faces outnumber attractive faces ($M = 76.98$, $SD = 9.26$). The effect of proportion supports the main hypothesis for this analysis, that as the majority of the group becomes more pronounced, accuracy increases from 69.69% ($SD = 13.86$) at a 5:4 proportion to 86.89% ($SD = 7.58$) at a 9:0 proportion. Thus, the non-fixated faces seem to also contribute to the assessment of the group.

3.2.3.2.4. 10AFC visual patterns

Using the same method as for the 2AFC, the distribution of fixations over time were calculated for each location. Again, split by timing, K-S comparisons of each location were made between the different levels of number of attractive faces, whether the face in that location was above or below the average attractiveness for the group, and the degree of error in the participants' response. None of these comparisons showed a significant difference between the distributions (all ps > .05), which follows the finding in the 2AFC that location of fixations was not affected by the attractiveness of the faces in the group, either the one being fixated or those surrounding it.

Table 2 lists the most commonly observed fixation patterns in each timing and for each level of degree of error. As can be seen, many of these patterns resemble those in the 2AFC, while an increase in error tends to show more sporadic, less inclusive patterns in the 1000ms and self-paced conditions.

Table 2: The most frequently observed gaze patterns (and their frequency) in trials with differing degrees of error in each timing condition.

	250ms		500ms		1000ms		Self-paced	
Error (no. attractive faces)	Seq.	(%)	Seq.	(%)	Seq.	(%)	Seq.	(%)
0	5	40.9	5-6	12.3	5-2-3-6	2.0	1-2-3-4-5-6-7-8-9	2.9
	5-2	15.1	5-2	11.1	5-4-7-9	1.7	1-2-3-6-5-4-7-8-9	1.9
	5-6	11.9	5-4	8.2	5-6-3-2-8	1.3	1-2-3-4-5-6-7-8-9-5	1.3
	2	9.5	5	2.9	5-6-3-1	1.3		
	5-4	6.8	2-5	2.9	5-6-8-7	1.3		
1	5	34.8	5-2	10.7	5-2-3	1.8	1-2-3-6-5-4-7-8-9	1.3
	5-6	16.3	5-4	10.1	5-6-3-2	1.8	5	1.1
	5-2	15.3	5-6	9.6	5-4	1.5	1-2-3-4-5-6-7-8-9	0.9
	5-4	9.2	5	4.8	5-4-1-2	1.5	5-2-1-2-3-6-5-4-7-8-9	0.9
	2	6.6	2-5	4.8	5-2-3-9-8	1.3	1-2-3-4-5-6-7-8-9-5	0.6
2	5	36.0	5-6	12.3	5-6-3-2	2.4	5-4	1.2
	5-2	16.3	5-4	10.8	5-2-3-6	2.1	5-4-5	1.2
	5-6	13.0	5-2	9.5	5-4-1-2-3	1.9	1-2-3-6-5-4-7-8-9	1.2
	5-4	9.5	5	4.9	5-6-9	1.6	5-8	0.9
	2	6.5	5-8	3.6	5-2-1-4	1.6	1-2-3-4-5-6-7-8-9-5	0.9
3	5	36.8	5-6	12.9	5	2.8	5	2
	5-6	13.4	5-2	9.7	5-2-3-6	2.8	1-2-3-6-5-4-7-8-9	1.2
	5-2	9.7	5-4	9.4	5-2-1-4	2.1	1-2-3-4-5-6-7-8-9-5	1.2
	2	9.3	5	7.6	5-4-1-2	2.1	1-2-3-6-5-4-7-8-9-5	1.2
	5-4	8.9	2-5	4.2	5-2-3-6-9	1.8	1-2-3-4-5-6-7-8-9	0.8
4	5	38.92	5-4	12.9	5	2.4	5	3.8
	5-2	10.3	5-2	9.1	5-6	2.4	5-6	1.1
	5-6	10.3	5-6	8.6	5-4-1-2	2.4	5-6-3-2	1.1
	5-4	8.7	5-8	5.2	5-2-1	1.9	5-6-9-8	1.1
	2	7.6	5-2-3	5.2	5-4-5	1.9	1-2-3-5-4-8-9	1.1

5	5	32.9	5-2	9.7	5	3.5	1-2-3-6-5-4-7-8-9	3.3
	5-6	16.8	5-4	9.7	5-4-7	2.1	5-4	1.7
	5-2	16.2	5-6	8.4	2-5-6-8	2.1	5-4-5	1.7
	5-4	10.2	5	6.5				
	2	3.6	2-5	6.5				
6	5	38.6	5-4	11.8	5-4-1-2	3.9	5-4	2.8
	5-2	11.4	5-6	11.8	5-4	2.6	5-6-9-8-7-4-1-3	2.8
	5-6	9.7	5	6.4	5-8	2.6		
	2	7.9	5-2	5.5	5-4-5	2.6		
	5-4	4.4	5-1-2	3.6	5-4-1-2-3	2.6		
7	5	33.9	5	9.1	5-4-1-2	4.8		
	2	16.1	2-5	9.1				
	5-2	14.5	5-6	9.1				
	5-6	9.7	5-2	6.8				
	5-4	4.8	2-5-2	6.8				
8	5	50.0	2-1	18.8	5-4-5	14.3	2-5-8	28.6
	5-6	14.3	2-4	18.8	1-2-3-6	14.3		
	2	7.1			5-2-1-6	14.3		
	5-2	7.1			5-2-3-6	14.3		
	5-4	7.1			5-6-3-2	14.3		
9	2	50.0	5-4	50.0	5-2-7	0.0	1-2-1-5-8	50.0
	2-5	25.0	5-8	25.0	5-2-1-5-6-8	0.0	2-3-6-5-8-7-9	50.0
	2-9	25.0	5-9	25.0				

† Missing cases or truncated lists indicate either no more cases in that condition (where the percentages sum to 100%), or that of the remaining cases, there were numerous different sequences, all at the same incredibly low percentage, meaning exemplary cases could not easily be selected and would not be wholly representative of the sequences observed.

It is not possible to consider the representativeness of fixation sequences in the 10AFC in the same way as done in the 2AFC. This is because for a sequence of

fixations to be truly representative, a participant needs to either fixate on every face in the group, or to fixate on every attractive face in the group, which would produce a representation that is accurate but also incomplete.

3.2.4. Discussion

The accuracy and RT data from the 2AFC task in Experiment 4 showed that there was no real impact of varying restricted display durations on the speed of responses, with 250ms, 500ms, and 1000ms exposures all having similar reaction times, suggesting a lack of hesitation, even in the shortest exposures. Whether this reflects a confidence in the decision, or an apathy given the restricted circumstances is perhaps slightly clarified by the accuracy of responses. That the 250ms and 500ms conditions did not differ in overall accuracy, but were in turn lower in accuracy than the 1000ms and self-paced conditions (which did not differ from each other), suggests perhaps a mix of both apathy (in the former) and confidence (in the latter). This suggestion of apathy does not necessarily mean that participants were apathetic about the task, more that they might simply have resigned themselves to a quick, estimated response, based on the restricted viewing conditions.

The results from the 10AFC are suggestive of the group setting having some impact on the way the faces are perceived. Despite removing the possible impact of individual, subjective ratings of attractiveness by having the groups composed of faces already rated for attractiveness by the participant, estimates of the number of attractive faces in the group still showed levels of inaccuracy, with a consistent overestimation at lower levels, and underestimation at higher ones. The fact that the least error appeared around the point of three attractive faces per group sits in line with previous experimental findings that the group format appears to skew perceptions (or at least

responses) towards the unattractive end of the spectrum. Even when given unrestricted viewing time, the same sorts of errors were occurring, though to a slightly lesser degree when there were no attractive faces (or no unattractive faces) in the group, but this still showed a slight overestimation (or a still substantial underestimation).

This skew in responses towards three attractive faces in the group might possibly reflect a tendency of responses, rather than a perceptual bias - participants might have been generally underestimating the number of the items for which they were looking, rather than necessarily underestimating the number of attractive faces in particular, these two things just happen to be one and the same in this scenario. The same could potentially happen when estimating the number of any target objects amongst other stimuli. The fact that Experiment 4 only asked participants to consider the number of attractive faces, and never the number of unattractive faces, is a flaw in its design that must be rectified with further study and limits the strength of the conclusions that can be drawn from its data.

Further, there may have been some uncertainty surrounding the instructions for the task, which did not clarify the meaning of the term “attractive face”. The intention of this instruction was to mean any face that would be considered a six or higher on a 10-point scale (while any face considered a five or lower would be an “unattractive face”), but participants may have mistaken “attractive face” to mean a “highly attractive” face, and thus have discounted some faces that might sit in the upper half of the scale, but still fall below the (albeit potentially arbitrary) threshold of “highly attractive”.

When comparing the results from the 10AFC with those from the 2AFC, participants were found to be less accurate in the 10AFC. While the tasks are not identical, and so any comparison is loose, at best, this does suggest a potential

differentiation between the general summary representations of the groups as “there are more attractive/unattractive faces in this group” and “I have counted the number of attractive faces in this group, and there are more/fewer than there are unattractive faces”. This also suggests a different approach to the task - if the approach was the same between the tasks, and an overall gist obtained, but participants were not able to accurately estimate the number of attractive faces, then a guess at that number should still reflect the detected majority, and results could reasonably be expected to be similar when making this comparison between the tasks, especially given that the stimuli in the 10AFC were tailored to reflect participants’ own ratings of attractiveness for the stimuli. In fact, on average, none of the majority attractive groups in the 10AFC were responded to with an estimate suggesting this majority.

These results continue to suggest that, when presented in a group setting, faces are perceived as being less attractive, whether it be some sort of singular gist representation (as appears to be the case in the 2AFC) or a representation of the actual composition of the group (as seemingly demonstrated in the 10AFC). However, it is still not clear from this data whether the bias stems from a perceptual or an attentional root. To answer this, the eye-tracking data must be explored.

It is clear from the analysis of gaze patterns and frequency distributions of fixations to each location in the group that the direction and location of gaze was not significantly directed by the attractiveness of the stimuli - there does not appear to be any sort of pop-out effect drawing visual attention to attractive (or to unattractive) faces, nor any suggestion of visual attention being held by either attractive or unattractive faces. In fact, the pattern of gaze across the groups was seemingly dictated by a systematic approach. In the two shorter timed conditions, there was only sufficient time for two fixations (perhaps occasionally three in the 500ms trials), and so any fixation

beyond the first was likely in an arbitrary direction, just to attempt to obtain some more information, but when given a full 1000ms to explore the group, participants begin to scan around the group in a circular motion, certainly in accurate/low-error trials, and trials in which this systematic approach is abandoned more frequently seem to lead to inaccurate/high-error responses. It appears that, when facing restricted viewing conditions, participants do still attempt to gather as much useful information about the group as they can, and that there is a certain degree of connection between making this attempt and performing better in the trial. It is unsurprising that in the self-paced condition, participants begin to adopt a left-to-right top-to-bottom reading style approach to viewing the groups, because the most effective way to gather information about all of the faces in the group is to fixate on each, and being a western culture, following the same pattern as reading makes a logical natural pattern to follow. It does, however, raise the question of whether a similar systematic approach might be adopted elsewhere in the world where the written word is read in a different direction (such as in Japan), but the pattern reflects this different reading pattern.

Performance on the 2AFC was also found to be partially contingent on which faces were fixated, and whether they were representative of the majority of the group. This is an unsurprising result, given that an informed decision is far more likely to be correct. What was perhaps slightly more surprising (but also somewhat in line with other findings so far) was that even when the fixated faces were representative of the majority of the group, accuracy was still only ~80% (and not far off chance levels when the faces were not representative, at ~45%), suggesting that non-fixated faces were still having some impact on the perception of the group. In representative trials, accuracy was higher when the majority of the group was unattractive than when it was attractive, which suggests a slight tendency to judge a group as having fewer attractive faces even when the majority of the fixated faces are actually attractive.

So, while performance on the 2AFC does appear to be influenced by the attractiveness of the fixated faces, the attractiveness of the faces doesn't seem to have any impact, in either task, on whether a given face is fixated in the first place. Further, in the 2AFC, the attractiveness of a fixated face seems not to have any impact on the duration of fixations to it either, whereas the general majority of the group does, with a majority of unattractive faces resulting in a small, but still significant increase in fixation duration of 10ms compared with a majority of attractive faces. So it seems participants were looking at faces for a fraction of a second longer when the majority of the faces in the group were unattractive, which might relate to the possible reduced risk of disease associated with attractive faces (Thornhill & Gangestad, 1993; Thornhill & Gangestad, 1999), and taking just a moment longer to assess any such potential threat when unattractive faces were more prevalent.

In the 10AFC, the duration of fixations to faces was modulated partly by the attractiveness of the majority of the group, as in the 2AFC, but only in cases where the fixated face was below the average attractiveness of the group. It seems that faces that were above the average attractiveness for their group drew fixations of a similar length, irrespective of the attractiveness of the majority of the group, while faces below the average attractiveness of the group were fixated for shorter periods when surrounded by a majority of unattractive faces, and for longer periods when surrounded by an attractive majority. Less attractive faces holding attention for marginally longer when they differ from the majority of the group might reflect the novelty of the stimuli, and if the risk of disease is indeed linked with unattractive faces, then such a face being interspersed with a number of faces seemingly not similarly afflicted might temporarily raise suspicion.

In summary, the attractiveness of individual faces appears to have had no impact on which faces were fixated, or the order in which they were fixated, and the only

impact seemingly imparted on visual attention is a slight modulation of fixation duration, to the order of 100th of a second. However, when the subset of the faces in the group that are fixated corresponds to the overall make-up of the group, performance in the 2AFC improved, but in all cases the response seems to have been affected by faces outside of fixations. Estimates of the number of attractive faces in the group are skewed towards seeing the group as more unattractive, as in previous experiments, but this is even more pronounced than when simply judging the majority of the group.

3.3. Experiment 5: Do frequency estimates differ when looking for attractive versus unattractive faces?

3.3.1. Introduction

Performance on the 10AFC in Experiment 4 showed error to be at its lowest when there were three attractive faces in the group, and that degree of error increased as conditions moved away from this point. This, in turn, highlights a higher likelihood of responding with a three. However, because participants were only counting the number of attractive faces in the group, it is unclear whether this obvious tendency to report a value closer to three was due to a general underestimation of number when performing a 10AFC of this sort, or whether it links to a specific underestimation of the attractive faces in the group. This point was clarified with Experiment 5, in which participants estimated both the number of attractive faces in one block, and the number of unattractive faces in another block.

If responses still show least error around three when estimating the number of unattractive faces, this would indicate a response bias, whereas less error centering around six would suggest a perceptual bias of more unattractive faces (or fewer attractive faces) in a group than there actually are.

Further, the instructions given to participants in the 10AFC task in Experiment 4 may have been a source of some of this shift in responses. Participants were only asked to estimate the number of attractive faces, and the meaning of this phrase was not qualified. It is entirely possible that participants took this instruction to mean "estimate the number of highly attractive faces in the group", when in actuality they were expected to estimate the number of faces on the upper half of the attractiveness scale (I.e. six or above on a 1-10 scale). As such, some moderately attractive faces (I.e. sixes and sevens, possibly eights) may have been discounted from the assessment of the

group by the participants, but still contributed to the number of attractive faces in the group when calculating error. To counter this point, in Experiment 5 the instructions for both tasks explicitly explained that an "attractive" face was one that was six or above on a 1-10 scale, while an "unattractive" face was five or below. Any effect of instructions should become obvious when comparing performance on the 10AFC in Experiment 4 with that when estimating the number of attractive faces in Experiment 5.

Thus, Experiment 5 is intended as a disambiguation of the results found in the 10AFC of Experiment 4. It aims to establish whether the existing skew in responses towards three is a bias in responding at this point on the 10AFC scale, or it is a perceptual bias for underestimating the number of attractive faces. Further, it aims to establish whether some degree of error can be attributed to an erroneous discounting of certain faces in the group due to a misinterpretation of the instructions.

3.3.2. Method

3.3.2.1. Participants

Twenty undergraduates from the University of Hull (18 females) participated in this experiment. Their age ranged from 18 to 42 years (M = 21.2 years, SD = 6.77), and all had normal or corrected-to-normal vision. None of the participants had participated in any of the previous experiments. The two tasks were performed in separate sessions, with the second session also containing a rating task, as listed previously.

3.3.2.2. Stimuli

For the 10AFC task, the stimuli were the same as in the 2AFC task in Experiment 4, arranged in the same way, and randomised into groups in the same manner. The ratings task was also identical to that used in Experiment 4.

3.3.2.3. Design and Procedure

The experiment was separated into two sessions, with participants completing a task similar to the 10AFC of Experiment 4 in each. The task differed from Experiment 4 in that in one session, participants were asked to estimate the number of attractive faces in the group (Attractive task), while in the other they were asked to estimate the number of unattractive faces (Unattractive task). In each, the instructions clarified that an attractive face was any that would be rated as a six or higher on a 1-10 scale, and that an unattractive face was any that would be rated as a five or lower, respectively. The order in which the two tasks were performed was counterbalanced across participants. At the end of the second session, participants also performed the same ratings task as in Experiment 4, and this was again used to verify participants' responses in the task. This task was conducted after the two experimental tasks, rather than before, in order for the ratings to be made in context, with participants having already seen the stimuli in their masked state, rather than in isolation, when the masked images might have appeared unusual, and thus skewed the judgements of the faces (especially for those observed earlier in the task, when the masking would have been unfamiliar).

3.3.3. Results

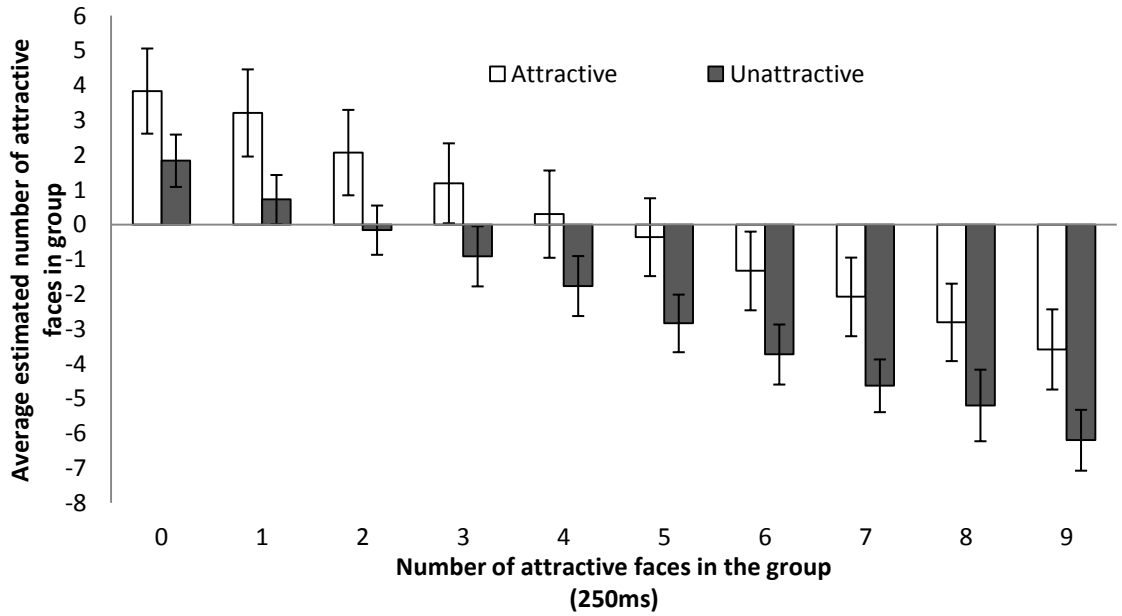
Because participants did not complete the ratings task until the end of the experimental sessions, the stimuli groups were categorised in terms of number of

attractive faces based on the original ratings, as in most of the previous experiments. This meant that the stimuli groups needed to be recategorised by the number of attractive faces in them as rated by each participant, with faces rated as a five or lower being considered unattractive and those rated as a six or higher being considered attractive (in line with the experimental instructions given to participants). As such, the number of trials of each adjusted condition varied. In all analyses presented here, the number of attractive faces in the group reflects this adjustment.

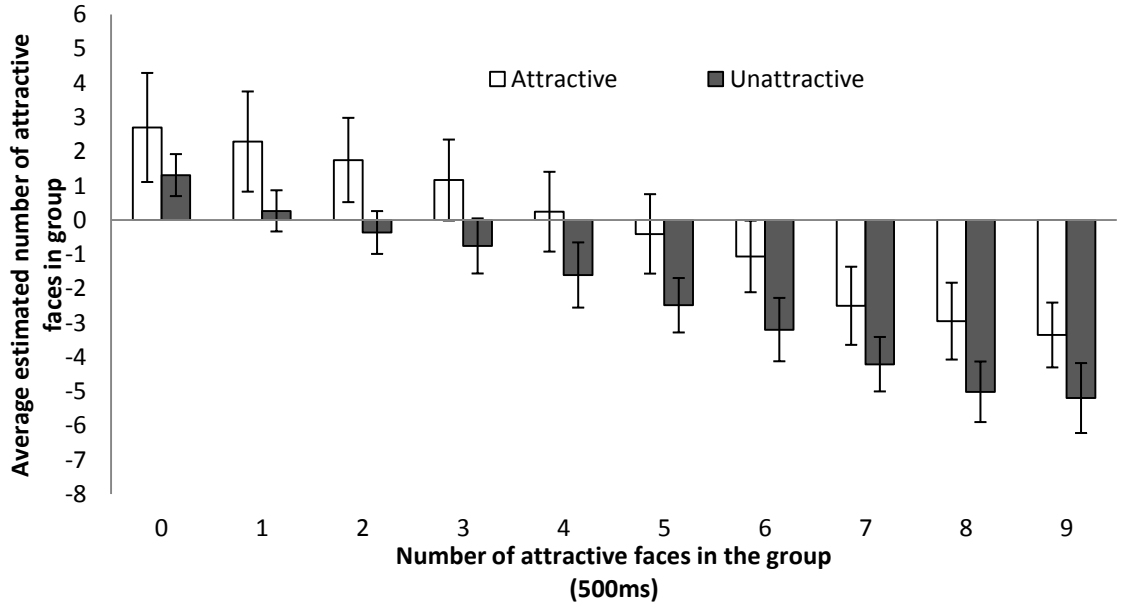
Responses for the unattractive task were re-coded as nine minus the response value, thereby representing the implied number of attractive faces in the group (E.g. a response of three unattractive faces would be analogous to six attractive faces). This meant that all responses could be considered in terms of the number of attractive faces estimated to be in the group, and thus the two tasks could be directly compared. A 2 (task) x 4 (timing) x 10 (number of attractive faces) repeated measures ANOVA on relative error in responses found a main effect of the number of attractive faces ($F(9,36) = 111.21$, $p < .001$, $\eta^2 = .97$), with an overestimation of the target faces when there were few of them, and an underestimation when there were many, similar to that seen in the 10AFC of Experiment 4, and illustrated in Figure 23. There was also a significant interaction between timing and the number of attractive faces ($F(27,108) = 8.62$, $p < .001$, $\eta^2 = .68$). Simple effects analyses showed an effect of timing when there were zero ($F(3,12) = 3.93$, $p < .05$, $\eta^2 = .50$), one ($F(3,21) = 4.89$, $p < .05$, $\eta^2 = .41$), six ($F(1.67,20.09) = 8.29$, $p < .001$, $\eta^2 = .409$, Greenhouse-Geisser adjusted), seven ($F(3,33) = 23.33$, $p < .001$, $\eta^2 = .68$), eight ($F(3,21) = 9.27$, $p < .001$, $\eta^2 = .57$), and nine ($F(3,12) = 7.25$, $p < .01$, $\eta^2 = .65$) attractive faces, with a general reduction in error as display duration increased. All other effects and interactions were non-significant (all ps $> .35$).

The interaction between timing and the number of attractive faces reflects a slight flattening of error levels as display duration was increased. That is, over- and underestimations of the number of faces of the given type in a group are generally lesser for groups with a more pronounced majority of attractive or unattractive faces when the display duration increases.

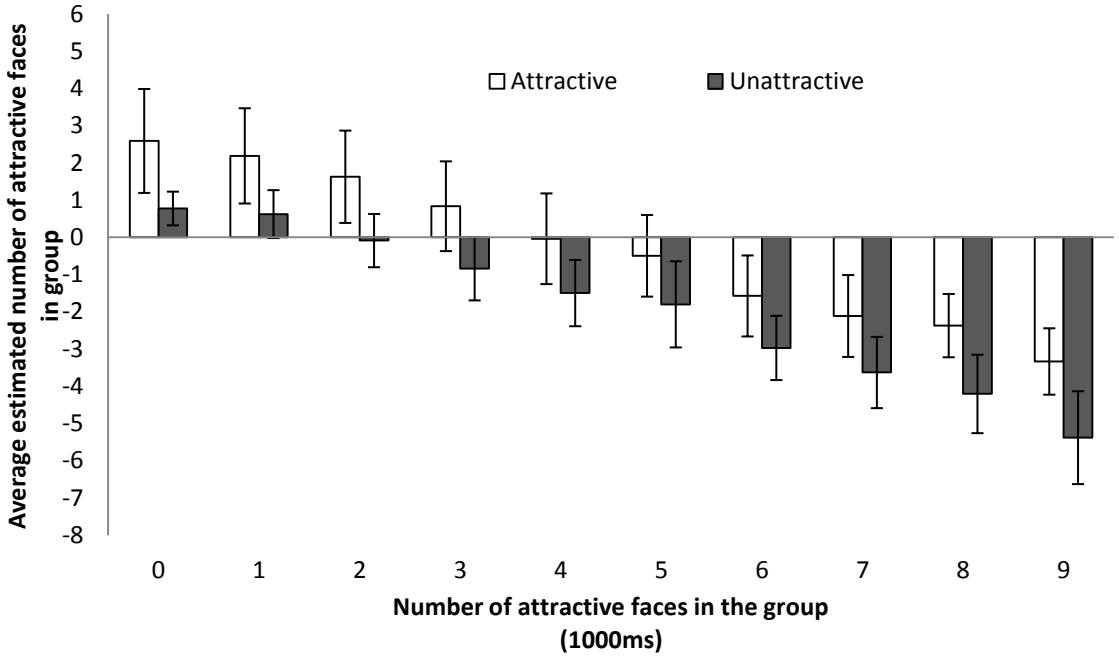
A.



B.



C.



D.

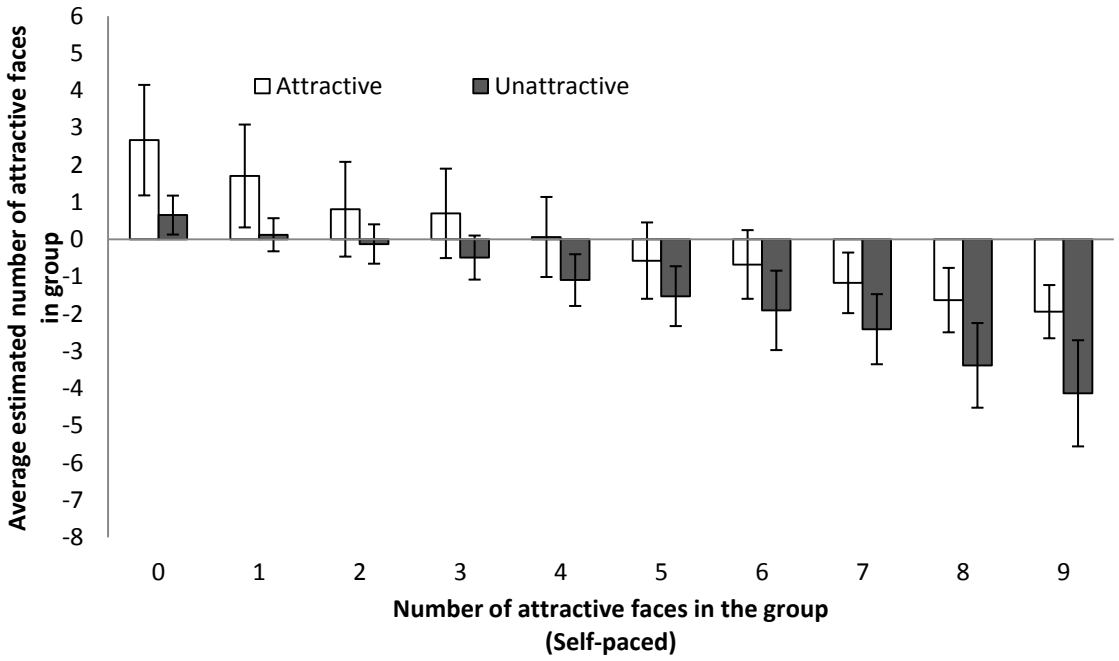


Figure 23: Degree of error in responses at each level of number of attractive faces in the group in the A. 250ms, B. 500ms, C. 1000ms, & D. Self-paced, condition, separated by task. Error bars indicate standard error.

Interestingly, the task (whether estimating the number of attractive or unattractive faces in the group) did not have any significant impact on error levels. The error is least when there were two attractive faces in the group during the unattractive task (M = -.18, SE = .63), but in the attractive task the error was least at four attractive faces (M = .14, SE = 1.13). This appears to differ slightly from the 10AFC of Experiment 4.

A comparison between the attractive task of Experiment 5 and the 10AFC of Experiment 4 using a 2 (experiment) x 4 (timing) x 10 (number of attractive faces) mixed design ANOVA found a main effect of number of attractive faces (F(1.53,33.63) = 322.99, p < .001, η^2 = .94, Greenhouse-Geisser corrected) and an interaction between this and timing (F(27,594) = 8.11, p < .001, η^2 = .27), which is unsurprising, given that both Experiment 4 and 5 found these effects individually. While there was no main effect of the experiment (p > .05), nor was the effect of the number of attractive faces modulated by the experiment (p > .99), the interaction between timing and the number of attractive faces was (F(27,594) = 1.83, p < .01, η^2 = .08). This interaction reflects an effect of increased display duration reducing error in more levels of the number of attractive faces in the attractive task of Experiment 5 than in the 10AFC of Experiment 4. Most important here is that the experimental instructions do not appear to have significantly altered the response trends between the two experiments overall.

3.3.4. Discussion

Experiment 5 was intended to disambiguate some of the findings of the 10AFC of Experiment 4; namely whether there was some impact on performance resulting from ambiguity of the experimental instructions, and whether the skew in responses centring

around three attractive faces was specifically due to an impression of there being fewer attractive faces or due to an impression of there being fewer of the target type of face.

The results showed that while error was lowest at two attractive faces in the unattractive task and at four attractive faces in the attractive task, this difference was not significant. As such, it would seem that the differing task did not have much impact on participants' perceptions of the number of attractive faces in the groups, and that responses were still slightly skewed towards there being more unattractive faces in the group, as in the 10AFC of Experiment 4.

Furthermore, the ambiguity in the instructions used in Experiment 4 does not appear to have impacted on performance in the task. In Experiment 5, the instructions clarified what constituted an "attractive face" and an "unattractive face", and yet when performance on the attractive task was compared with that in the 10AFC of Experiment 4, there was no significant difference between the tasks. This suggests that participants in the 10AFC were taking the instructions as intended, and their responses were just generally skewed towards unattractive faces being more numerous than they actually were.

These two points together lend further support to observations from earlier experiments, suggesting that there is something inherent about seeing the faces in a group setting that leads participants to judge them as being less attractive than when viewed individually (this idea is explored further in Chapter 5). This finding contradicts that of Maner, et al. (2003), who found that constrained viewing conditions increased estimates of the number of attractive faces.

3.4. Experiment 6: Can attractiveness of larger groups be accurately assessed?

3.4.1. Introduction

One issue with the three-by-three display of faces that was highlighted by the eye-tracking data from Experiment 4 was that each trial began with participants already fixating on the central face, which has the potential to impact on judgements of the group. As such, Experiment 6 aimed to replicate the experimental procedure of the timed condition of Experiment 3, but using a slightly larger group of faces, displayed in a way that should eliminate any impact of the initial fixation on the central face.

It was hypothesised that participants would still be able to perform the task with similar levels of accuracy as when there were only nine faces in the group, further demonstrating the likelihood of parallel processing and a gist summary, and as such to allow the potential for Experiments 7 and 8 to explore more about the perceptions of the group, without having the confounding issue of that first, central fixation.

3.4.2. Method

3.4.2.1. Participants

Twenty-four undergraduates from the University of Hull (20 females) participated in this experiment. Their age ranged from 18 to 38 years (M = 22.1 years, SD = 4.87), and all had normal or corrected-to-normal vision.

3.4.2.2. Stimuli

Experiment 6 used the same images as previous experiments, displayed at the same size on screen. However, the display configuration was changed such that each

group now contained 16 faces in a four-by-four grid (rather than nine faces in a three-by-three grid), thereby removing the centrally displayed face (see *Figure 24*).

Figure 24: An example of stimuli and their layout in Experiment 6.

3.4.2.3. Design and Procedure

The overall experimental paradigm was the same as that of the timed condition of Experiment 3, with participants being asked to estimate whether there were more attractive or more unattractive faces in a group, which was displayed for 500ms. Because of the change in the display pattern of the faces, the levels of difficulty needed to be augmented. The new levels of difficulty of the task were zero, two, four, six, 10, 12, 14, and 16 attractive faces, with zero and 16 being comparable in difficulty, two and 14, etc. A condition with eight attractive faces was not included as this is would be a 50-50 split of attractive and unattractive faces, and thus there would be no correct or incorrect answer to the task. There were 10 trials of each condition, presented in a

randomised order. Each trial began with a 500ms fixation cross, and was followed by a response prompt. Breaks were built into the experimental setup.

3.4.3. Results

A 2 (majority) x 4 (proportion) repeated measures ANOVA on accuracy data found a main effect of majority ($F(1,23) = 12.74$, $p < .005$, $\eta^2 = .36$), with higher accuracy in the majority unattractive cases ($M = 76.57$, $SE = 1.59$) than in majority attractive cases ($M = 66.47$, $SE = 1.95$), a main effect of proportion ($F(3,69) = 51.58$, $p < .001$, $\eta^2 = .69$) with accuracy decreasing as the proportion of the groups becomes more balanced, and no significant interaction between the two, as illustrated in *Figure* 25. These effects are overall as expected, based on the performance in Experiment 3. However, the questions of whether the increase in the number of faces in the group is impactful on performance requires a comparison between the results of this experiment, and those of Experiment 3.

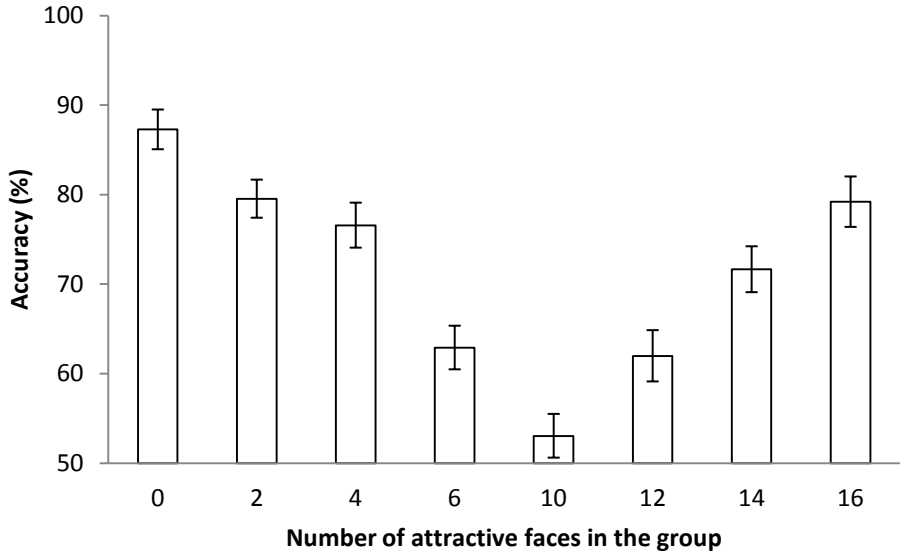


Figure 25: Accuracy at each level of number of attractive faces in the group. Error bars indicate standard error.

Because of the differing displays, a direct comparison between the two experiments is not possible. However, if the number of attractive faces in the group is taken as a percentage, each level of Experiment 6 has a corresponding level of Experiment 3, where the percentages are similar enough to make a loose comparison.

A 2 (majority) x 4 (proportion) x 2 (experiment) mixed design ANOVA found no significant effects involving the experiment factor (all *p*s > .35). This suggests that performance on the task in Experiment 6 was not significantly impacted by the increased group size.

3.4.4. Discussion

The primary purpose of Experiment 6 was to establish whether expanding the display from nine to 16 faces would have a detrimental impact on participants' ability to assess the attractiveness of the faces in the group. The results showed similar patterns to those observed in previous experiments using nine faces; namely that as the group became more balanced in the number of attractive and unattractive faces, performance on the task dropped, as expected, and that accuracy was higher when the majority of the group was unattractive.

Significantly, the comparison between Experiment 6 and the very similar difficulty levels in the timed condition of Experiment 3 showed that the increase in the group size had no impact on performance in the task. It did not impact overall performance, nor did it modulate the main effects, suggesting the possibility of parallel processing during this task. Overall, this can be taken as an indication that participants

are still able to judge the attractiveness of faces in a group to the same level when there are 16 faces in that group as when there are nine. As such, further experiments can use the 16-face groups, allowing for potentially finer manipulation of the difficulty levels of the tasks, but primarily removing any possible confound from having a face appear immediately in the point of fixation at the beginning of the trial.

4. Chapter 4: Identifying the extremes

4.1. Experiment 7: Can participants find the most or least attractive face in a group, and is this task impacted by brief exposures?

4.1.1. Introduction

While Experiment 4 found no real evidence suggesting a pop-out effect for attractive or unattractive faces over the other, or of either type of face holding attention more, it did also suggest that decisions were being made about the group based on information from faces that did not receive any direct visual attention in the form of a fixation. Experiment 7 sought to explore whether, when asked to select the most (or least) attractive face in a group, participants would select the face that would later be given the highest (or lowest) rating of attractiveness, despite the apparent lack of pop-out. Haberman & Whitney (2009) found that member discrimination from a group was poor, but their stimuli were highly homogeneous, as compared with those used here. This effect may not be present in more varied stimuli.

This experiment used two different display durations, one restricted and one unrestricted. The unrestricted condition simply required participants to methodically study the group and make a considered decision about which face was the most (or least) attractive, whereas the restricted condition was included to provide a comparison for when the target needed to be selected quickly and its location recalled for later response. Becker, Kenrick, Guerin, & Maner (2005) found that participants were better able to recall the location of attractive female faces than “average-looking” female faces within a group in order to match pairs of faces in a memory task, suggesting that this task should be possible for participants, but perhaps indicating that the selected face will be closer to the highest rated face when selecting the most attractive face than it might be to the lowest rated face when selecting the least attractive face.

It was hypothesised that performance would improve in the unrestricted display condition, as compared with that in the restricted. However, an apparent lack of a pop-out effect in Experiment 4 casts doubt on the general ability to perform the task, despite Becker et al.'s (2005) findings.

4.1.2. Method

4.1.2.1. Participants

Twenty two undergraduates from the University of Hull (19 females) participated in this experiment. Their age ranged from 18 to 38 years (M = 22.18 years, SD = 5.1), and all had normal or corrected-to-normal vision. None of the participants had participated in any of the previous experiments, in order to eliminate the confounds of familiarity with the task or the stimuli. The two tasks were performed in separate sessions, with the second session also containing a rating task, as listed below.

4.1.2.2. Stimuli

The stimuli were the same as those used in Experiment 6, arranged in a four-by-four grid on screen in the same manner, with the members of the group selected randomly for each trial, with no repetition of the same face in a given trial.

4.1.2.3. Design

The task in this experiment required participants to select one face from the group of 16 that they thought was either the most or the least attractive out of the group. The experiment was split across two sessions; in one the task was timed, and in the other, the task was self-paced, as in previous experiments. The order in which the sessions occurred was counterbalanced across participants. In each, there were four blocks,

separated by breaks. In a single block, participants were required to identify the face that they thought was the most attractive or unattractive from the group (two blocks of each task), and the order in which these four blocks occurred was counterbalanced across participants. Each block contained 80 trials, with ten of each level of number of attractive faces.

In both the timed and self-paced conditions, the groups were preceded by a fixation cross that appeared on-screen for 500ms, and the group was then displayed. In the timed condition, the group was displayed for 2000ms (as informed by pilot testing, in which multiple participants advised that at shorter displays, not only were they merely guessing blindly, but that they were becoming frustrated by the task), and then followed by a screen with 16 grey blocks in place of the faces. On this screen, participants were asked to use the mouse to click the location that had previously contained the most/least attractive face, after which the next trial began. In the self-paced condition, once the faces were displayed, they remained on-screen until the participant selected the most/least attractive face by clicking on it. In either case, a click outside of a box/face did not progress the experiment.

While the order of the trials, and the exact composition of each group was randomised within these confines for the first experimental session, this information was recorded and the sequence and composition of each trial was replicated exactly in the second session. This meant that in each timing condition, participants were seeing the same stimuli, in the same location, in the same order, thereby making a direct comparison between the two conditions more informative, because the decisions were being made based on exactly the same visual information, just varying in display duration.

After the second session, participants performed a ratings task the same as that used in previous experiments. This rating data was used to rank each selected face within the other faces used in the trial in terms of attractiveness.

4.1.3. Results

For each trial, the chosen face was assigned a rank within the group. The lower the value of the rank, the closer the face was to the target - that is, when identifying the most attractive face, a rank of 1 would indicate that the participant had selected the face from the group that they would later rate as the most (or joint-most) attractive, whereas when identifying the least attractive face, a rank of 1 would indicate that the participant had selected the face rated as the least attractive. This ranking was based on each participant's own ratings for the faces, as provided in the ratings task at the end of the second experimental session.

A 2 (timing) x 2 (majority) x 4 (proportion) x 2 (task) repeated measures ANOVA found three main effects: timing ($F(1,21) = 86.78$, $p < .001$, $\eta^2 = .81$), with there being less error in the self-paced condition ($M = 3.06$, $SE = .27$) than in the timed ($M = 4.42$, $SE = .23$); majority ($F(1,21) = 5.36$, $p < .05$, $\eta^2 = .20$), with less error when the majority was unattractive ($M = 3.56$, $SE = .29$) than when it was attractive ($M = 3.92$, $SE = .21$); and proportion ($F(7.74,36.5) = 51.34$, $p < .001$, $\eta^2 = .71$, Greenhouse-Geisser corrected), which showed less error as the group became more evenly balanced. The timing factor also interacted with each of the other three factors; majority ($F(1,21) = 11.33$, $p < .005$, $\eta^2 = .35$), proportion ($F(3,63) = 5.04$, $p < .005$, $\eta^2 = .19$), and task ($F(1,21) = 10.06$, $p < .01$, $\eta^2 = .32$). Proportion and task also interacted ($F(3,63) = 4.2$, $p < .01$, $\eta^2 = .17$).

There were also two significant three-way interactions, between timing, majority, and task ($F(1,21) = 7.94$, $p < .05$, $\eta^2 = .27$), and majority, proportion, and task ($F(3,63) = 26.00$, $p < .001$, $\eta^2 = .55$). However, all of these effects were modulated by a significant four-way interaction between timing, majority, proportion, and task ($F(3,63) = 3.90$, $p < .05$, $\eta^2 = .16$).

Exploring this interaction further reveals that the interaction between majority, proportion, and task was much stronger in the self-paced condition ($F(3,63) = 29.77$, $p < .001$, $\eta^2 = .59$) than in the timed condition ($F(3,63) = 9.77$, $p < .001$, $\eta^2 = .32$). Further refinement shows that the interaction between majority and proportion was not significant for the attractive task in the timed condition ($p>.17$), but was so in the self-paced condition ($F(2.14,45.00) = 8.46$, $p < .005$, $\eta^2 = .29$, Greenhouse-Geisser corrected) and for the unattractive task in both the timed ($F(3,63) = 9.58$, $p < .001$, $\eta^2 = .31$) and self-paced ($F(2.32,48.64) = 16.19$, $p < .001$, $\eta^2 = .44$) conditions.

These three interactions can be clarified by looking at the effect of proportion for each level of majority in each of these conditions. The similar *F*-values for the interactions in the self-paced condition for the attractive task (8.46) and the timed condition for the unattractive task (9.58) reflect a similar pattern (though in opposite directions) of a significant effect of proportion in one majority (($F(3,63) = 22.91$, $p < .001$, $\eta^2 = .52$) for majority attractive in the unattractive task in the timed condition and ($F(2.25,47.31) = 14.41$, $p < .001$, $\eta^2 = .41$, Greenhouse-Geisser corrected) for majority unattractive in the attractive task in the self-paced condition), and a non-significant effect of proportion in the other majority ($p>.1$ for majority unattractive in the timed condition for the unattractive task, and $p>.54$ for the majority attractive in the self-paced condition for the attractive task). Whereas for the unattractive task in the self-paced condition, the effect of proportion was significant for both majority unattractive

($F(3,63) = 3.26$, $p < .05$, $\eta^2 = .13$) and majority attractive ($F(1.87,39.16) = 43.65$, $p < .001$, $\eta^2 = .68$), but the difference in the *F*-values is clearly much more substantial.

Ultimately, these results show that the impact of proportion on the disparity between the selected most or least attractive face and the face actually ranked that way based on individual ratings is much greater when the majority of the group is attractive when selecting the least attractive face in the group, as compared with when the majority of the group is unattractive. Further, when selecting the most attractive face in the group, the impact of proportion is barely evident in the self-paced condition, and where it is evident in the timed condition it is not impacted by the majority of the group. These trends can be seen in Figures 26 and 27.

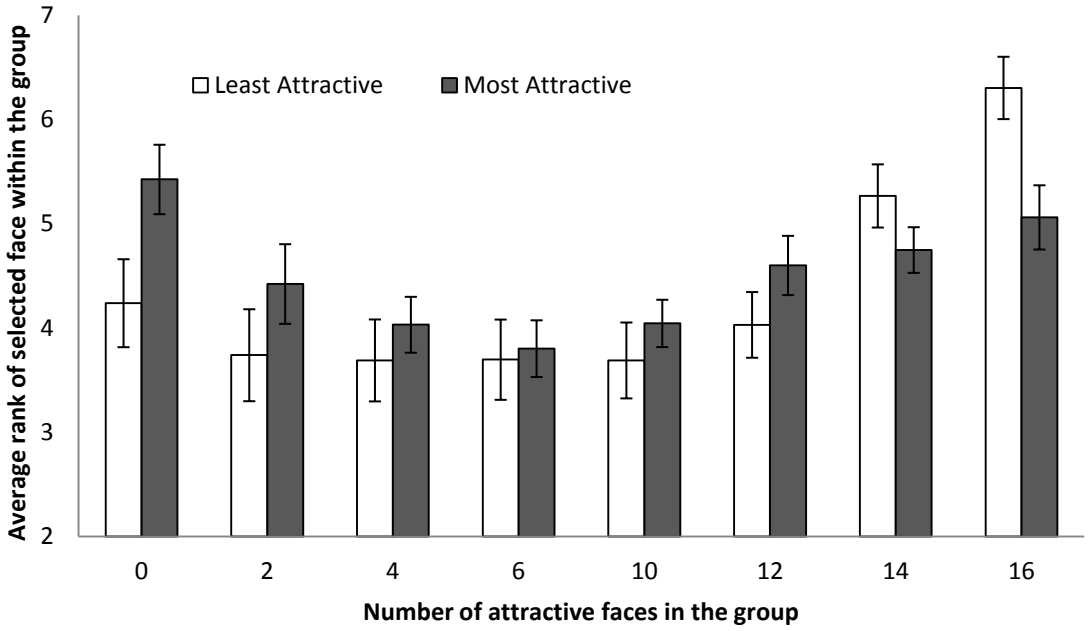


Figure 26: Average rank of the selected face within the group in the timed condition when selecting the most attractive face in the group, and the least attractive face in the group, separated by the number of attractive faces in the group. Error bars indicate standard error.

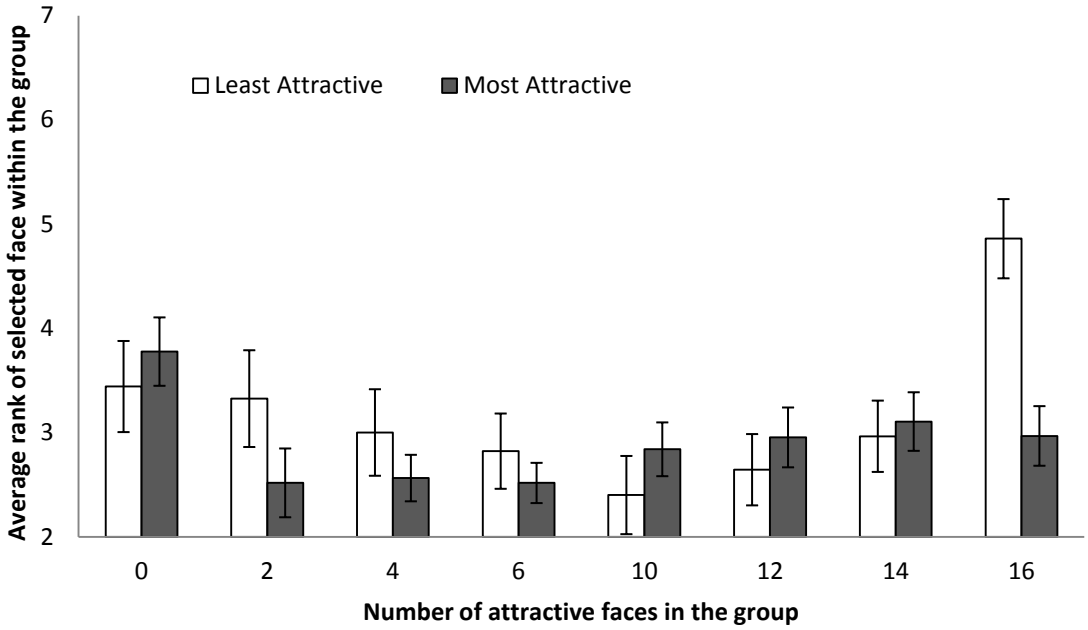


Figure 27: Average rank of the selected face within the group in the self-paced condition when selecting the most attractive face in the group, and the least attractive face in the group, separated by the number of attractive faces in the group. Error bars indicate standard error.

4.1.4. Discussion

Experiment 7 sought to explore whether participants would be able to select the most or least attractive face from a group, both under a restricted display duration and a self-paced condition. It was expected that responses would be more accurate in the self-paced condition, because participants would be able to commit as much time as required to fully assess each face in the group, and would also not be required to remember the location of the selected face when giving their response.

Previous research (Becker, Kenrick, Guerin, & Maner, 2005) suggested that participants might be able to better recall the location of attractive faces when presented in a group, and the results of this experiment certainly suggest that participants were better at selecting the most attractive face from a group than they were at selecting the least attractive, although this effect was stronger in the self-paced condition, when remembering the location of the face was not required.

Unsurprisingly, the highest errors occurred when selecting the most attractive face from a group composed solely of unattractive faces, or when selecting the least attractive face from a group composed solely of attractive faces (the latter having higher errors than the former). Overall, these results suggest that there is no real pop-out effect of attractive or unattractive faces, and that even when given as much time as desired to study a group of faces, participants are unable to consistently select the face that they would later suggest was the most (or least) attractive in the group. This adds further weight to findings in previous experiments that suggest that something inherent in the group setting modifies the perception of attractiveness of its members. The difference between the perceived attractiveness of faces in and out of the group context will be explored further in Experiment 8.

5. Chapter 5: Averaging methods

5.1. Experiment 8: Which model of ensemble representation most closely resembles overall judgements of groups?

5.1.1. Introduction

The results of previous experiments have suggested that participants generally rate a group as having fewer attractive faces in it than participants' own ratings of the faces would suggest. This seems to be true whether making a general statement about the group, or estimating a specific number of attractive or unattractive faces in the group. This latter effect is particularly pronounced when there are more attractive than unattractive faces in the group.

Experiment 5 in particular appears to demonstrate that there is something inherent in presenting faces in a group that generally makes either the faces themselves, or the group overall, appear less attractive than when the faces are viewed individually. This is further compounded by the results of Experiment 7, which highlight that even with no time restrictions, participants were not able to consistently select from the group the face that they would later be given the highest (or lowest) rating for attractiveness of all of the faces in the group.

Experiment 8 sought to explore the nature of the representation of the attractiveness of a group, and also the extent to which membership of a group alters the perceived attractiveness of a face. Participants were asked to rate the attractiveness of each face individually (to then be used to calculate an average for a group of faces), the attractiveness of each face when presented in a group of other faces, the overall attractiveness of the group of faces, and the attractiveness of a singular morphed face representing a visual amalgamation of all members of the group.

How the overall rating of the group compares to the other three measures will give some indication as to the way in which the attractiveness of a group of individuals is summarised and represented. If each face in the group is considered in isolation, represented in some way, and then the summation of these representations used as a representation for the group (though this seems unlikely), then one would expect the overall singular rating of the group to closely resemble the mean of the faces when rated independently. If each face is considered individually, but taken within the context of the group to then be summarised in a single representation, then the mean value when rating the individual faces in the group context should be similar to the singular group representation. If the visual information is all combined into a single representation to then be assessed, the rating of the morphed image should be more closely related to the singular rating.

Faces that more closely represent an average of a population are generally found to be more attractive, and morphing faces together tends to also increase their perceived attractiveness due to a negation of fluctuating asymmetries and blemishes, and a smoothing of skin tones and features (Perrett, Burt, Penton-Voak, Lee, Rowland, & Edwards, 1999; Valentine, Darling, & Donnelly, 2004). As such, it is reasonable to expect that the morphed image of a group would be rated as the most attractive, and because of these previous findings it seems unlikely that the singular rating of the group will be the same as the morph, but the degree of difference could give some indication of the level of visual summarisation that occurs when judging a group in such a way.

Further, given that participants seem to have largely shown a trend of responses suggesting there are fewer attractive faces (or more unattractive faces) in the group than a general consensus, or even their own ratings would later suggest, it seems to be a reasonable expectation that the average of ratings given in isolation should be higher

than the overall rating of the group. Even though a 1-10 rating of attractiveness does differ from an estimation of majority or number of attractive faces, a greater number of attractive faces should logically lead to a higher rating of the group.

Given that the group situation appears to impact upon the perception of attractiveness of faces, the ratings given to faces in the group context might vary depending on the faces surrounding them (I.e. the number of attractive faces in the group). As such, the results might demonstrate that at different numbers of attractive faces in the group, there might be a change in the degree to which this average value differs from the average of the ratings given in isolation. This might tend towards more unattractive ratings when there are fewer attractive faces in the group (and vice versa), with the overall impression of the group modifying the rating towards the majority, or it might result in a "flatter" average rating across the conditions, as the minority of the group have their attractiveness modulated away from the majority by comparison.

While much of this task sought to shed light on the possible methods of averaging, and thus very few formal hypothesis were made, it was still expected that the morphed image of the groups would consistently receive the highest ratings of attractiveness.

5.1.2. Method

5.1.2.1. Participants

Nineteen undergraduates from the University of Hull (16 females) participated in this experiment. Their age ranged from 18 to 33 years (M = 20.89 years, SD = 3.83), and all had normal or corrected-to-normal vision. None of the participants had

participated in any of the previous experiments, in order to eliminate the confounds of familiarity with the task or the stimuli.

5.1.2.2. Stimuli

This experiment used the same faces as the previous experiments. For two of the tasks, these were presented in a four-by-four grid, as in Experiments 6 and 7, in one task they were presented individually, as in the ratings task accompanying all previous experiments, and in the fourth task, multiple faces were morphed together.

The morphing of the faces was conducted using FantaMorph (FantaMorph, 2009). For the morphing process, 90 anchor points were placed onto each face, marking key structural points and features, corresponding to the same approximate place on each face. Using these points, the faces were combined together to create an amalgamation of both structure and texture. Each morphed image was informed equally by each face that was included in it.

5.1.2.3. Design

This experiment was conducted across two separate sessions, one in which the stimuli were presented in the four-by-four group format, and one in which they were presented as individual faces. The groups were randomly generated before the experiment, such that the same face did not appear more than once in a group. In total, 90 groups were created, 10 of each level of zero, two, four, six, eight, 10, 12, 14, & 16 attractive faces, with the rest of the faces being unattractive. These groups were the same for each participant. For one of the tasks, a morphed image was created for each of these groups. The morphed image was an amalgamation of all 16 faces in the group.

In the first session there were two tasks, the order of which was counterbalanced between participants. In one task, participants were asked to indicate the overall attractiveness of the group on a 1-10 scale, using the number keys across the top of the keyboard, with 0 being a substitute for 10. The group remained on-screen until the participant had responded. Each trial was preceded by a 500ms fixation cross. In the other group-presentation task, participants were asked to give a rating of attractiveness in a similar fashion, but for an individual face within the group. The face was highlighted with a box around it, and after each individual face was rated, the box moved to highlight a different face. The order in which the faces were highlighted was randomised within each trial.

In the second session there were also two tasks, the order of which was also counterbalanced between participants. In one task, participants were asked to rate each individual face for attractiveness, as in the ratings task included in each previous experiment. The faces were preceded by a 500ms fixation screen, and remained on-screen until a rating was given. In the other task, the premise remained the same, but the stimuli were made up of the morphed images of the groups.

5.1.3. Results

In order to compare the various rating methods used in this experiment, each pre-made group was assigned four singular values for each participant. The first value was simply the response given when asked to rate the overall attractiveness of the group, the second was the mean of the 16 ratings given when asked to rate each individual face within the group, the third was the rating given to the morphed amalgamation of the group, and the fourth was the mean rating of the 16 faces in the group when the faces were rated independently.

A four (task) x nine (number of attractive faces) repeated measures ANOVA found a main effect of task ($F(3,54) = 79.63$, $p < .001$, $\eta^2 = .82$), a main effect of number of attractive faces ($F(1.96,35.30) = 335.16$, $p < .001$, $\eta^2 = .949$, Greenhouse-Geisser corrected), and an interaction between the two ($F(24,432) = 1.64$, $p < .05$, $\eta^2 = .083$). Unsurprisingly, the average rating increased as the number of attractive faces in the group increased.

Planned contrasts compared the singular rating of the group with each other measure ($M = 4.73$, $SE = .18$). These showed that the lowest rating, the mean rating of the group when the faces were rated independently, was significantly lower ($M = 4.07$, $SE = .14$) than the singular rating of the group overall ($F(1,18) = 27.02$, $p < .001$, $\eta^2 = .60$). The mean of the ratings given to each face when presented in the group was also significantly lower ($M = 4.28$, $SE = .16$) ($F(1,18) = 14.13$, $p < .005$, $\eta^2 = .44$). The highest rating, that of the morphed amalgamation of the group ($M = 5.93$, $SE = .19$) was significantly higher than the singular rating of the group ($F(1,18) = 56.5$, $p < .001$, $\eta^2 = .76$). A further planned comparison showed that the mean rating of the independently rated faces was lower than that of the mean value of faces rated in the group context ($F(1,18) = 5.33$, $p < .05$, $\eta^2 = .23$), and that this effect was in no way modulated by the number of attractive faces in the group ($p > .69$).

Exploring the interaction, simple effects analyses showed that the effect of the number of attractive faces in the group was strongest when taking the mean of the faces when rated independently ($F(1.28,23.05) = 344.92$, $p < .001$, $\eta^2 = .95$, Greenhouse-Geisser corrected), and was also very strong when taking the mean of the ratings when rated in the group context ($F(1.52,27.41) = 279.42$, $p < .001$, $\eta^2 = .94$, Greenhouse-Geisser corrected). The effect was less pronounced in the overall highest rated task, the morphed image ($F(2.69,48.47) = 129.26$, $p < .001$, $\eta^2 = .88$, Greenhouse-Geisser

corrected), and was at its lowest for the second-highest rated task, the singular rating for the group ($F(3.25,58.58) = 81.10$, $p < .001$, $\eta^2 = .82$, Greenhouse-Geisser corrected). This effect can be seen in *Figure* 28.

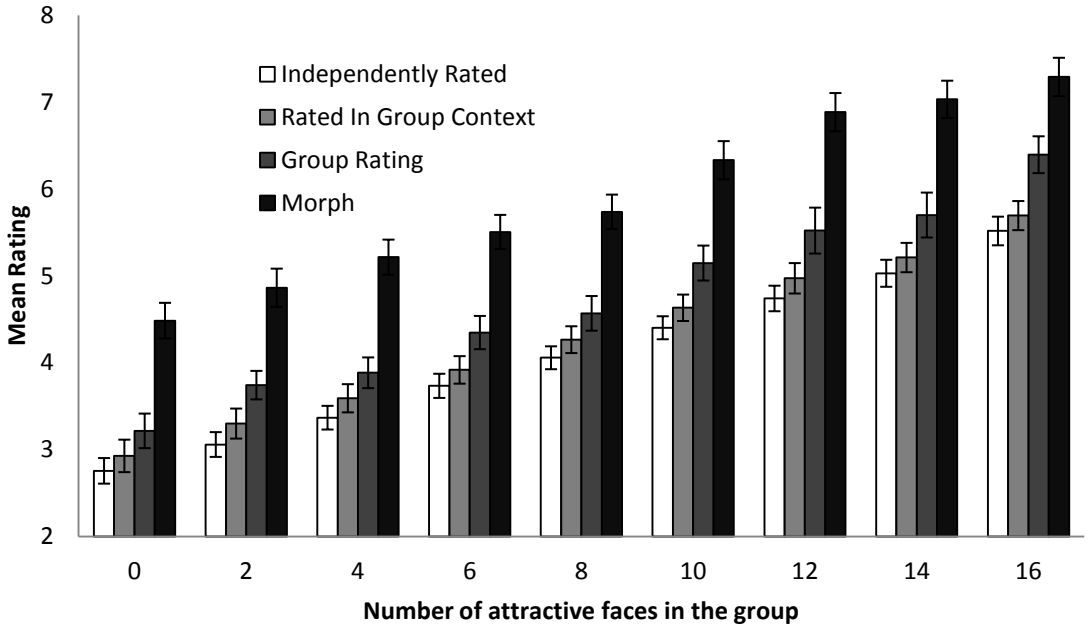


Figure 28: Ratings of attractiveness of groups of faces as either a mean value of all the members when rated independently, a mean value of all the members when rated in the group context, the overall singular rating given for the group, and the rating given to the morphed image of the group, across different numbers of attractive faces in the group. Error bars indicate standard error.

5.1.4. Discussion

The primary aim of Experiment 8 was to compare some possible methods of creating a summary representation of the attractiveness of a group. Using a singular overall rating of the group as a baseline, the accuracy of three summaries were assessed: the mean rating given to each of the members of the group when rated in isolation; the

mean rating given to each of the members of the group when rated in the group context; and the rating given to a morphed image of the group.

The results found that, on average, the morphed image was considered the most attractive, which was to be expected (Perret et al., 1999). They also found that this was the most significantly different from the singular overall rating. This would suggest that of the possible summarisation methods explored in the experiment, the idea of amalgamating visual information into a single representation seems the least likely.

Further, the mean rating of the group members when rated in isolation was, on average, the lowest rating of attractiveness, and was significantly lower than both the singular rating, and the mean rating of the individual faces rated in the group context. While it was hypothesised that the isolated ratings would likely differ from both the in-situ ratings and the group overall, previous results had suggested that the overall rating would actually be lower than the average of the isolated ratings, and the converse findings would seem to point to the difference between the rating of a group for attractiveness and assessing the majority or the number of attractive faces in that group; while the majority or number of attractive faces in a group might be deemed to generally contain more unattractive faces (even when the participants' own ratings of the attractiveness of the individual faces are taken into account), the group as a whole is given a higher rating of attractiveness than those ratings for the member faces. This could be that in the group situation, attractive faces carry more weight than others when formulating a summary representation for the group, but still only count as a single member when determining information about the composition of the group.

Also of interest is the fact that the average of the ratings given to the faces in the group context is consistently significantly higher than the average when the same faces were rated in isolation, and that this is in no way impacted by number of attractive faces

in the group. It was hypothesised that there would be a difference between these ratings, but that this difference might vary with the number of attractive faces in the group, as the ratings were affected by the composition of the surrounding group. That there is no variation in this difference suggests that it is merely the presence of the group that improves the attractiveness of the individual faces. This might perhaps link to a suggestion of acceptance and community, and thus likely an absence of disease, thereby giving some validation-by-proxy of the individual's worthiness (Perrett, 2010).

It would seem that neither a singular summary of the visual information, nor a consideration of each member of the group (with or without consideration of the other members) and an averaging of this information, represents an accurate reflection of the way in which the overall attractiveness of a group is represented.

6. General Discussion

Previous research has shown that the attractiveness of faces can be ascertained quite quickly (Olson & Marshuetz, 2005; Willis & Todorov, 2006), but that when they are presented in a group, the frequency estimates of attractive faces increases (Maner, et al., 2003). Research in emotional expression has found that the average expression of a group of faces can be accurately conglomerated into an ensemble representation (Haberman & Whitney, 2007; Haberman & Whitney, 2009). This thesis aimed to explore whether a similar ensemble representation is generated for the attractiveness of groups of different sizes with varying viewing constraints, and whether this representation reflected the mean attractiveness of all members of the group, a visual amalgamation of each member of the group, or some other value where different group members might contribute to the ensemble differentially. It also sought to explore whether attractive faces draw a greater duration of fixations, as predicted, and whether the pattern of gaze is influenced by them. Finally, it explored whether observers are capable of identifying the most or least attractive faces from a group.

6.1. The capacity to judge attractiveness from rapid presentations

It has been shown that humans are capable of making rapid judgements of facial attractiveness from brief presentations. Willis & Todorov (2006) demonstrated this from 100ms presentation times, where Olson & Marshuetz (2005) restricted presentation times to the minimum that the refresh rate of their presentation system would allow - 13ms. In both of these studies, the responses of participants were compared with the attractiveness of the target faces, and found to be generally reliable. However, these studies only considered the assessment of individual faces, where Haberman & Whitney (2007), among others, have demonstrated that information other than attractiveness can

be gleaned from groups of faces. Maner, et al. (2003) did study the perception of attractiveness from groups of faces, and while their results showed a higher estimate of the frequency of attractive faces when under restricted viewing conditions, as compared with much less restricted conditions, they did not consider whether these estimates were accurate reflections of the groups presented, which contained faces of varying attractiveness with a mean attractiveness of a middling value on a likert scale.

Experiment 1 was designed to explore how well participants could make rapid judgements of the attractiveness of briefly presented groups. Given that previous works did not consider the accuracy of rapidly generated representations of the attractiveness of groups, it was not clear how much impact both the group format and the restricted viewing conditions had upon the accuracy of responses. As such, rather than expecting participants to form an accurate and precise representation of a group by providing a judgement of a single group (E.g. “this group contains three attractive faces”), this task required that two groups be compared, and the one that contained the greater number of attractive (or unattractive) faces be identified. As such, the representations of the group need only be comparative (E.g. “this group contains fewer attractive faces than the previous group”). The results confirmed that participants were able to correctly identify which of the two groups contained the greater number of target faces at levels significantly above chance in both the 250ms and 500ms exposures, but, as was expected, accuracy decreased with the difference between the two groups. Thus suggesting that the previous findings of accurate judgements of single faces could be extended to groups of faces to some extent.

The results of Experiment 1 then informed the design of Experiment 2, where the group size was increased from four faces to nine. This served to further explore the impact of the group context on rapid judgements, and to investigate whether the groups

were being processed in a serial or a parallel fashion. Haberman & Whitney (2009) found no effect of set size on the capacity to generate ensemble representations, but their stimuli were all same-identity faces with not much variation in their visual information, which could potentially simplify parallel processing. Hansen & Hansen (1988), however, found that set size had a significant effect, suggestive of serial processing. With an increase in the number of faces in the groups, a decline in performance would suggest that the group members are being processed in a serial fashion. The results were not directly comparable, due to the different levels of difference between the groups in the two experiments, but there was little apparent difference in performance, suggesting that the group size was not impactful on performance. However, at very small differences between the two groups, in both experiments, performance dropped, and was down to chance level at the smallest difference in Experiment 2, which raised the question of whether performance was being affected by the group presentation, the brief exposure, or the fact that the task required two ensemble representations to be generated and recalled/compared.

In order to account for two of these possibilities, the experimental paradigm was shifted for Experiment 3, with each trial using only a single group of nine face faces, and the task changing to being a judgement of whether there were more attractive or more unattractive faces in the group. There were also two different display durations, one of 500ms, and the other an unlimited viewing period. The first of these changes sought to eliminate the potential confound of having to generate, and then recall, ensemble representations for two groups. Now, participants needed only to concentrate on the one group, but by evaluating the majority of that group, the nature of the representation did not need to change much, still not needing the level of specificity of the number of attractive faces in the group and only requiring a general gist. The

differing display durations were implemented to account for the possible impact of the brief exposure.

The results of Experiment 3 showed that participants could generally perform the task accurately, but, as with Experiments 1 and 2, this was less true as the group became more balanced (which is analogous to smaller differences between the two groups in Exps. 1 & 2). In fact, in the two most difficult conditions when the majority was attractive, performance was at or below chance level, suggesting a threshold of performance capacity. That this lack of accuracy was skewed slightly towards conditions when the majority of the group was attractive is in line with the overall effects of majority that the results showed, in which performance was pointedly better when the majority of the group was unattractive. These two points taken together show that participants are slightly more likely to respond that the group is unattractive, especially when the majority is less clear, and may reflect some erring on the side of unattractive when faces are presented in a group. Though this finding might seem counter to that of Maner, et al. (2003), it is one that pervades in the research presented here.

The impact of the differing display durations was only evident in tandem with the other factors of the experiment. First, there was no real difference between majority attractive and majority unattractive trials when display duration was restricted to 500ms. This suggests that when making less considered judgements, the erring on the side of unattractive does not occur, but that it is present when given time to peruse the group at leisure. This might reflect the additional opportunity to inspect such things as skin blemishes and asymmetries - which have previously been shown to be unattractive (E.g. Fink, et al., 2001; Perrett, et al., 1999) - that the shorter display duration did not afford. In particular, trials with five attractive and four unattractive faces in the self-paced

condition showed very poor performance, and the minor margin of 5:4 was obviously outweighed by the unattractive features of the four. This might point to some added weight being granted to unattractive faces when forming ensemble representations, or perhaps a capturing of visual attention, which might contribute to this additional weight.

Experiment 4 expanded on these findings by using two additional display durations for the same task. A 250ms duration, and a 1000ms duration, alongside the same 500ms and self-paced displays used in Experiment 3, showed that there was no difference in performance between 250ms and 500ms, or between 1000ms and the self-paced condition.

These experiments taken together suggest that it is possible to make rapid assessments of the attractiveness of groups of faces, and potentially form some sort of ensemble representation of these groups, with which to make later decisions. It does, however, raise the questions of the nature of the representation, and the way in which visual attention is distributed around the groups while processing the information for the representation (especially under restricted viewing conditions).

6.2. Eye movements and visual attention

There is research that suggests that attractive faces capture attention (Maner, et al., 2003; Maner, et al., 2007; Sui & Liu, 2009; Leder, Tinio, Fuchs, & Bohrn, 2010; Chen, Liu, & Nakabayashi, 2012), often to the detriment of other tasks. In particular, they tend to garner earlier, more frequent, and longer fixations (Leder, et al., 2010). However, in most of these studies, the faces are either task-irrelevant, used as masking for other stimuli, or are simply being observed in free-viewing conditions. Experiment 4 used eye-tracking technology to follow gaze patterns while judging either whether a group contained more attractive or more unattractive faces (2AFC), or estimating the number

of attractive faces in the group (10AFC). These tasks were performed with different display durations, as listed above. The 10AFC was analysed using each participant's own attractiveness ratings for the faces to account for any variation between their perceptions and those of the participants who originally rated the faces. This partially served to explore the nature of the ensemble representation, discussed later.

The analysis of the gaze data considered two primary measures; mean duration of fixations to faces, and the probability of fixating a given location on a given trial, accounting for various factors. Based on previous works, it would have been hypothesised that attractive faces drew longer fixations, and somewhat impacted on the locations within the group that were fixated. However, based on the findings of Experiments 1, 2, and 3, it would seem that unattractive faces exert more influence over the judgement of a group than attractive faces do, which would suggest the possibility that they are receiving more visual attention when the task at hand actually relates to their attractiveness.

The results certainly suggest that, in the 2AFC, fixations were longer when the majority of the group was unattractive, which would marry with the idea of unattractive faces garnering longer fixations, driving up the mean duration when they outnumber the attractive faces. However, fixation duration was largely not impacted by whether the fixated face was above or below the average attractiveness of the group, based on each participant's own ratings. This suggests little impact of the attractiveness of a given face on how long participants will fixate it during this task, but that the overall attractiveness of the group can drive fixation durations. Despite attractive faces pulling attention when they are task irrelevant (Sui & Liu, 2009; Leder, et al., 2010; Chen, Liu, & Nakabayashi, 2012), they have no real impact on visual attention when their attractiveness is actually relevant to generating a representation of a group.

Interestingly, performance on the task does not appear to be related to the duration of fixations directed to faces, irrespective of whether those faces were representative of the majority of the group.

These results were a bit different in the 10AFC, with increased display duration driving an increase in fixation duration, and longer fixations when the majority of the group was attractive. This immediately highlights some minor differences in the way groups are viewed between when judging an overall majority and when estimating the number of attractive faces. Indeed, in this task, faces that were above the average attractiveness of the group received fixations of similar lengths, regardless of the majority of the group, whereas faces below the average attractiveness of the group were given less fixation time when the majority was unattractive, and thus they were less out of place, as compared with when they were less representative of the groups of majority attractive faces, when they drew longer fixations. This suggests that perhaps the less attractive faces were categorised as such quicker when they were clearly not exceptional, perhaps in favour of studying the more attractive faces to better judge their numbers. This does show that, in this task, fixation durations to attractive faces were less variable than those to unattractive faces, which were modulated by the surrounding faces. However, more attractive faces did receive longer fixations when participants were free to study the groups for as long as they wished. In fact, in this condition, fixations were longer overall, hinting at the more relaxed approach taken when not constrained by limited viewing time, but in this situation, participants clearly took more time to study attractive than unattractive faces. This slight favouring towards attractive faces, particularly in the self-paced condition, might be partly reflective of the task at hand in the 10AFC - participants were only asked to estimate the number of attractive faces, with no comparable task for unattractive faces. This was addressed for the

response data in a later experiment, but without the eye-tracking element, where similar patterns might well have been observed, but favouring the target unattractive faces.

In sum, it seems that the duration of fixations to faces was largely unaffected by their relative attractiveness, and where this did take effect, it was modulated by the attractiveness of the other faces in the group. When the attractiveness of the face is relevant to the task at hand, the previously observed effects of attractive faces drawing longer fixations do not appear to manifest.

The locations that participants fixated also largely appeared to be unaffected by the attractiveness of the faces within or surrounding them, despite suggestions that attention should be drawn to attractive faces (Maner, et al., 2003, 2007). The probability curves of fixations occurring in a given location as a function of the fixation index in a trial did not significantly vary between: when the face in the location was above or below the average attractiveness of the group; the number of attractive faces in the group; whether participants responded correctly or incorrectly in the trial (in the 2AFC), or the degree of error in the trial (10AFC). This suggested that, instead of trying to seek out faces of a particular type, be that attractive, unattractive, or outliers in the group, participants simply tried to explore as much of the group as possible. This is supported by the most commonly observed fixation patterns, that generally show a methodical approach to fixating the faces in an order reflective of reading patterns, especially in longer display durations, and in trials with accurate responses/a lesser degree of error.

This further suggests that any special draw of attractive faces is largely negated by the task-relevance of that attractiveness and, while faces clearly enjoy a perceptual bias during unrelated tasks (Sui & Liu, 2009; Chen, et al., 2012) or in natural scenes (Fletcher, et al., 2008; Leder, et al., 2010) or among unrelated distractors (Hershler & Hochstein, 2005), having multiple faces present might mitigate that bias.

6.3. Contribution of non-fixated faces to the ensemble

Given the limited opportunity to visually explore the groups in three of the four display durations, it was a point of interest to explore the level at which the fixated faces impacted on performance in the 2AFC. As could be expected, when the fixated faces contained the same majority (attractive/unattractive) as the group, accuracy was pointedly higher than when they contained the opposite. What was of interest was that when the fixated faces showed a majority counter to that of the group, performance was still only slightly below chance, and in tasks where the fixated faces showed no clear majority, accuracy was pointedly above chance. These results strongly illustrate that the representation of the group is based on more than simply the faces that have been fixated. Whether this is through parafoveal vision, or through some details perceived during saccades is unclear at this stage, but there is obviously something other than fixated faces informing the ensemble representation, and thus the decision about the group.

6.4. Isolating individual group members

Haberman & Whitney (2009) found that, while participants were quite capable of discerning the mean emotion of a group, they were not so able to identify the individual group members. Their stimuli were, however, morphed images of a single identity, and the factor by which membership was to be determined was the intensity of the subtly varying expression. Experiment 7 examined whether participants could identify the most or least attractive face in a group, both from limited display durations, and in a free-viewing condition. The accuracy of responses was based on the ratings of attractiveness that each participant provided after the experiment. The results showed that even when given an unlimited time to view the group and make a decision,

participants were still not consistently selecting the face rated as the most/least (or joint most/least) attractive in the group. Understandably, the error was significantly less in the self-paced condition than in the timed condition, but it was still enough to suggest some impact of the group presentation on the way that attractiveness is represented. It could be that there is indeed a compulsory averaging of the group (Ariely, 2001), and any information about individual members that is subsequently required must be extracted back out of the ensemble representation, with less fidelity. While the self-paced condition allows the time to inspect the individuals further, their membership of the ensemble might have an indelible effect on any further assessments.

6.5. The nature of the ensemble

While there is a lot of research supporting the idea of an ensemble representation of some sort (Ariely, 2001; Haberman & Whitney, 2007, 2009; Alvarez, 2011), there is less clear evidence as to the exact form that the ensemble takes, and certainly not when considering the attractiveness of a group of faces. The 10AFC of Experiment 4, and more extensively its disambiguation with Experiment 5, combined with the results of Experiment 8 go some way to answering this question. Or, more accurately, demonstrate some of the methods that are less likely to reflect the ensembling process.

The 10AFC showed that not only were participants not particularly skilled at estimating the number of attractive (or unattractive) faces in the group, but also that these responses clearly reflected a different judgement from the 2AFC. This suggests that, when making a judgement of the overall majority of a group, this is done based on something other than a count of the number of faces of each type. Of course, this could potentially reflect the counting of a subset of the group to find a representation, but it could also reflect a level of inference from the ensemble - if the ensemble reflects a

generally attractive group, then it is a reasonable conclusion for the participant that the group contained a majority of attractive faces.

Experiment 8, however, actively compared the results of several different possible models of the ensembling process. Interestingly, while most of the results found elsewhere in this research seem to point towards the group context somehow reducing the perceived attractiveness of faces, the calculated average of the group from the ratings of the faces in isolation was the lowest value. This implies that the group context actually increased the perceived attractiveness of the faces, and that the rating of the group overall is certainly not simply a mean value of the attractiveness of each face as a number of singular entities. The morphed image of the group attracted the highest ratings, in line with previous research (E.g. Langlois & Roggman, 1990; Grammer & Thornhill, 1994; Valentine, et al., 2004), and this was actually the furthest from the overall rating of the group, which thoroughly discounts any suggestion of the ensemble representation being a visual amalgamation of all group members. The closest model to the overall ratings given to the groups was that of the mean rating when all members of the group were rated within the group context. However, even this model returned results significantly lower than the overall rating, again suggesting it is not an accurate reflection of the method used to generate the ensemble representation, and that a group does appear to be more attractive than the sum of its parts.

These results do not clarify the method used to generate the ensemble representation, but they do eliminate a few possibilities, some that previously seemed more likely than others. It might be that the ensemble representation and/or the judgements made in these tasks are only based upon a subsample of the group, which could explain why the calculated averages do not align with the ratings given. However, if this was the case, then one might reasonably expect the subsamples to contain a

mixture of attractive and unattractive faces across different trials (especially given earlier results that suggest there is no overt draw of visual attention to either attractive or unattractive faces during these sorts of tasks), and thus the resultant ratings to vary more widely, but still have a mean closer to that of the whole group.

The ensemble could, potentially, be a combination of statistical averaging (I.e. the mean rating of the group members, or the subsample thereof) and visual averaging. If some of the representation reflects a conglomeration of visual information, even if not of the complete set or a perfectly formed morph as used in Experiment 8, then it is possible that whatever visual information is used and combined could benefit from the same effects of averaging and symmetry as the morphed image, thereby increasing in attractiveness. This partial visual averaging, when combined with a baseline statistical averaging, could lead to an increased perception of attractiveness that sits somewhere between the two values, as did the overall rating in this task.

6.6. Implications of this research

The research in this thesis holds some important and intriguing implications for the study of processing attractiveness from groups, and the conventional wisdom of attractive faces drawing visual attention. Results suggested that the nature of the task at hand influenced perceptions of attractiveness and, counter to results from Maner, et al. (2003), when making judgements of the attractiveness of a group, estimates tend towards the unattractive.

When comparing overall judgements of attractiveness of a group with various other composite measures (be that a composite of ratings, or a rating of composites), the group appeared to be considered more attractive than the average of its members, but was still less attractive than an artificially averaged single member. The conclusion

from these results would suggest that whatever form the ensemble representation takes, it lies somewhere between a statistical average of the group members, and a visual average of them.

Finally, attractive faces did not draw any more visual attention than unattractive ones. This is counter to previous findings, suggesting attractive faces draw attention (Maner, et al., 2003; Maner, et al., 2007; Fletcher-Watson, 2008; Sui & Liu, 2009; Leder, et al., 2010; Chen, et al., 2012). It is hypothesised that this difference results from the task, which involves active assessment of attractiveness, rather than a task unrelated to attractiveness, or free-viewing conditions. When actively judging a group, the draw to attractive individuals appears to be mitigated by the need to survey a large number of stimuli.

6.7. Further research directions and limitations

The research presented here found that participants are generally able to assess the majority attractiveness in a group of up to 16 faces, even when presented for very limited periods. However, in this task, and when asked to estimate the number of attractive faces in the group, responses suggest that the faces appeared to be less attractive when presented in the group than when presented individually. The counterpoint to this was that attractiveness ratings of the groups were higher than the mean rating of its constituents, suggesting that the difference in task somehow modulates the way the faces are perceived. This is further supported by the eye-tracking data, which demonstrated, when judging the attractiveness of a group, a lack of the bias towards attractive faces that had previously been shown in free-viewing conditions (Leder, et al., 2010). There is already some evidence here of the nature of the task altering responses, even when still relating to the attractiveness of the faces, and this

finding could be explored further, by preceding the judgement of attractiveness of a group with another task relating to a different judgement of the group (E.g. gender, or emotional expression). These additional judgements might further modulate the judgements of attractiveness, and the eye movements during the trial.

One of the main limitations in this research is the fact that the 10AFC in Experiment 4 only asked participants to estimate the number of attractive faces, not unattractive, and that the instructions did not specify what was meant by an "attractive" face. While both of these were shown by the disambiguation to generally not have impacted the responses to the task, the lack of eye-tracking data for the unattractive task is regrettable. Given some of the differences observed in fixation durations between the 2AFC and 10AFC, there might prove to be further differences when the task changes, even slightly. Further research might look to address this, with a direct comparison between both tasks in the 10AFC.

The exploration of the nature of the ensemble ruled out a few possible suggestions for methods, but the results opened up some other possibilities. The obvious direction for further research from this point is to further investigate what form the ensemble representation might take. This could consider subsampling from the group to generate the ensemble. In particular, to explore the idea of a mix of the statistical and visual averaging, comparing the rating of the group overall with that of a smaller group in which each image is a morph of a subset of the original group. In this way, there is still some statistical averaging to perform, while some of the visual averaging has already been completed. The ratio of number of morphed faces to the number of faces in each morph could be varied, to find the closest approximation to the overall rating of the group.

6.8. Conclusion

This thesis explored participants' ability to assess the attractiveness of groups of faces from varying display durations and in varying group sizes, using eye-tracking technology in some tasks to investigate the patterns of fixations. Results showed that participants were able to rapidly judge the attractiveness of the groups, but that the nature of the task appeared to impact on these judgements. It seems that trying to make judgements about the members of a group causes an underestimation of attractiveness, whereas judging the group as a whole increases the apparent attractiveness. Further, the presence of an attractiveness-related task appears to negate previously observed tendencies to direct visual attention more to attractive faces. The main conclusion of this work is that the ensemble representation of the group does not appear to reflect either a purely statistical or a purely visual averaging of the information, but likely some combination of the two.

7. Biblography

Aharon, I., Etcoff, N., Ariely, D., Chabris, C., O'Connor, E., & Breiter, H. (2001). Beautiful faces have variable reward value: fMRI and behavioral evidence. *Neuron, 32,* 537-551.

Alvarez, G. A. (2011). Representing multiple objects as an ensemble enhances visual cognition. *Trends in Cognitive Sciences, 15*(3), 122-131.

Alvarez, G. A., & Cavanagh, P. (2004). The Capacity of Visual Short-Term Memory is Set Both by Visual Information Load and by Number of Objects. *Psychological Science, 15*(2), 106-111.

Alvarez, G. A., & Oliva, A. (2008). The representation of simple ensemble visual features outside the focus of attention. *Psychological Science, 19*(4), 392-398.

Ariely, D. (2001). Seeing Sets: Representation by Statistical Properties. *Psychological Science, 12*(2), 157-162.

Ariely, D. (2008). Better than average? When can we say that subsampling of items is better than statistical summary representations? *Perception & Psychophysics, 70*(7), 1325-1326.

Becker, D., Kenrick, D., Guerin, S., & Maner, J. (2005). Concentrating on beauty: Sexual selection and sociospatial memory. *PSPB, 31 (12),* 1-10.

Bindemann, M., Burton, A., Hooge, I., Jenkins, R., & De Haan, E. (2005). *Psychonomic Bulletin & Review, 12 (6),* 1048-1053.

Burt, D., & Perrett, D. (1996). Perceptual asymmetries in judgements of facial attractiveness, age, gender, speech and expression. *Neuropsychologia, 35 (5),* 685-693.

Cabeza, R., Bruce, V., Kato, T., & Oda, M. (1999). The prototype effect in face recognition: extension and limits. *Memory & cognition, 27*(1), 139-151.

Cavanagh, P. (2001). Seeing the forest but not the trees. *Nature Neuroscience,* 673-674.

Cerf, M., Frady, E., & Koch, C. (2009). Faces and text attract gaze independent of the task: Experimental data and computer model. *Journal of Vision, 9(12):10,* 1-15.

Chen, W., Liu, C., & Nakabayashi, K. (2012). Beauty hinders attention switch in change detection: The role of facial attractiveness and distinctiveness. *PLoS ONE, 7(2),* 1-7.

Chong, S. C., & Treisman, A. (2003). Representation of statistical properties. *Vision Research, 43*(4), 393-404.

Chong, S. C., & Treisman, A. (2005). Statistical procession: computing the average size in perceptual groups. *Vision Research, 45*(7), 891-900.

Coetzee, V., Re, D., Perrett, D., Tiddeman, B., & Xiao, D. (2011). Judging the health and attractiveness of female faces: Is the most attractive level of facial adipostiy also considered the healthiest? *Body Image, 8,* 190-193.

Crouzet, S., Kirchner, H., & Thorpe, S. (2010). Fast saccades towards faces: Face detection in just 100ms. *Journal of Vision, 10(4):16,* 1-17.

Fletcher-Watson, S., Findlay, J., Leekam, S., & Benson, V. (2008). Rapid detection of person information in a naturalistic scene. *Perception, 37,* 571-583.

De Fockert, J., & Wolfenstein, C. (2009). Rapid extraction of mean identity from sets of faces. *The Quarterly Journal of Experimental Psychology, 62*(9), 1716-1722.

DeBruine, L., Jones, B., Unger, L., Little, A., & Feinberg, D. (2007). Dissociating averageness and attractiveness: Attractive faces are not always average. *Journal of Experimental Psychology: Human Perception and Performance, 33 (6),* 1420-1430.

Diamond, R., & Carey, S. (1986). Why faces are and are not special: an effect of expertise. *Journal of Experimental Psychology, 115*(2), 107-117.

Fantamorph. (2009). Fantamorph 4. Abrosoft co.

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

Fei-Fei, L., Iyer, A., Koch, C., & Perona, P. (2007). What do we perceive in a glance of a real-world scene? *Journal of Vision, 7(1):10,* 1-29.

Fink, B., Grammer, K., & Thornhill, R. (2001). Human (homo sapiens) Facial attractiveness in relation to skin texture and color. *Journal of Comparative Psychology, 115 (1),* 92-99.

Fink, B., & Penton-Voak, I. (2002). Evolutionary psychology of facial attractiveness. *Current Directions in Psychological Science, 11 (5),* 154-158.

Galton, F. (1879). Composite portraits, made by combining those of many different persons into a single, resultant, figure. *Journal of the Anthropological Institute, 8,* 132-144.

Geldart, S. (2010). That woman looks pretty, but is she attractive? Female perceptions of facial beauty and the impact of cultural labels. *Revue europeenne de psychologie appliquee, 60,* 79-87.

Grammer, K., & Thornhill, R. (1994). Human (homo sapiens) facial attractiveness and sexual selection: The role of symmetry and averageness. *Journal of Comparative Psychology, 108 (3),* 233-242.

Guo, K., Liu, C. H., & Roebuck, H. (2011). I know you are beautiful even without looking at you: discrimination of facial beauty in peripheral vision. *Perception, 40*(2), 191-195.

Haberman, J., & Whitney, D. (2007). Rapid extraction of mean emotion and gender from sets of faces. *Current Biology: CB, 17*(17), 751-753.

Haberman, J., & Whitney, D. (2009). Seeing the mean: Ensemble coding for sets of faces. *Journal of Experimental Psychology: Human Perseption and Performance, 35*(3), 718-734.

Haberman, J., & Whitney, D. (2010). The visual system discounts emotional deviants when extracting average expression. *Attention, Perception, & Psychophysics, 72*(7), 1825-1838.

Haberman, J., & Whitney, D. (2012). Ensemble Perception: summarizing the scene and broadening the limits of visual processing. In J. Wolfe, & L. Robertson, *From Perception to Consciousness: Searching with Anne Treisman.* Oxford University Press.

Haberman, J., Harp, T., & Whitney, D. (2009). Averaging facial expression over time. *Journal of Vision, 9,* 1-13.

Halberstadt, J., & Rhodes, G. (2000). The attractiveness of nonface averages: Implications for an evolutionary explanation of the attractiveness of average faces. *Psychological Science, 11 (4),* 285-289.

Hansen, C. H., & Hansen, R. D. (1988). Finding the face in the crowd: An anger superiority effect. *Journal of Personality and Social Psychology, 54*, 917-924.

Henderson, J., & Hollingworth, A. (1999). High-level scene perception. *Annual Review of Psychology, 50,* 243-271.

Hershler, O., & Hochstein, S. (2005). At first sight: A high-level pop out effect for faces. *Vision Research, 45,* 1707-1724.

Holland, E. (2009). Limitations of traditional morphometrics in research on the attractiveness of faces. *Psychonomic Bulletin & Review, 16 (3),* 613-615.

Kampe, K. K., Frith, C. D., Dolan, R. J., & Frith, U. (2001). Psychology: Reward value of attractiveness and gaze. *Nature, 413*(6856), 589-589.

Kersten, D. (1987). Predictability and redundancy of natural images. *J. Opt. Soc. Am. A.*(4), 2395-2400.

Kowner, R. (1996). Facial Asymmetry and attractiveness judgment in developmental perspective. *Journal of Experimental Psychology: Human Perception and Performance, 27 (3),* 662-675.

Krysko, K., & Rutherford, M. (2009). The face in the crowd effect: Threat-detection, advantage with perceptually intermediate distractors. *Visual Cognition, 17 (8),* 1205-1217.

Langlois, J., & Roggman, L. (1990). Attractive faces are only average. *Psychological Science, 1 (2),* 115-121.

Langlois, J., Roggman, L., & Musselman, L. (1994). What is average and what is not average about attractive faces? *Psychological Science, 5 (4),* 214-220.

Leder, H., Tinio, P., Fuchs, I., & Bohrn, I. (2010). When attractiveness demands longer looks: The effects of situation and gender. *The Quarterly Journal of Experimental Psychology, 63 (9),* 1858-1871.

Levy, B., Ariely, D., Mazar, N., Chi, W., Lukas, S., & Elman, I. (2008). Gender differences in the motivational processing of facial beauty. *Learning and Motivation, 39,* 136-145.

Luck, S. J., & Vogel, E. K. (1997). The capacity of visual working memory for features and conjunctions. *Nature, 390,* 279-281.

Lundstrom, J., Boyle, J., Zatorre, R., & Jones-Gotman, M. (2008). Functional neuronal processing of body odors differs from that of similar common odors. *Cerebral Cortex, 18,* 1466-1474.

Maner, J., Gailliot, M., & DeWall, C. (2007). Adaptive attentional attunement: Evidence for mating-related perceptual bias. *Evolution and Human Behaviour, 28,* 28-36.

Maner, J., Kenrick, D., Becker, D., Delton, A., Hofer, B., Wilbur, C., & Neuburg, S. (2003). Sexually selective cognition: Beauty captures the mind of the beholder. *Journal of Personality and Social Psychology, 85 (6),* 1107-1120.

Marchant, A. P., Simons, D. J., & de Fockert, J. W. (2013). Ensemble representations: effects of set size and item heterogeneity on average size perception. *Acta Psychol (Amst), 142*(2), 245-250.

Martelli, M., Majaj, N. J., & Pelli, D. G. (2005). Are faces processed like words? A diagnostic test for recognition by parts. *Journal of Vision, 5*(1), 6-6.

Mealey, L., Ridgstock, R., & Townsend, G. (1999). Symmetry and perceived facial attractiveness: A monozygotic co-twin comparison. *Journal of Personality and Social Psychology, 76 (1),* 151-158.

Melacci, S., Sarti, L., Maggini, M., & Gori, M. (2010). A template-based approach to automatic face enhancement. *Pattern Anal Applic, 13,* 289-300.

Myczek, K., & Simon, D. J. (2008). Better than average: Alternatives to statistical summary representations for rapid judgments of average size. *Perception & Psychophysics, 70,* 772-788.

Norman, D. A., & Bobrow, D. G. (1975). On data-limited and resource-limited processes. *Cognitive Psychology, 7*(1), 44-64.

Nothdurft, H. C. (1993). Faces and facial expressions do not pop out. *Perception, 22,* 1287-1298.

O'Doherty, J., Winston, J., Critchley, H., Perrett, D., Burt, D., & Dolan, R. (2003). Beauty in a smile: The role of medial orbitofrontal cortex in facial attractiveness. *Neuropsychologia, 41,* 147-155.

Ohman, A., Juth, P., & Lundqvist, D. (2010). Finding the face in a crowd: Relationships between distractor redundancy, target emotion, and target gender. *Cognition and Emotion, 24 (7),* 1216-1228.

Olson, I., & Marshuetz, C. (2005). Facial attractiveness is appraised in a glance. *Emotion, 5 (4),* 498-502.

Parkes, L., Lund, J., Angelucci, A., Solomon, J., & Morgan, M. (2001). Compulsory averaging of crowded orientation signals in human vision. *Nature Neuroscience, 7,* 739-744.

Parkhurst, D., & Niebur, E. (2003). Scene content selected by active vision. *Spatial Vision, 16 (2),* 125-154.

Peelen, M., Lucas, N., Mayer, E., & Vuilleumier, P. (2009). Emotional attention in acquired prosopagnosia. *SCAN, 4,* 268-277.

Penton-Voak, I., Jones, B., Little, A., Baker, S., Tiddeman, B., Burt, D., & Perrett, D. (2001). Symmetry, sexual dimorphism in facial proportions and male facial attractiveness. *Proceedings of the Royal Society, 628,* 1617-1623.

Perrett, D. (2010). *In your face.* Basingstoke: Palgrave Macmillan.

Perrett, D., Burt, D. M., Penton-Voak, I., Lee, K., Rowland, D., & Edwards, R. (1999). Symmetry and Human Facial Attractiveness. *Evolution and Human Behaviour, 20,* 295-307.

Perrett, D., May, K., Yoshikawa, S. (1994). Facial shape and judgements of female attractiveness. *Nature, 368,* 239-242.

Peskin, M., & Newell, F. (2004). Familiarity breeds attraction: Effects of exposure on the attractiveness of typical and distinctive faces. *Perception, 33,* 147-157.

Potter, T., & Corneille, O. (2008). Locating attractiveness in the face space: Faces are more attractive when closer to their group prototype. *Psychonomic Bulletin & Review, 15*(3), 615-622.

PST. (2003). E-Prime 1.1. Sharpsburg, PA: Psychology Software Tools.

Rhodes, G., Hickford, C., & Jeffery, L. (2000). Sex- typicality and attractiveness: Are supermale and superfemale faces super- attractive?. *British Journal of Psychology, 91*(1), 125-140.

Rhodes, G., Jeffery, L., Watson, T. L., Clifford, C. W., & Nakayama, K. (2003). Fitting the mind to the world: Face adaptation and attractiveness aftereffects. *Psychological science*, *14*(6), 558-566.

Rhodes, G., Proffitt, F., Grady, J. M., & Sumich, A. (1998). Facial symmetry and the perception of beauty. *Psychonomic Bulletin & Review*, *5*(4), 659-669.

Rhodes, G., Sumich, A., & Byatt, G. (1999). Are average facial configurations attractive only because of their symmetry?. *Psychological Science*, *10*(1), 52-58.

Rhodes, G., & Tremewan, T. (1996). Averageness, exaggeration, and facial attractiveness. *Psychological science*, *7*(2), 105-110.

Schacht, A., Werheid, K., & Sommer, W. (2008). The appraisal of facial beauty is rapid but not mandatory. *Cognitive, Affective, & Behavioral Neuroscience*, *8*(2), 132-142.

Shepherd, J. W., & Ellis, H. D. (1973). The effect of attractiveness on recognition memory for faces. *The American journal of psychology*, 627-633.

Solso, R. L., & McCarthy, J. E. (1981). Prototype formation of faces: A case of pseudo- memory. *British Journal of Psychology*, *72*(4), 499-503.

Sui, J., & Liu, C. H. (2009). Can beauty be ignored? Effects of facial attractiveness on covert attention. *Psychonomic Bulletin & Review*, *16*(2), 276-281.

Thornhill, R., & Gangestad, S. W. (1993). Human facial beauty. *Human nature*, *4*(3), 237-269.

Thornhill, R., & Gangestad, S. W. (1999). Facial attractiveness. *Trends in cognitive sciences*, *3*(12), 452-460.

Thorpe, S., Fize, D., & Marlot, C. (1996). Speed of processing in the human visual system. *Nature, 381*(6582), 520-522.

Todd, J. J., & Marois, R. (2004). Capacity limit of visual short-term memory in human posterior parietal cortex. *Nature, 428*(6984), 751-754.

Todd, J. J., & Marois, R. (2005). Posterior parietal cortex activity predicts individual differences in visual short-term memory capacity. *Cognitive, Affective, & Behavioral Neuroscience, 5*(2), 144-155.

Treisman, A. (1964). Selective attention in man. *British Medical Bulletin, 20*, 12-16.

Treisman, A. (1988). Features and objects: The fourteenth Bartlett memorial lecture. *Quarterly Journal of Experimental Psychology, 40*(A), 201-237.

Treisman, A. (1998). Feature binding, attention and object perception. *Philos Trans R Soc Lond B Biol Sci., 353*(1373), 1295-1306.

Treisman, A. M., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology, 12*(1), 97-136.

Treisman, A., & Schmidt, H. (1982). Illusory conjunctions in the perception of objects. *Cognitive Psychology, 14*(1), 107-141.

Valentine, T. (1991). A unified account of the effects of distinctiveness, inversion, and race in face recognition. *The Quarterly Journal of Experimental Psychology, 43*(2), 161-204.

Valentine, T., Darling, S., & Donnelly, M. (2004). Why are average faces attractive? The effect of view and averageness on the attractiveness of female faces. *Psychonomic Bulletin & Review, 11*(3), 482-487.

Wallis, G., Siebeck, U. E., Swann, K., Blanz, V., & Bulthoff, H. H. (2008). The prototype effect revisited: Evidence for an abstract feature model of face recognition. *Journal of Vision, 8*(3), 1-15.

Webster, M. A., & MacLeod, D. I. (2011). Visual adaptation and face perception. *Philosophical Transactions of the Royal Society B: Biological Sciences, 366*(1571), 1702-1725.

Wickham, L. H., & Morris, P. E. (2002). Attractiveness, distinctiveness, and recognition of faces: attractive faces can be typical or distinctive but are not better recognized. *The American journal of psychology, 116*(3), 455-468.

Willems, R. M., Peelen, M. V., & Hagoort, P. (2010). Cerebral lateralization of face-selective and body-selective visual areas depends on handedness. *Cerebral Cortex, 20*(7), 1719-1725.

Willis, J, & Todorov, A. (2006). First impressions: Making up your mind after a 100-ms exposure to a face. *Psychological Science, 17 (7),* 592-598.

Young, A. W., Hellawell, D., & Hay, D. C. (1987). Configurational information in face perception. *Perception, 16*(6), 747-759.