Ensemble Perception of Hue

John Maule

Thesis submitted for the degree of Doctor of Philosophy

University of Sussex

September 2015

Declaration

I hereby declare that this thesis has not been and will not be, submitted in whole or in part to another University for the award of any other degree.

Signature:

……

Article Format Thesis Declaration

The thesis conforms to an ‘article format’ in which the first chapter presents an overview of the relevant literature, an outline of the empirical work of the thesis, and discussion of the overall contribution of the thesis to the field. The remaining five chapters consist of discrete papers written for publication in peer-reviewed journals. Two of these chapters have been published, two have been submitted for review, and one is prepared for submission.

Chapters and Author Contributions

Chapter 1 provides an introduction to the relevant literature, along with an outline of the empirical work of the thesis, and a discussion of the main findings, implications and conclusions.

Chapter 2 is published in the *Journal of the Optical Society of America: A* as:

> Maule, J., Witzel, C., & Franklin, A. (2014). Getting the gist of multiple hues: Metric and categorical effects on ensemble perception of hue. *Journal of the Optical Society of America: A*, 31(4), A93-A102, doi:10.1364/Josaa.31.000a93.

The author contributions are as follows: JM was responsible for writing the manuscript, designing the experiment, data collection and analysis. CW provided information on hue discrimination thresholds and provided feedback on the manuscript. AF provided feedback on the experiment design and manuscript.

Chapter 3 is published in the *Journal of Vision* as:

> Maule, J., & Franklin, A. (2015). Effects of ensemble complexity and perceptual similarity on rapid averaging of hue. *Journal of Vision*, *15*(4):6, 1-18, doi:10.1167/15.4.6.

The author contributions are as follows: JM was responsible for writing the manuscript, designing the experiment, data collection and analysis. AF provided feedback on the experiment design and manuscript.

Chapter 4 has been submitted for review to the *Journal of the Optical Society of America: A* as:

> Maule, J., & Franklin, A. Accurate rapid averaging of multi-hue ensembles is due to a limited capacity sub-sampling mechanism. *Journal of the Optical Society of America: A* (submitted).

The author contributions are as follows: JM was responsible for writing the manuscript, designing the experiment and simulation, data collection and analysis. AF provided feedback on the experiment design and manuscript.

Chapter 5 is written in a style appropriate for submission to *Autism Research:*

> Maule, J., Stanworth, K., Pellicano, E., & Franklin, A. Ensemble perception of colour in autistic adults.

The author contributions are as follows: JM was responsible for writing the manuscript, designing the experiment, data analysis. KS helped with recruitment and testing of participants. EP and AF provided feedback on the experiment design and the manuscript.

Chapter 6 has been submitted for review to the *Journal of Autism and Developmental Disorders* as:

> Maule, J., Stanworth, K., Pellicano, E., & Franklin, A. Colour afterimages in autistic adults. *Journal of Autism and Developmental Disorders* (submitted).

The author contributions are as follows: JM was responsible for writing the manuscript, designing the experiment, data analysis. KS helped with recruitment and testing of participants. EP and AF provided feedback on the experiment design and the manuscript.

Acknowledgments

First, and foremost, I am deeply indebted to my supervisor, Professor Anna Franklin, who has supported and encouraged me at every stage of the PhD process. Put simply, I could not have hoped for a better supervisor.

I would like to thank my collaborators: Christoph Witzel, whose advice and tutelage during the early part of my PhD was much appreciated; and Liz Pellicano, whose support in the design and interpretation of the autism experiments was invaluable.

I would also like to thank the undergraduate students who have assisted me in the collection of data during the past three years: Natasha Mckeeman, Jessica Banks, Diyana Ognyanova and Grace Morgan. I would especially like to thank Kirstie Stanworth for her hard work recruiting and testing so many participants for our autism projects.

Next I would like to thank my colleagues from The Sussex Colour Group: Alice Skelton, Lewis Forder, Gemma Catchpole, Xun He, Chris Racey and Marie Rogers, who have tolerated sharing an office and lab space with me during these three years, and have provided practical help and a scientifically simulating working environment, as well as cups of coffee, birthday cake, chatter and amusement. I would also like to thank all my peers in the School of Psychology for their solidarity, particularly Zoe Hopkins and Georgia Leith.

Finally, I thank my friends and family for their love and encouragement. I especially thank Becca, without whom I may not have applied for a PhD in the first place, and whose steadfast support and faith has kept me on an even keel during the more stressful times.

My PhD studentship was funded by the Economic and Social Research Council (ES/J500173/1) and the University of Sussex.

UNIVERSITY OF SUSSEX

JOHN JAMES MAULE – PHD PSYCHOLOGY

ENSEMBLE PERCEPTION OF HUE

SUMMARY

In order to rapidly get the gist of new scenes or recognise objects, the brain must have mechanisms to process the large amount of visual information which enters the eye. Previous research has shown that observers tend to extract the average feature from briefly seen sets of multiple stimuli that vary along a dimension (e.g., size), a phenomenon called ensemble perception. This thesis investigates ensemble perception of hue.

Paper 1 (Maule, Witzel & Franklin, 2014) demonstrates that human observers have memories biased towards the mean hue of a rapidly-presented ensemble of colours. Paper 2 (Maule & Franklin, 2015) further shows that observers are able to identify the mean hue from a distractor fairly reliably, provided the range of hues is manageable. Paper 3 provides evidence that, while observers' settings of the mean hue converge quite closely on the true mean across many trials, the precision of those settings is low and does not support claims that ensemble perception can surpass the limits of visual working memory. Paper 4 found that adults with autism have an enhanced ability to discriminate members from non-members of multi-hue ensembles, and a similar ability to extract the mean hue compared to typical adults, but are worse at averaging small sets. Finally, paper 5 investigated colour afterimages in adults with autism and whether they are affected by top-down gist of a scene. It was found that afterimages were no different in autism compared to a typical group.

Overall these studies provide the first comprehensive exploration of ensemble perception of hue, showing that observers can extract and estimate the mean hue of a rapidly-presented multi-colour ensemble with a small hue variance. The ability to average hue may be driven by a sub-sampling mechanism, but results from autistic adults suggests that it can be modulated by processing style.

Table of Contents

References

Chapter 1

Introduction & Thesis Overview

The overall aim of this thesis is to investigate and characterise ensemble perception of hue. This chapter will first review the key literature relating to this aim, summarising the main theoretical topics and issues relevant to the research conducted. Next, it will explain the rationale for the five empirical papers reported that form the body of the thesis. It will outline the main research questions which are addressed by each of the papers, and present an overview of the main findings. Finally, the overall scientific contribution of the thesis will be presented, and potential future directions will be outlined.

1.1 Ensemble perception

The human visual system receives vast quantities of information through the eye – from basic properties and features indicating the presence of edges and borders (e.g., defining the outer contours of a lemon), through to sets of objects (e.g., the lemons share a basket with some limes and oranges) to whole scene characteristics (e.g., the basket is in a marketplace). Generally we are able to rapidly, simultaneously and effortlessly understand and interact with new scenes, and have a rich and detailed sense of our surroundings. How the brain copes with the large amount of visual information available, how it processes and represents this without being overloaded, is a question which has been the topic of many scientific investigations.

One dominant theme in such research is the idea that the brain does not represent and process everything in the visual scene with high fidelity. Instead, the brain is expert at reducing the processing load by using summary representations – extracting the gist of a scene, rather than the precise nature of the elements within it. In order to provide the sensation of a rich and coherent visual experience with which we can interact with little delay or error, the brain must be an extremely powerful statistical processor, capable of extracting a summary, calculating correlations between features, integrating past experiences and constructing a coherent visual world for the beholder.

Ensemble perception describes the representation of a set of items by summary statistics, rather than as individual items (Haberman & Whitney, 2012). For example, when shown a set of differently-sized circles, observers are able to make accurate judgments about the mean size of the set, despite having poor knowledge of the individual sizes which they saw (Ariely, 2001). Since the early 2000s the phenomenon has attracted interest from vision and cognitive scientists due to the rapid speed with which ensemble representations appear to be formed, the large number of items which the process appears to be capable of integrating, and the seemingly automatic nature of ensemble perception. The perceptual abilities displayed led to the intriguing question of whether the mechanism responsible for ensemble perception could be accounted for within understood frameworks of visual attention and encoding, or whether it reflected an unknown mechanism of gist extraction and representation which might help explain how we come to experience the world with the richness that we do, in spite of the limits of our visual working memory (Alvarez, 2011).

This thesis is concerned with the extraction and representation of summary statistics from sets (“ensembles”) with local items or objects (“elements”) which can be individuated,

which have distinct identities in terms of the stimulus (such as noticeably different colours), and may otherwise be the subject of their own, singular representation in high-level visual cortex and hence have implications for the understanding of coding of colour and objects in visual working memory. For the purposes of this thesis, ensemble perception is distinct from global representations which could be explained by pooling across low-level neurons sensitive to the feature in question (e.g., in texture perception, see Dakin, 2012), and where averaging may be completely obligatory, e.g., orientation (e.g., Dakin, 2001; Parkes, Lund, Angelucci, Solomon, & Morgan, 2001) or motion direction (e.g., Watamaniuk & McKee, 1998; Watamaniuk, Sekuler, & Williams, 1989). Recent investigations into ensemble perception of such stimuli have found evidence for ensemble perception in size and in faces (identity, expression and gaze direction). These domains are focused on as they are most likely to be informative about ensemble perception for high-level representation of objects – size is a feature of which cannot be unbound from the object (Haberman & Whitney, 2012); likewise, faces are represented in the temporal cortex (e.g., Kanwisher, McDermott, & Chun, 1997) and are the result of the combination of features. A review of the key evidence for these domains will be presented in turn.

1.1.1 Size

Size is the visual feature which has received the most sustained focus for experiments investigating ensemble perception. Part of the reason for this is that there are not low-level size receptors across which pooling responses might explain the tendency toward average size (Haberman & Whitney, 2012; Marchant, Simons, & De Fockert, 2013; Simons & Myczek, 2008), leading to the intriguing prospect of an item attribute

pooling mechanism which surpasses the limits of visual working memory for discrete items (Alvarez, 2011).

Interest in the idea of perceptual averaging was prompted by Ariely's (2001) investigation of summary statistics for discs of different size. Ariely showed his observers ensembles consisting of 4, 8, 12 or 16 discs (see figure 1.1), each of which could be one of four different sizes. The ensemble was presented for 500ms, to minimise the time in which the observer could encode individual elements. Following the ensemble, the observer was shown one test disc, and was asked to indicate whether they thought the disc was a member of the original ensemble. The test disc could be either a match for one of the discs in the ensemble, intermediate to the size of the ensemble elements, or larger/smaller than the most extreme (largest/smallest) elements in the ensemble. Ariely also tested a two-alternative forced-choice (2AFC) version of this task, in which one disc was a member of the set (i.e. identical in size to one of the ensemble elements) and the other was not a member, but was intermediate to member sizes in the ensemble. The results indicated a tendency to attribute membership to any disc which was within the range of sizes present in the ensemble for the single test disc task, and observers made no distinction between the member and non-member sizes, performing at chance in the 2AFC task.

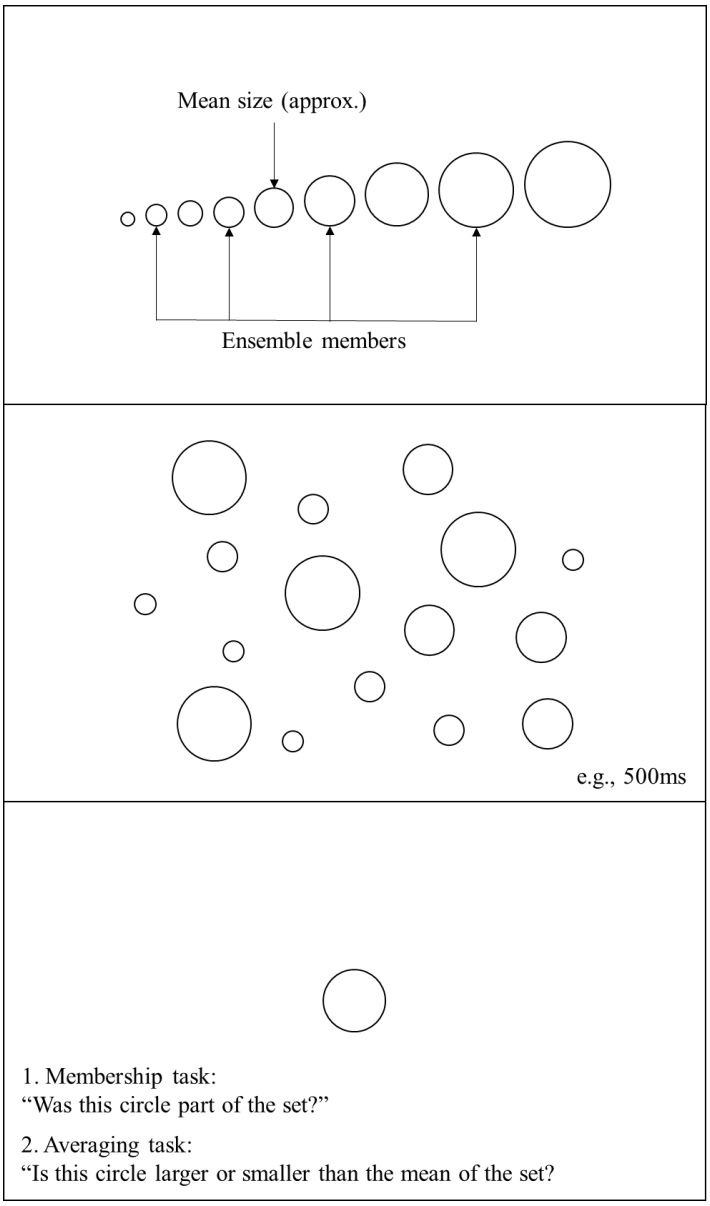

Figure 1.1. Upper panel: Stimulus array from which the ensemble and test spots are chosen. The ensemble contains four different sizes, each represented four times. In the membership task each exemplar from the stimulus range is presented. The spacing between ensemble member sizes allows for probing membership recognition for sizes within and outside of the range of sizes in the ensemble. Middle panel: Representation of a sixteen-element ensemble of the style used in Ariely's (2001) size averaging experiments. Lower panel: A singleton target, to which observers respond based on the task.

Ariely's second experiment used the same ensembles, but asked observers to assess whether they thought a single test disc was larger or smaller than the average size of the previous ensemble. Discrimination thresholds extracted from this task revealed that ensembles with average size differing by 6-12% were discriminable. This was in contrast to a lack of discriminability in the membership task for sizes which differed by 18% (about three times discrimination threshold for size). The data also showed that the accuracy of mean judgments were unaffected by changes in set size. Ariely claimed, with additional support from an ideal observer simulation, that the averaging of size must require a representation of the ensemble mean which is independent of representations of the individual elements, also suggesting that the mechanism operated holistically across the whole ensemble.

Over the following few years the claims of rapid averaging of size were replicated and extended in various ways (e.g., Allik, Toom, Raidvee, Averin, & Kreegipuu, 2014; Chong & Treisman, 2003, 2005b; De Fockert & Marchant, 2008; Marchant et al., 2013; Oriet & Brand, 2013; Robitaille & Harris, 2011; Sweeny, Wurnitsch, Gopnik, & Whitney, 2014; Utochkin & Tiurina, 2014). Chong and Triesman (2003) measured discrimination of average size using pairs of ensembles. Observers were asked to identify the ensemble of circles with the larger average size from the pair in a method of constant stimuli design. They found differing effects of exposure duration for single circles, homogenous ensembles (where every element is the same size) and heterogeneous ensembles (where a variety of sizes are represented). For homogeneous ensembles and single exemplars, average size discrimination thresholds increased with shorter exposure times, whereas thresholds for heterogeneous ensembles were relatively invariant to the duration of the stimuli. Furthermore, although thresholds were lower for the single circles and homogeneous ensembles when stimuli were presented for 1000ms, discrimination

thresholds were no different when shorter durations of 100ms or 50ms were used. This offered support for the rapid, holistic averaging mechanism proposed by Ariely (2001).

Other studies also replicated Ariely's finding that mean judgments are unaffected by set size (Chong & Treisman, 2005b; Marchant et al., 2013; Utochkin & Tiurina, 2014) and another found that reaction times for mean size judgments are actually reduced for larger sets (Robitaille & Harris, 2011). Ariely's finding of strong mean size representation in spite of a lack of knowledge about individual items has also been replicated (Allik et al., 2014; Corbett & Oriet, 2011), lending support to the idea that the representation of the mean size takes the place of individual representations in memory. Recently it has also been shown that 4-5 year-old children are able to select the group of 'oranges' with the largest size from two ensembles, and that this process cannot be based on simply picking the set with the largest individual orange (Sweeny et al., 2014). This study found that the mechanism appears to operate in a similar way in children as in adults, although is less efficient, implying that there may be a developmental trajectory of ensemble perception which is yet to be fully explored.

Latter studies have sought to characterise the role for summary statistics by using more indirect, implicit measures of their influence on perception and cognition. For example, it has been shown that a preceding ensemble can prime the detection of a low-contrast target circle if the mean size of the priming ensemble matches the size of the single target (Marchant & De Fockert, 2009). Similarly, Treisman (2006) reports that reaction times on a same-different task can be reduced by priming with an ensemble with a mean size equal to the target, and that observers make more false positive responses on a visual search task if the mean size of the array is equal to the target size. These observations

demonstrate that even when making a judgment about the ensemble is not the focus of a task, the summary statistics of size can still have an effect on the observers' responses.

It has also been demonstrated that observers can build up adaptation aftereffects due to average size. Following two minutes of adaptation to a pair of ensembles with different mean sizes (e.g., larger on the left, smaller on the right), subsequent displays will appear skewed away from the adapting ensemble (e.g., two equally-sized ensembles will appear smaller on the left than on the right) (Corbett, Wurnitsch, Schwartz, & Whitney, 2012; see also, Corbett & Song, 2014). Crucially, these aftereffects were reduced when the adapting ensembles contained more variance in size, indicating more imprecise mean representation and, hence, weaker aftereffects. This is concordant with data showing that ensembles of different sizes with greater range are averaged less precisely than those with a smaller range of sizes presented (Utochkin & Tiurina, 2014). Finding adaptation aftereffects of mean size indicates that ensemble representations of size may be encoded along a single dimension (i.e. small to large) by a distinct population of neurons at some point in the visual stream (Corbett et al., 2012). A subsequent study showed that mean size aftereffects affected judgments not just in the part of visual field adapted, but also transferred between eyes, spatial locations and retinotopic locations, suggesting that mean size is represented at multiple levels in the visual system (Corbett & Melcher, 2014a).

1.1.2 Faces

Ensemble perception of faces was first demonstrated by Haberman and Whitney (2007), who showed that observers were able to accurately say whether a single test face was happier or sadder than a preceding ensemble of faces from a happy-sad morphed continuum of emotional expressions. They found discrimination to be equally good for

heterogeneous ensembles (where a number of different faces are presented) as for homogeneous ensembles (where every face is identical), in spite of having very poor recognition of which specific faces were present in the ensemble, as indicated by at-chance performance on a membership task. Furthermore, the findings also generalised to faces which varied in perceived gender.

Latter explorations of the characteristics of ensemble perception of emotional expression found a pattern strikingly similar to that for size. Membership identification following ensembles of emotional expressions peaked at the mean, even though the mean was never a part of the set, and also showed invariance to set size (Haberman & Whitney, 2009). Likewise, observers are unable to identify the individual faces driving a change of gist in ensembles, yet can identify which of two ensembles is "happier/sadder" (Haberman & Whitney, 2011).

Variations in the shape of the distribution of the stimuli in ensembles have also been investigated for emotional faces. Haberman and Whitney (2010) used a method of adjustment, whereby observers were to adjust a single face to match their estimation of the mean face for a set which was presented for just 250ms. For each ensemble of 12 faces, two elements were emotional deviants – their expression was radically different from the rest of the set (e.g., two happy faces among ten sad). They found that observers made settings consistent with a systematic exclusion of the outlier faces from their chosen mean, arguing that this could reflect a discarding of the outlier information in order to maintain a manageable variance when averaging, or that other summary statistics, such as the mode or median, might be more informative, due to their robustness to the influence of outliers (Haberman & Whitney, 2010).

Further studies on ensemble perception of faces have demonstrated that as well as expression and gender, observers are also able to average identity (De Fockert & Wolfenstein, 2009), and that this ability is intact even when test identities are presented from a different view-point (Leib et al., 2014). Ensemble perception of a crowd's perceived direction of gaze has also been reported (Sweeny & Whitney, 2014).

The concept of computing the average face is also the basis for dominant norm-based coding accounts of face perception, in which the brain maintains a multi-dimensional face space which is centred on a mean face (see M. A. Webster & MacLeod, 2011 for a review), which does not represent any real face in the observer's environment. Therefore it is likely that a representation of the mean face does emerge from face processing, at least where faces are encoded and integrated individually over time. However, as for size, the formation of a rapid and holistic mean for a set in the absence of individual representations suggests that there must be early processes which represent the set by the average. However, this process cannot be mediated by lower-level averaging mechanisms, as it has been shown that inverted and scrambled faces are averaged far less accurately and efficiently than upright faces (Haberman & Whitney, 2009; Leib, Puri, et al., 2012; Sweeny & Whitney, 2014). The robustness of identity averaging to changes in viewpoint (Leib et al., 2014) also speaks to a high-level integration of the ensemble faces, perhaps using a mechanism related to the norm-based coding of 'face space'.

1.2 Mechanisms underlying ensemble perception

The mechanisms which underlie the rapid extraction of summary statistics have been the subject of intense debate, and have hinged around two closely related topics. One is the role of attention, specifically how focused and distributed modes of attention

can affect the extraction of summary statistics, or, indeed, whether summary statistics are pre-attentive. The second is whether the process is holistic – taking in all of the items in the ensemble and averaging them – or whether a sub-sample of items from each ensemble could suffice to explain the observer performance on the types of task used in the ensemble perception literature. Although these two issues are very closely related and are often conflated in discussions, the methods necessary to address each topic are quite different from one another. The first requires some manipulation or observation of attention during an ensemble perception task, while the second typically requires ideal observer simulations to provide baseline performance or demonstrate the gamut of possible patterns of performance given certain models of ensemble processing. Therefore the following sections will review the evidence pertaining to each of these topics individually.

1.2.1 The role of attention in ensemble perception

Understanding the role of attention is crucial to evaluating claims that summary statistics are extracted and represented automatically. There is consistent evidence that observers show better mean representation than individual item representation, that average judgments are unaffected by changes in number of elements and the ability to extract the mean applies to sets containing more elements than can be individually represented in visual working memory. These findings may imply that a pre-attentive summary is formed rather than the mean being computed from representations of individual items. Evidence for pre-attentive ensemble perception, or ensemble perception outside the focus of attention (see Alvarez & Oliva, 2008; 2009, for examples from orientation and location averaging) would be supportive of an automatic extraction

mechanism, and importantly would help to demonstrate that the representation of summary statistics occurs even when the observer is not tasked with judging an average or considering set membership. Priming (Marchant & De Fockert, 2009; Treisman, 2006) and adaptation (Corbett & Melcher, 2014a; Corbett & Oriet, 2011) are both suggestive of an automatic encoding of the mean size, but studies with manipulations of attention are better placed to explore the idea.

Some have claimed that size averaging is obligatory, occurring before the elements even enter our awareness. Invariance of averaging ability despite different distributions to sizes in ensembles (Allik et al., 2014) and intact size averaging despite object masking (Choo & Franconeri, 2010) both imply that averaging occurs before the objects reach attention. Furthermore, it has been shown that consistent summary statistics of size in the background can reduce saccade latencies and increase the efficiency of visual search over many trials, even when size is not the focus of the main search task and the summary statistical information is contained in the background (Corbett & Melcher, 2014b).

Evidence from several lesion studies has shown that patients with disorders of focal attention can still average sets. Patients with unilateral neglect, in which items presented in parts of the visual field are not consciously perceived, are nevertheless still able to average the size of items presented in their neglected area(s) of the visual field (Lanzoni, Melcher, Miceli, & Corbett, 2014; Pavlovskaya, Soroker, Bonneh, & Hochstein, 2015). These patients also show speeded visual search when global summary statistics of size are consistent across trials during a block (Leib, Landau, Baek, Chong, & Robertson, 2012) as has been shown for typical individuals (Corbett & Melcher, 2014b). There is also evidence for intact representations of average identity and emotion in patients with prosopagnosia (‘face-blindness’ causing difficulties in recognising individual faces)

(Leib, Puri, et al., 2012). These studies show that even where neurological conditions that cause focal attentional difficulties perceptual averaging can still occur, suggesting that ensemble perception may occur outside of conscious awareness. In contrast, averaging of facial identity has been found to be reduced in children and adolescents with autism (Rhodes, Neumann, Ewing, & Palermo, 2014). In a membership task typically developing children tended to think an unseen average face was a part of the ensemble with greater frequency than children with autism. This finding was not due to inattention, difficulty of task or difference in discriminability of the stimuli between groups. This is consistent with accounts enhanced processing of local detail in autism (Frith & Happe, 1994) and impaired global statistical representations (Pellicano & Burr, 2012), and implies that a global mode of processing is necessary for the extraction of summary statistics, even when the task does not require mean judgments.

On the other hand, the literature is also replete with evidence for the role that attention plays in the extraction and representation of summary statistics. For example, when an element is crowded out of visual awareness due to being in the periphery and surrounded by flanking elements, they do not appear to contribute to judgements of mean size (Banno & Saiki, 2012). De Fockert and Marchant (2008) suggested that attention could play a role in the selection of mean sizes, showing that when observers attention was cued towards particular elements of the ensemble their subsequent mean judgments were biased towards the size of that element. Similarly, Albrecht and Scholl (2010) investigated averaging of size over time, using a disc which increased in size (looming phases) and decreased in size (receding phases). Their findings suggested that, asked to judge the mean size across the duration of the dynamic display, observers were biased towards the looming phases. The authors suggested that this was because looming discs are more salient. Together these studies suggest that the encoded ensemble average may

not weight each item equally, but rather those that are more salient, whether due to cueing or a feature of the stimulus, attract a higher weighting in the computation of the average. This interpretation contrasts somewhat with findings of outlier exclusion (even when the outliers are highlighted to the observer) in facial expression mean judgments (Haberman & Whitney, 2010), possibly suggesting that the computation underlying the summary representation may differ for size and faces.

Individual item representations are also implied by a study looking at size averaging when stimuli are subject to the Ebbinghaus illusion (where identically-sized circles appear to be different sizes due to the size of surrounding objects) (Im & Chong, 2009). The results indicated that the mean judgments were based on the perceived size, rather than physical size. Although the Ebbinghaus illusion may be due to representations early in the visual stream, and so does not necessarily indicate that averaging is the result of attention, the phenomenon is based on the appearance of individual circles in context, and so challenges accounts of ensemble perception in which individual items are not represented.

Several studies have also considered the effect of task instructions on the mode of attention and processing. Observers performance on a mean judgment task is worse when trying to extract a summary of two different features (e.g., mean size and speed), compared to extracting just one (Emmanouil & Treisman, 2008). Also, given instructions, observers can extract separable mean size from two separate sets of items designated by colour, even if they are intermixed spatially (Chong & Treisman, 2005b), suggesting that at least general attention to the feature to be averaged and the task is needed to achieve efficient summary statistical representation. Chong and Treisman (2005a) have also shown that member identification is more accurate, and mean judgements less accurate, if a concurrent visual search task promotes serial processing (i.e. the target does not ‘pop-

out' from the distractors, so requires focused-attention to be paid to each element), while mean judgments are more accurate, and member judgments less accurate, if the visual search promotes parallel processing (i.e. the target 'pops-out' from the distractors, so search is very rapid and can be completed with attention distributed across the array).

Most recently, mean size has also been proposed as a possible "unit of selection" to which attention can be directed – Im, Park and Chong (2014) showed observers an ensemble of sizes containing four sub-sets, designated by colour. When asked to make a mean size judgment ("which was smaller" or "which was larger") about two of the sets from the ensemble (post-cued) there was a bias towards more accurate mean representation of the largest set of the four, even when the largest exemplar was part of a different set. This suggested that ensemble summary statistics can be the attribute upon which attentional selection acts – attention is drawn to the group with the largest mean size, even if the largest single exemplar is in the group with the smaller mean size. However, the experiment also showed that the bias can be reversed with different task instructions – when told to judge "which was smaller", mean judgments were most accurate for the smallest set.

In summary, the evidence for averaging with distributed attention, and outside of focal attention is quite strong, especially where the average is not the focus of the task. However, claims of automatic, effortless perceptual averaging may be overstating the strength of summary statistical representation. It appears that attention modulates ensemble representations to a significant extent, and the particular demands of an ensemble perception task may have a crucial effect on the responses given.

1.2.2 Holistic processing vs. sub-sampling

Ensemble perception is characterised by the accurate extraction of the mean despite poor knowledge of individual elements. The consistency of this pattern even with changes to the number of elements have led to a suggestion that the extraction of summary statistics is driven by a rapid, holistic mechanism, incorporating all elements, independently of representations of individual elements (e.g., Ariely, 2001). Rapid, holistic and obligatory perceptual averaging are well-established in the literature on texture perception, where orientation and spatial frequency signals are pooled across low-level representations leading to perception of a distinct texture (see Dakin, 2012). However, rapid and holistic perceptual averaging of objects, where the individual elements might also have their own representation is not as well established. Although the evidence for averaging is often from sets larger than can be held in visual working memory (Alvarez, 2011), claims of holistic ensemble perception must exclude the possibility that an observer could extract a small sub-sample of the ensemble from which they make their mean judgment. Particularly for size, which is a feature of objects not otherwise dissociable from the object itself, ensemble perception based on representation of individual elements is considered the most parsimonious explanation as it does not postulate a new mechanism of perception and cognition, whereas holistic averaging in the absence of individual representation does (Myczek & Simons, 2008; Simons & Myczek, 2008). For this reason, the focus of the ensemble perception literature has included both characterising the process in performance terms and attempting to account for the performance with models and simulations.

Early demonstrations of ensemble perception of size claimed that the performance observed was the result of a holistic averaging process which circumvents the limits of

visual working memory (Ariely, 2001; Chong & Treisman, 2003, 2005a, 2005b). Myczek and Simons (2008) offered the suggestion that the results could be driven by a limited-capacity sub-sampling mechanism which could be within those limits (around four items). They replicated the experiments of Ariely, and Chong and Treisman, and then demonstrated, using an ideal observer simulation, that performance on each of their tasks could be adequately simulated with a sub-sampling mechanism which represents only one or two items from each ensemble and then makes the mean judgment based on that subsample, or using a simple strategy such as identifying the largest and smallest items in each array. However, the simulations lacked any inclusion of judgment error, they represented noise-free observers which Ariely (2008) claimed was unrealistic, and suggested that the inclusion of internal noise in models might have a dramatic effect on the performance on the models relative to the real observers. A subsequent experiment by Chong, Joo, Emmanouil and Treisman (2008) confirmed that when real observers are given subsamples of one or two items from ensembles their mean accuracy is worse than when they are shown the whole ensemble. In reply, Simons and Myczek (2008) accepted the limitations of their simulations, but pointed out that even if the averaging does appear to exceed the capacity limits of working memory, this does not mean that the average is holistically computed.

Subsequent behavioural and simulation data have validated Ariely's (2008) call for the inclusion of internal noise in models of sub-sampling. The variance summation model used by Im and Halberda (2013) includes a parameter for internal noise (based on an estimate from discrimination threshold data) and external noise (variance of the ensemble). Simulations using this model imply that averaging performance would require a subsample of around seven items – far higher than the one or two items suggested previously, and outside the capacity limit of working memory. A sub-sampling model

including internal noise was also insufficient to explain the results of Haberman and Whitney's (2010) demonstration of expression averaging, showing that random sub-sampling could not account for the exclusion of outliers.

Invariance of mean judgments to changes in number of elements has been suggested to be the strongest behavioural evidence against sub-sampling (e.g., Ariely, 2008). If the sub-sample is a fixed size (e.g., two items) and random then adding more elements ought to reduce the accuracy of average judgments. Marchant et al. (2013) claimed that their (noise-free) sub-sampling model predicted invariance to set sizes in their observer data. However, it was subsequently shown that their experiment conflated the range of sizes with the number of elements (Utochkin & Tiurina, 2014). Another model, the Noise and Selection model, which includes internal noise but not external (Allik, Toom, Raidvee, Averin, & Kreegipuu, 2013), suggests that sub-sampling of between one and four items could account for average size discrimination and invariance to number of ensemble elements. These simulations found that the best fit for the data was achieved when the number of elements in the sub-sample was scaled with the number of elements, such that gist extraction for larger sets was based on larger sub-samples.

Although their implications may be equivocal with regard to the distinction between holistic averaging and sub-sampling, the attempts to simulate ideal observer performance have made important contributions to the understanding of the possible mechanisms underlying ensemble perception. Crucially, they also highlight the importance of considering competing explanations for the data from ensemble judgment tasks, and can provide a baseline against which to compare observers' performance.

1.2.3 The function of perceptual averaging

Whether the precise mechanism responsible is based on sub-sampling or is genuinely holistic, and questions about the role for attention notwithstanding, what purpose might the ability to extract the mean serve? Confronted with a complex world the visual system must efficiently process and represent those parts and features which are most salient or informative, and a rapid extraction of summary statistics may assist with this. In general terms, compressing the vast amount of visual information into summary representations may help reduce the load on working memory (Alvarez, 2011; Im & Chong, 2014; but see Baijal, Nakatani, van Leeuwen, & Srinivasan, 2013). By representing a set by its summary property the limited capacity of working memory can be more efficiently used (Attarha, Moore, & Vecera, 2014), allowing more information to be represented (albeit in summary form).

Faces are of particular interest for ensemble perception as there is a clear potential benefit to the efficient and rapid extraction of a set – social interaction with a group might be eased significantly if it is possible to quickly assess the general mood of the crowd, rather than have to process each individual face (Haberman & Whitney, 2007), and likewise, establishing a group's shared locus of interest from ensemble representations of gaze may also have social benefits (Sweeny & Whitney, 2014). For size, computing summary statistics may help maintain a stable visual world, allowing faster saccadic reactions and shorter fixations (Corbett & Melcher, 2014b). In short, the extraction of summary statistics may be crucial to our perception of a stable, rich and detailed visual world.

1.3 Interim conclusion

We have seen how ensemble perception can be demonstrated for size and faces – two domains with low-level attributes but high-level representations as individual objects. Following rapid presentation of an ensemble of different exemplars, observers are drawn more to an unseen mean when asked to identify ensemble members. Their knowledge of the individual elements present in the set is very poor, and yet they show an accurate representation of the mean value. The mechanisms behind this ability are debated. However, comparisons of behavioural and simulation data often imply that the precision of average representations outstrips that which could be expected given the known limits of working memory for individual items. Ensemble perception may help support our feeling of inhabiting a rich and detailed visual world, by allowing our visual system to trade knowledge of precise local detail for a sense of the global gist, represented by the average of the features seen. If this is the case then ensemble perception should also be found for other domains which are important to our sense of the gist of scenes. This thesis investigates ensemble perception in the domain of colour.

1.4 Ensemble perception of colour

1.4.1 Literature on colour averaging

Despite the depth of interest in ensemble perception for other visual domains, whether observers rapidly extract summary statistics of colour has not been adequately investigated. It is unclear whether an observer is able to get the gist of a set of coloured objects, and if so, how the colour might be integrated and represented in memory. Aside from being used as an indicator for grouping of inter-mixed sets (e.g., Chong & Treisman,

2003), before the current investigation, colour had only appeared as the feature to be averaged in one study on ensemble perception. Demeyere, Rzeskiewicz, Humphreys and Humphreys (2008) investigated whether perceptual averaging is affected by a specific neurological condition in one patient. The patient, GK, suffered from simultagnosia – showing almost no ability to attend to more than one object, and being unable to count objects shown to him. Demeyere et al. tested GK using ensembles of two colours (represented across 4, 6 or 8 elements) on a set membership task (e.g., "was this colour a part of the set?"). Surprisingly, they found that GK was more likely to guess that a colour was part of the set if it was intermediate to the two colours presented. They also tested GK's ensemble perception of size in the same way, with the same results. However, this case study offers limited conclusions about rapid averaging of hue, as the task did not involve rapid presentation (though GK's condition presumably restricts his serial access to the items). More importantly, the stimuli did not appear to be controlled in any colour space, being selected only on the basis of being reliably discriminated by the patient. Although this avoids the possibility that GK was simply confusing the colours this does not allow for replication of the stimuli. Furthermore, although all colour differences were above discrimination threshold this does not mean they were equally-spaced in perceptual terms – some neighbouring colours may have been easier to discriminate and remember than others. Equal perceptual spacing is important to ensemble perception as it ensures that inhomogeneity in the stimulus space do not affect the results by enhancing the distinctiveness of (and hence memory for) certain elements. Nevertheless the study suggests that an average colour computation from separable objects may be possible and may be similar to ensemble perception of other domains.

The question of whether multiple colours are perceived or can be represented by a single colour has been addressed before, but only in the context of texture-based perception from

arrays of adjacent, very small elements. For example, when given a mosaic arrangement of 400 small patches of colour, participants asked to adjust a uniform patch to a "representative colour" for the mosaic are able to make fairly reliable settings, but are biased towards the most saturated (intensely coloured) elements of the mosaic (Kuriki, 2004). This experiment involved constant presentation of the mosaic during the adjustment, allowing the observer time to compare their choice of uniform colour to the mosaic during the trial. A similar method was later used to investigate the appearance of the global colour of a mosaic composed of two hues which varied only in saturation (same hue and equal brightness) (Sunaga & Yamashita, 2007). The results revealed that settings were biased towards the more saturated elements, but also that they tended to stay on a curved lines representing the unique hues (see Kuehni, 2014). Although relevant to broad questions about how multiple colours are integrated in perception, these studies do not address how an observer might represent multiple colours that are presented rapidly in their memory. Similarly, the mosaic stimuli have small elements of different colours immediately adjacent and resemble a variegated surface. They therefore also do not address the question of what the ensemble impression is of a spatially sparse arrangement of larger elements, such the colours might belong to different objects and hence could be subject to an individual representation as well as a putative summary representation.

More recently, colour has become the subject of some investigations into the role of visual summary statistics in cognition. Asked to judge whether the mean colour of an ensemble presented for up to 1,500ms was more "blue" or more "red", observers reacted faster and made fewer errors when the ensemble contained less variance in colour (de Gardelle & Summerfield, 2011). A similar study showed that inserting a priming ensemble, visible for just 100ms, could reduce reaction times to the target ensemble (Michael, de Gardelle, & Summerfield, 2014). The priming effect was found to be dependent on congruency of

the colour variance of the prime to the target ensemble, while congruency of the mean colour of the prime and target ensembles had no effect. That is to say, an ensemble of blue colours (with a bluish mean) could prime responses to an ensemble of red colours (with a reddish mean) provided the variation in the blue ensemble is equivalent to the variation in the red ensemble. The authors proposed that this kind of summary statistical priming reflects the predictive nature of visual coding and perception, and that rapid extraction of variance (and not the mean) may be important to neural gain control. This suggests that the visual system is adapted to a visual world where the variance in features (e.g., colour) may be relatively stable, but the mean is a less reliable feature.

Variance has also been implicated as a crucial factor in the perception of mean hue in another recent study. Webster, Kay and Webster (2014) gave observers ensembles of dots containing two colours and allowed them to adjust the ensemble until the average of that ensemble represented their own internal standard for a blue-green boundary, a red-blue boundary, a green-yellow boundary, or unique red (red that appears neither bluish nor yellowish). They found that observers were able to judge average colour and made reliable settings when the hues of the ensemble were relatively similar, but that this ability deteriorated quickly as the hues became more dissimilar from one another. While clearly demonstrating some features of mean colour perception, the task in this study is not reflective of ensemble perception tasks generally. They used an unlimited exposure time for observers to make their selections, and relied on internal, categorical targets to which the observer made their adjustment. In short, it leaves open the question of whether averaging of hue occurs when ensembles are only seen very briefly, and does not address questions about the representation of average hue in memory.

In general, the previous literature has provided some indications of the perceptual and cognitive reality of summary statistics of colour extracted from multi-coloured arrays. However, a comprehensive exploration of ensemble perception for colour has not been carried out. The methods used have not addressed issues pertinent to the theoretical mechanisms behind ensemble perception, such as whether the mean colour is encoded rapidly, the role of distributed and focused attention, and whether the process may be holistic or the result of a limited-capacity strategy such as sub-sampling of items. Questions about how sets of coloured items (e.g., fruit at the market) are represented remain unanswered. When the items are visually separable and could be represented by the visual system individually, but are only visible for a short time, is average colour encoded at the expense of the individual items, in the way that it appears to be for size and faces?

1.4.2 The role of summary statistics in colour perception

The mean and variance of colours in a scene have been shown to have significance to the visual system, in terms of adaptation state, colour appearance and colour constancy. This section will discuss these areas in turn.

Adaptation is a basic function of the sensory system. The term refers to a reduction in neural activity in response to a persistent stimulus (Kohn, 2007). In the domain of vision it is possible to observe the effects of adaptation through aftereffects. For example, staring at a green shape against a white background for about half a minute will result in the beholder observing an afterimage of the shape when looking at a uniform white area. In this example the afterimage would appear pinkish, and occurs primarily as a result of adaptation of the retinal ganglion cells coding the L-M and S-(L+M) cone-opponent

colour axes (Demb & Brainard, 2010; Zaidi, Ennis, Cao, & Lee, 2012). Colour adaptation serves to keep the visual system calibrated to the immediate environment by providing a norm, the white point, to which object colours are relative (e.g., M. A. Webster & Leonard, 2008).

Maintaining a white point requires integration of information from across the visual field and in this process two summary statistics, the mean and the variance, appear to be most important. Webster and Mollon (1995) describe light adaptation (i.e. the approximation of the colour of the light illuminating the scene) as being based on the mean luminance and chromaticity, and contrast adaptation (i.e. the ability to discriminate colours) as being based on the variance of colour in the scene.

The role for variance in colour appearance has been demonstrated in several contexts. It has been shown that discrimination of colour textures (i.e. where elements of a mosaic are so small that they are spatially indistinguishable) is best for uniform textures (where all elements are the same colour) but decreases with greater variance in the colour of the elements (Hansen, Giesel, & Gegenfurtner, 2008; see also te Pas & Koenderink, 2004; Zaidi, Spehar, & DeBonet, 1998). Brown and MacLeod (1997) investigated the appearance of coloured patches displayed against a background composed of a patchwork of colours. They found that the appearance of the focal patches was significantly shifted in saturation by changes in the variance of the background colours, even when the mean chromaticity of the patchwork background was constant. When the background contained more variance (i.e. the background colours were more saturated) the focal patches appeared less saturated, compared to when the background contained less variance (i.e. background colours less saturated). The results could not the be result of a difference in global adaptation state, since the mean chromaticity of the background is consistent

between the conditions, and a secondary experiment showed that local contrast was not the cause of the difference as the effect persisted even when the target patches were surrounded by a grey border. Brown and MacLoed suggested that the phenomenon reflects a tendency for the gamut of perceivable colours to be related to the range visible in the scene – when the variance is higher there is a contraction of the gamut occupied by the focal patches, leading to the patches appearing less saturated and more similar to one another. Subsequent experiments have shown that this effect is strongest when the chromaticity of the elements in the background patchwork is modulated along the same opponent colour axis of MacLeod-Boynton colour space (i.e. cherry/teal or lime/purple colour directions) as target patches, and when the mean hue of the background matches the mean hue of the target patch (Ratnasingam & Anderson, 2015).

A similar effect has also been observed for the segmentation of a target region embedded within a chromatic texture, whereby segmentation is improved if the target texture and background texture vary in orthogonal directions of colour space (Li & Lennie, 1997). These studies demonstrate the influence that colour variance in a scene may have on appearance and discrimination in natural contexts, where surfaces may be variegated.

The distribution of colours in a scene can vary widely even in natural scenes and is an important cue as to the type of environment, (e.g. arid or lush) (Juricevic & Webster, 2009; M. A. Webster, Mizokami, & Webster, 2007). Indeed, perceptually-uniform colour spaces (i.e. spaces which attempt to equate distances in different parts of colour space by discriminability) tend to be skewed along the daylight axis (McDermott & Webster, 2012) – reflecting the calibration of the visual system to the most abundant and reliable illumination changes in the visual world – the change in the colour of light from the morning to evening.

Along with enhanced discrimination (M. A. Webster, 2011), a related role of maintaining an accurate white point is for colour constancy. Colour constancy describes the phenomenon that the colour appearance of a reflecting surface will remain roughly stable despite changes in illumination. The spectrum of light which is reflected from a coloured object is a combination of the surface properties of the object in terms of which wavelengths of light it absorbs and which it reflects, and the spectrum of light with which it is illuminated. Colour constancy mechanisms allow the brain to disentangle the surface colour from the illuminant colour, such that the observer can perceive the surface colour of the object (which is constant, and thus a more useful representation), rather than the reflected colour (which changes with illumination).

The reliability of colour constancy hinges on the estimation of the illuminant, based on cues from the colours present in a scene. One theory for how the visual system achieves this is that the mechanism responsible assumes that the mean colour of any given scene should be grey, and therefore by computing the mean chromaticity of the scene it is possible to estimate the illuminant (e.g., Buchsbaum, 1980). Such a "grey world assumption" is based on the idea that each colour in the scene has an equal probability of being the illuminant colour, and so a simple average is the best estimate (Smithson, 2005).

Although the mean colour across a scene does appear to contribute to colour constancy, when two scenes are manipulated to have the same mean despite being under different illuminants, some colour constancy (measured by their setting of a test patch to grey) is still evident (Kraft & Brainard, 1999). Therefore other cues to the illuminant must also be used by the visual system, perhaps in combination with the grey world assumption of mean colour, to achieve colour constancy. Basing illuminant assumptions on the brightest patch in a scene (an assumption that the brightest element is white), or specular highlights

from shiny objects (specular highlights are points on glossy surfaces where the illuminating light is reflected veridically) may provide alternative or additional cues as to the colour of the illumination, which could support colour constancy (see Smithson, 2005, for a review). Being able to extract summary statistics with which to set the white point is functionally very important for maintaining the relative colour differences present in a scene regardless of the colour of the illumination.

It is evident that the colour statistics of the natural environment, in particular the mean and variance of the colour in a scene, have an impact on the function of the visual system, and hence on colour perception. Colour averaging appears to affect processes at various spatial and temporal levels, from short term adaptation and constancy over local areas, to long-term adjustments in peak sensitivity to the spectral composition due to changes in natural illumination during the day. Colour summary statistics may be represented and have influence at multiple levels of the visual system, from retinal ganglion cells (for short term, local adaptation) to visual cortex (for longer term changes in sensitivity and aspects of colour constancy), and beyond (see Shevell & Kingdom, 2008, for a review).

1.4.3 The role of summary statistics in cognition

While there is evidence for low-level effects of colour statistics on perception, the question of whether summary statistics of colour are computed and have an influence on post-perceptual cognitive processes, such as colour memory, has not been adequately addressed.

There is evidence for a central tendency bias in the categorization of singular colours. Wright (2011) asked participants to individually name a series of rendered colour stimuli

from green to blue, presented in a random order and repeated several times. A range of nine chips from was used from which two stimulus sets were formed – one excluded the two bluest chips from the set, while the other excluded the two greenest chips. Therefore both sets contained five blue-green chips in common, but one had a range extending further into the blue category, and the other a range extending into the green category. When naming responses were gathered, the two groups reliably differed in their placing of the green-blue category boundary, both being biased to set the boundary near the centre of the set of stimuli they were exposed to.

In a similar vein, Olkkonen, McCarthy & Allred (2014) used overlapping stimulus sets to investigate central tendency bias in colour memory. Using a delayed-estimation task, in which two colours are presented successively and separated by a variable inter-stimulus interval (ISI), they found that the point at which a set of green colours were equally likely to be responded to as yellower/bluer was shifted in the direction of the centre of the stimulus set in that run of trials. They further found that the central tendency bias became stronger with longer ISIs, and with the addition of chromatic noise (small local variation in hue and lightness around a mean chromaticity to create a mosaic patch) to the stimuli. They suggested that this was because the responses were based on a combination of the stimulus seen and priors. Priors are a summary representation of the mean and variance of recent sensory experiences, are updated with new experiences and can influence current perceptions through predictive coding (e.g., Pellicano & Burr, 2012). Olkkonen et al.'s (2014) result suggests that increased internal noise (due to longer ISIs) and/or external noise (due to perturbations in the pixel chromaticity of the stimulus) enhance the influence of the prior on judgments, leading to a central tendency bias (see also Olkkonen & Allred, 2014).

Therefore, we have some evidence for a mean bias in colour memory judgments. These judgments appear to be modified by priors which are based on information integrated over time. A similar process is evident in children's learning of categories for novel objects. When ten-month-old infants are shown a series of cartoon 'deer' with several variable features (antler length, leg length, ear size etc.), they find the deer with features representing the average of the exemplars they have seen as more familiar, even though this average deer (the prototypical 'deer') was not part of the initial set (Younger, 1985). This process develops during infancy from, simple feature averaging at 4 months to showing sensitivity to correlations between features around 10 months (Younger & Cohen, 1986). This so-called 'shift-to-prototype' is common in perceptual tasks in adults (see Huttenlocher, Hedges, & Vevea, 2000) and reflects the integration of sensory information over time to form a summary representation. Whether observers are able to extract summary representations from exemplars presented simultaneously is a question which this thesis attempts to answer.

1.5 Interim summary

Although some studies have explored the concept of whether observers can average multiple colours, the methods and approaches taken have not directly addressed the question of whether humans show ensemble perception for separable coloured objects. The available evidence suggests that the colour variance, as well as the mean colour might be important psychologically. Colour averaging does occur at a lower level in the visual system, and is thought to underlie processes of adaptation and constancy over various temporal and spatial scales. Colour variance is also important to colour appearance, and to the precision of representation of colours in memory. Colour averaging over time is

evident in the central tendency bias in a delayed estimation task and in colour naming. However, whether observers are capable of reporting the average of a set of simultaneously-presented and discrete, spatially separable colours has not been investigated.

1.6 Thesis overview

Sets of similar objects can be found in many different parts of everyday life, whether leaves on a tree, fruit in a marketplace, or football fans in a stadium. Although such sets could present a significant perceptual load we are able to use colour cues to quickly decide whether the tree is in summer or autumn, the bananas are ripe or not and whether the football fans are supporting the home or the away team. Summarising the set by average colour may be an efficient and useful way to make such judgments without needing to represent individual elements of the group. Similarly it may support extraction of the gist from a whole scene where colour is a diagnostic feature, and it may also be more efficient to represent a single summary colour, rather than retain many individual elements in memory.

Although the average colour appears to be computed at various levels of the visual system, including playing a role in adaptation and constancy, whether people have a conscious access to a sense of average colour is another matter. Humans have three classes of colour receptor in the retina, maximally sensitive to long (reddish), medium (greenish) and short (bluish) wavelengths of light. Signals from these cones are combined in the retina and lateral geniculate nucleus into two independent ‘opponent’ channels. One channel compares the output of the long and medium cones (L-M), forming one axis of colour space which goes from cherry-red to bluish-green (although usually referred to as the

"red-green" axis). The other channel compares the output of the short wavelength cones to the long and medium cones (S-(L+M)), forming an orthogonal axis of colour space from chartreuse to violet (although commonly referred to as the "blue-yellow" axis). The relative responses on these two axes are used to code the colour of surfaces, creating a circular hue space, with an achromatic centre. The luminance of the colour (i.e. how light it appears) is derived from the total output of the long and medium wavelength cones (L+M), creating a third orthogonal axis, and a three dimensional colour space. Colour appearance can be described by the attributes of hue (the 'ink' of the colour), saturation (the intensity of the colour) and brightness (the amount of light coming from the colour). Hue is related to the wavelength of light (a linear property), but is perceptually organised into a circular continuum, such that no colour represents the 'end point' of our sensitivity to hue (as compared to saturation, where grey represents one end point, and monochromatic light represents the other). Hue is unusual in the visual domain as the relationships between hues cannot be described in terms of magnitude but rather are experienced as differences in the qualitative experience. Hue is also categorized using verbal labels. Yet we retain a continuous percept of hue variation. These factors make the prospect of hue averaging an intriguing prospect, since the average hue of a set would be a qualitatively different percept from the elements averaged, and may even be named differently. Averaging hue might be different to perceptual averaging in other domains because of these features of hue space. Investigating how a set of objects of various hues are represented enhances our understanding of the general encoding of colour, and whether people have an intuitive sense of the relationships between hues, and circular nature of hue perception.

1.6.1 Research questions

This thesis seeks to characterise and explore the nature of ensemble perception of hue using experimental methods based on those used in the literature on ensemble perception for size and faces. The primary research question is whether adult observers show a pattern of responses on ensemble perception tasks which is consistent with those found for other domains. Do observers show familiarity for an unseen mean hue? Can observers select a mean hue and how is this affected by changes in the number of elements, number of different colours and the range of colours in the ensembles? In addition to these questions characterizing ensemble perception for hue, the role of colour categories is also investigated – do observers show a shift-to-prototype when ensembles contain only one category? Can they average across categories if there is more than one category? Following this, the mechanism behind colour averaging is investigated – are ensemble means represented as precisely as single colours? Does the precision of observer mean settings imply that averaging is holistic, or could the performance be explained by a limited sub-sampling strategy? Finally, two aspects of colour perception in autism are investigated, addressing specific hypotheses about a local processing bias in autism and the use and establishment of perceptual priors in autism. Firstly, do autistic adults show reduced averaging of colour? Secondly, do autistic adults show reduced adaptation to colour generally, and do they show reduced effects of top-down knowledge on adaptation to scenes in which colour may support the extraction of the gist?

1.6.2 Paper 1 – Getting the gist of multiple hues: metric and categorical effects on ensemble perception of hue

Paper 1 (Maule, Witzel, & Franklin, 2014) presents two experiments using the membership task of Ariely (2001) and Demeyere et al. (2008) to investigate observers' memory for the particular colours present in an ensemble of hues. The first experiment attempted to address the question of whether observers encode the mean hue from a multi-coloured ensemble. It also asked whether a categorical difference in the ensemble (e.g., "blue" and "green" colours) disrupts averaging and whether memory is biased towards the prototype of the colour category represented in an ensemble containing colours of the same category (e.g., all "green"/all "blue"). The second experiment investigated whether a greater perceptual (metric) distance between ensemble elements inhibits encoding of the mean colour.

Ensembles of eight elements and containing two colours were displayed for 500ms, followed by a single test colour. Observers responded according to whether they thought the colour was a part of the ensemble or not (yes/no) The stimuli were selected based on measurements of just-noticeable differences (JNDs) for hue (Witzel & Gegenfurtner, 2013), at a regular spacing of 1.5 JNDs, to ensure that the hue range represented consistent perceptual differences and were discriminable from one another. The membership task was chosen as it is an implicit indication of mean encoding – if the familiarity (i.e. proportion of 'yes' responses to test hues) of the unseen mean hue is higher than for an unseen hue the same distance from the real ensemble members this implies some generalisation. A between-participants manipulation was also included to examine the effect of colour categories on the task. Three groups had partially overlapping stimulus sets (e.g., Wright, 2011), one containing only blue colours, one containing only green

colours, and one straddling the category boundary, such that ensembles contained “blue” and “green” elements.

The results indicated that observers generalise membership to the mean hue, but not to the outer hues in each range. This was evident in both the familiarity responses and in reaction times, where correct responses to the mean were slowest. There was an unexpected effect of categories, whereby observers in the same-category conditions found the boundary hue more familiar than the more prototypical end hue. This suggested that ensembles were not encoded by their categorical content.

In a second experiment the perceptual distance between elements in the stimulus range was expanded. This time there was no significant bias towards familiarity of the mean, suggesting that the generalisation causing the mean bias in the first experiment was limited by perceptual distance.

This study provided the first controlled and appropriate investigation of ensemble perception of hue. The results are interpreted as providing support for ensemble perception of hue, possibly based on the mean hue, or some other generalisation mechanism. Categories do not appear to influence ensemble encoding, but may influence subsequent judgments about the test colour. Crucially, this study also demonstrated that the mean bias has a limit, with the effect disappearing when the constituent colours of the ensemble are sufficiently different from one another.

1.6.3 Paper 2 – Effects of ensemble complexity and perceptual similarity on rapid averaging of hue

Paper 2 (Maule & Franklin, 2015) aimed to further characterise ensemble perception of hue, by establishing whether observers were able to reliably pick out the mean of a multi-coloured ensemble. The experiments in paper two addressed the question of whether mean hue was extracted through serial processing or whether a holistic process could be at work. They also asked whether more varied ensembles, were harder to average and what factor – the number of colours or the absolute range of colour – was responsible.

The experiment used an expanded range of stimuli and an explicit averaging task. Stimuli were taken from a full hue circle, with hues separated by 2 JNDs (Witzel & Gegenfurtner, 2013) and observers were asked to pick the average hue from a 2AFC following the presentation of an ensemble for 500ms. Experiment 1 of paper 2 varied the number of colours (2, 4, 8) and elements (4, 8, 16) present in ensembles. These manipulations enabled the comparison of the results with those for previous studies on ensemble perception of size (Ariely, 2001; Marchant et al., 2013). The results indicated that the accuracy of average selection was unaffected by changes in the number of elements, suggesting that the process might occur in a holistic fashion across the whole display, rather than through serial encoding. Increasing the number of colours in the ensembles had a deleterious effect on hue averaging, but it was noticed that this manipulation conflated the range (variance) in colour present in ensembles – ensembles containing more colours also had greater range (see also Utochkin & Tiurina, 2014).

Experiment 2a of paper 2 sought to disentangle the effects of range and number of colours from one another. Ensembles were set up to have fixed perceptual distances (12, 20, 28 JNDs) between their most extreme elements. The number of colours was manipulated

independently of this. The results showed that when range was fixed the number of colours did not affect mean selection, whereas the range itself did – wider ranges were harder to average. Therefore it appears that it is not the number of colours which affects the ability to average, but the perceptual difference between them.

An additional post-hoc analysis (experiment 2b, paper 2) also investigated whether there could be any effects of colour categories on the mean selection performance. Colour naming data were collected and trials from experiments 1 and 2a were re-coded by the number of colour categories present in them (1, 2 or 3). The analysis revealed no effects of the number of categories on averaging.

This study showed that observers were able to make selections of an average hue following rapidly-presented multi-coloured ensembles at above-chance levels. The lack of an effect of number of elements suggested that averaging might take place in the absence of serial encoding of elements. Findings from chapter two (paper one) were also corroborated – the ability to average appears to be range-limited, and there was no effect of categories on averaging. The break-down of averaging of hue ensembles with larger ranges may be reflective of the difficulty in averaging in a circular stimulus space, where widely-spaced hues may have two possible averages.

1.6.4 Paper 3 – Accurate rapid averaging of multi-hue ensembles is due to a limited capacity sub-sampling mechanism

Paper 3 aimed to build on the results of paper 2 by exploring the mechanism of hue averaging in more detail. The paper addresses the question of whether the precision of average hue representation suggests a role for holistic averaging across the entire

ensemble (as suggested in chapter three) or whether it may be achieved through sub-sampling of relatively few items from each ensemble.

Using the method of adjustment, observers' estimations of the mean hue from a 500ms ensemble of sixteen elements of four colours (heterogeneous ensembles) were compared to their estimations of the exact colour for ensembles with sixteen elements of one colour (homogeneous ensembles). Stimuli were selected from a 1 JND (Witzel & Gegenfurtner, 2013) hue circle, to allow finer measurement of observer precision and to ensure ensembles (elements spaced by 2 JNDs) were well within the manageable range for averaging identified in chapter 3.

The results showed that settings were heaped at the expected mean but with error distributed evenly either side. This error was greater for heterogeneous ensembles than for homogeneous, indicating that the mean hue of a set is not represented as precisely as a single hue. An ideal observer simulation was run to simulate the expected performance that each real observer might reach if they sub-sampled from the ensemble. The simulation perturbed sampled hues by an amount equivalent to each observer's variance from the homogeneous condition before estimating the mean, for 10,000 trials per sub-sample size. The results revealed that most observers' averaging performance could be equalled or bettered by a subsample of just one or two items from each ensemble.

The findings from this study depart from those found for some similar simulations of face and size averaging. The required sample size of 1-2 items is within the limits of focused attention and visual working memory. While this study cannot show that subsampling is the strategy used by the observers, it does suggest that the settings are so imprecise as to make postulating a holistic hue averaging mechanism untenable in the context of this task. However, just because the process may not be holistic does not imply that there is no use

or tendency to attempt to extract summary statistics from colourful sets. Further studies investigating implicit sensitivity to the average hue are needed (e.g., Michael et al., 2014).

1.6.5 Paper 4 – Ensemble perception of colour in autistic adults

The fourth paper attempts to ascertain whether the extraction of summary statistics of hue is different in autism, using the tasks from papers one and two. The study also sought to establish whether the distinction between local and global processing would be relevant to ensemble perception.

Autism is associated with perceptual atypicalities such as hypo- and hyper-sensitivity to stimuli, and intense fascination with particular sensory experiences (see Pellicano, 2013). In addition, autistic people show an advantage over typical individuals in visual tasks requiring processing of local detail, sometimes at the expense of their ability to extract the global gist (e.g., Frith & Happe, 1994). One recent account of visual atypicalities in autism has proposed that autistic people have attenuated updating and application of priors to their present experiences (Pellicano & Burr, 2012), which might be reflected in a number of atypicalities such as reduced adaptation and reduced extraction of summary statistics. There is already some evidence that autistic children have a reduced bias to the mean facial identity in an ensemble membership task, compared to typical children (Rhodes, Neumann, et al., 2014). However, given known difficulties and atypicalities in facial coding in autism (e.g., Rhodes, Pellicano, Jeffery, & Burr, 2007), further evidence from a non-social domain is needed.

A sample of autistic adults completed an ensemble membership task and a mean identification task with ensembles containing four colours across four, eight or sixteen

elements. It was predicted that the local processing bias and reduced extraction of global gist in autism would boost their performance on the membership task, but lower it on the averaging task, relative to an age-, IQ-, and gender-matched typical group. The results partially confirmed this prediction. There was a slight advantage associated with autism on the membership task. Post-hoc analyses revealed that this was specifically due to an improved rejection of unseen colours, while the acceptance of seen colours was unaffected by autism. The averaging task revealed that a disadvantage in the extraction of average hue for autism was restricted to the four-element condition – for eight- and sixteen-element ensembles the autism and typical groups were no different in averaging performance.

These results are interpreted in terms of the distinction between local and global processing – supporting recent accounts that suggest the local bias in autism is not necessarily accompanied by a deficit in global processing (Happe & Frith, 2006). A local bias enables the autism group to better identify which colours they haven't seen in the membership task. The result from the averaging task may also reflect the tendency to local, rather than global processing in autism – it is possible that with smaller sets the default mode of processing in autism is local, leading to worse averaging, whereas in larger sets, and in all set sizes for typical adults, a global processing mode is initiated, leading to better mean extraction.

1.6.6 Paper 5 – Colour afterimages in autistic adults

The fifth and final paper does not investigate ensemble perception but investigates another aspect of the relationship between colour and the gist of a scene, and also relates to the perceptual atypicalities in autism. The experiment is an attempt to replicate a

reported effect of scene content on the intensity of colour afterimages (Lupyan, in press), and to investigate whether this effect, and adaptation generally, is attenuated in autistic adults. Using a version of the Spanish Castle Illusion (see Sadowski, 2006), Lupyan found that colour adaptation afterimages are more intense for images with diagnostic colours (e.g., a grey castle with green grass and a blue sky) than for images without diagnostic colours (e.g., a shelf of books, which could contain any colours). Furthermore, he found that turning the image upside-down obliterated this effect. Lupyan claimed that the top-down knowledge about scene and object colours was responsible for the modulating effect on the perceived afterimage strength, but that turning the image upside-down disrupted the influence of top-down knowledge.

The hypo-priors account predicts reduced adaptation aftereffects in autism (Pellicano & Burr, 2012), for which evidence has been found in faces (e.g., Ewing, Pellicano, & Rhodes, 2013) and numerosity (Turi et al., 2015) for children with autism. A prior can be considered a kind of perceptual summary statistic – albeit one which is extracted and computed over time – reflecting repeated experience of a stimulus (e.g., Olkkonen & Allred, 2014). Therefore, attenuated priors may manifest as a reduction of top-down effects on perception. It was predicted that adults with autism would exhibit weaker afterimages overall, and an attenuated influence of top-down scene knowledge on afterimage strength, compared to a typical group.

Afterimages were measured in both groups, however the data showed no overall difference in afterimage intensity between groups. Lupyan's effect of top-down knowledge was not replicated – neither group exhibited the interaction between orientation and scene content, which is crucial to demonstrating the effect. The adaptation effects found are interpreted as reflecting the low-level nature of colour afterimages –

originating in the retina (Zaidi et al., 2012). The effects of hypo-priors on adaptation in autism may be restricted, then, to stimuli which are adapted at a cortical level. There was no strong evidence in favour or against the possibility that attenuated formation of summary statistical priors, or top-down knowledge relevant to object and scene colours could impact the appearance of colour afterimages in autism.

1.7 Overall contribution

1.7.1 Contribution to research on ensemble perception

This thesis makes several novel contributions to the understanding of ensemble perception and has a number of important methodological and theoretical implications.

First, the thesis has shown the potential value and importance of using a stimulus set which is carefully and precisely controlled for discriminability (papers 1, 2, 3). Using a stimulus space defined in terms of discriminability, with supra-threshold steps between hues appearing together in an ensemble, rules out the possibility that individual ensemble exemplars were indistinguishable in the first place. This approach of equating ensemble elements in supra-threshold discriminability has provided a number of advantages. Firstly, it has been possible to approximate the perceptual distance at which hue averaging breaks down (paper 2; experiment 2a) – this was between the 12-JND range and the 28-JND range of hues. Secondly, it provided a precise and perceptually-scaled measure of the noise present in estimations of the ensemble mean and single hues to feed into the ideal observer simulation (paper 3). Finally, and most notably, the JND is a potentially useful metric for comparing ensemble perception of hue to other domains. Due to these multiple

advantages, it is recommended that future ensemble perception studies adopt this approach.

Second, paper 4 has demonstrated that the membership and averaging tasks may not elicit the same cognitive processes, even though they are based on the same initial stimulus (a colourful ensemble). It was found that autistic adults demonstrated an interaction between performance and number of ensemble elements in the averaging task, but not in the membership task. This likely reflects an influence of the demands of the task and a difference in local or global strategy in response to the ensembles. This suggests that performance on these two common ensemble perception tasks may not be as closely related as one would expect. Prior research has not considered the implications of task differences, however the impact of possible strategies for completing tasks should be considered carefully when comparing the results from different tasks, or where instructions differ.

Third, similarities in ensemble perception of colour, size and faces have been found. Like size and faces, colour averaging ability is invariant to changes in number of ensemble elements (paper 2, experiment 1; paper 4) and is limited by the range of stimuli (paper 2, experiment 2a). These similarities occur in spite of the apparent difference in the nature of the mechanism underlying the averaging process (paper 3). It may be that simultaneous averaging (or averaging-like performance) is an emergent property of many sensory systems. That is to say that rather than being a specific, evolved mechanism for coping with the demands made on the visual system, it may simply be that the noisy representation of individual stimuli tend to the average, over many trials or many exemplars. Thus, the load on working memory is not reduced by averaging per se, but rather is minimal anyway due to the imprecision of stimulus representation.

Fourth, although similarities have been found across domains, the research has also identified differences. The simulation data from paper 3 suggest that ensemble perception mechanisms are neither domain-general, nor served by a common feature-averaging mechanism. Unlike similar simulations for size (Im & Halberda, 2013) and faces (Haberman & Whitney, 2010), it was found that rapid averaging of hue may be served by focused attention paid to a very small sub-sample of elements. Therefore, perceptual averaging may be served by different mechanisms across domains, in the case of colour the average hue may not be automatically computed at all, but rather is an effortful process. The circularity of hue space is a possible explanation for the lack of holistic averaging of colour – as hues become more distant in colour space there are competing possible average hues (see paper 2, discussion). The nature of the stimulus domain may be crucial in considering the ability or tendency to average.

1.7.2 Contribution to research on colour cognition

This thesis also makes a number of novel, important and timely contributions to the understanding of colour cognition.

First, it has been shown that observers can form an idea of the average hue from a rapidly-presented ensemble (paper 2). Furthermore, their responses tend to heap around the expected mean hue, i.e. the point on the hue arc which is intermediate to all of the hues displayed (paper 3). Whether this reflects knowledge from experience of colour mixing and colour wheels, or intuition based on a sensory signal is an open question, however it is clear that observers do show awareness of how hues are related to each other in continuous terms. Therefore, this thesis has demonstrated the viability of using ensembles to investigate observers' intuitive knowledge of colour space and colour relations. From

the experiments presented in this thesis it appears that observers are able to represent hue relations in terms of the hue circle. However, gathering averaging data from ensembles manipulated along multiple directions and dimensions of a perceptual colour space may help answer questions about whether the perceptual dimensions of hue, saturation and brightness, are integrated into a contiguous colour model as is assumed in perceptual colour spaces, or whether individual dimensions might be represented separately (see Burns & Shepp, 1988).

Second, the thesis has shown that memory for multiple colours is poor (paper 1, paper 4). The representation of multi-hue sets appears to be based on generalisation across the range of hues which were seen, reducing discrimination within the range to near zero. This is not due to a limit in representing more than one colour, however, as when ensemble hues are more widely-spaced, discrimination of members from non-members improves (paper 1, experiment 2). Paper 2 (experiment 2a) has also demonstrated that range plays a limiting role in the representation of the average hue. Together the results suggest that the hue averaging mechanism may not be functional for hues which are very different. This result makes some ecological sense – visual sets which are similar in colour (e.g. leaves on a bush) do not need demand individual representations, whereas elements which are grouped, but different in colour (e.g., ripe fruits amongst foliage) are better attended to individually.

Third, the results show that multi-hue ensembles are not encoded by their colour categories (paper 1, experiment 1; paper 2, experiment 2b), but that colour categories might influence later judgments about ensemble membership (paper 1, experiment 1). This suggests that processes mediating categorisation may occur after the extraction of summary statistics from an ensemble (at least for colour), corroborating the prevailing

recent evidence that colour categorisation affects only post-perceptual processing (e.g., Bird, Berens, Horner, & Franklin, 2014; He, Witzel, Forder, Clifford, & Franklin, 2014), rather than being early and pre-attentive (e.g., Athanasopoulos, Dering, Wiggett, Kuipers, & Thierry, 2010; Siok et al., 2009; Thierry, Athanasopoulos, Wiggett, Dering, & Kuipers, 2009). Furthermore, it implies that the continuous, metric relationships between colours are preserved when represented in short-term memory.

Fourth, paper 5 adds some doubt to the possibility that top-down knowledge plays a role in colour afterimages (Lupyan, in press). Nevertheless this paper demonstrates a method to apply a perceptual colour space to measuring the strength of 'Spanish Castle' aftereffects (the original paper used a device-dependent colour space). The question of whether top-down knowledge can effect afterimages is certainly worthy of further research, particularly given the parallel evidence for a memory colour bias in achromatic settings of diagnostically-coloured objects (Hansen, Olkkonen, Walter, & Gegenfurtner, 2006) and brain imaging evidence for representation of colour even for achromatic versions of diagnostically-coloured objects (Bannert & Bartels, 2013).

1.7.3 Contribution to research on autism

Finally, papers 4 and 5 make contributions to the theoretical understanding of perception and cognition in autism.

First, there is no difference between the strength of colour adaptation in autistic adults and typical adults (paper 5). This is contrary to the predictions of the hypo-priors account of perception in autism (Pellicano & Burr, 2012). Given that reduced adaptation in children with autism has been found for faces (e.g., Ewing, Pellicano, et al., 2013) and

numerosity (Turi et al., 2015), this suggests that the results of paper 5 are either reflective of a developmental difference – whereby children 'grow out' of the hypo-priors effect on adaptation, or that the low level of colour adaptation spares that process from the influence of hypo-priors.

Second, the extraction of summary statistics of colour is atypical in autism (paper 4). The effect is rather mild, but is manifest as an advantage in detecting non-members of an ensemble, and a disadvantage in averaging small sets. However, these results may be accounted for by considering the local processing bias in autism and evidence for intact, but less readily used, global processing (e.g., Mottron, Burack, Iarocci, Belleville, & Enns, 2003). In the light of the demonstrated benefit for rejecting non-members and the relatively intact ability to extract the mean (for larger sets), previous results demonstrating reduced familiarity of the average face from the ensemble in autistic children, compared to typically developing children (Rhodes, Neumann, et al., 2014) may be equally interpreted as reflecting an advantage in rejecting an unseen exemplar, rather than a deficit in encoding the mean. A local processing bias and the weakened integration of information to form priors do not seem to prevent the extraction of summary statistics in autism, rather these mechanisms appear to be intact, but are not as readily called upon.

1.8 Future research

Several important research questions have emerged from this thesis which future research should seek to address.

First, how does ensemble perception vary across domains? At present the studies on ensemble perception from different visual domains are relatively disparate. There are also

discrepancies in the methods, stimulus specification, and simulation models which make it difficult to compare results across domains. Testing ensemble perception for multiple domains with the same methods, participants and a stimulus space which can be compared across domains (e.g., JNDs) will allow more direct comparison of the mechanisms and characteristics of ensemble perception for different domains. This will better explore the possibility of a domain-general ensemble perception mechanism, and may reveal individual differences in the general ability to extract summary statistics. In particular, a comparison of facial expression with colour should be made in order to establish whether the shift-to-boundary found in paper 1 (experiment 1) and the lack of categorical effects on averaging (paper 2, experiment 2b) can be replicated for categories of expression.

Second, is the average hue computed automatically from rapidly-seen sets? The experiments in this thesis have shown that observers *can* estimate an average hue from a briefly-presented ensemble, but cannot demonstrate that they *do* habitually, i.e. where there is not an experimental task probing for mean or member representations. Investigations using implicit measures of sensitivity to ensemble statistics of colour are needed. Previous studies using ensemble perception of size have successfully shown that the mean size of a background or priming array, can affect measures such as the speed of visual search (Lanzoni et al., 2014), target detection (Marchant & De Fockert, 2009), or search saccade latencies (Corbett & Melcher, 2014b). The prospect of priming by summary statistics of colour found by Michael et al., (2014) is a promising place to start, although that experiment used a categorical decision task and stimuli defined only in terms of their display equipment. For example, brief presentation of a prime ensemble followed by a simple target detection or visual search task, with perceptually-defined stimuli. If detection of the target is can be primed (in speed, accuracy or a signal detection measure) specifically by matching the mean hue of the prime ensemble with the target

hue, this would suggest that the mean hue of multiple objects is automatically computed, and plays an active role in our perception of colour.

Third, how does ensemble perception develop? If ensemble perception is evident in very young infants it may support proposals that it is a fundamental property of visual coding, whereas if a developmental trajectory for the ability to visually average can be detected it may reflect a process rooted more in cognitive development, depending on when it emerges. To date there has been only one investigation of ensemble perception with young children (Sweeny et al., 2014). This study found that 4-5-year-olds are sensitive to mean size. It is known that infants as young as 10 months tend to average features across successive presentations to form prototypes (Younger, 1985), while infants aged 4-7 months do not (Younger & Cohen, 1986), but it is not known whether they are able to do this across a simultaneously-presented set. If ensemble perception emerges at the same stage of development as successive feature averaging the two may share a mechanism, pointing to a role for gist extraction in prototype formation. Measuring infants' habituation (i.e. a decrease in looking time to repetitive stimuli) to a series of ensembles with either consistent or inconsistent mean hue (or size, face etc.) would be one possible experiment. If habituation occurs more quickly in the consistent-mean condition compared to the inconsistent-mean condition, this suggests the infants are sensitive to the ensemble statistics of the scene. By testing several age groups (e.g., 4, 7 and 10 months-old (following Younger & Cohen, 1986)) the developmental trajectory of mean encoding can be revealed.

Fourth, can the effects of hypo-priors affect low-level vision in autism? If there are general neurological differences which result in reduced adaptation and pooling across signals then there may be effects detected in low-level vision. On the other hand, if the

effects of hypo-priors are related to a local processing bias in autism, as papers 4 and 5 suggest, we would not expect to find effects on low-level visual phenomena. One way to extend the findings from paper 5 would be to investigate colour constancy in autism. Autistic individuals (children and adults) show reduced shape constancy, being better able to match the apparent two-dimensional shape of a slanted circle than typical individuals (Ropar & Mitchell, 2002). Autistic people may also be less likely to integrate information from the illumination context, and thus be more likely to select a reflectance match than a surface match in a colour constancy task. Furthermore, since colour constancy is based on estimating the illuminant, perhaps based on a combination of low- and high-level cues, different influences of prior experience (e.g., knowledge of canonically coloured items) and low-level stimulus differences on constancy may be detectable. Individual differences in the perception of "the dress" appear to be due to assumptions about the illuminant fixing ones perception either towards a blue/black percept or a white/gold percept (Brainard & Hurlbert, 2015; Gegenfurtner, Bloj, & Toscani, 2015; Lafer-Sousa, Hermann, & Conway, 2015; Winkler, Spillmann, Werner, & Webster, 2015). However, some people report experiencing bi-stable perception of this image, being able to 'switch' between one interpretation and another. Perhaps autistic individuals (or typical individuals with more autistic-like symptoms) are better able to switch their perception of the dress, due to the attenuated influence of priors on their perception.

Finally, do hypo-priors in autism affect children and adults in different ways? Paper 5 finds no reduction in colour adaptation afterimages in adults with autism. Previous demonstrations of reduced adaptation in autism have been restricted to children (e.g., Ewing, Pellicano, et al., 2013; Turi et al., 2015), and although reduced face adaptation has also been found in the parents of children with autism (Fiorentini, Gray, Rhodes, Jeffery, & Pellicano, 2012), the effect was not found in autistic adults (Cook, Brewer,

Shah, & Bird, 2014). Further studies should seek to comprehensively establish the differences in adaptation across domains and ages in autism. Similarly, paper 4 finds only mild effects in ensemble perception due to autism in a sample of autistic adults and the only previous study to address the question of perceptual averaging in autism was conducted with children (Rhodes, Neumann, et al., 2014). Establishing the developmental trajectory of ensemble representation in autism compared to the typical population will tell us more about the underlying differences in perception and cognition in autism, and also the mechanism behind ensemble perception generally. Whether autistic adults and children have different patterns of adaptation and perceptual averaging relative to typical members of their age group has implications for the understanding of the development of compensatory cognitive strategies in autism.

1.9 Conclusion

Through a series of experiments this thesis provides the first wide-reaching demonstration and characterisation of ensemble perception of hue. It has been shown that the gist of a multi-coloured set can be represented by the mean hue, with reasonable accuracy. Unlike ensemble perception of size and faces, the mechanism behind average hue judgments does not appear to be holistic and parallel. Instead it may be the result of serial processing of a small number of ensemble elements. A local processing bias is revealed in ensemble perception for autistic adults, but does not prevent the extraction of summary statistics. Finally, in autism attenuated establishment and updating of summary representations (i.e., priors) is not found for colour adaptation afterimages. Overall, the findings of this thesis suggest that the topic of ensemble perception of hue is fertile ground

for research into ensemble perception, colour cognition, and the role of priors in typical and atypical perception.

Chapter 2

Paper 1: Getting the gist of multiple hues: Metric and categorical effects on ensemble perception of hue

Maule, J., Witzel, C., & Franklin, A. (2014). *Journal of the Optical Society of America: A, 31*(4), A93-A102, doi:10.1364/Josaa.31.000a93.

2.1 Abstract

This study investigated the perception of colourful ensembles and the effect of categories and perceptual similarity on their representation. We briefly presented ensembles of two hues, and tested hue recognition with a range of seen and unseen hues. The average hue was familiar, even though it never appeared in the ensembles. Increasing the perceptual difference of ensemble hues inhibited this mean bias, and the categorical relationship of hues also affected the distribution of familiarity. The findings suggest that there is ensemble perception of hue, but this is affected by the categorical and metric relationships of the elements in the ensemble.

2.2 Introduction

2.2.1 Ensemble perception

One proposal for how the brain copes with the encoding of the visual world suggests that statistical regularities in the environment allows for information to be compressed, leading to more efficient coding of scenes (Alvarez, 2011; Haberman & Whitney, 2012). As such, the gist of the scene may be extracted via a statistical summary of the features of the objects in a scene – for example, a forest can be summarized by the mean height of the trees visible, rather than requiring serial processing and individual representation of each tree. For colour, a similar process would not need a precise representation of any single leaf on a tree, but instead that the hue of all the leaves is summarized as a mean, providing a cognitive representation of the gist of that scene. Therefore, without focused attention and in the absence of knowledge about any single item (Alvarez & Oliva, 2008, 2009; Marchant et al., 2013), the observer may be able to report with striking accuracy some of the mean characteristics of a scene.

A lively literature has built up around ensemble perception since Ariely (2001) demonstrated that observers had a reasonably accurate impression of the mean size of a group of circles, despite having poor knowledge of the sizes of the individual elements. His experiment involved the brief (500ms) presentation of a group of 16 circles of varying size, followed by a single test circle either from the set seen previously, a circle not from the set seen but within the range of sizes, or a circle outside of the range of sizes just shown. When participants were asked to identify whether the circle was a part of the set, results demonstrated that they were sensitive to the range of sizes, making errors where test circles were within the range of sizes first shown, but tending to correctly reject circles which were larger or smaller than the circles seen in the original ensemble.

Secondly, Ariely directly probed the knowledge of the mean size of the set by asking whether the test circle was larger or smaller than the mean of the set and found that participants had a reasonably precise idea of the mean size of the set.

It has since also been shown that accurate summary statistics of size can be drawn independently for different groups of stimuli presented together simultaneously but designated by colour or location (Chong & Treisman, 2003), and for displays of stimuli which increase and decrease in size over time (Albrecht & Scholl, 2010). The explanation offered for these effects is that the gist of the scene is encoded, rather than the details of individual items, and that it is the mean size which is stored as the representation of that ensemble. Although there is some debate over the possibility of subsampling strategies (see Ariely, 2008; Baijal et al., 2013; Marchant et al., 2013; Myczek & Simons, 2008; Simons & Myczek, 2008), it is generally accepted that this cannot account for the accuracy of mean perception.

Ensemble perception has also been demonstrated for more complex visual stimuli. Shown an ensemble of faces of differing expressions for a very short time (250ms), participants were able to successfully extract the mean expression, with the additional tendency to discount emotional deviants (expressions vastly different from the majority, e.g., one happy amongst eleven angry faces) from that mean (Haberman & Whitney, 2010). Accurate means are also extracted for crowds of faces of different identity, regardless of viewpoint (Leib, Fischer, Liu, Whitney, & Robertson, 2013), and from inverted faces viewed by prosopagnosiacs (Leib, Puri, et al., 2012).

2.2.2 Ensemble perception of colour

It is not clear how colour features are integrated so as to form a meaningful representation of a multi-coloured display. Only two studies have investigated the cognitive representation of multicoloured displays (Demeyere et al., 2008; Kuriki, 2004). One study asked participants to select a 'representative colour' for a mosaic of 400 small patches, presented continuously until a decision was reached (Kuriki, 2004). The choice of representative colour tended to be biased slightly towards the most saturated (intense) element colour of the mosaic. This study therefore failed to find convincing evidence for a mean bias in how participants represented the multicoloured ensembles. However, it remains possible that if ensemble hues were equally saturated then such a mean bias would be found (as in ensemble perception for size).

Demeyere et al. (2008) presented a case study of the perceptual abilities of GK – a patient living with simultagnosia. This condition makes GK unable to reliably count visual objects and he has a very limited capacity for attention to more than one thing. Nevertheless, when shown an array of objects consisting of two different colours or sizes, GK demonstrated familiarity to subsequent test items which were intermediate to the items in the original set. The findings of this study suggest that a mechanism for ensemble perception and encoding of size may survive, even when serial processing is impaired. Unfortunately, the conclusions regarding ensemble perception of colour are limited by a failure to adequately control the colours used. The colours were not controlled for differences in luminance or perceptual difference in any relevant colour space. Discriminability is crucial in ensemble perception experiments in order to rule out confusion due to similarity of test items to elements of an ensemble, rather than due to encoding as summary statistics.

Apart from these two studies, colour has been left out of the literature on ensemble perception, or has been used solely as a marker for perceptual grouping of other stimulus attributes (e.g., Chong & Treisman, 2003; Huang, Treisman, & Pashler, 2007). Therefore, at present the question of whether there is ensemble perception for colour has not been adequately addressed. However, colour is an excellent candidate for investigations into ensemble perception. Colour is applicable to almost any natural stimulus and is considered a fundamental element of visual perception (Shevell & Kingdom, 2008). Furthermore, the phenomenon of colour constancy, whereby the distribution of light reflected from objects in the environment is used to estimate the colour of the illuminant, and hence allows surfaces to be perceived as a constant colour under different lighting conditions (e.g., McDermott & Webster, 2012; M. A. Webster et al., 2007; M. A. Webster & Mollon, 1995, 1997) may be related to a colour-averaging process.

2.2.3 Categories and ensemble perception

Like facial expression, colour is a continuous percept that is categorized (e.g., Calder, Young, Perrett, Etcoff, & Rowland, 1996; Kay, Berlin, Maffi, Merrifield, & Cook, 2011). The effect of categories on ensemble perception has never before been investigated, and therefore further investigation in the domain of colour offers an excellent opportunity to establish whether the categorical relationship of stimuli in the ensemble affects the ensemble representation. Many have argued that colour categories and their prototypes affect the perceptual or cognitive representation of colour. For example, studies have argued for categorical perception of colour – where colours from different lexical categories are discriminated (e.g., Pilling, Wiggett, Ozgen, & Davies, 2003), identified in search (e.g., Roberson, Pak, & Hanley, 2008)) and remembered (e.g., Davidoff, Davies,

& Roberson, 1999) more easily than colours from within a category. Colour categories have been found to cause different-category targets to ‘pop-out’ from the search array, enabling faster detection than same-category search (Daoutis, Pilling, & Davies, 2006). Prototypical colours, i.e. the best examples of a category, may be remembered more accurately than boundary or non-focal exemplars of categories (Heider, 1972; but see Roberson, Davies, & Davidoff, 2000, for a study showing that this advantage is not present for unusual linguistic colour categorical distinctions). Additionally, it has been suggested that prototypical colours (those specific colours which are chosen as the ‘best example’ of a lexical category) show greater consistency in their subjective appearance across different illuminants (Olkkonen, Hansen, & Gegenfurtner, 2009; Olkkonen, Witzel, Hansen, & Gegenfurtner, 2010). Although the effect of colour categories and prototypes on the perceptual and cognitive representation of colour has been thoroughly examined, there has been no investigation of whether categories and prototypes affect the representation of multiple hues. Our understanding of colour categories and their effect on cognition may also be improved from investigating this issue. Since ensemble perception is now understood to operate under distributed attention (Alvarez & Oliva, 2008), in the absence of serial processing (Demeyere et al., 2008), and compulsorily (Oriet & Brand, 2013; Parkes et al., 2001), if linguistic colour categories have an effect on the encoding of a briefly presented colourful ensemble it will provide further evidence for the depth of the effect of colour categories on the perceptual or cognitive representation of colour (Athanasopoulos, 2011; A. Clifford et al., 2012; Daoutis et al., 2006; Gilbert, Regier, Kay, & Ivry, 2006; Kay et al., 2011; Pilling et al., 2003; Regier, Kay, & Cook, 2005; Roberson & Davidoff, 2000; Roberson et al., 2008; Thierry et al., 2009; Winawer et al., 2007; Witzel & Gegenfurtner, 2013).

2.2.4 Current study

The current study aims to establish whether there is ensemble perception for hue by following the protocol of Ariely (2001) and Demeyere et al. (2008). In two experiments we investigated observers' knowledge of ensemble membership for exemplars from within and outside of the range of hues in a briefly presented ensemble. Test colours from a range of hues encompassing the exemplars in the ensembles were used to probe for the familiarity of individual elements. As in Demeyere et al. (2008), participants were simply required to indicate whether they thought the test colour was a part of the ensemble they had just seen. In experiment 1, there was careful manipulation of the categorical relationship of the ensemble colours in order to investigate the influence of colour categories and prototypes on ensemble perception. The experiment also offers a crucial improvement to the control of the colour stimuli used in the Demeyere et al. paper through the equating of hue difference in just noticeable differences (JNDs). Equating colour differences using JNDs was chosen over alternative models of perceptual uniformity, such as CIE spaces, as these have been shown to be limited in the accuracy of their estimation of cognitive colour similarity judgments (e.g., Komarova & Jameson, 2013).

Discriminability is rarely discussed in the context of ensemble perception but may be a crucial factor in how ensembles are encoded. For example, ensembles where the differences between exemplars are too small to be reliably discriminated cannot provide evidence for ensemble perception, since the ensemble contains no variation to be summarized. Likewise very large differences between the exemplars may cause a break-down of the averaging process, due to the requirement for the visual system to average across a large perceptual gap. Manipulations of perceptual difference in ensembles have not been widely investigated in the ensemble perception literature (whereas perceptual

similarity is commonly manipulated in the literature on colour cognition). Therefore, in experiment 2, the perceptual difference of the colours in the ensemble is increased to establish whether metric manipulations affect the degree of ensemble perception and mean bias via comparison to the data from experiment 1.

2.3 Experiment 1: Ensemble perception of hue and influence of colour categories

The first experiment aims to establish whether there is ensemble perception of hue, investigating how briefly-presented ensembles of two different hues are represented, and aims to establish whether colour categories and prototypes affect this representation. Given the effects of categories described above, it is possible that there would be some impact of categorical manipulations in ensemble perception.

If ensemble perception occurs and the ensembles are encoded by their mean hue it is expected that participants would find the mean hue the most familiar (i.e. falsely believe it was a part of the ensemble). If, for same-category ensembles, the colour category label which could be applied to the stimuli (i.e. “green” or “blue”) influences the retrieval of the ensembles, we expect to see a shift-to-prototype (Huttenlocher et al., 2000) manifest as a tendency towards mistaking colours closer to the category prototype as being part of the ensemble. Where ensembles contain colours of two categories it is possible that a bimodal distribution – corresponding to shift-to-prototype for both categories at once – could occur. This would imply that the multi-category ensemble is encoded as having two categories present, rather than a single mean. Alternatively, another possibility is that the presence of a categorical boundary would disrupt the averaging process, leading to consideration of the hues by their metric relationship (see Komarova & Jameson, 2013 for models including both categorical and metric differences in colour similarity

judgments). Therefore, we may see a greater tendency towards the mean hue for different-category ensembles, as the categorical relationship between the colours in the ensemble means that colour category is no longer diagnostic of ensemble membership. Figure 2.1 shows some predicted patterns of results given these hypotheses.

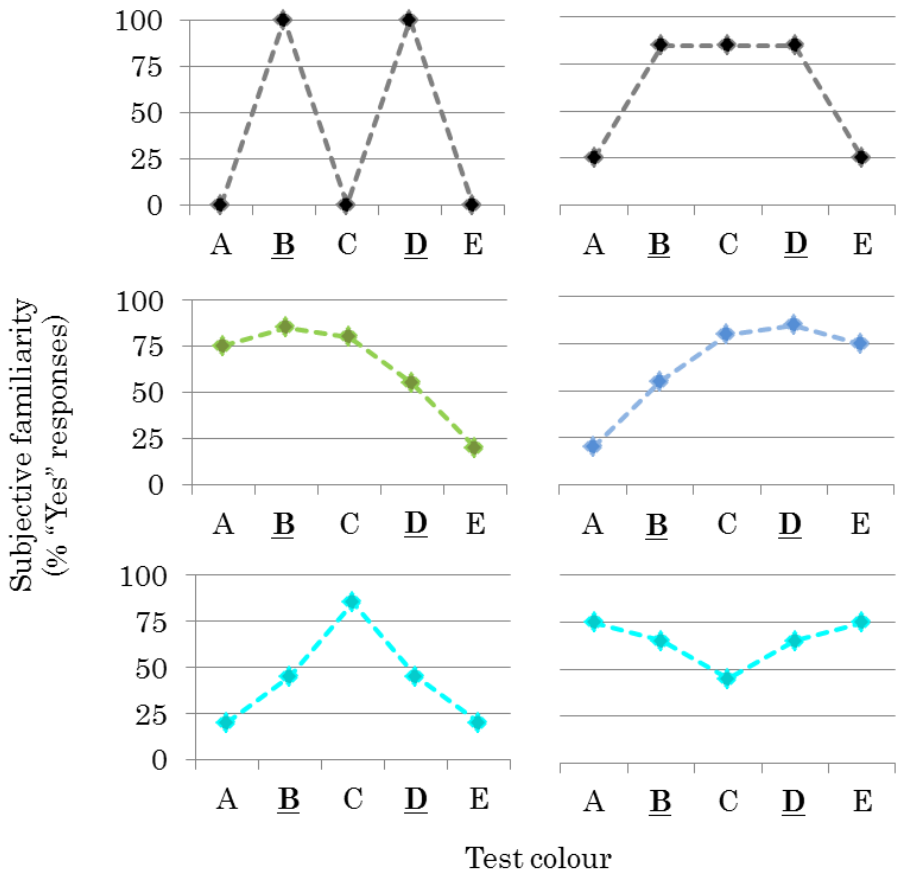

Figure 2.1. Possible patterns of results for the ensemble perception task. These graphs represent broad patterns of results for an ensemble where hues B and D (underlined) are present and hues A-E are tested with the question "Is this colour part of the ensemble?" **Top panels:** The left panel represents a perfect observer with 100% accurate memory for which hues are present and which are not in the ensemble. The right panel shows a general sensitivity to the range of hues present, but a failure to

reject the middle hue (as seen for size in Ariely, 2001). **Middle panels:** Both represent shift-to-prototype in the green (left) and blue (right) category conditions, where in the green condition, test hue A is the closest to prototype and in the blue E is closest to prototype. Familiarity responses are biased towards the more prototypical colour and away from hues closer to the boundary. **Bottom panels:** These represent results which may occur in the different-category condition. The left panel shows results if the mean is encoded, resulting in a tendency to feel the mean hue is familiar even though it was not in the ensemble. The right panel shows simultaneous shift-to-prototype towards the prototypical hue of both categories in the ensemble.

2.3.1 Methods

2.3.1.1 Participants

Twelve native English speakers (six females, mean age 22.74 years, SD = 3.0 years) took part in a preliminary colour naming experiment. Thirty-nine native English speakers (22 females, mean age 24.25 years, SD = 4.7 years) took part in the ensemble perception experiment. All were tested for colour vision deficiencies using Ishihara plates (Ishihara, 1973) and The City University Colour Vision Test (Fletcher, 1980).

2.3.1.2 Apparatus

The task was carried out on a Dell PC running Windows XP with stimuli displayed on a 22” Mitsubishi DiamondPlus 230SB Diamondtron CRT monitor, with a resolution

of 1600x1200, 24-bit colour resolution and a refresh rate of 85hz. The experiment was programmed in MATLAB (The MathWorks Inc., 2007) using the Psychtoolbox 3 extension (Brainard, 1997; Kleiner, Brainard, & Pelli, 2007; Pelli, 1997). A viewing tunnel lined with black felt was used to eliminate any effects of peripheral objects or colours and the experiment took place in a blacked-out room, with the monitor the only source of light. Viewing distance was not constrained, but participants were asked to view from the opening of the tunnel (approx. 57cm).

2.3.1.3 Stimuli and design

Stimuli were selected from the blue-green region of colour space. Measurements of hue discrimination collected by Witzel and Gegenfurtner (2013) were used to specify the colour range and ensure equal separation of colours in just-noticeable differences (JNDs). Following Witzel and Gegenfurtner, colours approximately equal in saturation (i.e. at a constant distance from the white point) were sampled from an isoluminant hue circle in Derrington-Krauskopf-Lennie (DKL) colour space, (Derrington, Krauskopf, & Lennie, 1984; Krauskopf, Williams, & Heeley, 1982) and the background was set to grey with CIE xyY (1931) chromaticity x=0.310, y=0.337, Y = 30cd/m^2. First, a set of nine colours equiluminant to the background were selected based on measurements of hue discrimination from Witzel and Gegenfurtner, with adjacent hues separated by 1.5 JNDs. The middle colour in the stimulus range corresponded closely to the anticipated blue-green category boundary, and the end colours fell very close to the prototypical “green” and “blue” at that lightness and saturation (Witzel & Gegenfurtner, 2013). To attempt to disrupt any long-term averaging which might occur in ensemble perception tasks (so called ‘regression to the mean’ (Bauer, 2009a; Crawford, Huttenlocher, & Engebretson,

2000)), the nine hues were also produced at four other lightness levels (Y = 20, 25, 35 and 40 cd/m^2) to provide darker and lighter versions of the JND-controlled hues. It should be noted that the stimuli are "isoluminant" for the standard observer, based on having an equal CIE 1931 Y value. Table 2.1 provides the CIE (1931) xy chromaticities for all colours used in the experiment. A ColourCal colourimeter (Cambridge Research Systems) was used to identify the correct RGB values by systematic adjustment of monitor primaries. Rendered colours were all very close to their desired values, with no monitor gamut issues.

Table 2.1. CIE (1931) xy chromaticity values for stimuli used in experiments

Prototype/boundary	x	y
(Background)	0.310	0.337
Green	0.256	0.464
-	0.241	0.430
-	0.231	0.400
-	0.227	0.374
Boundary	0.225	0.350
-	0.225	0.330
-	0.229	0.308
-	0.235	0.288
Blue	0.244	0.272

Table notes: These chromaticities were used at five lightness levels during the experiments, in which luminance (Y) varied from 20-40cd/m^2 in 5cd/m^2 intervals. The background was always 30cd/m^2.

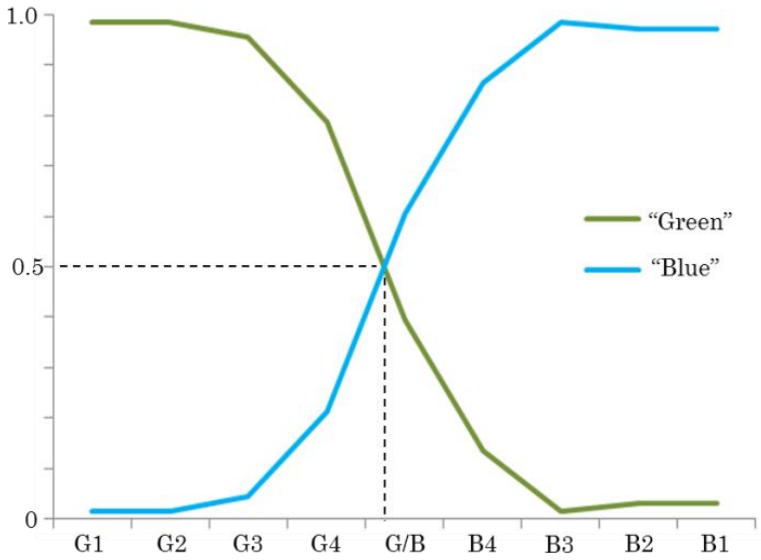

Figure 2.2. Results of the 2° colour naming experiment. These curves show the proportion of responses giving each hue the name either "blue" or "green" in a 2AFC task. The 50:50 point indicated by the dotted lines falls very close to the hue identified as the boundary by Witzel & Gegenfurtner (2013).

Three stimulus templates, each with five hues labeled A-E were formed. Based on the isoluminant hues, templates for the 'same category' conditions ('green' and 'blue') each had the boundary hue at one end and the prototype of the category at the other, such that the same lexical label would apply to all five hues in the set (except for the boundary hue). A third template was formed for the 'different-category' condition and centered on the boundary hue and included the two adjacent hues either side of the boundary. Each template was applied to five different lightness levels, creating 15 same-lightness stimulus sets, derived from the 45-hue matrix in figure 2.3.

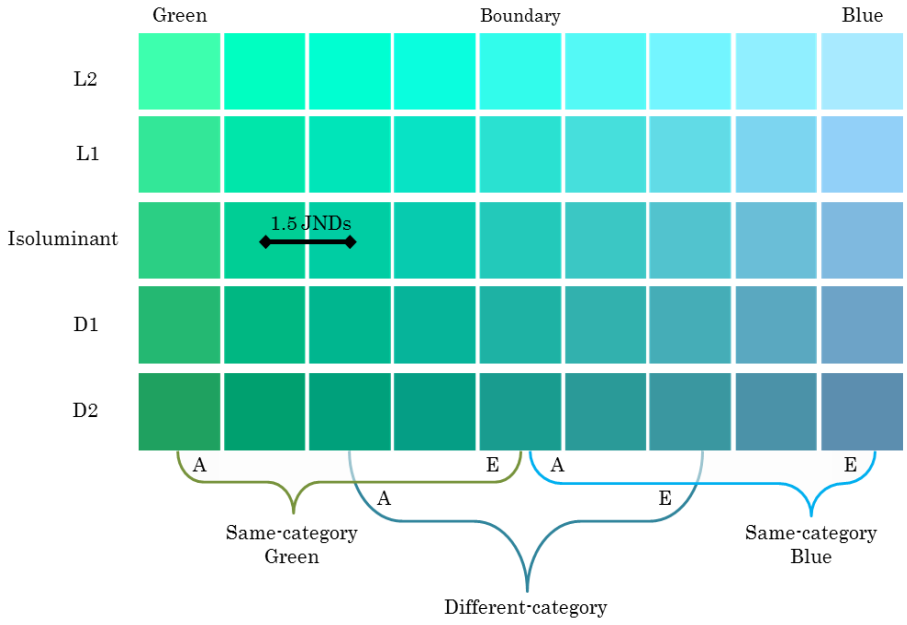

Figure 2.3. Range of colours used in the study. Boundary and best example labels apply only to the row labeled 'isoluminant' (i.e. of the same luminance as the background – 30cd/m^2). Horizontally adjacent hues on the isoluminant row are each separated by 1.5 JNDs (Witzel & Gegenfurtner, 2013). Sets of colours for each category condition are indicated by the brackets. In each set A and E indicate the ends of the range used, with the other three hues in between being B, C and D (see text).

2.3.1.4 Procedure

Participants were sequentially allocated to a colour condition – 'same-category green', 'different-category', or 'same-category blue'. Participants were blind to this condition allocation. Instructions were presented on the same grey background as used in the rest of the experiment – the time spent reading ensured adaptation to the background prior to the beginning of trials. Prior to the main blocks the participant completed ten practice trials (using colours selected to make the trials easy, but involving no blue or

green) which provided feedback to reinforce the task and the meaning of the response keys.

The structure of a trial is summarized in figure 2.4. A white fixation point appeared on the background for 1000ms, followed by the appearance of the ensemble for 500ms. Ensembles consisted of eight coloured patches. These eight patches were randomly allocated to a cell in an invisible 4x4 grid positioned in the center of the screen, and randomly jittered within that cell to remove the appearance of uniform arrangement. The underlying grid subtended approximately 4 degrees visual angle (at 57cm), and patches were squares each subtending approximately 0.5 degrees. Patches never came within 0.25 degrees of one another in any direction.

In experimental trials, ensembles had four patches of colour B and four of colour D taken from the stimulus set for that condition and lightness level. Test colours could be any from A-E (see figure 2.3 and caption). This method probes both for knowledge of the colours present in the ensemble, as well as detecting false alarms for colours both within and outside the range of hues shown. Control trials were also included where only one colour was presented in the ensemble. All the patches were of the same hue – either B or D, while test colours could be any from A-E. To limit the number of control trials and the impact of the single-colour trials on experimental trials, control trials always used the darkest stimulus sets (20 cd/m^2). These control trials provide a baseline against which to compare the performance in the experimental trials.

After a 1000ms ISI with a black fixation cross, a test patch appeared. Test patches were the same size as patches in the ensemble and were always presented in the center of the screen. At this point participants were required to respond as to whether they thought the colour of the test patch matched one of the two colours in the ensemble just seen or not.

Participants were instructed to press the ‘z’ key for “yes” or the ‘m’ key for “no” and to respond as quickly as they could.

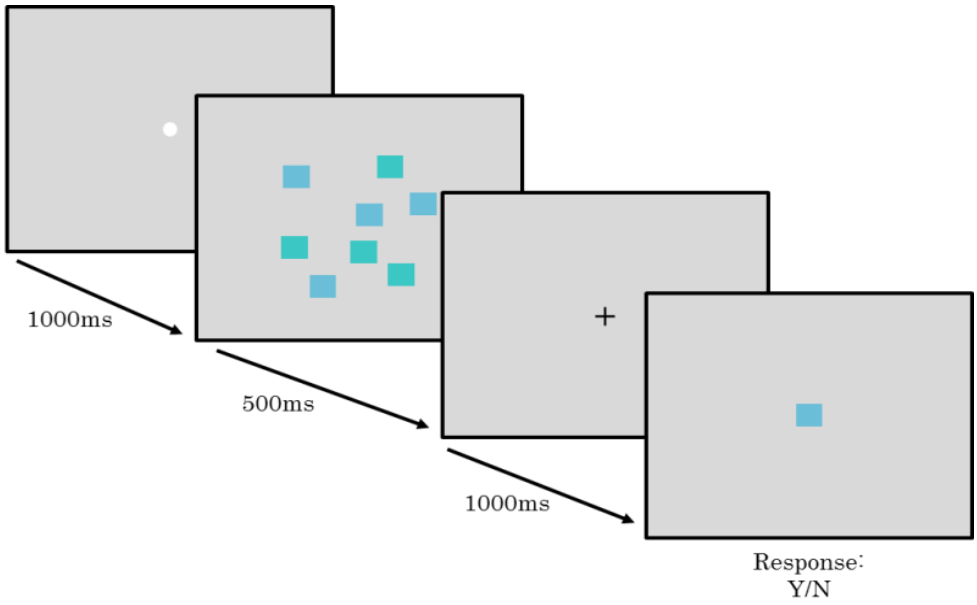

Figure 2.4. Structure of a single trial in the ensemble perception task. This example shows an ensemble in the different-category condition (containing blue and green elements), followed by a test colour which is not part of the ensemble. Participants were required to respond on each trial whether the test colour presented was part of the ensemble just seen or not (yes/no). Instructions re-appeared on screen at the start of each block.

In the main experiment each participant took part in six blocks of 55 trials (330 total), with each block containing each of the 25 (5 lightness levels x 5 test hues) experimental trials and 30 (3 repetitions x 2 ensemble types (B or D) x 5 test hues) control trials.

2.3.2 Results

Prior to analysis, trials with a reaction time (RT) of less than 200ms were removed, along with trials where the RT was greater than 3 standard deviations above the mean RT for each participant. This resulted in the removal of 115 trials (out of 5850, ~ 2%) from the experimental trials and the removal of 142 control trials (out of 7020, ~2%). This RT screening ensured that the trials submitted to analysis were valid responses.

Analysis of the ensemble perception data was carried out in two main ways. Firstly, the possibility of mean extraction for multi-coloured ensembles was examined by comparing responses to the test colour which is the unseen mean of the ensemble (C), to the test colours actually present in the ensemble (B and D) and to the unseen, non-mean test colours (A and E). Both the familiarity (i.e. percentage of trials answered "yes") and RT data for correct trials are discussed. Secondly, the effect of categories was examined by comparing the distribution of familiarity to the test colours across conditions.

A mixed 5 (Test Colour: A-E) x 5 (Lightness Level: 1-5) ANOVA on data from the experimental trials showed a significant main effect of Test Colour on familiarity (F (4, 760) = 212.90, $p < .001$), but no interaction between the pattern of responses and the Lightness Level of the ensemble stimuli (F (16, 760) = .88, $p = .59$). Further analysis, therefore, will be averaged across Lightness Levels. Another mixed 5 (Test Colour: A-E) x 6 (Block) ANOVA on these data showed no interaction between the pattern of responses and the block (F (20, 912) = .91 , $p = .58$), indicating that performance was consistent through the experiment, with no evidence for 'regression to the mean' occurring over the course of the experiment.

Figure 2.5 shows the mean familiarity of each test colour where ensembles included colours B and D (the exact hues in the ensemble depends on the condition). A mixed 5

(Test Colour: A-E) x 3 (Category: same-category blue, same-category green, different-category) ANOVA showed a significant main effect of Test Colour on familiarity (F (4, 144) = 34.94, $p < .001$), and also a significant interaction between the pattern of responses and Category (F (8, 144) = 10.35, $p < .001$).

If ensembles are encoded by their mean hue then it is expected that test colour C (the mean of the set) would be falsely recognized more readily than the other unseen test colours A and E. Following up on the main effect of test colour on familiarity, a paired-samples t-test found familiarity for C (M_C = 81.3%, SE = 2.9%) to be significantly higher than for A and E combined (M_{AE} = 59.0%, SE = 3.1%) (t (38 = 6.73, $p < .001$). Furthermore, if mean encoding occurs, then the mean hue should be as familiar as the hues which were truly present in the ensemble (B and D). A paired t-test found that the familiarity of the mean hue (C) was not significantly different from familiarity for the two seen hues, B and D (M_{BD} = 80.9%, SE = 1.9%) (t (38) = .18, p = .86). These results highlight that the unseen mean hue is as familiar as the seen colours and more so than the other unseen colours.

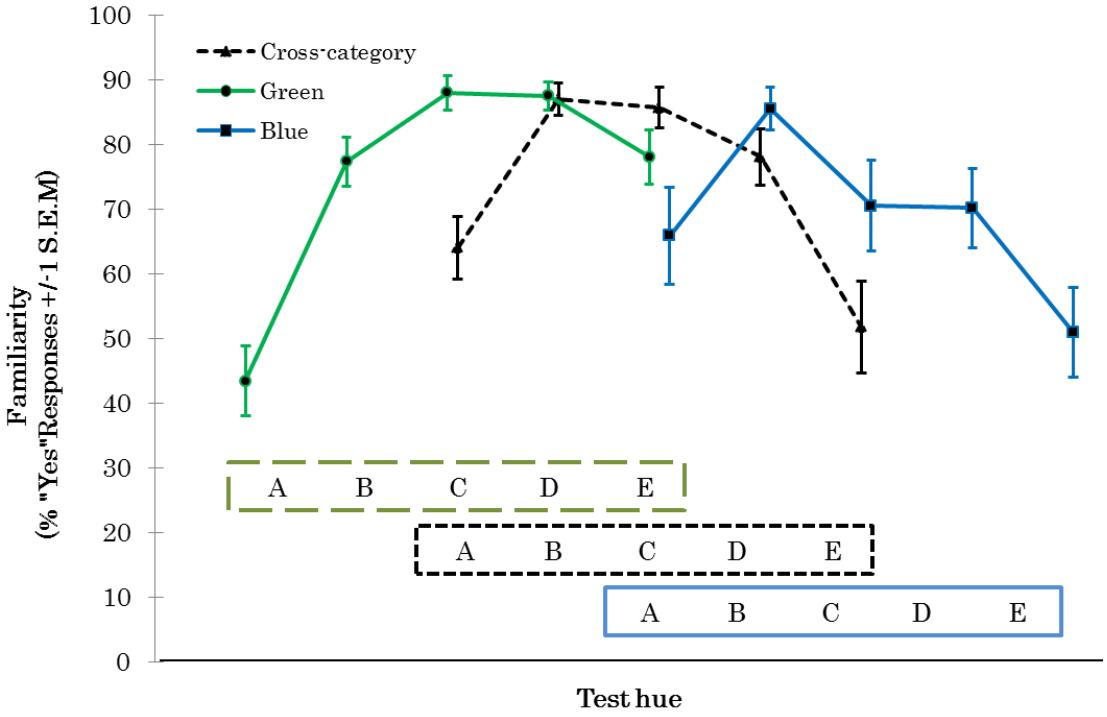

Figure 2.5. Familiarity data from experiment 1. These curves show the proportion of "yes" responses to each test hue following presentation of two-colour ensembles. Ensembles always contained hues B and D, therefore a "yes" response is correct for these hues, but is incorrect for A, C and E, which do not appear in the ensembles.

The data from the single-colour control trials carried out on the darkest colour set confirmed the mean effect – the difference in familiarity for test hue C in the experimental trials (darkest colour set) (79.5%) compared to familiarity in the control trials (60.8%) (M_{diff} = 18.6% (SE = 3.3%)) was significantly greater than the difference in familiarity for either of the other unseen test hues (A or E) (smallest t = 3.35, largest p = .002). This indicates that the presence of a multi-coloured ensemble depletes accuracy for the mean hue to a greater extent than for the other, non-mean hues.

Related to the main effect of Test Colour on accuracy, RTs on the correct trials also reflected the greater difficulty in correctly rejecting the mean hue (figure 2.6). A mixed 5 (Test Colour: A-E) x 3 (Category: same-category blue, same-category green, different-category) ANOVA showed a significant main effect of test colour on correct RTs (F (4, 128) = 17.42, $p < .001$), and no interaction between Test Colour and Category (F (8, 128) = .842, $p = .57$). Follow-up paired-samples t-tests indicate that the mean RT for correct responses to the mean hue C (M = 1.18s, SE = .09s) were significantly slower than correct responses to any other test hue (smallest $t = 2.83$, largest $p = .01$). For incorrect RTs a similar (although inverse) pattern is observed (see figure 2.6) – a mixed 5 (Test Colour: A-E) x 3 (Category: same-category blue, same-category green, different-category) ANOVA showed a significant main effect of test colour on incorrect RTs (F (4, 128) = 16.66, $p <.001$), and no interaction between Test Colour and Category (F (8, 128) = 0.78, $p = .614$). Follow-up paired-samples t-tests indicate that the mean RTs for incorrect responses to the mean hue C (M = 0.72s, SE = .03s) were significantly faster than incorrect responses for test hues A, B, or D (smallest $t = 2.43$, largest $p = .02$), but for hue E (M = 0.76, SE = .04s) the difference was marginal (t (38) = 1.97, $p = .06$).

It is important to note that the RTs for A, C and E are reflecting different processes for correct (correct rejections – "no" responses) and incorrect responses (false alarms – "yes" responses) from those for B and D correct (hits – "yes" responses) and incorrect responses (misses – "no" responses). Crucially, of the three unseen hues A, C and E, it was the ensemble mean hue (C) which had the slowest RT for correct rejections, and the fastest RT for false alarms, implying more doubt over the familiarity of this hue and a tendency to incorrectly recognise this hue more readily than the other unseen test hues.

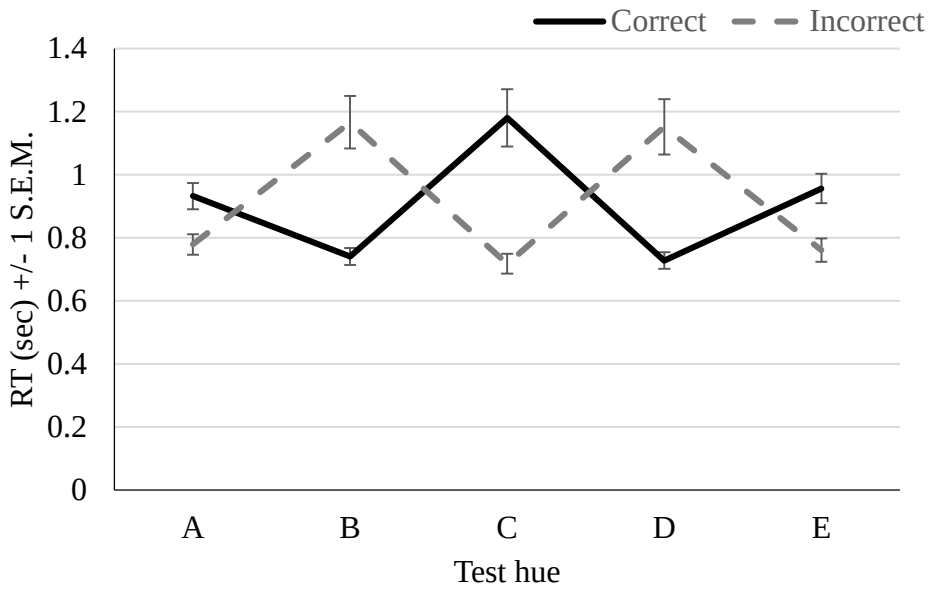

Figure 2.6. Mean RT for correct and incorrect responses to each test hue.

Returning to the accuracy data, and considering the interaction of Category with Test Colour for familiarity, it appears from figure 2.5 that there is asymmetry in the distribution of familiarity, particularly in the two same-category conditions. To assess this, the familiarity of the test colours A and E was compared across conditions (see shape of curves for different conditions in figure 2.5). A 2 (Test Colour: A, E) x 3 (Category: same-category blue, same-category green, different-category) repeated-measures ANOVA showed no significant main effect of the different test colours (F (1, 36) = .57, p = .46), but there was a significant interaction between test colour and condition (F (2, 36) = 23.7, p < .001). Paired t-tests found a significant difference in the familiarity of A and E in both same-category conditions, but not in the different-category condition (Green: t (12) = 7.22, p < .001 | Blue: t (12) = 2.73, p = .02 | Different-Category: t (12) = 1.81, p = .10) (figure 2.7). This confirms the presence of an asymmetry in the distribution of familiarity in the same-category conditions which is directed towards greater familiarity

of a test colour at the category boundary (i.e. E in the green condition, A in the blue), compared to the prototype of the category (A for green, E for blue).

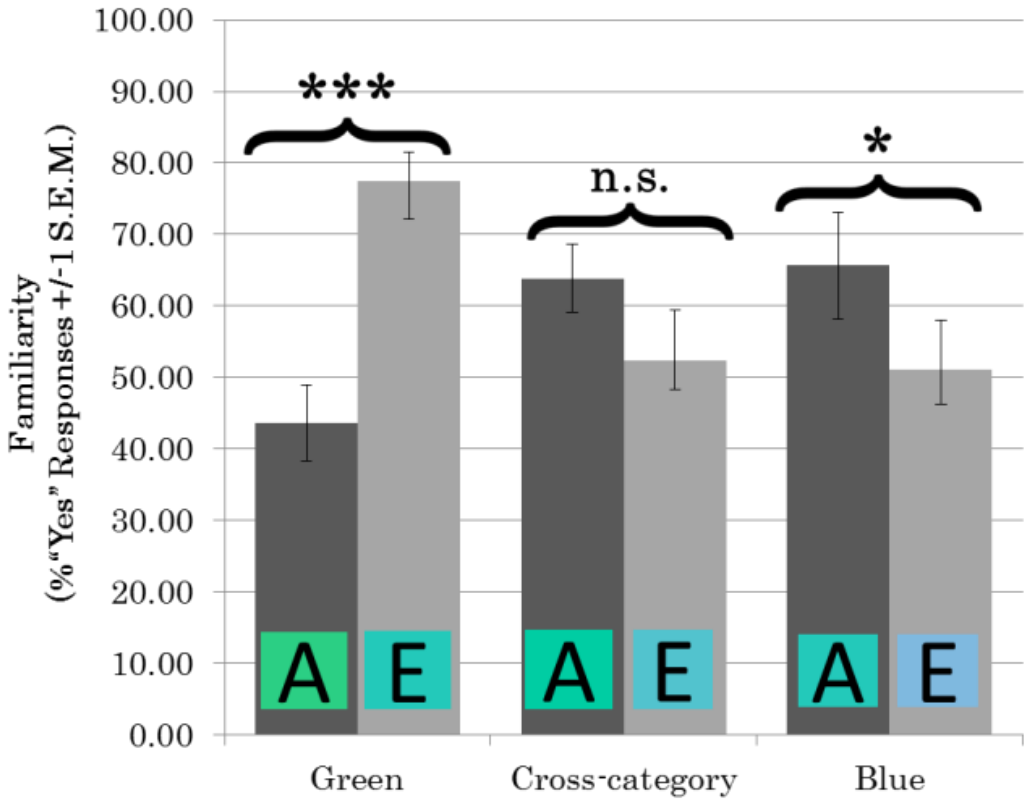

Figure 2.7. Mean familiarity for the end-point test hues for each condition showing the unequal response to the prototype compared to the boundary. *** denotes $p < .001$; * denotes $p < .05$; n.s. – not significant.

2.3.3 Discussion

The results of experiment 1 suggest that the mean hue of an ensemble is more likely to be incorrectly 'recognized' as part of the ensemble, than a hue which falls outside of the range of hues in the ensemble. Following the brief presentation of an ensemble containing two colours, the unseen mean of these hues is as familiar as the colours which were present, while other unseen colours which are equally discriminable from the colours in the ensemble, but are not the mean, are far less familiar. This result is very similar to that observed in Ariely's experiment on size (Ariely, 2001), and therefore it seems that ensemble perception of hue may be similar to that for size. Single-colour control trials confirmed that the effect was not a mere failure of colour discrimination due to the brief viewing time as the mean hue was significantly more familiar in the two-colour ensembles than in control trials, when just a single-colour ensemble is shown.

Analysis of the manipulation of colour category found an interaction of category with the familiarity of the unseen test colours. For same-category ensembles, the boundary test hue was found to be more familiar than the prototypical test hues. The familiarity of colours in the different-category ensemble was more symmetrical, with no difference between the two end colours. Contrary to the expectation of a shift-to-prototype (e.g. due to the covert categorical labeling of the colours in the ensembles), both conditions show a higher tendency for correct rejection of the most prototypical test colour, and greater false recognition of the boundary hue. We return to further discuss this categorical effect on ensemble perception in the General Discussion.

2.4 Experiment 2: The effect of perceptual difference on ensemble perception

Having identified an ensemble effect in experiment 1, experiment 2 sought to better understand the effect of colour similarity on the encoding of ensembles using ensembles with a larger perceptual distance separating the element colours. This was achieved by using a set straddling the blue-green category boundary, but with the perceptual distance twice that used in experiment 1. This allows a direct comparison with the different-category condition in experiment 1, and will indicate whether perceptual similarity mediates the ensemble effect.

2.4.1 Methods

2.4.1.1 Participants

Participants were thirteen native English speakers, none of whom had taken part in the first experiment. All were tested for colour vision deficiencies using Ishihara plates (Ishihara, 1973) and The City University Colour Vision Test (Fletcher, 1980).

2.4.1.2 Apparatus

The apparatus was as used in the first experiment.

2.4.1.3 Stimuli and Design

The stimuli were selected by taking every other hue from the rows of stimulus colours (figure 2.3), such that the closest any pair of hues could be was 3 JNDs. The range

was centered on the boundary hue and extended to the prototype in both directions. Therefore, there was only one condition – all participants saw experimental trials which displayed different-category ensembles (i.e. 4 “blue” patches and 4 “green”).

2.4.1.4 Procedure

The procedure was identical to that in the first experiment.

2.4.2 Results

Trials were screened for very short or very long RTs as in experiment 1, with a similar rate of removal of trials.

A repeated measures ANOVA revealed that, as expected, there was a significant effect of Test Colour on familiarity (F (4, 48) = 23.78, $p < .001$, see 3 JND condition of figure 2.8). Unlike in experiment 1, the familiarity of the three unseen colours (A, C and E) were broadly similar. Follow-up t-tests showed that the mean hue was not recognized any more frequently than the other unseen hues (t (12) = 1.60, $p = .14$), and was less familiar than the seen hues (t (12) = 4.62, $p = .001$).

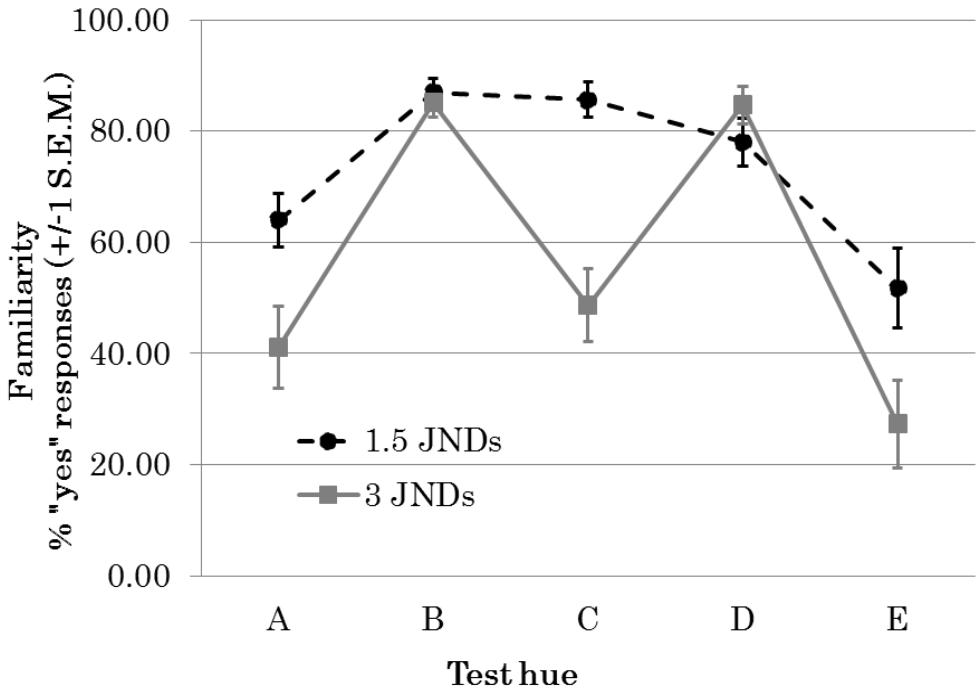

Figure 2.8. Familiarity of test hues for experiment 2 where adjacent hues are separated by 3 JNDs (solid grey line) and for the different-category condition of experiment 1.

To quantify the effect of perceptual similarity on ensemble perception, a comparison was made between these data and those from experiment 1. The most comparable condition from experiment 1 is the different-category condition since it also presented ensembles containing two categories and was centered on the boundary hue. If increasing perceptual distance between elements has an effect on ensemble perception, then we expect to see differences in the familiarity of the mean hue across the two experiments. A 5 (Test Colour: A-E) x 2 (Perceptual Distance: 1.5 JNDs, 3 JNDs) repeated measures ANOVA confirmed a significant interaction of the familiarity of Test Colour and the Perceptual Distance between stimuli (F (4, 96) = 7.65, p < .001). Figure 2.8 plots the familiarity data from experiment 2 and the different-category condition of experiment 1. From this figure, it is clear that increasing the perceptual distance decreases the rate at which unseen

colours are perceived as being familiar. Familiarity of the mean hue (C) was significantly lower when there was greater perceptual distance between hues (M = 48.7%, SE = 6.5%) than experiment 1 when the distance was less (M = 85.6%, SE = 3.1%) ($t = 5.10$, $p < .001$). The seen colours (B and D) do not differ in their familiarity across experiments (highest $t = 1.21$; lowest $p = .24$).

2.4.3 Discussion

Experiment 2 investigated the effect of increasing the discriminability of hues. When colours were separated by 3 JNDs, the unseen mean was as familiar as the other unseen hues, and was less familiar than the seen hues. Therefore, when hues in the ensemble were more discriminable, the bias for familiarity of the unseen mean hue that was found in Experiment 1 disappeared. The findings could suggest that the increased discriminability of the colours removes the ensemble effect, suggesting a role for colour similarity in the ensemble perception of hue.

2.5 General discussion

2.5.1 Ensemble perception of hue

The findings of the current study suggest that if people view a briefly presented ensemble of two hues, then under some conditions they will partly represent that ensemble as the mean of the two hues. For example, in experiment 1, the unseen mean of the ensemble was as familiar to the participants as the seen ensemble hues, and more familiar than other unseen colours. Furthermore, when they responded correctly and

rejected the mean hue, it took participants longer to make that decision compared to correct responses to the other unseen test hues. In experiment 2, when the perceptual difference of the exemplars in the ensemble is increased, this mean effect was abolished.

The findings of the current study for colour are similar to prior findings for size (Ariely, 2001). For example, for size there is also high familiarity for unseen objects which are within the ensemble range. This result also corroborates the findings from simultagnosia patient GK (Demeyere et al., 2008), who showed a similar failure to reject hues intermediate to those in an ensemble of two hues (notwithstanding the stimulus issues in that study discussed in the introduction).

Ensemble perception is understood to occur in the absence of serial processing (Haberman & Whitney, 2012), where knowledge of the specific features present is either not encoded or is sacrificed in exchange for a summary representation (Corbett & Oriet, 2011). Thus, if ensembles are represented and encoded only by their mean then we would expect familiarity to peak at the mean. In experiment 1, the seen hues and the mean hue are found to be equally familiar to observers, suggesting that there is some encoding of which hues are present, in addition to the encoding of the gist of the ensemble. Indeed, if colour ensemble perception is similar to that for ensembles with objects of varying size (Albrecht & Scholl, 2010; Ariely, 2001), or ensembles of different facial expressions (Haberman & Whitney, 2010), then the possibility that the mean hue is automatically encoded and guides the perception of the gist of that scene should not be ignored.

The present design was intended to replicate the early findings from other visual modalities. As such, the simplicity of the ensembles means that we cannot certainly rule out sub-sampling or encoding of all the hues present as explanations for the data. Likewise, the tightly-packed distribution of hues used in the present experiment does not make it

possible to separate the effect of the mean from the median, or range of hues. Nevertheless, our data does provide a strong suggestion that some summary statistic of hue from the briefly-presented ensembles may be extracted in cognitive representation of that scene. Furthermore, given the similarity in the pattern of responses to the equivalent study for size (Ariely, 2001), it is predicted that given a more complex ensemble with a greater number of different hues, the ability to serially encode element hues would be reduced and the mean effect would appear more strongly with poorer recognition performance for the seen hues (Brady & Alvarez, 2011; Marchant & De Fockert, 2009).

2.5.2 Categorical effects on ensemble perception

The hues in the current study were specifically controlled to manipulate categorical relationships between the colours seen across the conditions, with some specific hues being present in more than one condition but differing in their categorical relationship to the other colours used in that condition. Differences in the pattern of responses across the conditions can therefore be linked to the structure and characteristics of colour categories. Given the categorical nature of colour perception reported for many memory and search tasks, it was expected that in the single-category conditions some shift-to-prototype would occur – with representations of green and blue ensembles being drawn towards the prototypes such that familiarity would be higher for prototypical test colours compared to the boundary. This could imply that the encoding and/or retrieval of ensembles involves the application of linguistic labels. The evidence from experiment 1, however, does not support this prediction, with the asymmetry analysis revealing the opposite bias – a shift-to-boundary for familiarity – in both of the same-category conditions. Despite

being of equal perceptual distance from the ensemble hues, the boundary hue was more familiar than the prototype, not less.

By definition, boundary colours have an unstable categorical identity, being the point at which each answer is given with equal frequency in a 2AFC naming task. In contrast, prototypes have a much more stable categorical identity (e.g., across illuminations – Olkkonen et al., 2009; Olkkonen et al., 2010), are more easily named (Heider, 1972), and also appear to cluster around similar points in colour space across languages (Collier, 1976; Kay & Regier, 2003; Lindsey & Brown, 2009; Regier et al., 2005). One explanation for the bias in responses towards the boundary hues, is that the distinctive identity of the prototype serves to make it easier to reject when given as a test colour. If the test colour is of a strong categorical identity, as a prototype is, then it may be more easily rejected, whereas if a test colour is at a category boundary, and so is of a less certain categorical identity, errors of false recognition are made more easily.

2.5.3 Metric effects on ensemble perception

Most of the experimental work on ensemble perception has not considered the role of perceptual discriminability (metric differences) in estimations of the mean. This study sought to address this by using JND-equated hue differences and, in the second experiment, comparing the responses to ensembles of elements separated by a larger hue difference. This comparison revealed that the strong familiarity for the mean of the ensemble observed in experiment 1 disappears when the element colours are more discriminable from each other, as in experiment 2. In the case of stimuli which are more discriminable, the familiarity responses are more similar to what would be expected if

observers have an accurate representation of each of the element colours, with no evidence for a summary statistical representation.

The different pattern of responses in experiment 2 indicates that metric (as opposed to categorical) perceptual difference may have a role to play in ensemble perception of hue. One possibility is that when an ensemble is composed of fairly similar hues (as in experiment 1), the mean is extracted in addition to accurate representations of the element colours, whereas when perceptual distances are larger and the ensemble has higher variance in hue the mean is a less diagnostic representation of the scene. This would require a mechanism sensitive to the variance in hue present in a scene (e.g., Brown & MacLeod, 1997; M. A. Webster & Mollon, 1997).

This manipulation of perceptual similarity also raised another question which may apply to the whole ensemble perception literature. Although adjacent hues used in experiment 1 are discriminable when viewed simultaneously, it is known that successive viewing decreases the acuity of hue discrimination compared to simultaneous viewing (Romero, Hita, & Delbarco, 1986; Uchikawa & Ikeda, 1981). This raises the possibility that the greater familiarity of the mean hue is due to a failure in colour discriminability in memory. Therefore, the mean hue may be a familiar signal due to its simultaneous colour similarity to both element hues, while the other unseen hues bear resemblance to only one element and so do not provide such a strong signal. This explanation requires no ensemble perception, nor any summary statistics.

In the two-colour trials, there is a deficit in accuracy for the mean hue (compared to the control trials) which is not present for the other unseen hues. If it were the case that this effect is driven by the simultaneous similarity of C to both B and D in the ensemble, this would entail some additive effect of having two adjacent familiar stimuli rather than just

one. Imagining that both ensemble elements are encoded accurately, it is not clear why this information would be discarded in favor of an additive generalization mechanism like that described. Nevertheless, the result would still be that the observer exhibits a set of responses which are sensitive to some summary statistics, especially the range. Alternatively if there is ensemble perception, then this already implies some gestalt processing in which the summary representation is encoded rather than precise stimulus information.

Although both explanations are compatible with the results, both can be argued to invoke summary statistics either at the point of encoding (ensemble perception), or at the point of retrieval (generalization).

2.5.4 Summary and future directions

The current study sought to establish whether ensemble perception for hue occurs in a similar way to that for size. It found that the unseen mean hue of a two-coloured ensemble was more familiar than other unseen hues controlled for perceptual similarity. This replicates the findings for ensemble perception of size (Ariely, 2001). One explanation would be that the mean hue is automatically extracted and encoded for ensembles of this type, causing the mean hue to appear familiar, in spite of it never being present. Experiment 2 found that the mean effect is attenuated when ensemble exemplars and test hues were separated by a larger perceptual distance. To some extent the results of experiment 2 may challenge the mean extraction view, as it appears that the mechanism is active only when there is a small amount of hue variance in the ensemble. Another explanation, therefore, is simply that the observers make systematic mistakes around the actual hues that were present, causing the mean to have a higher rate of familiarity.

However, it is not clear that this can fully explain the selectively enhanced familiarity of the mean hue for two-colour ensembles, compared to single-colour presentations, without a reference to some summary statistics.

Experiment 1 also found that hues near the category boundary were more familiar than prototypical hues where ensembles contained exemplars from only one category. This result is contrary to the shift-to-prototype that would be expected if colour is automatically categorically encoded in such situations. Instead it appears that the presence of the categorical boundary has an effect only when deciding on the familiarity of a test hue. One explanation offered for this is that the strong categorical identity of the prototype suppresses any effect of familiarity due to similarity to the ensemble hues. This implies that colour categories affect decision-making about perceptual tasks rather than perception itself (Witzel & Gegenfurtner, 2013).

This study presents a number of new avenues for investigation of the perception of multi-coloured ensembles. The generalization and ensemble perception accounts of the data can be resolved by using more complex ensembles. As well as increasing the visual short-term memory demanded by the task, this will also allow the manipulation of different statistics of the ensemble – the mean, median, range or variance – in order to see which appears to be most important to the visual system. Secondly, if linguistic identity is driving the shift-to-boundary effect (rather than any particularity of the colours used in this study), we expect the same phenomenon to be found for other colour category boundaries. The same effect may also be found for other stimuli, such as categorical facial expressions, and correlation may be found between how easily namable a colour is (Heider, 1972) and it's familiarity following ensemble presentation.

2.6 Conclusion

When viewing a multi-coloured ensemble, extraction of the gist in the form of the mean colour, may be necessary when opportunity for focused attention and serial encoding of features is not possible. When observers in our study viewed briefly presented ensembles of two hues they had a tendency to incorrectly recognize the mean hue (despite it never being present), finding it as familiar as the seen hues. This familiarity of the unseen mean was attenuated when discriminability of exemplars was increased, potentially suggesting that averaging is only applied when the variance in hue is small. Furthermore, observers tended to also find hues at the blue-green category boundary more familiar than prototypical hues when exposed to same-category ensembles, contrary to the predictions of a 'shift-towards prototype' account. Taken together, the findings could suggest that multi-coloured ensembles are partly represented as their mean hue when hues are similar, but future research should seek to further determine the mechanism responsible.

Chapter 3

Paper 2: Effects of ensemble complexity and perceptual similarity on rapid averaging of hue

Maule, J., & Franklin, A. (2015). *Journal of Vision, 15*(4):6, 1-18, doi:10.1167/15.4.6

3.1 Abstract

The ability to extract the mean of features from a rapidly-viewed, heterogeneous array of objects has been demonstrated for a number of different visual properties. Few studies have previously investigated the rapid averaging of colour, those that did had insufficient stimulus control or inappropriate methods. This study reports three experiments which directly test observers' ability to extract the mean hue from a rapidly-presented, multi-element colour ensemble. In experiment 1 ensembles varied in number of elements and number of colours. It was found that averaging was harder for ensembles with more colours, but that changing the number of elements had no effect on accuracy, supportive of a distributed-attention account of rapid colour averaging. Experiment 2a manipulated the hue range present in any single ensemble (varying the perceptual difference between ensemble elements), while still varying the number of colours. Range had a strong effect on ability to pick the mean hue. Experiment 2b found no effect of colour categories on the accuracy or speed of mean selection. The results indicate that perceptual difference of elements is the dominant factor affecting ability to average rapidly-seen colour ensembles. Findings are discussed both in the context of perception and memory of multiple colours and ensemble perception generally.

3.2 Introduction

It has been claimed that humans have the ability to extract summary statistics from a briefly-viewed visual scene for a number of visual properties, including orientation (e.g., Parkes et al., 2001), motion speed (e.g., Watamaniuk & Duchon, 1992), motion direction (e.g., Watamaniuk et al., 1989), brightness (e.g., Bauer, 2009b), size (e.g., Ariely, 2001), emotional expression of faces (e.g., Haberman & Whitney, 2009), facial identity (e.g., De Fockert & Wolfenstein, 2009; Leib et al., 2014) and direction of biological motion (Sweeny, Haroz, & Whitney, 2012). Much of the literature is focused on perceptual averaging – whether ensembles are encoded by their mean properties.

To date, relatively little published research has investigated ensemble perception of colour (Demeyere et al., 2008; Kuriki, 2004; Maule et al., 2014; J. Webster et al., 2014). This is despite colour being a good candidate for investigation, both from the point of view of better understanding how colour is perceived and encoded, but also for understanding the perceptual averaging mechanism. For colour scientists, perceptual averaging experiments provide a paradigm which could help elucidate questions about the shape and organization of perceptual colour space (J. Webster et al., 2014). For those interested in ensemble coding mechanisms and functions more generally, colour is an ideal substrate for investigation. It is well-described and characterized in terms of human perception, is continuous yet also subject to categorization (e.g., Bird et al., 2014; He et al., 2014; Roberson et al., 2008), and can help answer ecologically valid questions about the appearance of surface colours (Giesel & Gegenfurtner, 2010; Sunaga & Yamashita, 2007). Furthermore, hue (the subjective experience of which is qualitative), along with the other dimensions of colour, saturation and lightness (which are matters of magnitude)

provide an opportunity to understand how ensemble coding deals with the integration of multiple perceptual dimensions.

3.2.1 Ensemble coding of colour

Various papers have addressed the question of colour averaging, and while in most cases their findings hint at a meaningful colour averaging mechanism, most have not employed the methods used commonly in investigations of ensemble perception, or investigated averaging under rapid viewing conditions. It has been shown that when colourful ensembles are presented for a short time (500ms), observers tend to find the unseen mean hue as familiar as the hues that were present in the original ensemble (Maule et al., 2014), a pattern reminiscent of Ariely's (2001) finding for ensembles of different sizes. This same pattern appears for ensembles of colour in a patient with simultagnosia, a condition rendering the patient unable to reliably count more than one or two items (Demeyere et al., 2008), suggesting that the colour averaging, like face averaging in cases of prosopagnosia (Leib, Puri, et al., 2012), can survive in spite of cognitive deficits limiting the encoding of individual items.

In addition to the indications of mean encoding there is also evidence that other summary statistics might also play a role in the extraction of gist from a rapidly-viewed colourful ensemble. For example, high variance colourful ensembles tend to elicit slower reaction times and less accurate responses when judging whether a mean of another ensemble is "blue" or "red" (de Gardelle & Summerfield, 2011). Variance is further implicated as an important determinant of the accuracy of summary statistics by evidence that priming with a colourful ensemble can speed a judgment about another ensemble of the same variance, even when the mean changes from prime to target (Michael et al., 2014).

Likewise, the tendency to find the mean hue familiar in a membership identification task disappears when the perceptual difference between ensemble hues was increased (Maule et al., 2014), suggesting that there may be a functional limit to the amount of variance which can be rapidly encoded by summary statistics. These findings support suggestions that summary statistics serve to help tune the visual system to the environment and support the stability of the visual world (Corbett & Melcher, 2014a, 2014b; Corbett et al., 2012; Lanzoni et al., 2014).

Although promising, results from membership identification tasks (Maule et al., 2014) provide only an indirect measure of the encoding of the mean following brief ensemble presentations. A more direct test of colour averaging is required to establish whether observers demonstrate ensemble coding for colour by the mean when presented with an ensemble very briefly. If a mean hue can be accurately extracted from a rapidly-presented ensemble this will add weight to claims that ensemble coding is a pervasive feature of the visual system, possibly driven by a mechanism common to many different visual domains (e.g., Alvarez, 2011; Haberman & Whitney, 2012). Nevertheless, the evidence of priming by the mean and variance of colourful ensembles (Michael et al., 2014) is a strong indication that summary statistics of colour may be encoded by the visual system.

Previous research has also found that observers can approximate the colourimetric mean when adjusting a homogenous patch to represent a continuously-presented multi-colour mosaic, but that these estimates are biased towards the salience or saturation of the mosaic elements (Kuriki, 2004) and the position of the unique hues (Sunaga & Yamashita, 2007) – i.e. the red, green, blue or yellow which appears pure, unmixed with any other hue (see Kuehni, 2014). Such adjustments are also more variable when the perceptual distance between element colours is greater (J. Webster et al., 2014). These studies allowed

observers to view ensembles for an unlimited amount of time while making their settings. Therefore, it remains unknown whether these averaging mechanisms are the same when observers have limited time to view the stimuli – conditions in which ensemble coding would be most beneficial (Alvarez, 2011).

In summary, although some research on mean colour perception has been carried out, there is no equivalent of the studies investigating ensemble perception of other features. Studies which have probed directly for representations of the mean colour (e.g., Kuriki, 2004; Sunaga & Yamashita, 2007; J. Webster et al., 2014) have allowed observers unlimited time to view the ensemble stimuli, rather than using rapid presentation to encourage the deployment of distributed attention (e.g., Alvarez & Oliva, 2008; Baijal et al., 2013; Treisman, 2006). Where rapid presentation has been used (e.g., de Gardelle & Summerfield, 2011; Maule et al., 2014; Michael et al., 2014) tasks have not probed directly for representations of the mean colour. Some studies have also failed to adequately control, using colour spaces, the perceptual differences between the colours used in their experiments (e.g., de Gardelle & Summerfield, 2011; Demeyere et al., 2008; Michael et al., 2014). This study seeks to address this gap in the literature as well as further explore the nature of rapid mean colour encoding (if it exists).

3.2.2 Theoretical questions

This study also addresses four important theoretical questions relevant to both ensemble perception generally and colour cognition. Firstly, any claim of robust, rapid colour ensemble coding also requires attention to be paid to the mechanism and limits of this process. As highlighted in the literature on size averaging (e.g., Ariely, 2001, 2008; Marchant et al., 2013; Simons & Myczek, 2008), a crucial distinction regarding the

mechanism of ensemble perception is between holistic processing with distributed attention (exhaustive ensemble processing) and sub-sampling of relatively few elements (limited capacity ensemble processing). Exhaustive ensemble processing implies the parallel integration of all of the objects or items in an ensemble to provide a summary representation, often with inaccurate or no representation of the individual items. This has been a controversial suggestion, as it postulates a mechanism which exceeds the limit of visual short-term memory (Alvarez, 2011), with some studies appearing to demonstrate exhaustive processing (e.g., Ariely, 2001, 2008; Chong & Treisman, 2003, 2005a, 2005b) while others propose that models which use sub-sampling can account for the accuracy of mean judgments, without the need for a holistic mechanism (e.g., Marchant et al., 2013; Myczek & Simons, 2008; Simons & Myczek, 2008; Whiting & Oriet, 2011). Varying the number of elements in ensembles will provide an indication as to whether averaging is affected by an increase in the number of objects to average. If the process requires focused attention to a sub-sample of elements then the averaging process should be subject to more error when there are more elements. In contrast, if the process occurs across the whole ensemble the rate of error will be unaffected.

Secondly, given the evidence for the importance of variance and/or range of colours in extracting a mean (de Gardelle & Summerfield, 2011; Maule et al., 2014; Michael et al., 2014; J. Webster et al., 2014) it is pertinent to investigate the limit of averaging for ensembles varying more or less widely in colour. Some functional limit for perceptual averaging of colour is likely to be present (larger variances in size of ensemble elements is also detrimental to averaging performance (Utochkin & Tiurina, 2014)), but whether this a limit dependent on the number of different colours present in an ensemble or the perceptual difference between the elements of the ensemble is an open question.

Thirdly, the perception and cognition of hue has features which are not shared by other visual domains investigated in the ensemble perception literature. Variation in hue tends to be represented (in colour spaces) as forming a circular perceptual continuum. It is possible that the circularity of hue perception will interfere with the encoding of the mean colour, as any pair of hues has two angular differences (clockwise and counterclockwise) which could describe their perceptual relationship (these two angular differences adding up to 360°). As these angular differences approach parity (180°), these competing interpretations of the mean colour become equally likely, and may make the extraction of the mean colour more difficult. For this reason it might be expected that hue may be unsuited to ensemble coding, in which case averaging colour would be very difficult or effortful.

Fourthly, colour is also subject to verbal labels (e.g. "green", "yellow"), which (at least in English) divide the hue circle into more-or-less discrete categories. The position of linguistic boundaries have been claimed to have effects on colour memory (e.g., Roberson & Davidoff, 2000), visual search (Daoutis et al., 2006), neural representation (A. Clifford et al., 2012; A. Clifford, Holmes, Davies, & Franklin, 2010), and colour discrimination (Drivonikou, Clifford, Franklin, Ozgen, & Davies, 2011; but see Witzel & Gegenfurtner, 2013). Notably, however, the effects of categories appear to be post-perceptual in origin (Bird et al., 2014; He et al., 2014; Roberson et al., 2008). Given these effects we might predict that ensembles containing multiple categories would prove harder to average. This would have implications for ensemble perception in other domains with categorical labels, such as facial expression/identity. If the categorical content of an ensemble does not make any difference this will provide some support for the early and automatic computation of mean colour from rapidly-presented ensembles.

3.2.3 The present study

The present study attempts to address these questions, building on and rectifying the methodological constraints identified in the past literature. Using a 2-alternative forced-choice (2AFC) task observers were tested on the correct identification of the mean hue following the rapid presentation of a multi-hue ensemble. Following Maule, et al. (2014) the stimuli are controlled to ensure equality of difference between adjacent colours in terms of just-noticeable differences (JNDs) (Witzel & Gegenfurtner, 2013). This ensures that ensembles of different colours can be used with reasonable assurance of their equivalence in perceptual terms and allows us to use a broad stimulus set which reduces the probability of trial-by-trial learning of the stimuli (Bauer, 2009a). This stimulus control also supports the validity of any findings regarding ensemble perception as errors due to failures of discrimination will be minimized.

Experiment 1 aimed to establish the basic conditions in which observers can reliably extract an average hue from a multi-hue ensemble, and how hue averaging is affected by the number of elements and number of colours in the ensemble. The design was similar to that of the Marchant et al. (2013) study on the effect of set size and heterogeneity on estimations of mean size, with manipulations of the number of elements in ensembles and number of different colours in ensembles. Where that study used limited combinations of these parameters for their analysis, we have included independent manipulations of levels of both number of colours and number of elements.

Experiment 2a built on the results of experiment 1, investigating the effect of varying the range of colours in ensembles on the accuracy of mean extraction. The overall design and aim was similar to that of Utochkin and Tiurina's (2014) replication and extension of the work of Marchant et al. (2013). Utochkin and Tiurina attempted to parse the effect of

number of elements, the effect of number of different sizes and the effect of difference between ensemble elements on estimations of mean size. They found evidence for the variance of sizes being a strong determinant of accuracy when estimating mean size.

Experiment 2b reanalyzed the data from experiments 1 and 2a on the basis of colour naming data in order to observe the influence of categories on the extraction of mean hue and further investigate the mechanism of hue averaging.

Overall, the experiments aim to further characterize the conditions under which rapid averaging of colour occurs, explore whether rapid averaging appears to be a result of focused or distributed attention, and demonstrate the effect of hue range and categorization on mean estimation.

3.3 Experiment 1

3.3.1 Methods

3.3.1.1 Participants

Eighteen observers (14 female, mean age = 22.4 years (SD = 2.2 years)) naive to the purpose of the experiment took part. All reported normal or corrected-to-normal visual acuity and were assessed as having normal colour vision using Ishihara plates (Ishihara, 1973) and the Lanthony test (Lanthony, 1998). All were undergraduate students at the University of Sussex and were paid £4.50 for their time. The research protocol was approved by the university ethics committee.

3.3.1.2 Stimuli

A stimulus range consisting of 24 hues was specified from a circle on an equiluminant plane in Derrington-Krauskopf-Lennie (DKL) space (Derrington et al., 1984; Krauskopf et al., 1982) (see figure 3.1). These hues were spaced such that each differed from its neighboring hues by 2 just-noticeable differences (JNDs), as measured by Witzel and Gegenfurtner (2013). Since the stimuli are recreating those from Witzel and Gegenfurtner's measurements, their position in colour space describe a circle in DKL-space, however the experiment uses the JND-scaled version of this circle, so the discriminability of neighboring hues is consistent throughout the circle and is not warped by inhomogeneities in the hue specification of that, or any other, colour space. A grey background (xyY (1931): 0.310, 0.337, 30.039) was used throughout the experiment.

3.3.1.3 Apparatus

Stimuli were displayed on a 22-inch Mitsubishi DiamondPlus 2070SB Diamondtron CRT monitor, with a resolution of 1600 x 1200 pixels, 24-bit colour resolution, and a refresh rate of 100 Hz. Monitor primary values (RGBs) for all of the colours used in the experiment were selected manually using systematic adjustment of monitor primaries to output the correct xyY values as measured by a ColourCal colourimeter (Cambridge Research Systems).

The experiment took place in a blacked-out room, with the monitor the only source of light. A cardboard viewing tunnel lined with black felt was used to obscure peripheral objects and colours which could otherwise be illuminated by the light from the monitor,

from the participant's field of vision. A chin rest was used to constrain viewing distance at 57cm, and responses were given using the keyboard.

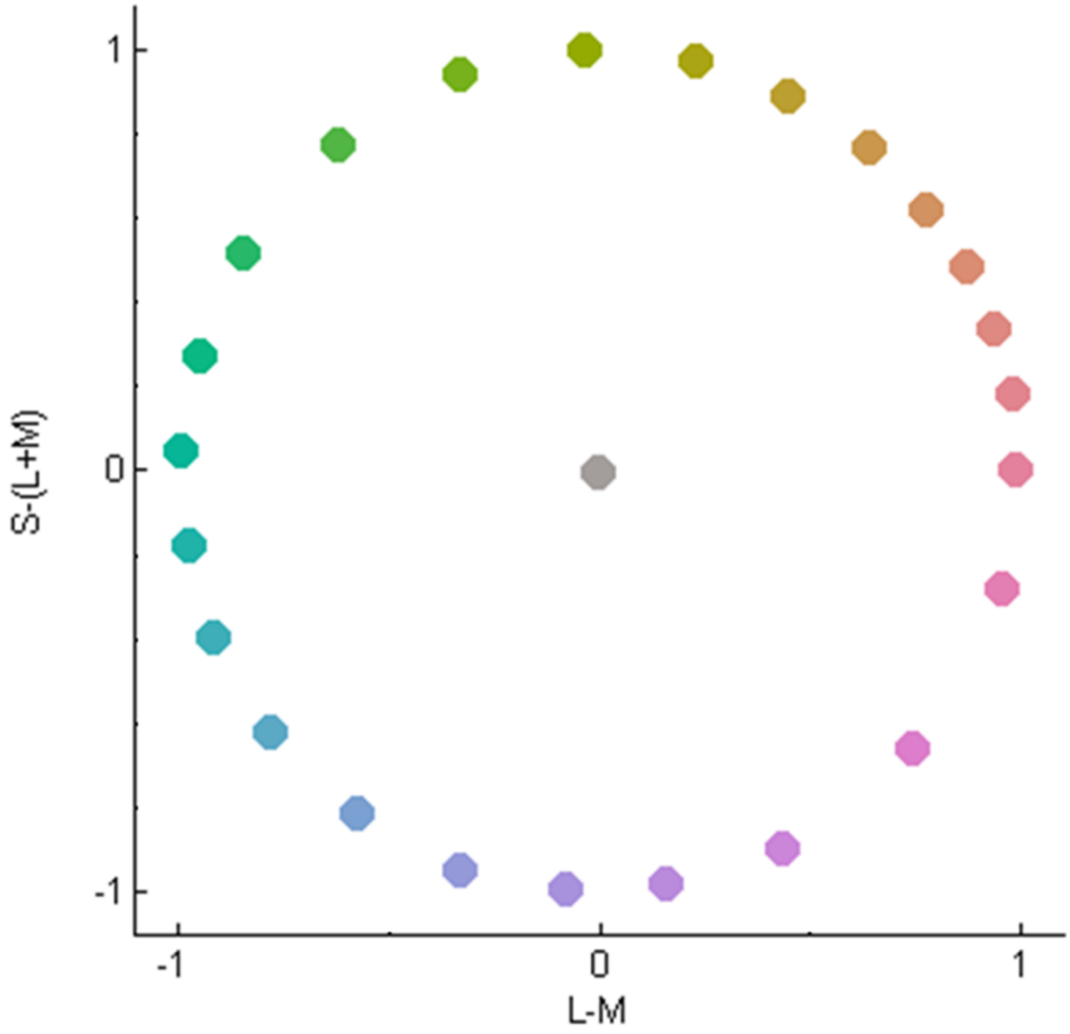

Figure 3.1. Arrangement of stimulus colours on an isoluminant plane in DKL space. Axes correspond to activation of the cone-opponent processes and are scaled arbitrarily according to the maximal monitor output described by Witzel & Gegenfurtner (2013). Adjacent hues are spaced by 2 JNDs (ibid.), resulting in some inhomogeneity in the angular distance – discriminability is higher in some areas of this colour space than in others. The gray point in the center represents the

background colour. In this (and all figures) colours are intended to approximate those used in the experiment, but should not be taken to represent those colours precisely due to differences in printing and display equipment. This arrangement of hues was never shown to the observers in any experiment.

3.3.1.4 Design

Ensembles consisted of four, eight or sixteen circular, uniformly-coloured patches (‘elements’). Elements were allocated at random to a cell in an invisible 4-by-4 grid subtending approximately 8° of visual angle. Each element patch subtended approximately 1.2°, with a spatial location jittered randomly by up to 0.6° from the center of the allocated cell in the vertical and horizontal directions to remove the appearance of a regular arrangement of elements.

At the beginning of each trial one hue was randomly selected from the 24-hue stimulus array. This ‘seed’ hue, along with the required number of hues for the trial, was used to calculate the ensemble hues. As such the particular segment of the hue circle represented by the elements in the ensemble varied randomly on each trial, reducing the possibility of trial-by-trial averaging affecting responses, as can be caused by using a limited set of stimuli (Bauer, 2009a).

Each ensemble contained two, four or eight different hues (see figure 3.2), always represented in equal number across the elements. These two parameters (number of hues and number of elements) were varied independently such that each level was combined with the others where possible. This resulted in eight within-participant conditions (see figure 3.3).

Following research into ensemble perception in other perceptual features, we tested recognition of average hue using a two-alternative forced-choice (2AFC) task. In each case the target patch was the 'mean' hue of the preceding ensemble. Mean is defined as the centroid of the distribution of hues in the ensemble. The distractor hue was spaced 4 JNDs from the mean in either the clockwise or anti-clockwise direction in DKL space, counterbalanced across trials. The 2AFC patches were the same size as the elements in the ensemble, arranged along the horizontal midline of the monitor, to the left and right of the vertical midline and spaced by 3.5° of visual angle. The location of the target on the left or right was counterbalanced across trials.

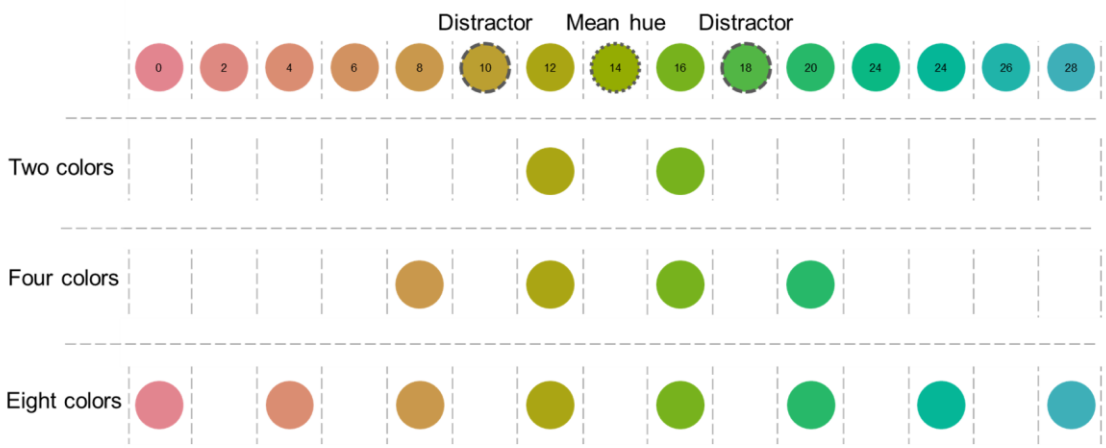

Figure 3.2. An example of stimulus selection for ensembles in experiment 1. Numbers at the top refer to number of JNDs from the seed hue (left-most, numbered 0). In this example each distribution has a mean colour corresponding to number 14, which would be paired with a distractor chosen from 4 JNDs either side for the 2AFC task. It should be noted that this is an example section of the hue continuum used, but on each trial the seed colour varied freely around the whole hue circle (figure 3.1). Ensembles also varied in number of elements, but each hue was always present in equal number.

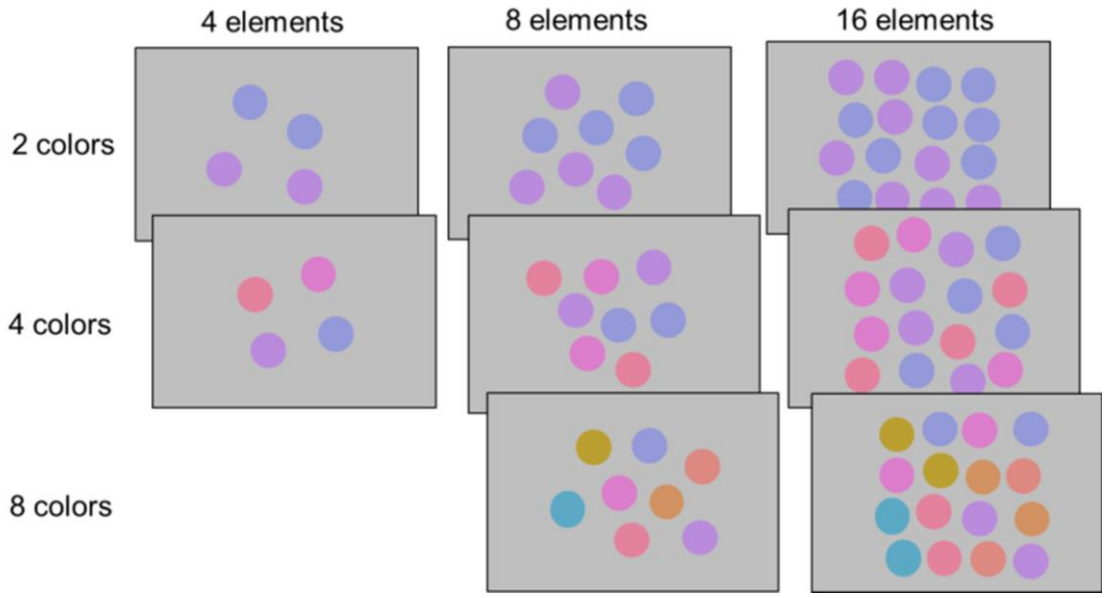

Figure 3.3. Variations in ensembles presented in experiment 1. All of these ensembles have the same mean hue but vary in the number of elements and number of different colours.

Each trial began with a black fixation point which appeared in the center of the display for 1000ms, this was followed by the onset of the ensemble, present for 500ms. After a blank screen inter-stimulus interval (ISI) of 1000ms the 2AFC patches appeared. The 2AFC patches remained onscreen until the participant indicated their choice by pressing a key (‘Z’ for the left patch, ‘M’ for the right patch). A 1000ms inter-trial interval (ITI) followed. This is summarized in figure 3.4.

Each participant completed four blocks of 96 trials. The eight element-colour conditions were interleaved pseudo-randomly within each block, with the constraint that each block would present each condition 12 times. Thus each participant provided responses to a total of 384 trials, 48 for each condition.

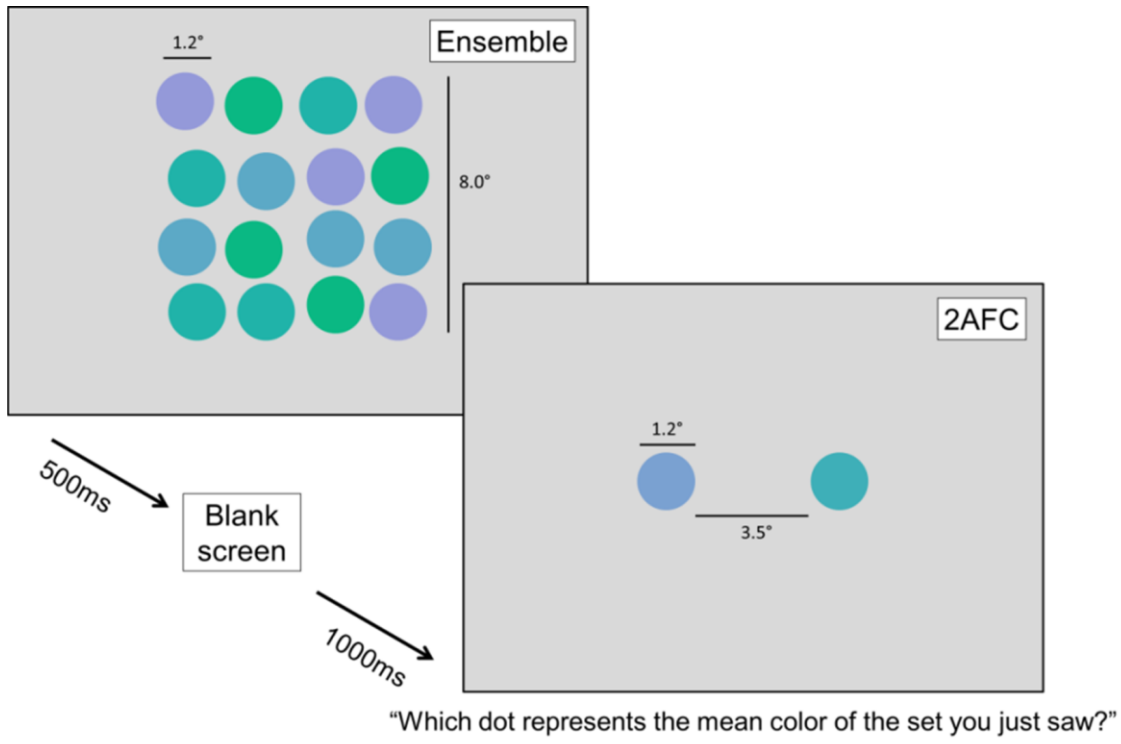

Figure 3.4. Trial structure, timing and stimulus size and arrangement for experiment 1 and 2. The 2AFC patches remained present until the participant responded.

3.3.1.5 Procedure

Participants were briefed on the basic task of the study. On-screen instructions emphasized that participants should "try to get the gist of the set of colours" when shown the ensemble and then, when given the 2AFC patches, to "decide which you think is the mean colour of the set of dots". If the participant asked for clarification of how to interpret "mean colour" the experimenter prompted them to "choose the colour which you think best represents the whole set". Participants completed a short set of eight practice trials before beginning the experimental task. The time spent on reading and practice trials ensured adaptation to the white point. No feedback on performance was given at any point during practice or experimental trials (Bauer, 2009a).

3.3.2 Results

Data were screened for reaction time (RT) outliers prior to analysis. Trials with an RT less than 200ms were removed, along with trials where the RT was more than three standard deviations above the participant's mean RT. This resulted in the removal of 143 trials (out of 6912) or approximately 2% of trials. This rate of trial removal is comparable to the same RT screening process applied in a previous study on ensemble perception of hue (Maule et al., 2014). The interpretation of the statistical analysis which follows is the same whether these trials are removed or not.

A trial was coded correct if the participant selected the mean hue (i.e. the hue falling in the center of the ensemble distribution in terms of our 2JND hue circle) rather than the distractor in the 2AFC. The main dependent variables for analysis are the proportion of correct trials and reaction times (RTs) on correct trials, with number of elements and number of colours as factors. Figure 3.5 presents the across-participant mean proportion and RT of correct trials for each of the eight ensemble types. Inspection suggests that selecting the mean for an ensemble containing fewer colours is easier (more correct responses and faster RTs) than where there are more colours (the lines are "stacked" vertically), but there is no or little effect of number of elements.

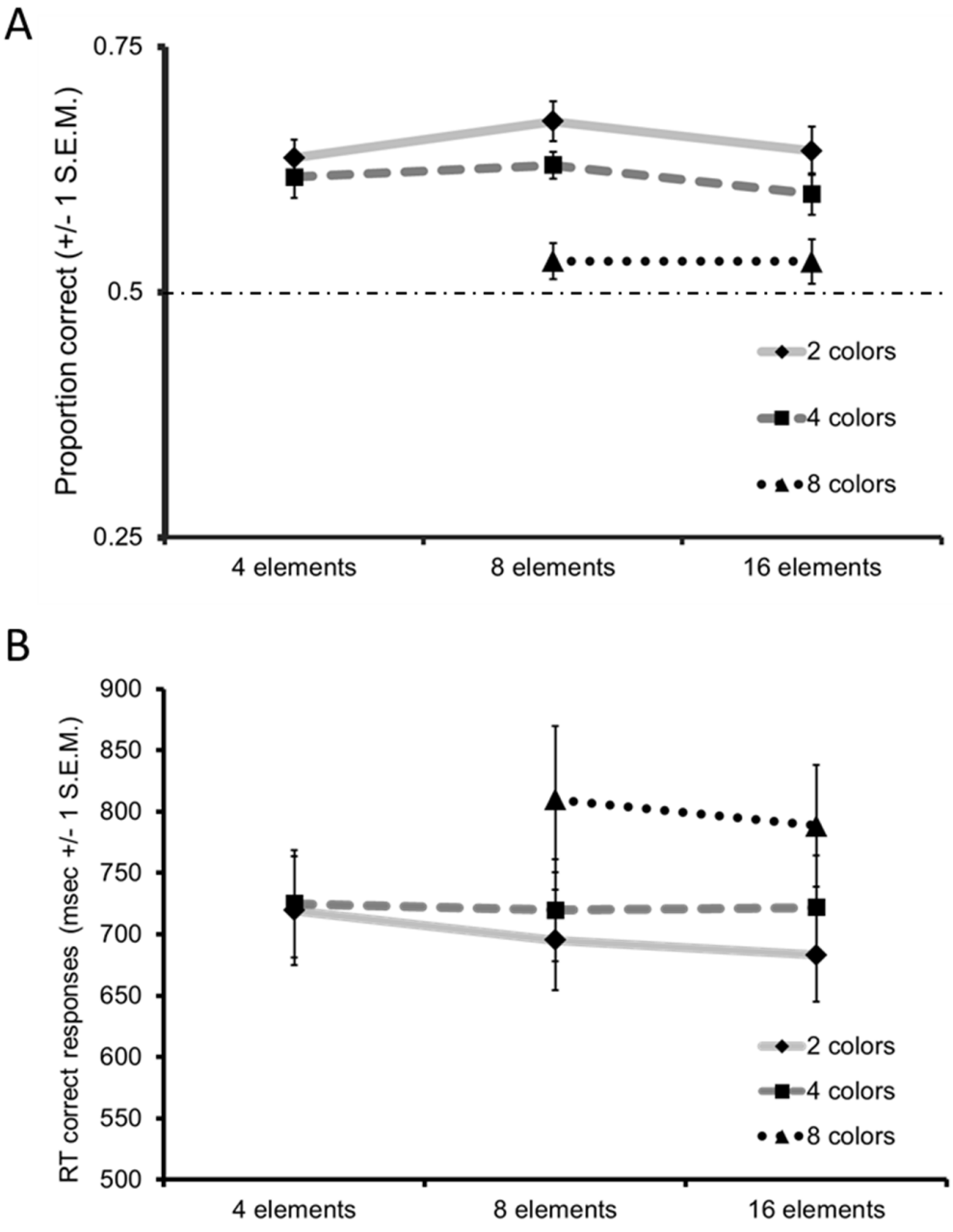

Figure 3.5. Results of experiment 1. (A) Accuracy: Mean proportion of trials where the mean colour was correctly chosen. The dotted line represents chance (0.5). (B) RTs: Mean response latency for correct trials. All error bars represent 1 S.E.M.

Due to the 'missing' cell in the elements x colours matrix (the impossible 4-element, 8-colour – see figure 3.3) the analysis was divided into two repeated-measures analyses of variance (ANOVA). Reported effect sizes are generalized eta-squared (η2G), as recommended for use with repeated-measures designs by (Bakeman, 2005).

A 3 (number of colours: 2, 4, 8) x 2 (number of elements: 8, 16) repeated-measures ANOVA revealed a significant main effect of number of colours on accuracy ($F(2, 34) = 19.48$, $p < .001$, $\eta^2_G = .30$) no main effect of number of elements ($F(1, 17) = 1.68$, $p = .21$, $\eta^2_G = .01$) and no interaction between the factors ($F(2, 34) = 0.43$, $p = .66$, $\eta^2_G = .01$). These results were mirrored in a parallel analysis; a 2 (number of colours: 2, 4) x 3 (number of elements: 4, 8, 16) repeated measures ANOVA found a marginally non-significant main effect of number of colours ($F(1, 17) = 4.36$, $p = .05$, $\eta^2_G = .04$), no main effect of number of elements ($F(2, 34) = 1.86$, $p = .17$, $\eta^2_G = .02$), and no interaction ($F(2, 34) = 0.31$, $p = .74$, $\eta^2_G < .01$). Analysis of reaction times (RTs) on correct trials found a similar pattern. A 3 (number of colours: 2, 4, 8) x 2 (number of elements: 8, 16) repeated-measures ANOVA found a main effect of number of colours ($F(2, 34) = 15.75$, $p < .001$, $\eta^2_G = .06$), with no effect of number of elements ($F(1, 17) = 1.28$, $p = .27$, $\eta^2_G < .01$) and no interaction ($F(2, 34) = 0.73$, $p = .49$, $\eta^2_G < .01$). Figure 3.5 suggests that this is due to longer RTs being associated with more colours, particularly the 8-colour ensembles.

After collapsing individual mean accuracy across elements, one-sample t-tests (all two-tailed) revealed that performance was significantly above chance (50%) on 2-colour ensembles (M = 0.65, SD = 0.07, $t(17) = 9.19$, $p < .001$) and 4-colour ensembles (M = 0.62, SD = 0.05, $t(17) = 9.41$, $p < .001$), but was not above chance (albeit very marginally) for the 8-colour ensembles (M = 0.53, SD = 0.06, $t(17) = 2.06$, $p = .06$).

3.3.3 Interim discussion

The results from experiment 1 suggest that the number of elements in an ensemble has no effect on participants' ability to pick the mean hue of a briefly-presented colour ensemble, whereas increasing the number of colours does have a deleterious effect on performance on this task. Likewise, RTs were longer for ensembles with more colours, but insensitive to changes in number of elements, supporting the implication that more variegated ensembles are harder to average. The main effect of number of colours could be interpreted in a couple of ways. Firstly it could suggest that while perceptual averaging of a stimulus attribute is robust when there are a manageable number of unique stimulus values in an ensemble, the averaging process becomes less accurate when this number increases (≥ 8). This would have implications for the suggestion that ensemble perception circumvents the limits of focused attention (e.g., Alvarez, 2011) (e.g. Alvarez, 2011). An alternative explanation is that it is not the number of colours per se but the difference between them which makes averaging difficult. In this sense it is the range in hue in the 8-colour ensembles which depletes the ability to accurately choose the mean. It can be clearly seen in figure 3.2 that range varies with number of colours in the ensembles. It is not possible to resolve the difference between these two explanations of the data from experiment 1. Therefore, in experiment 2a we used fixed ranges for ensembles with different numbers of colours (similar to Utochkin & Tiurina, 2014). By varying each factor independently we can examine which is having the stronger effect on accuracy in the task.

3.4 Experiment 2a

3.4.1 Methods

3.4.1.1 Participants

Nineteen observers (12 female, mean age = 22.7 years (SD = 2.8 years)) took part. Observers were naive to the purpose of the experiment and none had taken part in experiment 1. Visual acuity, colour vision tests and payment were as for experiment 1.

3.4.1.2 Stimuli & Apparatus

Range of hues and apparatus were as described for experiment 1.

3.4.1.3 Design

The design of ensembles was similar to experiment 1. Number of elements was fixed to eight throughout. Ensembles again consisted of 2, 4 or 8 different hues, with a range (i.e. the distance in JNDs from the hues at the extreme ends of the distribution for a given ensemble) fixed at 12, 20 or 28 JNDs. These conditions are summarized in figure 3.6.

Fixing the range meant that the spacing between element hues was variable depending on the range used. In the case of 20-JND, 4-colour ensembles, the resolution of our stimulus set (hues in 2 JND steps) meant that some ensembles had intermediate elements close to the mean (‘tight’), while others had the intermediate elements closer to the extrema (‘loose’ – see figure 3.6). These trials were counterbalanced through the experiment and

are pooled together for analysis. Averaging these two arrangements results in an approximately equal perceptual spacing between elements, without the need to present a distractor which matches an ensemble member. The main conditions for analysis were the 2- and 4-colour ensembles across the three ranges, but an additional 8-colour, 28-JND condition was included for comparison.

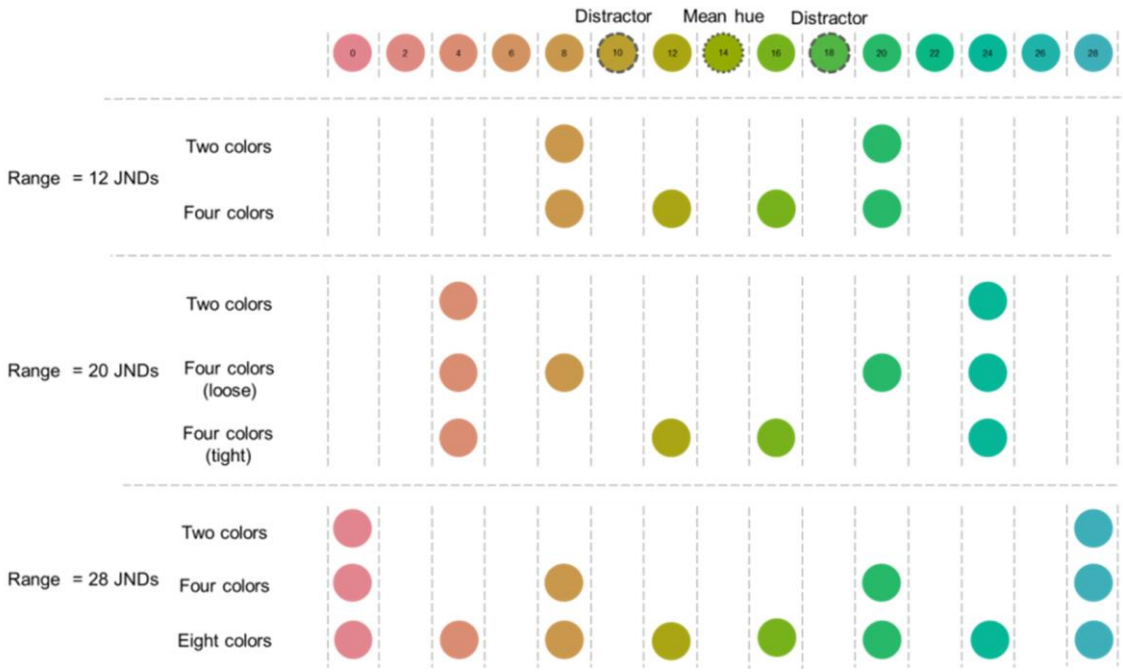

Figure 3.6. An example of stimulus selection for ensembles in experiment 2. In contrast to the design in experiment 1 (shown in figure 3.2), the outermost hues are a fixed distance apart, regardless of the number of intermediate colours which are included. Note the difference between 'tight' and 'loose' arrangements for the 20-JND, 4-colour ensembles. In experiment 2 all ensembles had eight elements.

3.4.1.4 Procedure

Procedure, instructions and the total number and arrangement of trials, conditions and blocks were as described for experiment 1.

3.4.2 Results

Data were screened using the same procedure as described for experiment 1, resulting in the removal of 191 trials (out of 7296, approximately 3%). The interpretation of the statistical analysis which follows is the same whether these trials are removed or not.

Figure 3.7 presents the mean proportion and RT of correct responses for the 2- and 4-colour ensembles across the different ranges. A 3 (range: 12, 20 or 28 JNDs) x 2 (number of colours: 2 or 4) repeated-measures ANOVA found a significant main effect of range on accuracy ($F(2, 36) = 17.99$, $p < .001$, $\eta^2_G = .35$), and a significant range x colours interaction ($F(2, 36) = 8.08$, $p = .001$, $\eta^2_G = .13$). In contrast to the results of experiment 1 there was no main effect of number of colours ($F(1, 18) = 1.78$, $p = .20$, $\eta^2_G = .01$).

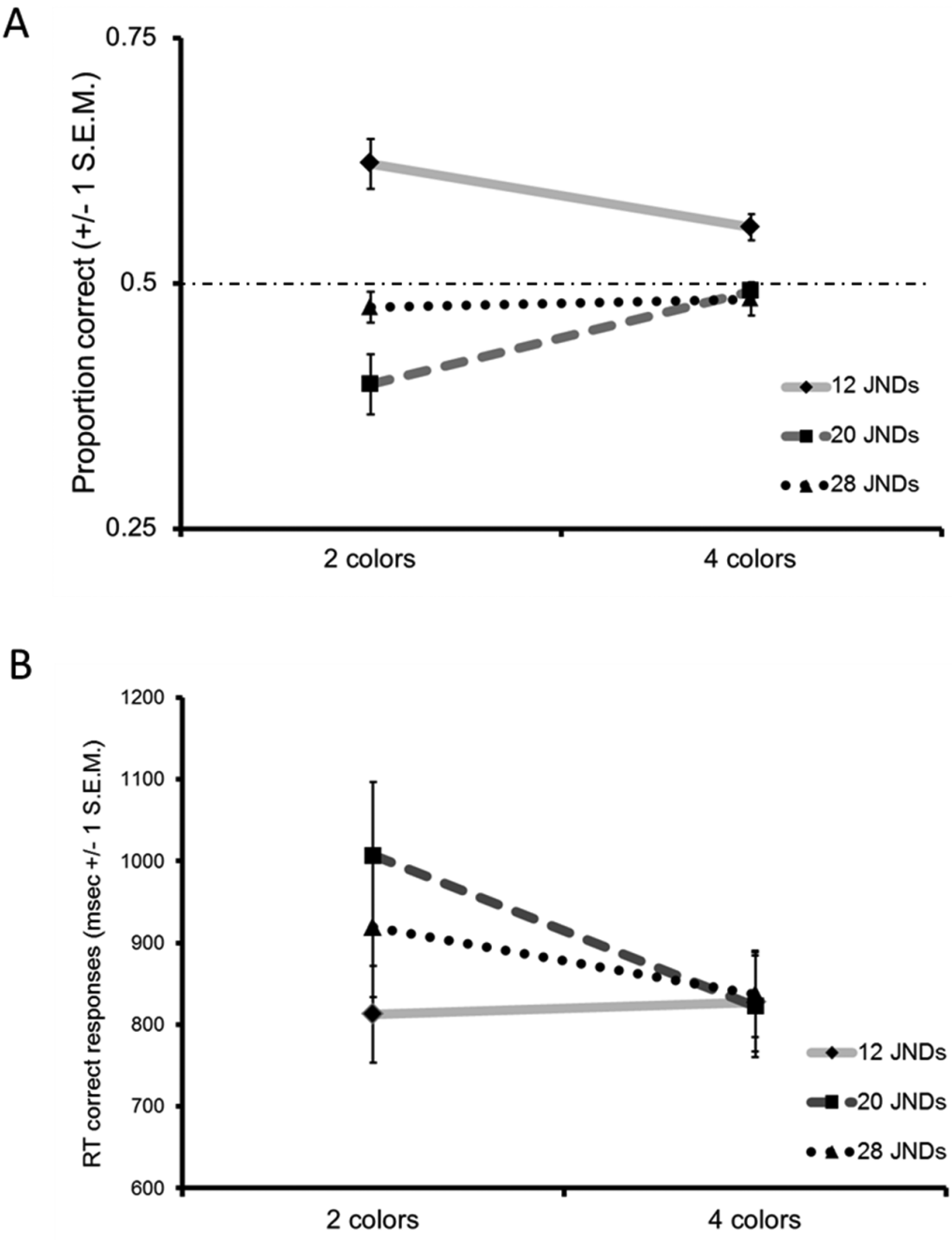

Figure 3.7. Results of experiment 2a. (A) Accuracy: Mean proportion of trials where the mean colour was correctly chosen from the 2AFC. The dotted line represents chance (0.5). (B) RTs: Mean response latency for correct trials. All error bars represent 1 S.E.M.

Performance is generally worse than observed in experiment 1, with several conditions appearing at or below chance. While it appears that participants were most successful in the 12-JND condition, the relationship between performance and range in the 20- and 28-JND conditions is less clear. One-sample t-tests (two-tailed) confirmed that, when collapsed across number of colours, only performance in the 12-JND condition was significantly above chance (50%) (M = 0.59, SD = 0.08, $t(18) = 5.04$, $p < .001$). Performance was significantly below chance for 20-JND ensembles (M = 0.44, SD = 0.07, $t(18) = 3.29$, $p = .004$) and at chance for 28-JND ensembles (M = 0.48, SD = 0.05, $t(18) = 1.93$, $p = .07$). A one-sample t-test on the 8-colour, 28-JND condition (included for comparison, not shown in figure 3.7) was also significant (M = 0.53, SD = 0.07, $t(18) = 2.12$, $p = .049$).

The below-chance performance for ensembles with a range of 20 JNDs warrants further attention as it indicates a systematic bias away from the colourimetric mean in this condition. The lack of a consistent effect found in the more wide-ranging ensembles suggests this could be an aberration caused by some feature of that condition (a potential explanation for this will be given in the discussion). From inspection of the figure 3.7 it seems possible that this condition could be masking a main effect of number of colours and causing the appearance of an interaction. Therefore in an attempt to verify the strength and nature of the main effects found, a further analysis was performed, this time excluding the below-chance 20-JND conditions. A 2 (range: 12 or 28) x 2 (number of colours: 2 or 4) repeated-measures ANOVA found a significant main effect of number of colours ($F(1, 18) = 6.31$, $p = .02$, $\eta^2_G = .03$), a significant main effect of range ($F(1, 18) = 31.64$, $p < .001$, $\eta^2_G = .50$) and a non-significant interaction ($F(1, 18) = 3.60$, $p = .07$, $\eta^2_G = .05$).

A 2 (range: 12 or 28) x 2 (number of colours: 2 or 4) repeated-measures ANOVA on the RTs for correct trials also found a significant main effect of range ($F(1, 18) = 9.33$, $p = .007$, $\eta^2_G = .01$) but no effect of colours ($F(1, 18) = 1.52$, $p = .233$, $\eta^2_G < .01$), and no interaction ($F(1, 18) = 1.84$, $p = .192$, $\eta^2_G < .01$). From figure 3.7 it would appear that selecting the mean for the smallest range ensembles (12-JND) was quicker, at least for the 2-colour condition.

Finally, the 'tight' and 'loose' variations of the 20-JND, 4-colour condition were compared. Accuracy in the 'tight' condition (mean = 0.55, SD = 0.08) was significantly higher than that in the 'loose' condition (mean = 0.44, SD = 0.10), ($t(18) = 2.89$, $p = .010$). There was no difference in reaction times for these two conditions ('tight': M = 824ms, SD = 347ms; 'loose': M = 828ms, SD = 266ms; $t(18) = .06$, $p = .956$).

3.4.3 Interim discussion

The results of the second experiment indicate observers are able to extract the mean for small ranges (12 JNDs), but unable to do so accurately for larger ranges (20 JNDs or more). This suggests that the effect of colours observed in experiment 1 was not due to the difficulty of averaging a greater number of unique exemplars, but rather is a result of greater perceptual difference between the most extreme elements of the ensemble.

Due to the large inter-element perceptual differences in this experiment there is an additional variable which may be affecting the representation of the mean. Wider ranges may be more likely to be associated with ensembles containing elements from multiple colour categories, and this may have a detrimental effect on the accuracy of hue averaging.

Since categories seem to affect post-perceptual processes, rather than early, unconscious processing (Bird et al., 2014; He et al., 2014; Roberson et al., 2008) it will be of interest to see whether, for ensembles of equal range and number of colours, the presence of one or more categorical boundaries separating the elements affects the accuracy of mean hue encoding. If ensemble coding occurs early in visual processing the categorical complexity of the ensemble (number of categories in the ensemble) should have no effect on the accuracy of mean encoding. If there is an effect of categories on mean accuracy this may challenge views that perceptual averaging is automatic and compulsory (e.g., Parkes et al., 2001), instead reflecting cognitively effortful averaging, at least for colour.

3.5 Experiment 2b

To address the question of the impact of categories on mean extraction, a sub-experiment was conducted, applying colour-naming data from an extra sample of observers to a reanalysis of the trials from experiments 1 and 2a.

3.5.1 Method

3.5.1.1 Participants

Ten observers (8 female, mean age = 29.5 years (SD = 8.2 years)) took part in the colour naming task. None had taken part in experiment 1 or 2. Visual acuity and colour vision tests were as for experiments 1 and 2a, and participants were paid £2 for their time. All were native British English speakers.

3.5.1.2 Stimuli & Apparatus

The range of hues, background and apparatus were as described for experiment 1, but responses were given using a number pad.

3.5.1.3 Design

Each trial involved the presentation of a single, uniform, circular colour patch, one of the 24 hues used in experiments1 and 2a, positioned centrally on the monitor and the same size as elements from ensembles (1.2° visual angle). The colour patch was displayed for 500ms, after which the patch disappeared and a key legend indicating which key on the number pad corresponded to which colour name appeared in the bottom left-hand corner of the monitor. On every trial observers had a choice of any of the eight English basic colour terms – "green" (1), "brown" (2), "yellow" (3), "blue" (4), "orange" (6), "purple" (7), "pink" (8), "red" (9), and gave their response on a USB number pad.

3.5.1.4 Procedure

Observers were briefed on the task, with instructions emphasizing that they should respond based on their first reaction to the colour, since the patch would not be visible for long. Each observer completed 96 trials. These trials were split in to four blocks (although there was no break between blocks for observers) with each block presenting each of the 24 hues once, in a pseudo-random order.

3.5.2 Results

Figure 3.8 presents the arrangement of the consensus colour category boundaries from the naming data collected. By taking the modal naming response to each hue across all trials and participants a consensus naming map was determined. Across participants, agreement over the name given to each hue was high (mean agreement = 86%).

Following the establishment of the consensus position of category boundaries, the data from experiment 1 and 2a were reanalyzed. First the ensemble elements from every trial were recoded according to their consensus colour category from the participants of experiment 2b. The number of different categories present in that ensemble was then calculated, for use as an independent variable in the analyses that follow. Conditions which were at or below chance performance in the original non-categorical analyses were excluded from the categorical analysis. Ensembles in the 12-JND, 2-colour condition of experiment 2a were all two-category so have not been included. In both experiments, the 4-colour condition also yielded a small number of 4-category ensembles, however, this represented too few trials to provide a reliable indication of performance (on average 7 trials per observer during experiment 1, and 2 trials per observer during experiment 2a). Therefore, the remaining conditions upon which the category analysis was performed were as follows: 2-colour ensembles with 1 or 2 categories (exp 1); 4-colour ensembles with 2 or 3 categories (exp 1); 4-colour ensembles with a range of 12-JNDs and 2 or 3 categories (exp 2a). Due to the main effects of number of colour and range already established, comparisons across these parameters should be avoided, so the focus of the analysis is on the difference between levels of category only where other factors (number of colours, range, experiment) are the same.

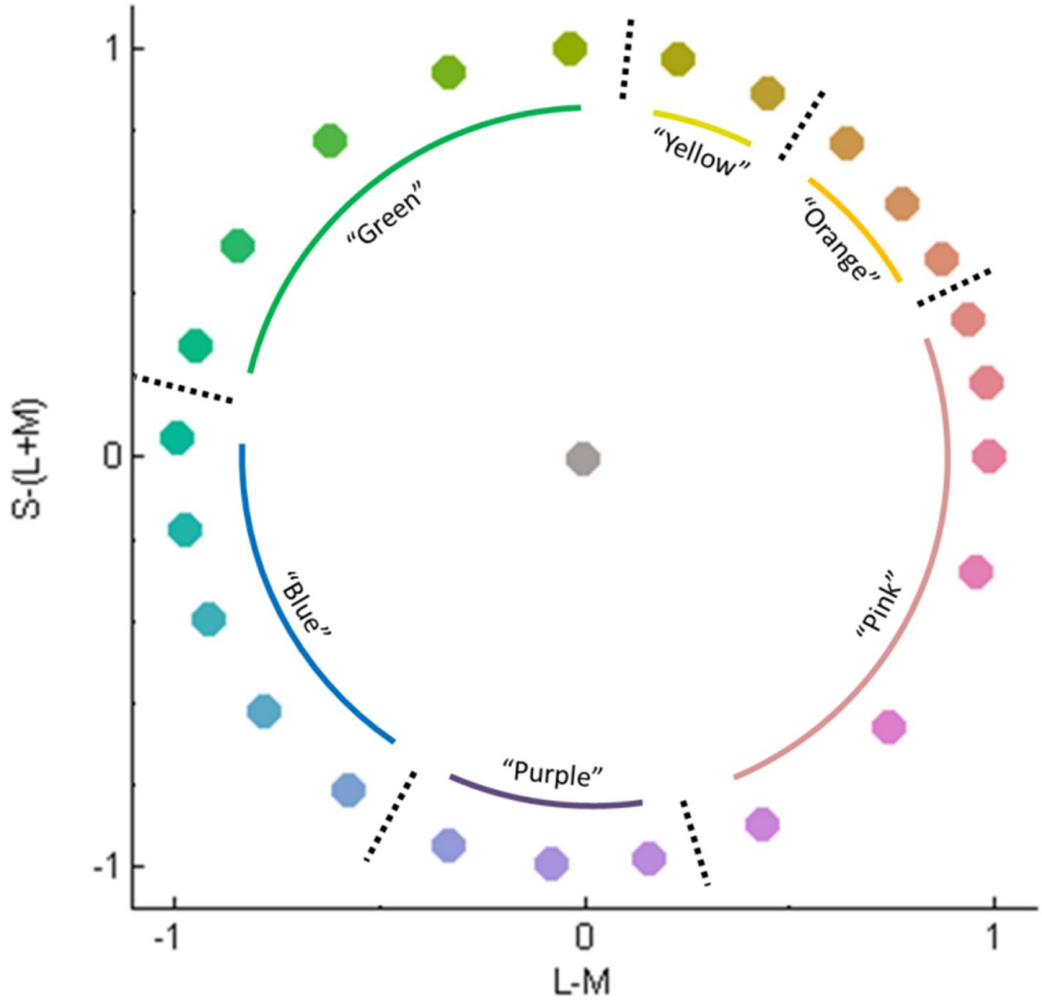

Figure 3.8. Reproduction of stimuli in DKL space (see figure 3.1 caption for details) with consensus colour categories, obtained from the naming task of experiment 2b, labeled for the 24-hue circle (see approximate colour rendering) used throughout the experiments.

Figure 3.9 shows the mean proportion and RTs of correct trials for ensembles with 1, 2 and 3 categories, across the 2- and 4-colour conditions from experiment 1 and the 12-JND, 4-colour condition from experiment 2a. Contrary to the expected disadvantage of averaging across a category boundary, in the 2-colour condition the proportion of correct

responses to ensembles with two categories (mean = .68, SEM = .02) was slightly more accurate than responses to ensembles with one category (mean = .64, = .02), though this difference was not significant ($t(18) = 1.70$, $p = .106$). Likewise, in the 4-colour condition of experiment 1, there was no significant difference between accuracy for two-category (mean = .61, SEM = .02) and three-category (mean = .61, SEM = .01) ensembles ($t(18) = .39$, $p = .699$); nor was there any difference between category levels in the 12-JND, 4-colour condition of experiment two (two-category mean = .54 (SE = .03), three-category mean = .55 (SE = .02), $t(18) = 0.263$, $p = .795$). Reaction times revealed the same pattern – with no significant differences due to number of categories (largest $t = .424$, smallest $p = .677$). These results were replicated even when trials containing any colour with low agreement on naming (<75%) is removed from the analysis.

3.5.3 Interim Discussion

The results of the post-hoc category analysis on the data from experiment 1 and 2a show that the number of colour categories represented by the elements of a multi-hue ensemble affects neither the accuracy of mean selection nor the time taken to identify the mean. This finding is concordant with the both the post-perceptual role for colour categories in cognition (He et al., 2014) and the early and automatic nature of ensemble encoding (e.g., Allik et al., 2014; Corbett & Oriet, 2011; Im & Chong, 2014), and also supports our previous findings of the effect of categories on mean colour familiarity (Maule et al., 2014).

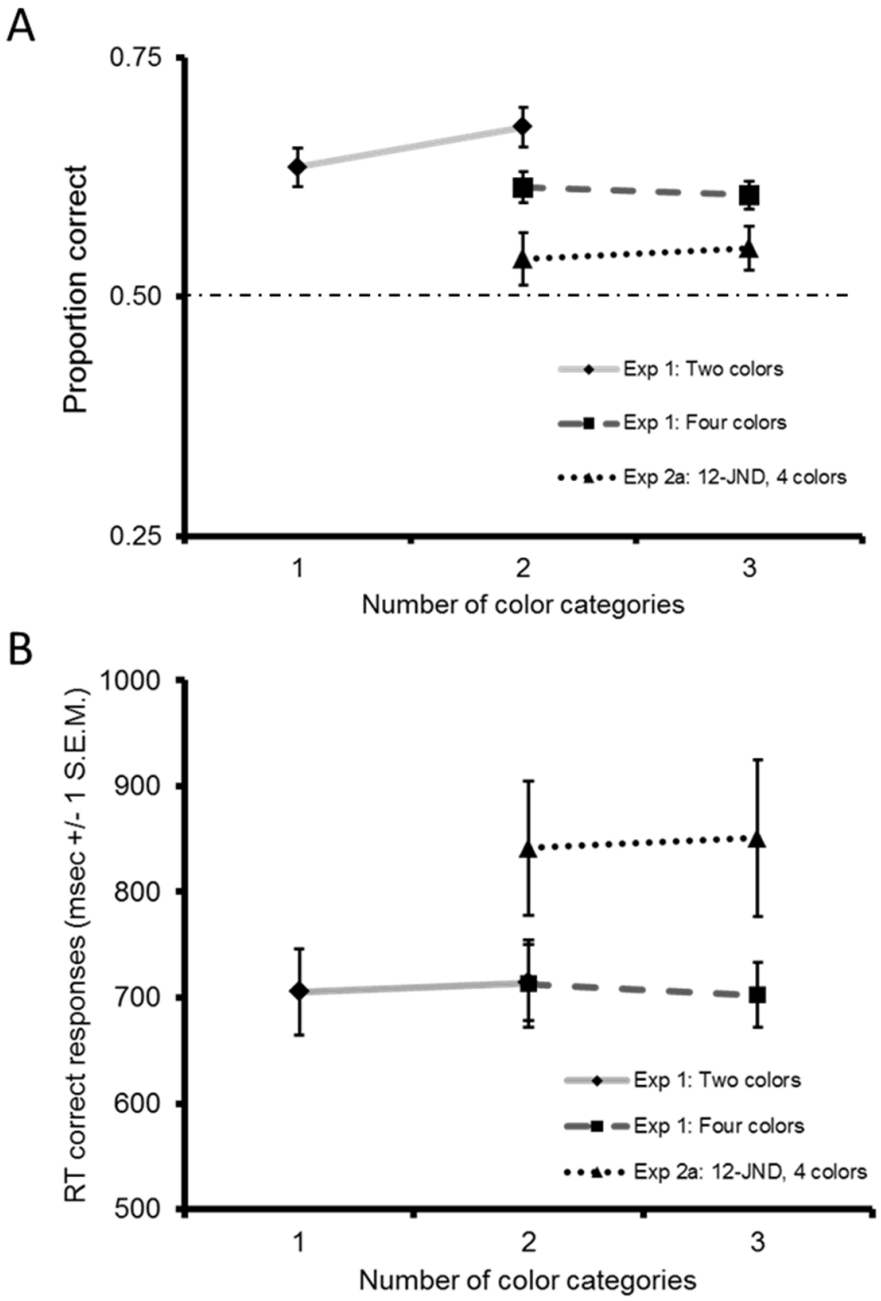

Figure 3.9. Results of experiment 2b – reanalysis of selected conditions from experiment 1 and 2a, with trials re-coded by number of colour categories present in each ensemble. (A) Accuracy: Mean proportion of trials where the mean colour was correctly chosen from the 2AFC. The dotted line represents chance (0.5). (B) RTs: Mean response latency for correct trials. All error bars represent 1 S.E.M.

3.6 General Discussion

3.6.1 Overview of findings

The present study aimed to establish whether observers can extract an accurate mean hue from a rapidly-presented multi-colour ensemble and how this ability is affected by ensembles of varying number of elements, number of colours and the amount of variation in hues in the ensemble.

In experiment 1 both number of elements and number of different colours in ensembles were varied in order to explore the effects of these parameters separately, as well as any interaction between them. The results revealed that observers are able to reliably extract the mean hue from the multi-coloured ensemble, and that increasing the number of elements (essentially giving the observer more exemplars, more of the same information) had no effect on observers' ability to identify the mean hue following rapid exposure to an ensemble of different hues. However, increasing the number of colours in the ensemble was detrimental to the observers' ability to identify the mean hue. This could be due to there being more unique stimuli to average or, alternatively, could be caused by the stimuli being more different from one another. In experiment 2a the range of hues (i.e. the perceptual difference between the most extreme elements of an ensemble) and the number of colours were disassociated, making it possible to observe the effects of variation in stimuli and number of unique stimuli independently. The results demonstrated a crucial role for hue variation in observers' ability to extract the mean, with selection for ranges greater than 20 JNDs being made at chance levels. Even with range of hue controlled, increasing the number of colours still affected observers' ability to select the mean hue, although this effect was small. Finally, in experiment 2b, based on colour naming data from an additional sample, a post-hoc category classification of

ensemble elements was performed in order to assess the accuracy of averaging across colour category boundaries, compared to ensembles consisting of elements of the same category. Results demonstrated no effect of categories on the accuracy of mean selection, suggesting that the averaging process occurs prior to the influence of categories on perception and cognition of colour. The findings of this study have a number of implications for the literature on both ensemble perception and colour cognition. These implications will each be discussed in turn.

3.6.2 Exhaustive processing vs sub-sampling

In experiment 1, the observed consistency in averaging ability despite changes in the number of elements in ensembles supports the proposition that mean hue may be extracted from colourful ensembles using distributed attention. This is contrasted with the prediction that serial, focused attention on individual elements. A basic prediction of a focused-attention account is that, assuming the number of elements sampled is constant (but see Allik et al., 2014), averaging accuracy should decline with more elements as the sub-sample of elements to which attention is paid is more prone to bias. Marchant et al. (2013) found that adding more elements decreased accuracy of mean settings for ensembles where each element had a unique size. Comparing to simulation data, they interpreted this finding as supporting a focused, rather than distributed mode of attention and, hence, a sub-sampling mechanism guiding observer judgments. However their design meant that ensembles with more elements also had more sizes present. In the current study, when number of elements and number of colours varied independently this effect was not observed. Thus we find better support for a distributed-attention account of the ability for averaging hue. Our results combined with results from similar

manipulations with ensembles of different size elements (Utochkin & Tiurina, 2014) suggest that Marchant et al.'s (2013) finding may have been an artifact of the conflation of number of elements and number of unique sizes in their ensembles. It should be noted that there was no evidence for the improvement in performance, or reduction in RT, associated with having more elements, as previously reported for ensembles of different size elements (Robitaille & Harris, 2011). Whether this is a result of a difference in the ensemble processing of these types of stimuli, or some experimental factor is an area for further research.

It is important to note that while focused attention would entail a sub-sampling mechanism, distributed attention is not incompatible with a sub-sampling explanation for the mechanism in general. Although the observed insensitivity to number of elements lends weight to holistic processing explanations (e.g., Robitaille & Harris, 2011), a sub-sampling process immune to increasing elements, is still possible. For example, attention might be distributed across all elements, but only a subsample is included when encoding summary statistics (see also Utochkin & Tiurina, 2014, p. 17). Indeed such a mechanism might help account for down-weighting or extraction of outlying exemplars in other studies of ensemble perception (e.g., de Gardelle & Summerfield, 2011; Haberman & Whitney, 2010; Myczek & Simons, 2008).

3.6.3 Importance of range/variance to ensemble representations

From the current and previous studies it is clear that the mean is not all-important in ensemble perception. Although many studies allude to summary statistics other than the mean (e.g. variance, range, presences of outliers), few have systematically varied those characteristics of ensembles (Haberman & Whitney, 2010; Im & Halberda, 2013;

Maule et al., 2014; Utochkin & Tiurina, 2014; J. Webster et al., 2014). Summary statistics which represent the variance (or other measure of spread around the mean, such as standard deviation) are clearly of relevance to ensemble perception. While it is now relatively uncontroversial to suggest that the mean characteristics (or an estimate of the mean) are often encoded in response to rapidly-presented ensembles of various types of stimuli, it also seems likely that information pertaining to the variation around that mean is also encoded and used to guide perceptual judgments.

The present study has also added weight to the argument that the variance, as well as the mean, has a crucial role in the extraction of summary statistics – demonstrating that the range of colours present in an ensemble has a strong effect on the accuracy of mean judgments. Experiment 2a adds to growing evidence for the importance of range in summary statistical visual processing, and the advantage for accuracy in the 'tight' compared to 'loose' variations of the 20-JND, 4-colour ensembles give a further indication of the role that variance (i.e. inter-element difference) might have in perceptual averaging, independently of range. Similarly it has been shown that adaptation to mean size is weaker when ensembles contain more variance (Corbett et al., 2012), and models of mean size judgments are closer to actual observer performance when internal (i.e. judgment error) and external (i.e. ensemble variance) noise are included as factors affecting the judgment (Im & Halberda, 2013).

The results of the present study suggest that although multiple hues can be represented by their mean hue, this is subject to modulation by the variance (external noise) of the hues in the ensemble. As such, the pattern of results is similar to that expected from models representing ensembles (holistically and their elements individually) as a probability distribution or set of probability distributions, each subject to internal noise

(e.g., Alvarez, 2011; Haberman & Whitney, 2012). Similarly, it has been shown that visual judgments of which of two groups of bars has the greater mean height involves assessment of the relative variance in each set as well as the mean difference, the process of which appears to follow that of Student's t-test (Fouriezos, Rubenfeld, & Capstick, 2008).

3.6.4 Threshold of segmentation

The idea that variance, effectively the perceptual similarity of elements, drives the bias, accuracy and strength of mean representations has previously been expressed in terms of the 'threshold of segmentation' (Utochkin & Tiurina, 2014). According to this theory, the variance of an ensemble is crucial in the coding of an accurate mean representation. When the variance is high, elements are very different perceptually, and if no intermediate elements are included, segmentation occurs, the averaging mechanism fails to form a unifying mean (e.g., J. Webster et al., 2014), and hence alternative decision processes take over when selecting a mean.

Our experiment is well-placed to estimate the threshold of segmentation for ensemble perception of hue as our stimuli are controlled in JNDs. In this case it appears that the threshold at which segmentation occurs and the mean can no longer be extracted from a two-colour ensemble is between 12 and 20 JNDs. This is also reflected by the difference observed between the 'tight' and 'loose' variations of the 20-JND, 4-colour condition. The superior performance in the 'tight' version (where inner colours are separated by 4 JNDs, compared to 12 JNDs in the 'loose' version, see figure 3.6) of this ensemble suggests an important role for variance in whether segmentation or averaging occurs – when the 'inner' colours are more similar to each other averaging is stronger compared

to when they are more different. This suggests that, at least in ensembles with four colours, the extrema may have less impact on the accuracy of average representation – a process which, if true, would imply a mechanism identifying which exemplars differ greatly from the mean (see also outlier exclusion in face ensemble representation in Haberman & Whitney, 2010) serving to down-weight the extrema (e.g. "precision-weighted averaging", Alvarez, 2011; see also de Gardelle & Summerfield, 2011). Understanding which elements contribute most strongly to the representation of the set is certainly a matter that further ensemble perception research should seek to address through systematic manipulation of the inter-element perceptual difference.

It should be noted that in experiment 2a accuracy was slightly above chance in the 28-JND, 8-colour condition while it was at chance for the 2- and 4-colour conditions at the same range. The low accuracy in the 28-JND conditions limits the conclusions somewhat, but such a pattern would point to further support for the idea that smaller inter-element differences support more accurate mean representation, regardless of total range.

Part of the results of experiment 2a run contrary to the findings and theories of mean size perception. In the 12-JND condition, rather than supporting mean encoding, the presence of intermediate colours (compare 4-colour to 2-colour conditions) was somewhat detrimental to performance. The cause of this decrement is unclear since, holistic averaging, weighted-averaging, threshold of segmentation and sub-sampling models would all predict an improvement in the accuracy of mean representation given exemplars which fall closer to the mean. It is possible that this is an effect of the decision-making process when offered the mean and a distractor, both of which bear a quite close resemblance to some of the ensemble elements. As such it may be that the responses to 2-colour ensembles reflect different processes due to the simplicity of the ensemble, while

the more challenging 4-colour ensemble shows a pattern which is better explained by models entailing ensemble coding. In any case, the performance is still above chance, and the main finding, that range has a stronger effect than number of colours on the accuracy of mean encoding, is unaffected.

3.6.5 Below-chance performance

The below-chance level of accuracy for 20-JND, 2-colour ensembles deserves further attention. In this condition the range and number of colours interact to bias the observer systematically to selecting the distractor hue over the mean hue. As the experiment uses a full hue–circle, with a mean colour for ensembles sampled randomly on each trial and counterbalanced distractor locations and perceptual relationships to the mean, irregularities in the salience, perceived lightness or the perceptual spacing of the hues cannot account for a tendency towards choosing the distractor (and particularly it would not account for this phenomenon occurring in this condition only). A speculative explanation for this finding is that, if the colours are sufficiently different, deciding on a mean colour is made more effortful, and, since there are only two element colours to integrate, judgments are subject to be being biased more strongly towards the individual colours in the ensemble itself. Since the distractor is always more similar to one of the two element colours than the mean colour is to either one individually, it may be that the distractor is erroneously chosen due to a bias towards the individual representations of the ensembles hues, rather than the ensemble mean. The fact that this effect disappears when there are intermediate hues also present (i.e. in the 20-JND, 4-colour condition) could be taken to suggest that individual representations no longer bias the choice, perhaps due to the constraints of visual working memory inhibiting the encoding of

individual items (e.g., Alvarez & Oliva, 2008; Attarha et al., 2014; Baijal et al., 2013; Chong & Treisman, 2005a; Corbett & Oriet, 2011; De Fockert & Marchant, 2008). Where the hue range is even greater (i.e. 28-JND, 2 colour) we suggest that the difference between distractor and element hues is sufficiently large that similarity to element colours is not strong enough to bias responses towards either option of the 2AFC, so mean accuracy is near chance.

3.6.6 Circularity

An additional problem for the visual system in the averaging of such radically different hues is the circularity of hue space. In a perceptually-uniform hue circle, averaging the angular position of any two colours (i.e. averaging along the perimeter of the hue circle) has two possible solutions – the 'clockwise' mid-point between those hues, and the 'anti-clockwise' mid-point. In experiment 2a, the 28-JND, 2-colour ensembles recreated this problem – although the hues were separated by 28-JNDs in the direction of the colours presented for the 2AFC mean selection task, they were actually separated by only 20-JNDs in the other direction. It would be expected that, given only two hues, any hue-averaging mechanism would average across the smaller distance and thus the options given in the 2AFC is unlikely to have reflected the mean encoded. The fact that observers were at chance for this condition may reflect this. However, the presence of intermediate hues (in the 4-colour condition of the 28-JND range) should anchor the direction of hue averaging, providing a better guide as to which direction the mean should be approached from. Observers were equally unsure in this condition though – with accuracy again falling at chance levels. The circularity of hue space, therefore, is probably partly

responsible for the breakdown of averaging at wider ranges. However the present data show that averaging along the hue circle is still reliable where the range is small.

3.6.7 Common mechanism of ensemble perception

In spite of the circular representation of hue and application of categorical labels to colour perception, the results of this study suggest a similarity between the mechanism responsible for ensemble coding of colour and ensemble coding of other features, such as size, length, orientation and facial expression. Colour averaging appears to be rapid, is insensitive to changes in number of elements indicating distributed attention, and is range-limited but with a sensitivity to variance, or inter-element difference. We find no evidence that the categorical relationships of ensemble elements affects mean encoding. When stimuli are appropriately controlled for perceptual difference, colour categories only affect post-perceptual processing (He et al., 2014), so the lack of a category effect is further support for early encoding of the mean.

Although the circularity of colour space may be partly responsible for the breaking down of averaging at higher ranges, the results can still be compatible with models incorporating both internal and external noise, such as the threshold of segmentation (Utochkin & Tiurina, 2014). Therefore, in spite of the differences between colour perception and that of other, more linearly-represented features, it does appear that ensemble coding for colour shares a common mechanism. Whether this mechanism is located in a specific part, or parts, of the brain which deals with all summary statistics, or whether it is an emergent property of the organization of the visual cortex is a subject for further research.

3.6.8 Purpose of colour averaging

Evidence that consistent summary statistics help speed visual judgments (Michael et al., 2014) and facilitate visual search (Corbett & Melcher, 2014b) support the idea that the function of an ensemble coding mechanism able to rapidly extract summary statistics is to tune the visual system to the characteristics of the environment. This proposition is also supported by evidence for adaptation to the mean of an ensemble (Corbett et al., 2012), evidence that ensemble means may be used as 'units' of working memory (Im & Chong, 2014; Im et al., 2014) and are represented across multiple frames of reference (Corbett & Melcher, 2014a). Recent evidence also suggests that ensemble means can be used by 4-5 year-olds to guide their perceptual judgments (Sweeny et al., 2014), adding weight to the argument that ensemble coding is a pervasive feature of the visual system and, in some cases, averaging appears to be an obligatory process (e.g., Allik et al., 2014; Parkes et al., 2001) requiring little or no attention to the features in question (Alvarez & Oliva, 2008; Baijal et al., 2013).

For colour perception, ensemble coding may also help tune the visual system to the environment. It has long been proposed that the average colour of a scene could be used by the visual system to estimate the characteristics of the light illuminating the scene, and thus support colour constancy (e.g. "gray-world" hypothesis; Buchsbaum, 1980). It has also been demonstrated that modulation of the variance of colours surrounding another is sufficient to induce a change in colour appearance of that focal patch (Brown & MacLeod, 1997), therefore representation of the mean colour of a scene may play a role in colour constancy. Colour summary statistics could also be relevant to the perception and memory of surface colours (e.g., Milojevic, Ennis, & Gegenfurtner, 2014). It is clearly inefficient and unnecessary to recall all of the precisely and subtly varying colours of an

object in order to recognize it in future. Instead a summary representation could provide an adequate prior to aid these processes (Olkkonen & Allred, 2014; Olkkonen et al., 2014). Additionally, although categories do not appear to affect ensemble coding of colour, it is plausible that ensemble coding may inform our categorical structure. Infants as young as 10 months are known to be able to form category prototypes based on the mean features of a series of successively-presented stimuli (Younger, 1985). Similarly, adult colour category boundaries have been shown to shift towards the center of the range of stimuli offered (Wright, 2011). Therefore it may be that ensemble statistics are used within the context of learning colour categories, and tuning them to the environment.

3.7 Conclusion

The present study has shown that a mean colour can be selected with above-chance accuracy from rapidly-presented ensembles containing multiple hues. The veracity of the process appears to be limited by the range, or variance of the colours shown in the ensemble. Further research into ensemble coding (in any domain) should focus not solely on the mean (as has been common), but should also consider the role of the range and variance of the scene in the extraction of summary statistics, with particular view to its contribution to segmentation of ensembles and the contribution of ensemble coding to visual stability. Future research on ensemble perception of colour will present a unique opportunity to better understand the extent to which ensemble coding might be influenced by categorical relationships between ensemble elements and how ensemble coding might operate in the real world with regard to surface colour perception and memory.

Chapter 4

Paper 3: Accurate rapid averaging of multi-hue ensembles is due to a limited capacity sub-sampling mechanism

Maule, J., & Franklin, A. *Journal of the Optical Society of America: A* (submitted)

4.1 Abstract

It is claimed that the extraction of average features from rapidly-presented ensembles is holistic, with attention distributed across the whole set. We investigated whether observers' extraction of mean hue is also holistic or could reflect sub-sampling. Analysis of selections for the mean hue revealed a distribution which peaked at the expected mean hue, but with a degree of error. An ideal observer simulation showed that a sub-sampling mechanism incorporating just two items from each ensemble would suffice to reproduce the performance of most observers. The results imply that averaging of hue does not occur as efficiently as for other domains.

4.2 Introduction

Ensemble perception describes the extraction of summary statistics from a set of items varying in some stimulus dimension, typically in the absence of representation of the individual items and with very short stimulus presentation time (Haberman & Whitney, 2012). For example, observers can extract the mean size from a set of circles of different sizes seen for 500ms, but are relatively poor at identifying individual members of the set (Ariely, 2001). Sensitivity to summary statistics has been demonstrated in a variety of perceptual domains, including size (e.g., Ariely, 2001; Chong & Treisman, 2003; Corbett & Oriet, 2011; Marchant & De Fockert, 2009), orientation (e.g., Dakin, 2001; Parkes et al., 2001), facial expression (e.g., De Fockert & Wolfenstein, 2009; Haberman & Whitney, 2010; Leib, Puri, et al., 2012), facial identity (e.g., De Fockert & Wolfenstein, 2009; Fiorentini et al., 2012; Leib et al., 2014), and colour (e.g., Maule & Franklin, 2015; Maule et al., 2014; J. Webster et al., 2014). Much of this research has focused on whether the mean value of an ensemble has a special perceptual salience. The encoding of a mean in spite of a lack of individual item representation has led to suggestions that the ensemble perception mechanism could operate outside of the limits of focused attention, instead using distributed attention to process sets holistically (see Alvarez, 2011; Treisman, 2006). However this mechanism is subject to debate, with various researchers pointing out that a mechanism combining focused attention with sub-sampling might be adequate to explain observers' performance on perceptual averaging tasks, without the need to postulate a new holistic processing mechanism (De Fockert & Marchant, 2008; Marchant et al., 2013; Myczek & Simons, 2008; Simons & Myczek, 2008).

Summary statistics of colour are likely to be of relevance to the visual system. For example, the colour variance in surrounds is known to modulate the appearance of individual colours (Brown & MacLeod, 1997; Ratnasingam & Anderson, 2015), and priming by the variance of colour present in a rapidly-presented ensemble has also been reported (Michael et al., 2014). The mean colour of a scene may also play a role in the estimation of the illuminant, necessary for colour constancy ("gray world hypothesis", Buchsbaum, 1980), and in colour memory (Olkkonen et al., 2014). We have previously shown that, when observers are presented with an ensemble of two different hues for a short time (500ms), they tend to have a bias in their memory of which hues were present in the ensemble towards the mean hue even if that mean hue was not present (Maule et al., 2014). We have also shown that observers can reliably identify the unseen mean hue of an ensemble when that hue is paired with a similar distractor hue (Maule & Franklin, 2015). Both of these studies found an effect of the range of hues in the ensemble where the mean bias and mean identification accuracy were both reduced when the range of ensemble hues was increased (see also J. Webster et al., 2014). We further found that there is no impact of increasing the number of elements in an ensemble – observers were able to identify the mean equally reliably whether required to average 4, 8 or 16 patches of colour (Maule & Franklin, 2015). The robustness of mean identification ability to changes in number of elements has also been demonstrated for ensemble perception of size (e.g., Ariely, 2001; Chong & Treisman, 2005b; Marchant et al., 2013; Robitaille & Harris, 2011; Utochkin & Tiurina, 2014) and faces (Haberman & Whitney, 2009; Leib et al., 2014), and is suggestive of an efficient mechanism where processing occurs in parallel, across the whole display and all items (Treisman, 2006).

Although we have shown that mean identification is above chance on a two-alternative forced-choice (2AFC) task (Maule & Franklin, 2015), no study has directly investigated

the accuracy of mean representation following rapidly-presented ensembles. Kuriki (2004) has shown that adjustments to mosaics with many tiny elements were not reflective of the colourimetric mean, being biased towards the most saturated element (see also Sunaga & Yamashita, 2007). However this was under continuous viewing conditions, rather than the rapid-exposure of the ensemble perception paradigm.

We know that number of elements has no effect on the identification of the mean colour given a 2AFC (Maule & Franklin, 2015) and Ariely (2001, 2008) reasoned that a sub-sampling mechanism with a fixed sample size, should extract the mean with an accuracy proportional to the set size. However, ideal-observer simulations have suggested that the performance of actual observers in a number of experiments showing rapid extraction of mean size (Ariely, 2001; Chong & Treisman, 2003, 2005a, 2005b) could be explained by a limited sub-sampling mechanism with a sample size as small as just one or two items from each set (Marchant et al., 2013; Myczek & Simons, 2008). Similarly, Marchant and de Fockert (2013) showed that their finding that mean size estimates are affected by set size, for irregular sets (ensembles in which all elements have a unique size) but not regular sets (where some elements are the same size), can be predicted by a limited sub-sampling model. Other simulations (e.g. Haberman & Whitney, 2010; Im & Halberda, 2013) have attempted to better characterize the process of mean estimation by including internal noise into simulations – i.e. the "judgment error" (Ariely, 2008) present in all psychophysical measurements. Simulations of sub-sampling which incorporate or estimate internal noise as part of the model tend to perform less well compared to real observer data and suggest that larger sub-samples (around seven items), would in fact be required to simulate the averaging performance of human observers (Im & Halberda, 2013, but see also Allik et al., 2013). Likewise, experimental evidence also suggests that observers still outperform subsampling expectations even when explicitly instructed to use such strategies in

ensemble perception tasks (Chong et al., 2008), and when ensembles contain a manageable range of stimuli (Utochkin & Tiurina, 2014). Such simulations have generally shown that a sub-sampling mechanism or strategy cannot account for the level of performance observed on ensemble perception tasks in these domains (faces and size).

The present study investigated whether adjustments to the mean hue for rapidly-presented ensembles are equivalent to adjustments to a single hue. A distinction is drawn between accuracy and precision in this study. Accuracy describes the tendency for settings to average at the expected value (the particular hue shown for single hues, or the mean hue for heterogeneous ensembles) and is indicated by the position of the peak in the distribution of settings relative to the expected mean/actual colour. Precision describes the amount of error in the settings, and is indicated by the width, or standard deviation, of the distribution of responses. If both the accuracy and precision of settings of mean hue is similar to the accuracy and precision for single hues it will be a strong indicator that the mean hue is encoded rapidly and as strongly as individual hues, suggesting that the ensemble is represented by a single average hue. These measurements of accuracy and precision were also used to address the question of whether the observed hue averaging precision could be the result of a limited sub-sampling mechanism or whether the performance could support the proposal of a holistic mechanism, integrating attention from the whole ensemble and circumventing the limited capacity of visual working memory (Alvarez, 2011; Attarha et al., 2014; Baijal et al., 2013; Im & Chong, 2014). Measurements of internal noise (based on adjustments to single hues) were incorporated into a simulation estimating the precision of mean adjustments based on a random sub-sample of *n* elements. By comparing the simulation results to the precision of real observers in the ensemble adjustment task it was possible to estimate the sample size required to explain their performance by sub-sampling.

4.3 Methods

4.3.1 Participants

Fifteen observers (three males) of average age 20.5 years (SD = 2.97) took part in the experiment. All reported normal or corrected-to-normal visual acuity and were assessed as having normal colour vision using the Ishihara plates (Ishihara, 1973) and City University test (Fletcher, 1980). All spoke English as their first language. Participants received either payment at a rate of £7.50 per hour or course credit. The research protocol was approved by the University of Sussex Sciences and Technology Cross-Schools Ethics Committee.

4.3.2 Stimuli

All colours were taken from a set of 48 hues specified from a circle on an equiluminant plane in Derrington-Krauskopf-Lennie (DKL) space (Derrington et al., 1984; Krauskopf et al., 1982). In order to ensure that the colours were approximately equally discriminable, and thus provide uniform perceptual differences between the hues presented in ensembles, hue discrimination threshold data from Witzel and Gegenfurtner (2013) were used to space the selected hues by 1 just-noticeable difference (JND) (see figure 4.1). Throughout the experiment the background was a uniform grey (xyY (1931): 0.310, 0.337, 30.039), as used by Witzel and Gegenfurtner.

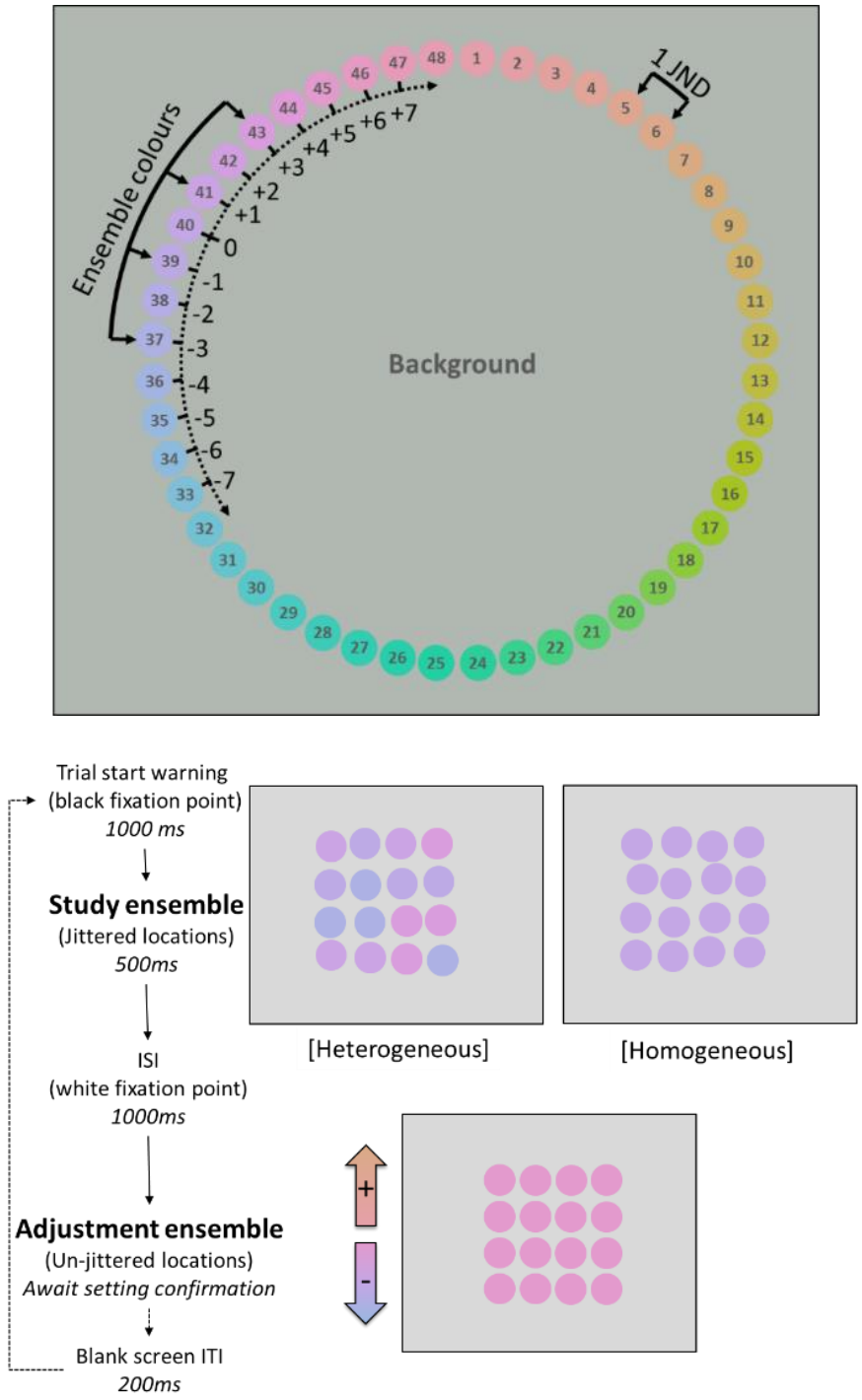

Figure 4.1. Upper panel – an approximate rendering of the 48 hues, and the background, as used in the experiment, arranged in a continuous hue circle. Adjacent hues are separated by 1 JND. The solid black line towards the top left indicates the selection of hues for ensembles – each ensemble had four different hues, drawn from a 6-JND span with 2-JNDs between each member. This arrangement moved at random

around the hue circle on each trial to present ensembles with different mean hues, but with the spacing of the member hues yoked in the way shown. The dotted line inside represents the adjustment phase at which participants could select any hue from the circle, moving in single JND steps in the positive (clockwise) or negative (counterclockwise) direction. These responses are coded according to their JND-distance from the ensemble mean, which was assumed to fall at the middle of the distribution of ensemble colours. Lower panel – the order and timing of events in a single trial of the ensemble task. JND = Just-noticeable difference; ISI = inter-stimulus interval; ITI = inter-trial interval.

4.3.3 Apparatus

A 22-inch Mitsubishi DiamondPlus 2070SB Diamondtron CRT monitor, set to a resolution of 1600 x 1200 pixels, 24-bit colour resolution, and a refresh rate of 100 Hz was used. A Cambridge Research Systems ColourCal colourimeter was used to measure the monitor gamut and primary outputs, gamma correction applied, and look-up tables generated to automatically estimate the RGB primary values required to render each desired stimulus colour. The experiment took place in a blacked-out room with the monitor the only source of light. A cardboard viewing tunnel lined with black felt obscured peripheral objects from the participants' view and a chin rest was used to maintain a viewing distance of approximately 57cm. Participants gave their responses using the keyboard.

4.3.4 Design

Ensembles consisted of sixteen coloured circles (elements) each allocated to a cell in an invisible 4-by-4 grid centered on the screen. Elements subtended 1.75° visual angle and were spatially jittered by up to 0.25° visual angle horizontally and vertically from the center point of the cell to remove the appearance of a regular structure in the ensemble. Ensembles contained either elements all of one hue (homogeneous trials) or four hues, i.e. four elements of each hue, arranged randomly (heterogeneous trials).

The task used the method of adjustment, in which participants first saw an ensemble and then attempted to match the average colour of the ensemble in a subsequent display. The adjustment display was an ensemble of 16 elements arranged in a 4-by-4 grid (un-jittered), and all elements of the adjustment display were the same colour. Trials began with the presentation of a black fixation point in the center of the display for 1000ms, immediately followed by the presentation of a 'study' ensemble for 500ms. An inter-stimulus interval lasting 1000ms was indicated by a white fixation point and then replaced by the adjustment ensemble. The initial colour of the elements of the adjustment ensemble was selected at random from a range +/-7 JNDs from the actual mean of the ensemble. By pressing the left and right arrow keys participants were able to adjust the colour of the elements in the ensemble, around the hue circle in 1 JND steps. The space bar was used to confirm their selection for that trial.

Participants took part in five blocks of trials. Each block presented trials from a list comprising 48 heterogeneous ensembles (i.e. ensembles with a mean corresponding to one of the 48 hues in the stimulus set) and all 48 homogeneous ensembles, in a random order.

4.3.5 Procedure

Participants read instructions on the screen prior to the task. The instructions stated that participants should pay attention to the initial ensemble and then "adjust the dots until they match the average colour of the first set" for heterogeneous ensembles, or "match the colour exactly" for homogeneous ensembles. Before beginning the main task, participants completed 10 practice trials using ensembles of achromatic stimuli, varying in lightness relative to the background (8 – 48 cd/m^2, in 4 cd/m^2 steps). In order to help the participant understand the nature of perceptual averaging the practice included feedback to indicate if the participants selection was "correct" (at the mean lightness, also the mid-point of the range shown in the ensemble), "close" (within one step of the correct response) or "incorrect". Participants were required to be "correct" or "close" on each of the last three practice trials in order to proceed to the main task, otherwise the practice was repeated. No feedback on performance was given during the colour task. The time spent reading instructions and completing practice trials ensured adaptation to the white point.

4.4 Results

4.4.1 Homogeneous vs heterogeneous ensembles

Observer settings were coded by their absolute error from the actual mean of ensembles, in terms of 1 JND steps around the hue circle. For heterogeneous ensembles this was assumed to be the mid-point of the distribution of hues which were present in the ensemble, for homogeneous ensembles this was the hue matching those used in the ensemble. Mean absolute error (i.e. error in either hue direction from the correct mean)

was significantly greater for the heterogeneous ensembles (M = 2.02, SD = 0.25) than the homogeneous (M = 1.34, SD = 0.21) ($t(14) = 9.44$, $p < .001$). This can be seen in the data presented in figure 4.2 (selected individuals) and figure 4.3 (average observer) – the distribution of selection errors around the mean is greater (a wider normal curve with a greater standard deviation) in the heterogeneous condition compared to the homogeneous.

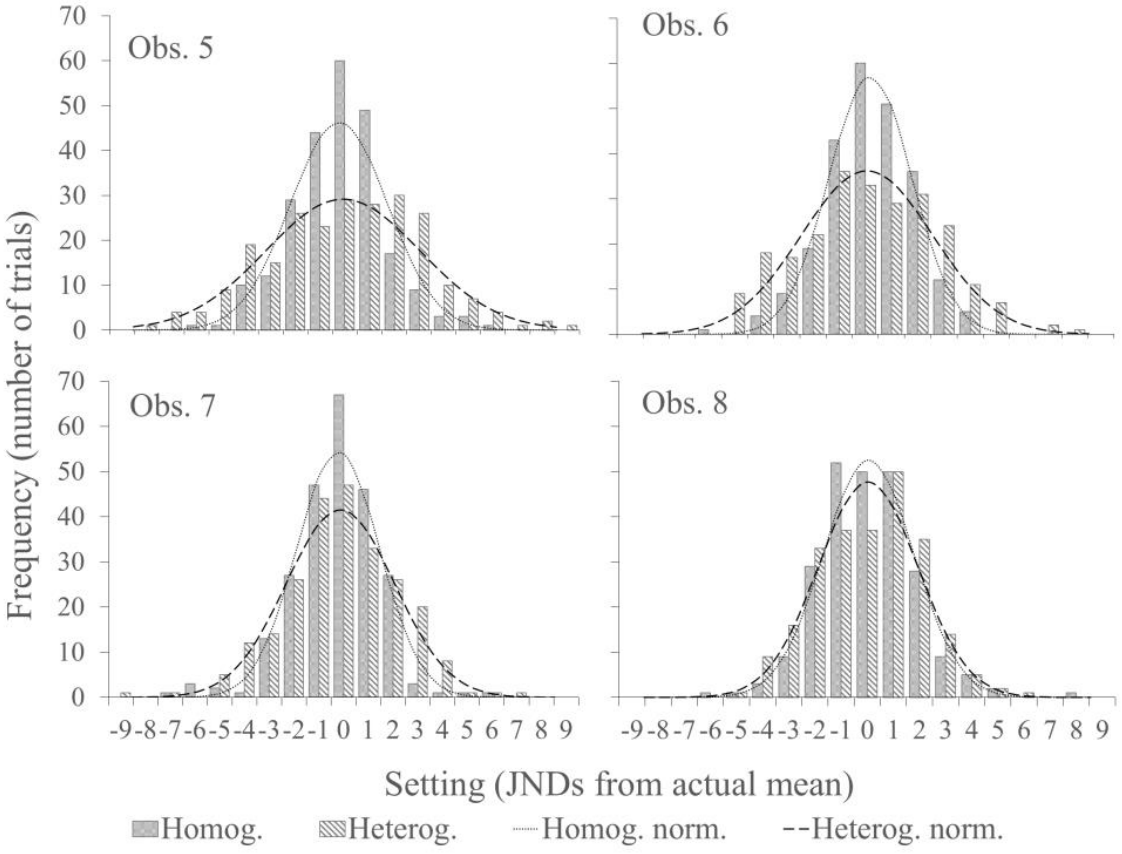

Figure 4.2. Error distribution histograms for homogeneous and heterogeneous conditions for four example observers typical of the whole sample. Dashed curves indicate normal distributions with a mean and standard deviation (SD) equal to that for each observer and condition. N.B. Settings outside the range of +/-9 JNDs are not displayed by the histogram, but do contribute to the mean and SD of the normal curves.

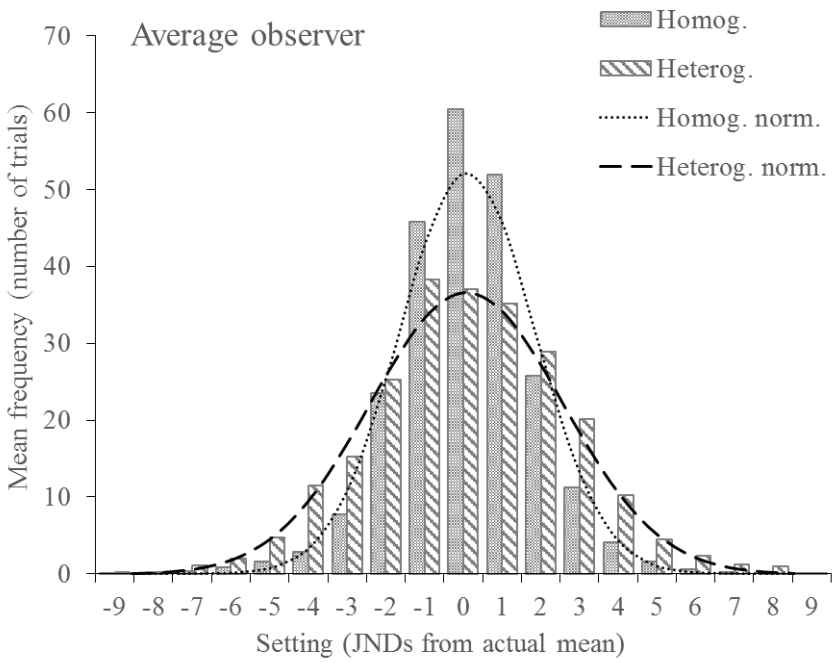

Figure 4.3. Error distribution histograms for homogeneous and heterogeneous conditions for the average observer. Bars are based on mean frequency of response across observers. Dashed curves indicate normal distributions with a mean and standard deviation (SD) equal to the mean for each observer and condition. N.B. Settings outside the range of +/-9 JNDs are not displayed by the histogram, but do contribute to the mean and SD of the normal curves.

4.4.2 Simulation of limited-capacity sampling strategies

In order to evaluate observers' performance in the heterogeneous condition (when they are required to pick a single colour to represent a multi-colour ensemble), relative to the homogeneous condition (where they needed simply to match the single colour present in the ensemble), an ideal observer simulation was carried out. This analysis sought to ascertain how many single ensemble elements an observer would have to sample in order to achieve performance at the level observed in the heterogeneous condition.

Two models were used for the simulation (figure 4.4). Both involved a *sampling* of elements from an ensemble composed exactly as in the adjustment experiment, followed by *averaging* of that sample, and finally *selection* from the available hues. The *early noise* model (Haberman & Whitney, 2010) applied noise to the representation of the colours at the sampling stage, such that each sampled element would be represented by a value selected from a normal distribution with a mean equal to the true element value and a standard deviation (SD) equal to that observed for settings in the homogeneous condition. The *late noise* model (Haberman & Whitney, 2010) applied noise after the averaging stage, such that the colour representing the whole ensemble was subject to noise prior to selection. In both models, selection involved rounding to the nearest integer. Simulations were run for each observer, using their individual SD from the homogeneous condition, for 10,000 trials.

The results of the simulation are considered in terms of accuracy of performance, summarized by the standard deviation of error from the true mean in the simulated adjustment settings. By comparing these to the standard deviation of adjustment settings in the observer data it is possible to evaluate, given a limited-capacity sampling strategy or mechanism, how many elements an ideal observer would need to sample in order to reach or exceed the level of performance exhibited by the observers during the mean adjustment task.

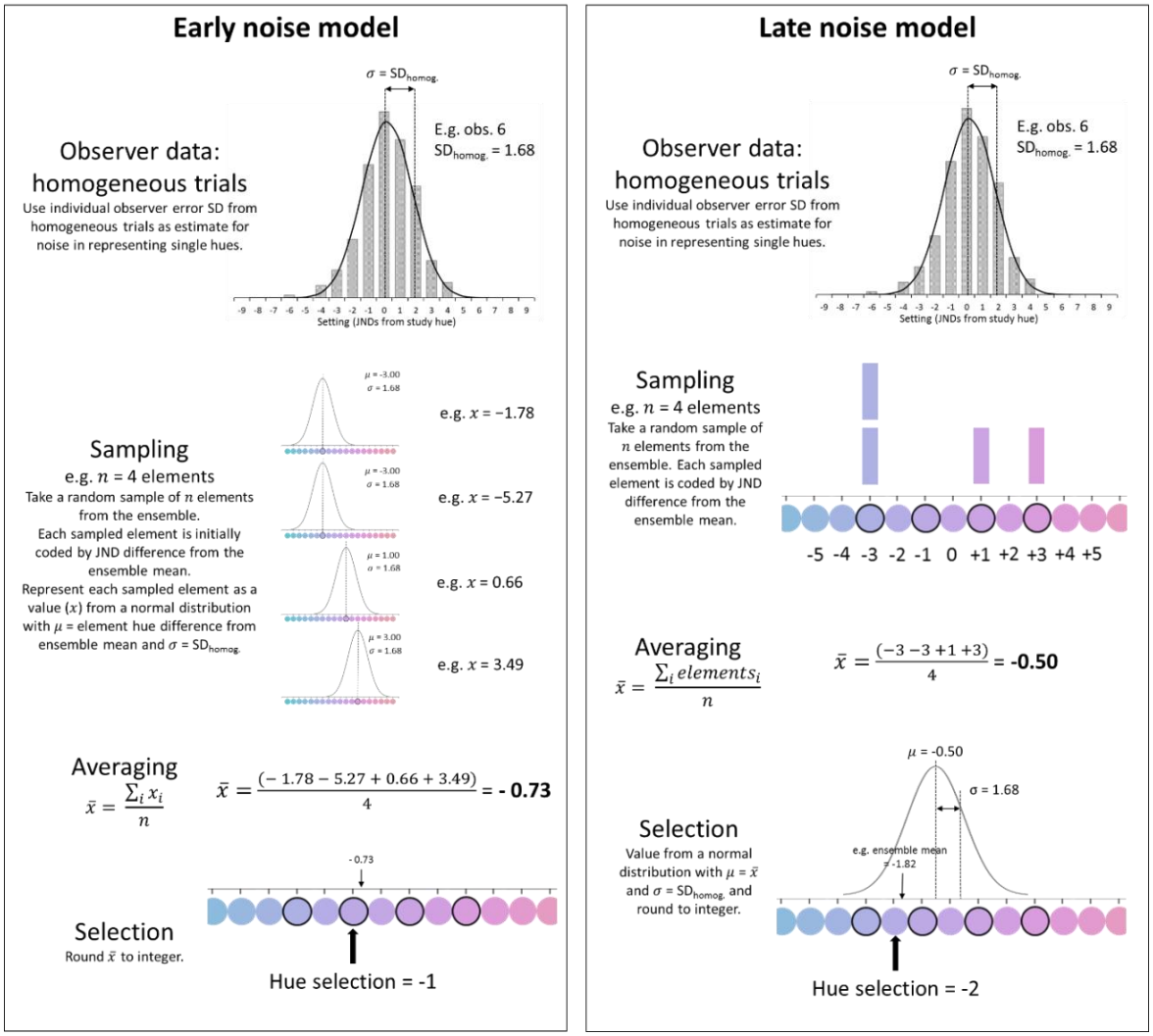

Figure 4.4. Schematic representation of the early and late noise simulations. In the early noise model noise is added to each sampled element prior to averaging. In the late noise model each element is first averaged, after which noise is added to the mean representation. Selection of the eventual mean hue response requires rounding to the nearest integer. Noise is equivalent to the observed standard deviation of settings from each real observer's responses to homogeneous (i.e. single-hue) ensembles. Both panels represent a single exemplar trial where the same 4 elements are sampled from an ensemble and noise is based on observer 6. Note that the simulation was run for each observer and at sample sizes of 1-16 elements, for 10,000 trials each. n = number of samples; μ = mean of normal distribution

indicating a noisy representation of a hue; σ = standard deviation of normal distribution; x = value assigned to a sampled element prior to averaging; $\bar{x}$ = calculated value for mean hue. All values given are in terms of JNDs from the true mean hue of the ensemble.

The simulation revealed that most observers were performing at a level equivalent to sampling between one and two elements from each ensemble. This was true of both the early and late noise models (see table 4.1). Figure 4.5 shows the simulated data for four observers with the actual performance also plotted for comparison. The simulation data show that there are diminishing returns from taking more and more samples, and in the case of the late noise simulation, an optimum number of samples is reached at around six or seven elements. Importantly, however, only one observer (obs. 8) exhibited performance near this optimal level.

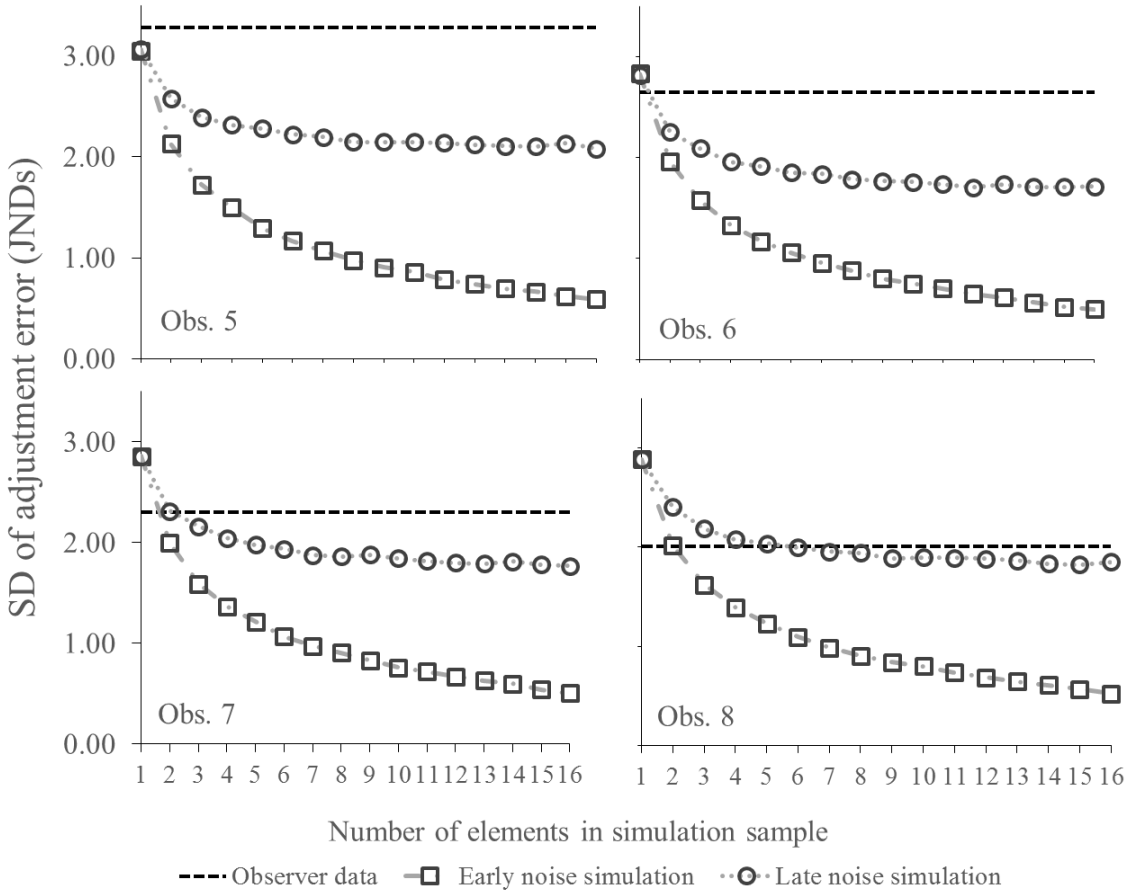

Figure 4.5. Simulation results and actual data for four observers. The top left panel shows observer 5, who performs worse in the ensemble condition than would be predicted if they made their settings based on sampling just one ensemble element. Observer 6 (top right panel) and 7 (bottom left panel) perform at a level equivalent to sampling 1-2 ensemble elements (based on either late or early noise). Observer 8 (bottom right panel) performs at a level equivalent to sampling two elements in the early noise model, or between 5 and 6 in the late noise model.

Table 4.1. Actual observer performance vs. limited-capacity simulation performance.

	Actual performance (SD of error in JNDs)		Simulation performance – number of elements in sample								
			Early noise			Late noise					
Obs.	Homog.	Heterog.	1	2	3	1	2	3	4	5	6
1	1.73	2.48	2.85	**1.98**	1.57	2.83	**2.36**	2.13	2.04	1.96	1.90
2	1.55	2.74	**2.72**	1.91	1.51	**2.74**	2.19	1.99	1.86	1.77	1.76
3	1.46	2.08	2.65	**1.88**	1.49	2.69	2.14	**1.93**	1.78	1.72	1.68
4	1.95	2.69	3.02	**2.08**	1.69	2.98	**2.50**	2.30	2.20	2.15	2.08
5	2.07	3.28	**3.05**	2.13	1.72	**3.06**	2.58	2.39	2.32	2.28	2.22
6	1.68	2.64	2.83	**1.96**	1.58	2.82	**2.25**	2.10	1.96	1.92	1.85
7	1.76	2.31	2.86	**2.00**	1.59	2.86	**2.31**	2.16	2.05	1.98	1.94
8	1.82	2.01	2.89	2.02	**1.62**	2.89	2.40	2.19	2.08	2.04	**1.99**
9	1.55	2.27	2.73	**1.88**	1.54	2.75	**2.17**	1.98	1.90	1.78	1.74
10	1.36	2.52	2.63	**1.81**	1.47	2.63	**2.06**	1.84	1.71	1.64	1.58
11	2.51	3.16	3.39	**2.35**	1.91	3.39	**2.94**	2.82	2.70	2.68	2.63
12	1.76	2.83	2.84	**1.98**	1.59	**2.82**	2.34	2.14	2.04	1.96	1.92
13	2.27	2.94	3.20	**2.22**	1.80	3.20	**2.78**	2.57	2.53	2.48	2.39
14	2.13	2.40	3.06	**2.16**	1.75	3.08	2.64	2.45	**2.40**	2.32	2.28
15	2.02	2.98	3.01	**2.13**	1.69	3.01	**2.53**	2.38	2.25	2.24	2.14

Table notes: SD = standard deviation; JNDs = just-noticeable differences; Obs. = observer; Homog. = homogeneous condition; heterog. = Heterogeneous condition. Values in table correspond to the SD of the error in the selected mean for ensembles. Simulation columns correspond to how many elements were included in the sample taken from the ensemble prior to averaging, with values in **bold type** indicating the point at which the simulation performance equals or exceeds (i.e. lower SD) that taken from the actual observers settings for heterogeneous ensembles (compare with column marked "Heterog.").

4.5 Discussion

This study had two main aims. The first was to compare the precision and accuracy of settings of the mean hue of a rapidly-presented ensemble of different hues to the settings for a single hue. The data show that, on average, observer settings tended to the mean colour highly accurately. Settings peaked around the expected mean hue, with error

distributed symmetrically either side indicating no bias or skew to the settings. The same pattern was found for the homogeneous condition, and analysis revealed the mean setting of the two conditions to be no different, indicating equal accuracy for mean hue reproduction as for single hue reproduction. However the variance of settings was greater in the heterogeneous condition than in the homogeneous, indicating a difference in precision between these conditions. Settings of a mean hue were less precise than for single hue, indicating that reproducing the mean hue was subject to more error than reproducing a match for a single hue.

The second aim of this study was to establish whether a limited sub-sampling mechanism could explain the observers' performance on the hue averaging task. The results of the simulation suggest that a sub-sampling mechanism where attention is devoted to encoding and averaging no more than two elements would be sufficient to provide estimates of the mean with precision equal to, or better than, most observers. Therefore, while the observers are clearly able to pick a mean hue following a rapidly-presented ensemble, and those selections converge on the true mean across trials, our data do not provide support for the proposal of a holistic hue averaging mechanism using distributed attention, or a mechanism with a capacity beyond the limits of visual short-term memory (Alvarez, 2011).

Our early noise simulation is most similar to that used by Haberman and Whitney (2010), using measurements of error in setting a homogeneous ensemble as noise applied to each sampled element. We also included a late noise simulation, where the internal noise was applied after averaging had taken place. In reality, noise is present at both of these stages, but the measure of internal noise taken from the homogenous condition conflates these two sources of error, meaning that it would not be possible to include an accurate estimate

of the noise at both stages in a single simulation. As can be seen from figure 4.5, with increasing set sizes, the late noise simulations asymptote at a higher level of error than the early noise simulations. This is because when noise is applied independently to each sampled element it is then subject to noise cancellation, where noise in the positive direction for one element is cancelled out by noise in the negative direction for another. In contrast there is no noise cancellation in the late noise model. As the sample size approaches the whole set, the late noise simulation predicts that precision will improve to equal that observed in the homogeneous condition, while the early noise simulation predicts that averaging precision will be higher than for the homogeneous ensembles. Therefore, if it was necessary to prefer one of the two, the late noise simulation would seem to make more realistic predictions than the early noise simulation. This difference notwithstanding, the conclusion with regard to observer performance is similar for both simulations – observers are sufficiently imprecise in their mean hue settings that the difference between simulations is trivial.

Our finding is somewhat at-odds with other simulations incorporating internal noise, which have found sub-sampling models of ensemble perception of faces (Haberman & Whitney, 2010) and size (Im & Halberda, 2013) underperform on averaging accuracy relative to real observers with sample sizes fewer than seven elements. Notably, Haberman and Whitney (2009) report that discrimination for the mean emotional expression from an ensemble was at least as good as discrimination for individual expressions – a trend not evident in our data, where precision for homogeneous ensembles (single colours) was better than for heterogeneous.

It is unlikely that a sub-sample would be taken from an ensemble at random, i.e. some elements may contribute disproportionately to the mean estimation (weighted-averaging,

e.g., de Gardelle & Summerfield, 2011), or be more likely to be selected for a sub-sample. Attention cued to individual items has been shown to affect mean size estimates (De Fockert & Marchant, 2008), and averaging of size over time has been shown to be biased towards looming phases, perhaps because these are more salient (Albrecht & Scholl, 2010). Variations in the salience of individual hues would not affect the overall accuracy of mean adjustment in this experiment as the position of the ensemble elements rotates through every possible colour in the stimulus circle (in a random order), however it could exaggerate deviations from the expected mean if those colours are assigned higher weights when averaging. The salience of each hue in the present study should be approximately equal in this study (stimuli are equated for luminance and equally distant from the white-point in DKL space). As DKL space is not scaled to equate saturation there may be some residual variation in salience of hues around the hue circle. However this variation is gradual around the hue circle, meaning that local saturation differences (i.e. the difference between neighboring hues) is subtle.

It is possible that hue is not as apt to be averaged using holistic sampling as size or faces. There are some differences between hue and other domains which may be responsible for differences in ensemble processing, but also several similarities. Unlike size, hue is a matter of qualitative experience, rather than magnitude. Saturation and lightness may both be described in terms of magnitude or intensity, so one colour can be said to be "more saturated" or "lighter" than another. In contrast, hue is a circular dimension, requiring reference to colour categories to describe relationships. Therefore, given highly distant exemplars (e.g. red and green) it may not be easy to imagine what the mean should look like. As the angle (e.g., in DKL space) between hues to be averaged approaches 180 degrees averaging could become increasingly difficult, or impossible, as the elements now represent opponent colours with qualitatively different sensations which do not blend

into a meaningful average. Although the qualitative and circular nature of hue perception seems a plausible reason that averaging would be harder, these do not necessarily preclude hue from rapid, holistic averaging and there remain similarities with other domains. For example, hue averaging ability is reduced by increased ranges of hue in ensembles (Maule & Franklin, 2015; J. Webster et al., 2014), but this is also the case for size (Utochkin & Tiurina, 2014), and the stimuli in the present experiment were within the range at which mean selection from a 2AFC is reliable (Maule & Franklin, 2015). Hue is subject to categorization (e.g. Bird et al., 2014), however our previous study has demonstrated that there is no effect of colour categories on mean selection (Maule & Franklin, 2015). Face perception is also somewhat qualitative (in terms of emotions and identity), and is widely understood in terms of norm-based coding accounts, which rely on extraction of the mean (for a review see Tsao & Freiwald, 2006). Like norm-based models of face coding, colour perception is subject to white-point adaptation which supports colour constancy (M. A. Webster & MacLeod, 2011). Nevertheless it appears that efficient averaging is evident for faces, and not hue, at least under the conditions tested here. In short, there do not appear to be particular features of hue perception which can be said to account for the poor precision of average hue settings, particularly given the evidence for holistic ensemble perception in other domains.

It should be noted that just because sub-sampling could explain the results in this experiment, it does not necessarily imply that holistic averaging of hue cannot or does not take place (Haberman & Whitney, 2009). Evidence from other domains suggests that averaging may be most reliable when the summary statistics are incidental to the main task. Summary statistics can have effects on response times and accuracy even when observers are not instructed to judge the mean or extract the gist at all. For example, response times for ensemble classification (“red” or “blue” average) can be reduced when

a prime ensemble with the same variance is presented beforehand, even when the prime ensemble has a different mean colour (Michael et al., 2014). There are many other examples of tasks in which implicit processing of the summary statistics of sets appears to influence responses (Allik et al., 2014; Alvarez & Oliva, 2008, 2009; Ariely, 2001; Corbett & Melcher, 2014a, 2014b; Corbett et al., 2012; Lanzoni et al., 2014; Maule et al., 2014; Oriet & Brand, 2013; Parkes et al., 2001). It may be that instructing observers to consider and retain an average hue results in the use of a sub-sampling strategy, whereas observers may perform better relative to sub-sampling when the encoding of mean hue is an implicit part of the task.

4.6 Conclusion

The present study has shown that observers are able to reproduce the average hue following a rapidly-presented multi-hue ensemble, and their settings tend to the expected mean over many trials. However, these settings are distributed noisily around the mean, showing imprecision in the representation of mean hue. This imprecision is far greater than observed for reproduction of a single hue setting for single hues presented in the same way. The ideal observer simulation suggests that a sub-sampling mechanism integrating just two items from the set would outperform most observers on the task. This implies that holistic averaging of the whole set may not occur for ensembles of hue and that our percept of colour gist may be biased towards the particular sub-sampled colours of an ensemble. Further research is needed to clarify what factors drive certain elements to be included in the sub-sample (e.g. salience, spatial position/fixation), and whether holistic ensemble representations of hue can be promoted during tasks where colour summary statistics are not the focus of attention. The present study suggests that average

hue may not be a summary statistic which is automatically and efficiently encoded by observers, and that the perception of a rich world of colour may be biased by the hue of individual elements in a scene.

Chapter 5

Paper 4: Ensemble perception of colour in autistic adults

Maule, J., Stanworth, K., Pellicano, E., & Franklin, A. (to be submitted to *Autism Research*)

5.1 Abstract

Dominant accounts of visual processing in autism posit that autistic individuals have an enhanced access to details of scenes (weak central coherence; Frith & Happe, 1994), which is reflected in a general bias toward local processing. Furthermore, the hypo-priors account of autism (Pellicano & Burr, 2012) predicts that the updating and use of summary representations is attenuated in autism. Ensemble perception describes the extraction of global summary statistics of a visual feature from a heterogeneous set (e.g. of faces, sizes, colours), often in the absence of local item representation. The present study investigated ensemble perception in autistic adults using a rapidly-presented (500ms) ensemble of four, eight or sixteen elements representing four different colours. It was predicted that the autism group would be less accurate when averaging the ensembles, but more accurate in recognizing individual ensemble colours. The results were consistent with the predictions. Averaging was impaired in autism, but only when ensembles contained four elements. Ensembles of eight or sixteen elements were averaged equally accurately across groups. The autistic group also showed a corresponding advantage in rejecting colours which were not originally seen in the ensemble. The results demonstrate the local processing bias in autism, but also suggest that the global perceptual averaging mechanism is intact under some conditions. The theoretical implications of the findings and future avenues for research on summary statistic in autism are discussed.

5.2 Introduction

Autism is diagnosed on the basis of difficulties in social interaction, communication and rigid patterns of behaviour, preference for sameness, and intense and restricted interests (American Psychiatric Association, 2013). Sensory atypicalities, such as hyper- and hypo-reactivity and differences in the processing of sensory information are also increasingly recognised as being associated with autism (Pellicano, 2013). Atypical visual processing has received particular attention in the literature (for reviews, see Dakin & Frith, 2005; Simmons et al., 2009), and has led to the generation of various accounts of autistic perception. The Weak Central Coherence account (Frith & Happe, 1994) posits that autistic individuals have superior access to local detail, but at the expense of the ability to extract the global gist or see "the big picture". Several studies have shown advantages for autistic individuals in visual tasks supported by attention to local detail over global (e.g., Bolte, Holtmann, Poustka, Scheurich, & Schmidt, 2007), advantages in visual search (e.g., Gliga, Bedford, Charman, Johnson, & Team, 2015), and enhanced low-level discrimination (e.g., 'Enhanced Percetpual Functioning' account, Mottron, Dawson, Soulieres, Hubert, & Burack, 2006). However, disadvantages specifically in global processing are less easily established, and evidence is mixed (for a review, see Happe & Frith, 2006).

Pellicano and Burr's (2012) Bayesian account of sensory differences in autism builds on the predictions of Weak Central Coherence, suggesting that the autistic individuals have an attenuated ability to establish, maintain and/or use priors to inform their current perception such that the distribution of prior expectations is relatively flat (i.e. has greater variance) compared to that for typical individuals. The hypo-priors account has been followed by other predictive coding theories of autism (e.g. Lawson, Rees, & Friston,

2014; Sinha et al., 2014; van Boxtel & Lu, 2013; Van de Cruys, de-Wit, Evers, Boets, & Wagemans, 2013). The updating of priors is reliant on the integration of visual information from current and recent experiences with past experience. Some support for the account has been found in evidence for attenuated adaptation to faces in autistic children (Ewing, Leach, Pellicano, Jeffery, & Rhodes, 2013; Ewing, Pellicano, et al., 2013; Fiorentini et al., 2012; Rhodes et al., 2007). Face adaptation is the result of the norm-based coding, in which the visual "diet" of faces to which the observer is exposed is integrated into a continually-updated average face against which new exemplars can be compared (e.g. M. A. Webster & MacLeod, 2011). The hypo-priors account also makes other predictions about perception in autism. For example, the ability to integrate information from a large number of sources may be crucial to the formation and maintenance of priors across the visual domain (Pellicano & Burr, 2012), therefore, autistic people may be relatively weaker at extracting summary statistics from scenes.

Ensemble perception describes the rapid extraction of summary statistics from a set containing items which vary along some stimulus dimension (Haberman & Whitney, 2012). Ensemble perception has been demonstrated for many different visual domains, including size (e.g., Ariely, 2001), orientation (e.g., Parkes et al., 2001), facial expression (e.g., Haberman & Whitney, 2009), facial identity (e.g., De Fockert & Wolfenstein, 2009), brightness (e.g., Bauer, 2009b) and hue (e.g., Maule & Franklin, 2015; Maule et al., 2014; J. Webster et al., 2014). In many of these studies the ability to extract the average appears to exceed the limited capacity of visual working memory for representing individual items (Alvarez, 2011). This has led to the suggestion that the extraction of summary statistics takes place in the absence of individual item representation, and requires holistic, parallel processing with attention distributed across the whole ensemble (e.g., Allik et al., 2014; but see Myczek & Simons, 2008).

Rhodes and colleagues found that autistic children and adolescents were less likely to falsely recognise an unseen average face than those without autism (Rhodes, Neumann, et al., 2014). The task involved the presentation of an ensemble of four different faces (for 2000ms), followed by a test face which the participant had to decide whether or not they thought the face was in the initial ensemble. They found that while typical participants tended to endorse a morphed mean face as part of the set, autistic participants did not. This suggests that automatic extraction of the mean face in the typical group causes the false-positive familiarity of the mean face, whereas the mean representation is not as strong in the autistic group. This interpretation is concordant with the predictions of the hypo-priors account, suggesting that autistic individuals would exhibit weaker extraction of summary statistics from rapidly-presented ensembles (Pellicano & Burr, 2012).

Rhodes et al.'s procedure is also somewhat unusual for studies of ensemble perception in using a relatively long exposure time of 2000ms and a relatively small ensemble of just four faces. Previous studies of face averaging have used presentation times as low as 250ms and sets of up to 12 faces (e.g., Haberman & Whitney, 2010). Rapid presentation and large set sizes help reduce the possibility that serial processing of individual items is responsible for subsequent judgments about the set or about test items (Alvarez & Oliva, 2009). Furthermore, variations in set size may be able to help establish whether judgments could be based on a small sub-sample of items rather than the whole set (Ariely, 2001), since an average based on a fixed sub-sample should become increasingly inaccurate with larger set sizes (Ariely, 2008). The small sets and long presentation time may mean that the averaging mechanism is not required to encode the group. It is also known from adaptation studies that face coding is atypical in autism (e.g., Pellicano, Jeffery, Burr, & Rhodes, 2007; Rutherford, Troubridge, & Walsh, 2012). Likewise, autistic individuals

show various deficits in emotion, gender, identity and gaze discrimination (for a review, see Behrmann, Thomas, & Humphreys, 2006). Thus, the result of the Rhodes study may be specific to faces, rather than a general property of autistic visual processing. In order to explore whether the extraction of summary statistics is impaired in autism generally it is therefore necessary to investigate other domains of visual processing, using tasks which present larger sets in a shorter time to reduce the extent to which the serial representation of individual elements could influence the responses.

The present study investigated ensemble perception of colour in autistic and typical adults. We sought to better replicate the paradigms used in the ensemble perception literature by using a shorter exposure time (500ms). We also included variation in the number of elements in ensembles (four, eight and sixteen) and using two different tasks – a membership identification task (Maule et al., 2014) providing an indication of local knowledge of individual items in the set, and an averaging task (Maule & Franklin, 2015) providing an indication of knowledge of the global gist from making an explicit judgment about the mean colour.

We predicted that autistic adults would have superior performance to typical adults in the correct recognition of individual colours from ensembles (membership identification task), reflecting better representation of local detail, but worse performance when selecting a colour to represent the mean of the ensemble (averaging task), representing the difficulties in extracting summary statistics predicted by the hypo-priors account.

5.3 Methods

5.3.1 Participants

Twenty-one adults (8 males) with autism spectrum disorders (ASD) took part. All were recruited through two local autism charities to which only individuals with an independent clinical diagnosis of autism (n=9) or Asperger's syndrome (n=9) may be referred. Three participants who did not meet cut-off criteria on at least one of the adult Social Responsiveness Scale II (SRS-II; Constantino & Gruber, 2012) (T-score < 60) or the Adult Autism Quotient (AQ; Baron-Cohen, Wheelwright, Skinner, Martin, & Clubley, 2001) (score < 30) were excluded from the analysis. One was excluded due to low IQ (72 on Wechsler Abbreviated Scale of Intelligence – Second Edition (WASI-II; Wechsler & Psychological Corporation, 2011) and another excluded due to a fault during the testing session. A final sample of 16 adults (6 male) formed the ASD group.

Twenty-one adults were recruited from community contacts for the typical group. Data from one participant was excluded due to a fault during testing and another did not complete the WASI-II. Two further participants were excluded to bring mean IQ, mean age and gender proportion to match the ASD group (see table 5.1). A final sample of 16 adults (6 male) formed the typical group.

All reported normal or corrected-to-normal visual acuity and were assessed as having normal colour vision using Ishihara plates (Ishihara, 1973) and the Lanthony tritan test (Lanthony, 1998). Participants were paid £7.50 per hour. The research protocol was approved by the local university ethics committee.

Table 5.1. Descriptive statistics for each participant group.

	Group				
	Autistic adults		*Typical adults*		
Measure	*Mean* (SD)	*Range*	*Mean* (SD)	*Range*	*Group difference*
Age (years)	24.9 (4.4)	19-34	24.5 (4.2)	19-33	$t(30) = 0.25$, $p = .807$
IQ[a]	105.5 (13.7)	82-133	111.3 (10.7)	94-131	$t(30) = 1.32$, $p = .195$
AQ[b]	38.6 (5.6)	29-49	15.8 (5.8)	7-28	$t(30) = 11.24$, $p < .001$
SRS-II[c]	78.8 (6.45)	68-90	50.0 (9.0)	26-62	$t(30) = 10.44$, $p < .001$

Table notes: [a]IQ = Intelligence Quotient, as measured by the Wechsler Abbreviated Scale of Intelligence-II (WASI-II; Wechsler & Psychological Corporation, 2011); [b]AQ = Adult Autism Quotient (Baron-Cohen et al., 2001); [c]SRS-II = Adult Social Responsiveness Scale 2 (Constantino & Gruber, 2012).

5.3.2 Stimuli

Coloured stimuli were chosen to represent a continuous hue circle, approximately at the gamut of the monitor but avoiding the extreme corners. Slight variations in luminance were allowed (see Table 5.2), creating a stimulus set with colours varying in all three components of colour perception – hue, saturation and lightness. A uniform grey background was used throughout.

Table 5.2. CIE (1931) *xyY* chromaticity values for the colours used in the experiment.

	CIE (1931)				*CIE (1931)*		
Colour	*x*	*y*	*Y*	Colour	*x*	*y*	*Y*
Background	0.310	0.337	30.04				
1	0.488	0.319	14.54	**13**	0.237	0.428	13.96
2	0.501	0.342	14.05	**14**	0.221	0.361	13.43
3	0.509	0.365	13.48	**15**	0.208	0.299	12.88
4	0.507	0.390	12.81	**16**	0.197	0.243	12.20
5	0.496	0.423	12.14	**17**	0.198	0.202	11.70
6	0.457	0.460	11.72	**18**	0.208	0.176	11.55
7	0.414	0.503	11.55	**19**	0.226	0.169	11.86
8	0.360	0.547	11.87	**20**	0.249	0.171	12.38
9	0.313	0.585	12.51	**21**	0.286	0.182	13.34
10	0.282	0.592	13.48	**22**	0.347	0.213	14.68
11	0.267	0.556	14.00	**23**	0.419	0.259	15.23
12	0.252	0.494	14.10	**24**	0.463	0.294	14.93

Table note: The numbering of the colours 1-24 is arbitrary, since the complete set represents a continuous hue circle.

5.3.3 Apparatus

The ensemble perception tasks were completed on a 22-inch Mitsubishi DiamondPlus 2070SB Diamondtron CRT monitor, with a resolution of 1600 x 1200 pixels, 24-bit colour resolution, and a refresh rate of 100 Hz. Responses were given using a button box connected through the parallel port. A ColourCal colourimeter (Cambridge Research Systems) was used to measure the monitor and calibrate the primary values (RGB) for the stimulus colours. The tasks took place in a blacked-out room, with the monitor the only source of light. A cardboard viewing tunnel lined with black felt was used to eliminate effect of peripheral objects and colours and a chin rest was used to

constrain viewing distance at 57cm, ensuring consistency of the perceived size of the stimuli.

5.3.4 Design

The experiment involved two ensemble perception tasks: (1) membership and (2) averaging. In both tasks, ensembles comprised four different colours (‘members’) taken from a segment of the 24-colour stimulus circle. In terms of the colour circle, ensemble members were always flanked on both sides by non-members (see figure 5.1), and the segment of the colour circle from which the members were taken was varied at random on each trial. Ensembles contained either four, eight or sixteen elements, resulting in three within-participant conditions for both tasks.

In both tasks, each trial began with a black fixation point displayed for 1000ms. A multi-colour ensemble was displayed for 500ms, followed by a black fixation cross for 1000ms (figure 5.2). In the *averaging task*, a 2-alternative-forced-choice (2AFC) display followed, with two colour patches displayed. The ‘middle’ colour was always the mid-point from the segment of the colour circle from which the ensemble was generated. The ‘distractor’ colour was two colour steps away from the middle, either in the clockwise or anticlockwise direction (see figure 5.1); this was counterbalanced across trials. The positions (left or right) of the middle and distractor patches were assigned at random for each trial. The 2AFC colours remained on-screen until the participant responded by pressing a button to indicate which they thought best represented the average colour. There were six types of trials, including three levels of number of elements (4, 8, 16) in combination with clockwise/anticlockwise 2AFC distractor colour. Each trial type was

repeated eight times per block and each session had four blocks, yielding a total of 192 trials in the averaging task.

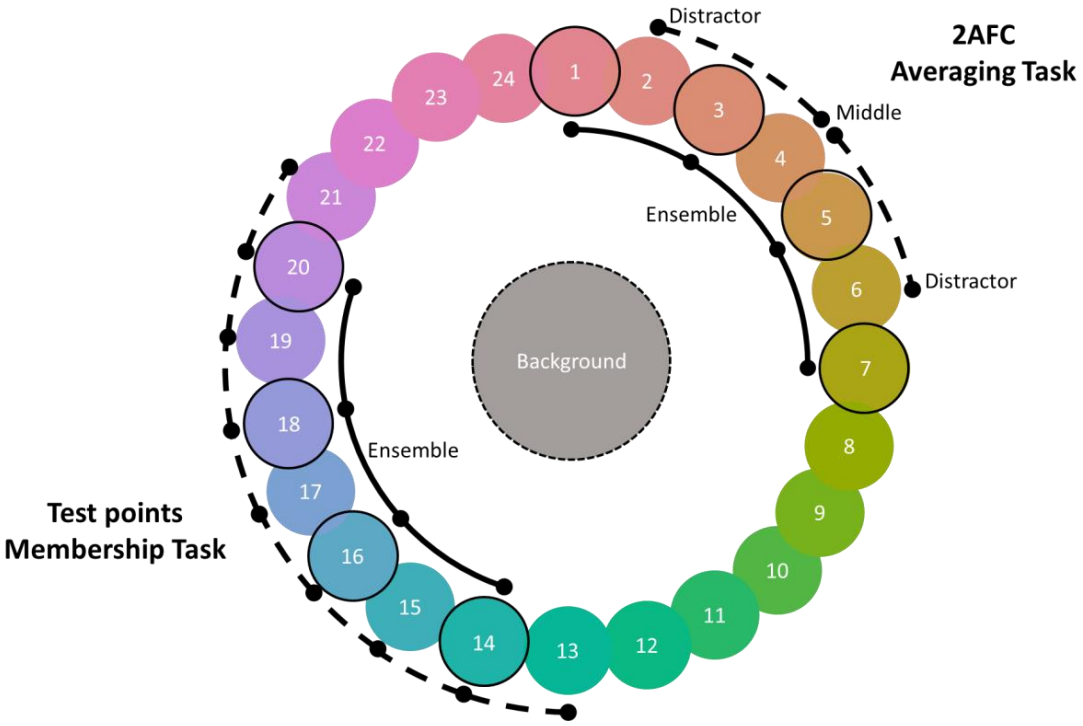

Figure 5.1. Circular arrangement of stimulus colours. The top-right annotations indicate the arrangement of stimuli for the averaging task. The initial ensemble would contain four colours (indicated by a dark border), while the subsequent 2AFC would consist of the middle colour and one of the distractors. Note that neither the middle nor distractor colour was ever present in the ensemble. The annotations to the bottom left indicate the structure of the stimuli for the membership task. Ensembles also comprised four colours but the single test point colours presented could be any of the colours spanning the ensemble range +/-1. In both the averaging and membership tasks the starting point for ensembles was selected at random from this 360° circle. Colours rendered are an indication of those used, but are not intended to reproduce the stimuli, in print or on readers' monitors.

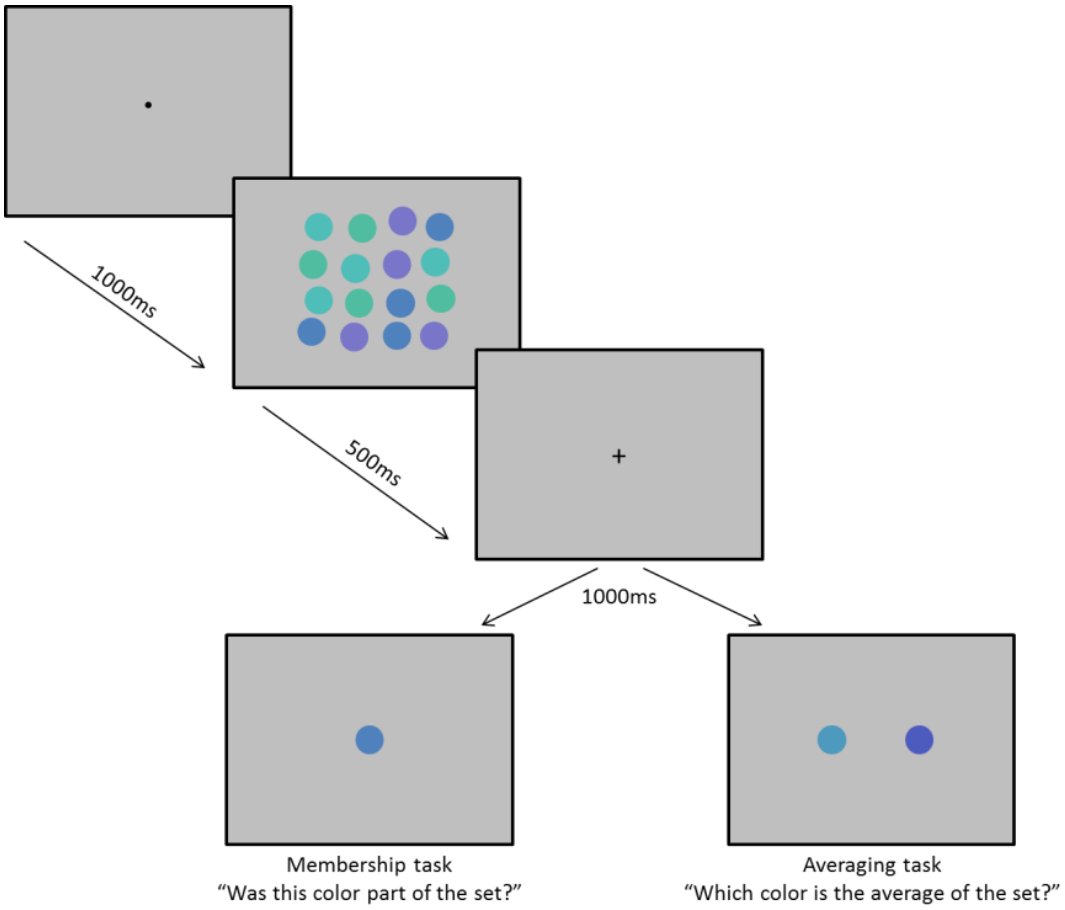

Figure 5.2. Trial procedures for the membership and averaging tasks.

In the *membership task*, a single colour patch was presented on the screen, until the participant responded according to whether they believed the patch was a part of the set. The button mapping (e.g. left = member, right = non-member or *vice versa*) was counterbalanced across participants. The colour presented could be any of the four colours from the ensemble, the three colours between the ensemble colours, or the two colours immediately adjacent to the outer colours of the ensemble (see figure 5.1). These nine conditions, multiplied by the three levels of number of elements in the initial ensemble resulted in 27 unique trial types, of which each participant completed 8 times, yielding a total of 216 membership trials for this task.

5.3.5 Procedure

Participants completed a battery of tests in a single session lasting approximately two hours, or in two shorter (1 hr) sessions. The order of the two experimental tasks (averaging and membership) was counterbalanced across participants. Within the session, the two experimental tasks were separated by an interval during which the participant completed the AQ (Baron-Cohen et al., 2001) and SRS-II (Constantino & Gruber, 2012) self-report questionnaires. Once the second experimental task was complete, the WASI-II IQ test (Wechsler & Psychological Corporation, 2011) was administered. Before each task, participants were briefed with instruction sheets, which explained the trial procedure and the participant's task. These instructions encouraged the participants to try to 'respond as quickly and accurately' as they could. The averaging task also included an additional instruction sheet, which showed a demonstration of visually averaging a group of black lines of different lengths.

5.4 Results

5.4.1 Averaging task

A response was coded as accurate if the participant chose the colour falling in the middle of the range of colours in the ensemble (see figure 5.1), rather than the distractor colour from the 2AFC. Participants in both groups tended to select the middle colour over the distractor colour, such that overall mean accuracy on the task was significantly above chance (0.5) (ASD: mean (M) = 0.57 (SD = 0.05), $t(15) = 5.12$, $p < .001$; Typical: M = 0.58 (SD = 0.06), $t(15) = 5.55$, $p < .001$). Unexpectedly, however, there was no difference between the groups on overall accuracy ($t(30) = 0.70$, $p = .490$). A 3 (number of elements:

4, 8, 16) x 2 (group: ASD, typical) repeated-measures ANOVA on accuracy found no main effect of number of elements ($F(2, 60) = 0.44$, $p = .645$), no main effect of group ($F(1, 30) = 0.49$, $p = .490$) and no interaction between group and number of elements ($F(2, 60) = 2.70$, $p = .075$).

Although this raw accuracy provides some insight, it does not take advantage of, or account for, the variations in luminance and saturation present in the stimuli. When example ensembles are plotted in perceptual colour space (CIE L*u*v*, 1976) it becomes clear that for some ensembles the middle colour is very close (i.e. similar) to the colourimetric mean (defined as the Euclidean mean of the four ensemble elements in perceptual L*u*v* colour space), but for others the middle colour is further away (i.e. dissimilar). There are even a small number of possible ensemble-distractor combinations in which the distractor colour is closer to the colourimetric mean than the middle colour.

We therefore coded each trial with the colourimetric mean of the ensemble in CIE L*u*v* space in order to get a better estimate of the accuracy of mean encoding and selection. Then we calculated the 3-dimensional Euclidean distance (ΔE) of the chosen 2AFC colour (regardless of middle/distractor status) from the ensemble's colourimetric mean. Lower values of mean ΔE suggest more accurate mean encoding and selection as this implies that the participant 2AFC choices cluster more closely to the colourimetric mean. Figure 5.3 shows these data for both groups, across different levels of number of elements.

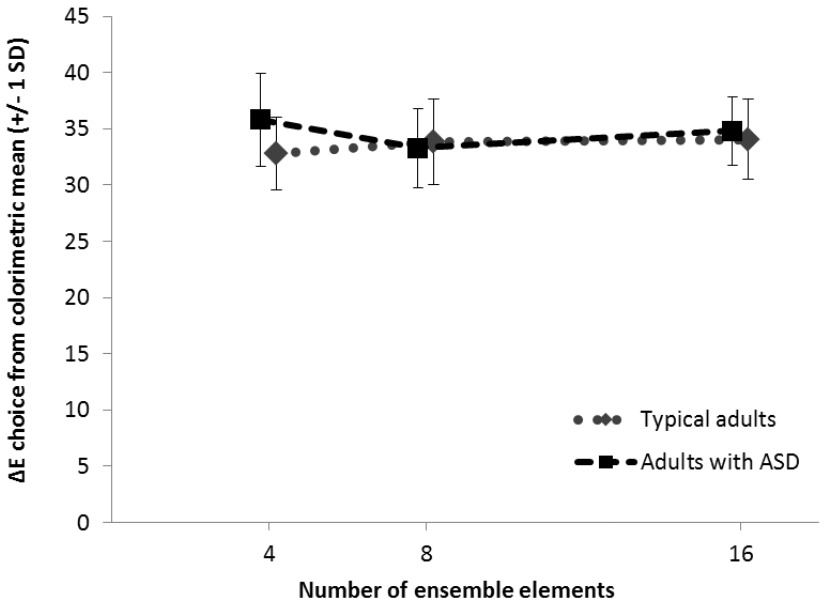

Figure 5.3. Mean distance in perceptual colour space (CIE L*u*v* Euclidean distance (ΔE)) between the chosen colour and the ensemble colourimetric mean for each group and by number of elements. Higher values indicate selections that were more perceptually-distant from the colourimetric mean of the ensembles (i.e. less accurate choice of average).

A 3 (number of elements: 4, 8, 16) x 2 (group: ASD, typical) repeated-measures ANOVA on ΔE found no main effect of number of elements ($F(2, 60) = 1.10$, $p = .339$), no main effect of group ($F(1, 30) = 1.10$, $p = .304$), but a significant group x elements interaction ($F(2, 60) = 3.83$, $p = .027$). Independent t-tests comparing the groups on each condition revealed no difference in the 8-element ($t(30)=0.41$, $p = .684$) and 16-element conditions ($t(30)=0.63$, $p = .532$), but the mean ΔE in the ASD group was significantly higher than in the typical group for the 4-element condition ($t(30)=2.27$, $p = .031$)[1].

[1] This pattern of results was consistent when the ensemble mean and distractor distances were calculated based on perceptual hue difference (angle in CIE L*u*v* space).

Similarly, regression of distance from the colourimetric mean (ΔE) for each condition on SRS-II scores revealed that these measures were significant predictors of ΔE, but only in the 4-element condition (Table 5.3).

Table 5.3. Linear regression of SRS-II on perceptual distance of selected mean from colourimetric mean.

	SRS-II			
	R^2	B (SE)	*F*	*p*
4-elements	.187	.104 (.039)	6.92	.013*
8-elements	.010	-.021 (.040)	0.29	.592
16-elements	.028	.033 (.036)	0.88	.356
All conditions	.049	.039 (.031)	1.54	.224

Table note: B – unstandardized slope coefficient; * indicates $p < .05$; overall N = 32.

5.4.2 Membership task

Trials were coded as correct when the participant responded ‘no’ to the test colour on trials where the test colour was not originally in the ensemble, or ‘yes’ when the test colour was originally in the ensemble. With responses collapsed across conditions, the ASD group (M = .53, SD = .03) performed significantly better than the typical group (M = .48, SD = .03) ($t(30) = 4.22$, $p < .001$), demonstrating better recognition for which colours were present in the initial ensemble.

In order to better understand the source of this group difference it is most informative to analyse the pattern of responses not in terms of their correct/incorrect status, but whether the participant thought that the test colour was a part of the ensemble – the proportion of "yes" responses, which we refer to as 'familiarity'. Veridical memory would result in a pattern of high familiarity for the seen colours (test points 2, 4, 6, 8) and low familiarity for the unseen colours (test points 1, 3, 5, 7, 9). Therefore the familiarity of seen and unseen test points was pooled for a 2 (test point: seen, unseen) x 3 (number of elements: 4, 8, 16) x 2 (group: ASD, typical) ANOVA. This revealed a main effect of seen/unseen test point ($F(1, 60) = 14.90$, $p = .001$) and significant interactions between elements and group ($F(2, 60) = 4.62$, $p = .014$) and between seen/unseen test points and group ($F(1, 60) = 12.77$, $p = .001$). No other main effects or interactions were significant (all ps $> .05$).

Paired *t*-tests comparing seen to unseen test points for each group separately revealed that, while the autistic adults found seen colours significantly more familiar than unseen colours ($t(15) = 4.83$, $p < .001$), the typical adults did not ($t(15) = 0.23$, $p = .830$). Independent *t*-tests comparing the groups confirmed that the ASD group found unseen colours significantly less familiar than the typical group ($t(22.62) = 2.40$, $p = .023$ (corrected for unequal variances)), but there was no group difference in familiarity of the seen colours ($t(24.18) = 0.24$, $p = .811$ (corrected for unequal variances)) (see Figure 5.4). Regression of familiarity for seen and unseen colours on SRS-II scores revealed that these measures were significant predictors of familiarity for unseen but not seen colours (Table 5.4).

These data can also be summarised using signal detection measures. Using the rate of hits (responding "yes" to seen colours) and false alarms (responding "yes" to an unseen colour) we calculated d′ (Brophy, 1986), as a measure of the observers' sensitivity to the

difference between familiar and unfamiliar colours. The autistic group (M = .23, SD = .19) were significantly more sensitive than the typical group (M = .02, SD = .16) (t(30) = 3.45, p = .002) (see Figure 5.4). Scores on the SRS-II were also a significant predictor of sensitivity measured as d′ (see Table 5.4), with higher SRS-II scores being associated with higher sensitivity to the difference between seen and unseen test colours. The tendency for observers to respond "yes" given the uncertainty of the task (the criterion) was also calculated (C – Brophy, 1986). Both groups exhibited a liberal criterion (values of C below zero), indicating a tendency towards responding "yes" (Typical group: M = -0.41, SD = 0.18; Autistic group: M = -0.30, SD = 0.34). There was no significant difference in mean criterion between the groups (Levene's test indicated inequality of variance between groups (F = 1.09, p = .001), so adjusted degrees of freedom are used: t(22.94) = -1.24, p = .227), indicating that neither group had a more stringent criterion, and the differences in task performance are due to better sensitivity to seen vs. unseen hues in the autistic group.

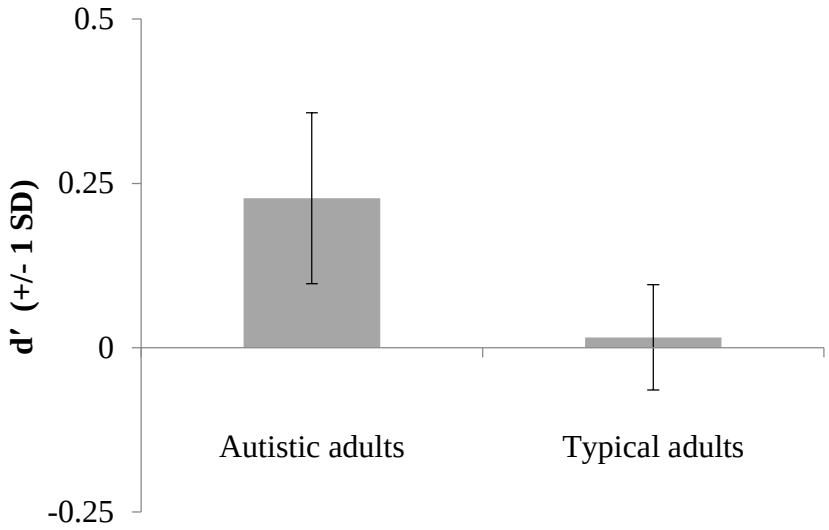

Figure 5.4. Sensitivity (d′) to the difference between seen and unseen test colours, by group. Higher values of d′ indicate higher sensitivity.

Table 5.4. Linear regression of SRS-II on membership task responses.

	SRS-II			
	R^2	B (SE)	*F*	*p*
Seen colours	.015	-.001 (.001)	0.47	.499
Unseen colours	.208	-.003 (.001)	7.90	.009*
Overall correct	.348	.001 (<.001)	16.04	< .001*
d′	.240	.006 (.002)	9.49	.004*
Criterion (C)	.085	.005 (.003)	2.80	.105

Table notes: B – unstandardized slope coefficient; * indicates $p < .05$; overall N = 32.

5.5 Discussion

This study sought to establish whether autism is associated with reduced ability to extract summary statistics from a rapidly-presented ensemble. Following the hypo-priors (Pellicano & Burr, 2012) and Weak Central Coherence (Frith & Happe, 1994) accounts of autism, we predicted that autistic adults would be less accurate when picking the average colour of an ensemble, but have an enhanced ability to remember the specific colours present in the ensemble, relative to a group of typical adults.

In the averaging task we found that autistic adults were no different to typical adults in raw accuracy on a 2AFC rapid averaging task. Furthermore, although they did tend to make selections slightly further from the colourimetric mean in perceptual colour space

this difference was restricted to the 4-element condition – when ensembles contained 8 or 16 elements there was no difference due to autism. These results suggest that the extraction and representation of average colour from a rapidly-presented ensemble is intact in autism in response to larger sets, but that for small sets this mean extraction and representation process is less accurate. In the membership task we found that autistic adults were slightly better than typical adults at recognising colours which were not part of the original set. While typical adults did not appear to distinguish seen from unseen colours, finding all test colours familiar at roughly a similar rate, the autistic adults were better at detecting unseen colours, and rejected them more often.

It appears that the autistic adults may exhibit attenuated global representations of the ensemble in the averaging task. The autism group showed impaired averaging only in the 4-element condition, but improved averaging of larger sets of 8 and 16 elements, being equal to the typical group in these conditions. This finding replicates and extends that of Rhodes et al., (2014), who found reduced averaging of a set of four faces in autistic children, and provides some support for the suggestion that the extraction of summary statistics is impaired in autism (Pellicano & Burr, 2012) and for a local processing bias at the expense of global gist in autism (Frith & Happe, 1994), but finds that this is restricted only to small sets.

One hallmark of ensemble perception is invariance in averaging performance to changes in the number of elements in the ensemble (Ariely, 2001; Chong & Treisman, 2005a, 2005b; Haberman & Whitney, 2010; Leib et al., 2014; Marchant et al., 2013; Maule & Franklin, 2015; Robitaille & Harris, 2011; Utochkin & Tiurina, 2014), or even improvement in averaging with larger sets (Robitaille & Harris, 2011). Such findings are often interpreted as suggesting that rapid averaging is underpinned by a global gist-

extracting mechanism, occurring in parallel across the whole ensemble (e.g., Ariely, 2008; Treisman, 2006). The effect of the number of elements on the averaging of the autistic group suggests that the global representations are impaired for autistic adults, but only in the four-element condition.

The suggestion that invariance of averaging performance to set size implies a rapid, global averaging mechanism is not uncontroversial. The possibility that observers could sub-sample items from the ensemble has been suggested by several researchers (e.g., Marchant et al., 2013; Myczek & Simons, 2008; Simons & Myczek, 2008; Whiting & Oriet, 2011) as an explanation for perceptual averaging which does not entail a mechanism of visual processing with processing capacity above that of visual working memory of around three to four items (Alvarez, 2011). Recent simulations and observer data suggest that to account for the accuracy of averaging, sub-samples from ensembles of size (Im & Halberda, 2013; Utochkin & Tiurina, 2014) or faces (Haberman & Whitney, 2010), would need to incorporate individual representations of at least seven items from each ensemble – far beyond the limit for visual working memory. For colour, however, it appears the representation of the average hue may be based on sub-sampling as few as one or two elements from an ensemble (Maule & Franklin, under review – see chapter 4). However, creating, maintaining and deploying a global representation does not require that the information is gathered in a holistic way. Sub-sampled information still requires integration into a coherent average for the averaging task. Therefore, if the mechanism underlying averaging of colour is sub-sampling, the differences observed are still consistent with attenuated combining of information into a coherent global representations in autism.

We suggest that the results may indicate that a perceptual averaging mechanism is intact in autism, at least for colour, but that this mechanism is not as effectively deployed for small sets as it is in typical adults. The local processing bias in autism may cause autistic adults to apply a local strategy to small sets, but shift to a global strategy for larger sets with more items than can be represented in visual short-term memory (Alvarez, 2011). On the other hand, typical adults may be more likely to apply a gist-based global strategy to extract the average, regardless of the number of elements in a set. Experiments manipulating local and global attention provide some support for this idea. For example, average judgments are better when combined with a concurrent task requiring global attention, compared to a concurrent task requiring local attention (Chong & Treisman, 2005a). Similarly, average judgments are less accurate when attention is cued locally to individual elements or when some elements are more salient than others (Albrecht & Scholl, 2010; De Fockert & Marchant, 2008). Therefore, a bias towards local processing can explain impaired averaging performance, and furthermore, this bias is most evident in the group difference for four elements – the approximate limit for visual working memory (Alvarez, 2011).

We propose that a tendency to apply a local strategy is present for small sets in autism but that a shift to a global strategy occurs for larger sets and that this coincides with the capacity limit of working memory for individual item representation. In contrast, the typical group exhibit a global bias even for the relatively small sets, but may also exhibit a similar switch in strategy if averaging of smaller ensembles (i.e. three or two elements) was tested.

The data from the membership task also reflects the bias to local processing in autism, as described in the Weak Central Coherence account (Frith & Happe, 1994). The better

overall performance of the autistic adults on the membership task is attributable to an advantage in rejecting unseen test colours. This implies that the local bias in autism enhances the representation, and hence recognition of, individual elements. One possible enhancement could be improved precision of the representation of individual elements, so that given an unseen test colour there is a reduced likelihood of a false alarm, leading to better rejection of colours which are not ensemble members. This may help overcome the limited capacity of visual working memory. It has been suggested that the variance in item representation is the limiting factor in working memory capacity, rather than there being discrete "slots" for items (Bays, 2015). As such, having more precise representation (i.e. less variance in representation) of ensemble elements may help boost the number of elements which can be encoded and recognised in the test phase.

However, this does not explain why the advantage for autistic adults did not generalise to seen colours, as well as unseen – better representation of the individual colours present in an ensemble ought to lead to ensemble members also being more familiar. There is not a clear explanation why this pattern of results might occur. If the autism group had a more stringent criterion – i.e. required a stronger positive signal before being willing to respond "yes" to a test colour, this might result in more correct rejections of unseen colours. However the data do not support this as the groups did not differ in criterion. Both groups do, however, exhibit a very liberal criterion – responding “yes” very readily. For the typical group in particular, all of the test stimuli feel familiar, as evidenced by d' values for this group being near zero – indicating that these observers were insensitive to the difference between seen and unseen colours. Since participants are aware that in the experiment some colours are members and some are non-members they may respond “no” on occasion simply to redress the balance of responses, and as a result the hit-rate is reduced (as well as the rate of false alarms). Whether this type of conscious bias-reducing

process is also limiting the hit-rate in the autism group is unclear since this group also shows some sensitivity to the difference between the seen and unseen stimuli. This may be investigated in future studies by asking for a confidence rating for each yes/no response. If typical participants are truly insensitive to member/non-member colours, we may expect them to report low (or equal) confidence across all types of response (hit, false alarm, miss, correct rejection), whereas confidence may be more closely related to accuracy for the autism group (i.e., high confidence for correct responses, low for incorrect).

The results from the membership task have implications for the interpretation of the previous finding of reduced set averaging in children with autism by Rhodes et al. (2014). They used a membership task and did not directly assess the representation of the average, finding that the average face was rejected as a member of the ensemble more frequently by the autism group. In their experiment the average was never a part of the set – it was an 'unseen' face. Therefore, based on our findings in the membership task, their result may be driven by better rejection of unseen items by the autism group, rather than reduced averaging. As Dakin and Frith (2005) point out, tasks designed to test global processing should attempt to preclude the use of local processing strategies. The membership task is not sufficient to make claims about extraction of global summary statistics.

Our study is the first to directly probe the representation of the average of an ensemble in autism and the results are broadly complementary with the membership task. However, there is a slight difference in the effect of the number of elements in an ensemble from the two tasks used in this study – the averaging task found that autistic adults had weaker averaging in the 4-element condition but not for larger sets, whereas the membership task finds no effect of the number of elements on the pattern of responses to seen and unseen

colours. This is potentially a challenge to our explanation of a shift from local to global processing style with increasing numbers of elements. If the processes underlying the two tasks are the same we would expect that the autism group would show a distinctive higher accuracy on the membership task for four-element ensembles, but be similar to the typical group for larger sets. Instead the membership advantage in autism is uniform across set sizes.

We propose that this difference reflects the fact that each task is actually invoking slightly different processes. While the requirements of the averaging task encourage a global approach, as the observer tries to integrate all of the information, the membership task is more local, as the observer tries to store the specific exemplars present. There is also some redundancy of information in the sets, as there are always four different colours, which are repeated in the larger sets. If an observer is able to identify the four different colours this may support a better representation of individual colours, regardless of the increased set size. Since autistic individuals demonstrate a local bias (e.g. Mottron et al., 2006) they may tend to employ a global approach less readily than typical individuals, but can do so when the task demands it, i.e. when averaging is required and sets are large (see the "first principle of autistic perception" in Mottron et al., 2006). Previous studies have also found global processing to be intact, while local processing is enhanced in autism (Dakin & Frith, 2005; Mottron et al., 2003), similarly, ensemble statistical representations have been shown to be intact in individuals with various conditions which impair focal attention (Demeyere et al., 2008; Lanzoni et al., 2014; Pavlovskaya et al., 2015). We recommend that future research should use at least a membership and an averaging task to assess the strength of summary representation and the level of local knowledge about ensembles.

Although we find averaging to be less accurate for small ensembles, we cannot say whether the average is encoded automatically. Further research is needed to establish whether there is a difference in the automatic extraction of summary statistics in autism, as has been suggested by averaging element which are outside the focus of attention, both for size (Oriet & Brand, 2013) and location (Alvarez & Oliva, 2008). Such work would help establish whether differences in summary statistical representations in autism are due to reduced extraction of summary statistics generally, or whether gist-based representations are encoded but not accessed as easily in explicit averaging tasks, perhaps due to the strength of local element representations superseding the gist-based information. Implicit measures of the extraction of summary statistics may be desirable to avoid possible effects of task on local/global bias. For example, saccades to a visual target have been shown to be faster when the mean orientation of background elements is constant, compared to when the mean changes (Corbett & Melcher, 2014b). It has also been shown that responses on a categorization task are faster when a preceding prime ensemble has the same variance as the target array, even when the actual mean is different (Michael et al., 2014). Tasks such as these, in which the participant is not required to respond or consider the mean or members, but in which effects can be demonstrated, help to indicate the automaticity of the extraction of summary statistics. If it is the extraction of summary statistics which is impaired in autism we might expect these implicit effects to be reduced or absent, compared to a typical group.

Understanding how the visual system processes simple ensembles can provide insight into how it copes with the vast amount of information it receives in the real world. Aspects of autism symptomatology such as hyper-sensitivity and sensory overload imply that the integration of information is impaired (Pellicano, 2013), while perceptual talents such as highly accurate recall of a scene, or superior visual search demonstrate the benefits of

maintaining representations of the details present in the visual world (Frith & Happe, 1994; Happe & Frith, 2006). The present study supports the suggestion that although autism may be characterised by a cognitive style which enhances local processing, the advantages of this are not necessarily traded-off against global processing ability (Dakin & Frith, 2005; Happe & Frith, 2006; Mottron et al., 2003). The group differences which were found do not suggest a complete lack of summary representation, but do appear to reflect a difference in broad cognitive style. Our finding of typical summary representations of colour for larger sets leads to further questions about whether the difficulties in the integration of visual information associated with autism reflect low-level differences at encoding and storage (e.g., Mottron et al., 2006), high-level differences in the integration of information (e.g., Pellicano & Burr, 2012), and/or differences in meta-cognition (e.g., Friston, Lawson, & Frith, 2013; Lawson et al., 2014).

5.6 Conclusion

In conclusion, we find that the pattern of responses to tasks requiring perceptual averaging and summary representation are consistent with a local bias in autism (Frith & Happe, 1994), and consistent with attenuated use of summary statistics in autism (Pellicano & Burr, 2012), but not a complete absence of their representation. The advantage for autistic adults in recognising that they haven't seen a stimulus before is not always accompanied by a disadvantage in averaging, but may be when sets are small. Instead it appears that a global averaging mechanism is intact under some conditions, but that autistic adults tend to use local information by default.

Chapter 6

Paper 5: Colour afterimages in autistic adults

Maule, J., Stanworth, K., Pellicano, E., & Franklin, A. *Journal of Autism and Developmental Disorders* (submitted)

6.1 Abstract

It has been suggested that attenuated adaptation to visual stimuli in autism is the result of atypical perceptual priors (e.g., Pellicano & Burr, 2012). This study investigated adaptation to colour in autistic adults, measuring both strength of afterimage and the influence of top-down knowledge. We found no difference in colour afterimage strength between autistic and typical adults. Effects of top-down knowledge on afterimage intensity shown by Lupyan (in press) were not replicated. This study suggests that colour adaptation is intact in autism. This is in contrast to findings of attenuated adaptation to faces and numerosity in children with autism, and suggests that the effect of attenuated priors is restricted to higher-level cortical adaptation, or may be manifest primarily in childhood.

6.2 Introduction

Autism is characterized by difficulties in social communication and behavioral traits including rigid patterns of behavior, preference for sameness, and intense and restricted interests (American Psychiatric Association, 2013). Yet, a range of sensory symptoms, including both hyper-reactivity and hypo-reactivity to external stimuli, have more recently been recognized as characterizing the sensory experience, and relating to some of the behavioral traits, of individuals with autism (Pellicano, 2013). Pellicano and Burr's (2012) Bayesian account of the sensory traits of autism posits that an under-weighting (relative to typical individuals) of the strength of *prior* information when interpreting sensory information may underlie sensory differences in autism.

One sensory process which could be affected by under-weighting of prior information is adaptation (Pellicano & Burr, 2012). Adaptation is a fundamental property of neural networks and describes a reduction in neural activity in response to a persistent stimulus (Kohn, 2007). Adaptation is thought to serve a crucial function for sensory systems, tuning neural responses to maximize sensitivity to the range of stimuli present in the immediate environment (e.g. M. A. Webster, 2011). Adaptation to visual stimuli can result in the experience of aftereffects – distortions in perception which tend to bias perception in the opposite direction to that adapted (e.g. adaptation to a grating tilted to the left, causes a vertical grating appears to tilt slightly to the right) (C. W. G. Clifford, 2012). Since adaptation is based on the sensory history of the beholder, whereby recent sensory input is used to calibrate sensory systems and hence bias current perception, the *hypo-priors* account of autism therefore predicts that adaptation should be attenuated in autism (Pellicano & Burr, 2012). This suggestion is supported by evidence that children with autism show less adaptation to facial identity than typical children (Ewing, Pellicano,

et al., 2013; Rhodes, Ewing, Jeffery, Avard, & Taylor, 2014), reduced set averaging of facial identity (Rhodes, Neumann, et al., 2014), and evidence for reduced face identity aftereffects in parents and siblings of children with autism (Fiorentini et al., 2012). Autistic adults also show different patterns of afterimage to emotional expressions, being more likely to report an afterimage biased in the direction of the adapting stimulus (i.e. a neutral face appears "sad" following adaptation to a sad face) than typical adults, suggesting the organization of emotion coding in autism may be different (Rutherford et al., 2012), possibly a result of atypical use of priors in the interpretation of sensory signals.

Furthermore, it has recently been shown that children (age 7-14 years) with autism show less adaptation to numerosity (i.e. number of items present in a part of visual field) than IQ- and age-matched typically developing children (Turi et al., 2015). This finding extends the evidence for reduced adaptation in autism to a non-social visual stimulus, lending weight to the suggestion that this is a sensory feature of autism which generalizes beyond of the social domain of face perception, and supporting Bayesian theories of autism (Pellicano & Burr, 2012) and predictive coding accounts of perception (e.g. Kersten, Mamassian, & Yuille, 2004; Knill & Pouget, 2004). The finding for numerosity also raises the question of whether other forms of perceptual adaptation are reduced in autism as well. Here, we address this issue with regards to adaptation to colour.

Adaptation to colour is demonstrated by the phenomenon of colour afterimages – an observer who stares at a coloured patch for a few seconds will experience an afterimage of the patch in the opponent colour (e.g. afterimage of cyan following adaptation to magenta) if they look at a uniform white field (Wheatstone, 1838). The locus of adaptation which causes colour afterimages has been the subject of discussion, and the current evidence suggests that colour afterimages are instantiated by adaptation of

photoreceptors in the retina and retinal ganglion cells, but may then be subject to further modulation by cortical processes (Zaidi et al., 2012). For example, if an observer is alternately shown a vertical grating on a red background and a horizontal grating on a green background, for around two minutes, they will subsequently perceive a reddish tinge to a horizontal grating on a white background, and a greenish tinge on a vertical grating on a white background (McCollough, 1965). This is known as the McCollough effect, and is most likely mediated by processes in early visual cortex (area V1; Vul & MacLeod, 2006), not just at the level of the retina.

The current study investigated whether colour adaptation is reduced in autism, as has been shown for faces and numerosity. There is already evidence that children with autism differ from their typically developing peers in both colour discrimination and memory at least under some conditions (Franklin, Sowden, Burley, Notman, & Alder, 2008; Fujita, Yamasaki, Kamio, Hirose, & Tobimatsu, 2011; Heaton, Ludlow, & Roberson, 2008; but see also Koh, Milne, & Dobkins, 2010). There is also anecdotal and case study evidence for enhanced colour associations (both phobias and intense interests) in autism (Ludlow, Heaton, Hill, & Franklin, 2014), together suggesting that colour may be processed differently by autistic people. If perception in autism is associated with a generalized reduction in adaptation – from retinal to cortical levels – as the hypo-priors account predicts, then we would expect to find weaker afterimages for autistic participants compared to typical participants. It is important to note, however, that the initial neural locus of colour afterimages is the retina (Zaidi et al., 2012), whereas the prior evidence for attenuated adaptation in autism has all been for higher-level stimuli (faces, numerosity) which are represented in cortical areas (Kanwisher et al., 1997; Piazza & Izard, 2009).

In addition to investigating the strength of colour afterimages, this study also investigates whether the effect of top-down knowledge on colour afterimages is reduced in autism. Lupyan (2013, in press) has claimed that colour adaptation is subject to influences from top-down knowledge of a scene or object's typical colour, using a task based on the 'Spanish Castle Illusion'. The 'Spanish Castle Illusion' is a colour adaptation illusion where exposure to a highly-saturated adapting image, in the complementary colours to the original, causes the observer to see a subsequent greyscale image in the original colours (albeit desaturated). The effect is most striking when the adapting and greyscale image are created from a photograph of a landmark or natural scene. The version which gave the illusion its name used an image of the castle of Manzanares el Real in Spain, which included the castle, with a clear sky forming most of the background, and vegetation in the foreground. The adapting image is formed by inverting the colours of the base image, flattening the luminance profile and boosting the saturation of the resulting "negative" image (Sadowski, 2006). After a short period of fixation (e.g. 20 seconds), the adapting image is replaced by a greyscale version of the original. The superimposition of the colour afterimages with the greyscale contours of the image results in the percept of an image presented in the full, original colour (Daw, 1962) – an effect which is also observed in illusions of "filling-in" of colour afterimages when coupled with luminance contours (e.g. Anstis, Vergeer, & van Lier, 2012; Powell, Bompas, & Sumner, 2012; van Lier, Vergeer, & Anstis, 2009).

Using a nulling method, whereby participants were asked to adjust the image until it appeared greyscale following adaptation to measure the strength of the afterimage, Lupyan found that afterimages were stronger for images of natural scenes (which have typical colours) compared to images of man-made objects (which might be any colour). Yet when the same images were turned upside-down in order to disrupt the ease with

which the image content could be extracted, the difference between the natural and man-made scenes was abolished. This effect of image orientation on the natural scenes cannot be accounted for by low-level differences in the stimuli used, since it is the same image, rotated through 180 degrees. Lupyan attributed this interaction between image orientation and content as being reflective of top-down influence on the perception of the afterimage – scenes with colour-diagnostic content (i.e. typically containing particular colours) are subject to modulation by colour knowledge, whereas scenes with non-colour-diagnostic content receive no top-down influence. Turning a diagnostic image upside-down disrupts the mechanism responsible for predicting the scene colours, and thus diminishes the effect of colour knowledge on the afterimage.

In the current study, we sought to replicate Lupyan's study with samples of autistic and non-autistic adults. We also attempted to improve on Lupyan's methods by using his two original stimuli, in addition to two new images with colour-diagnostic and non-colour-diagnostic content. Furthermore, we used a perceptually-valid, display-independent stimulus space to measure the strength of afterimages. The effect of top-down knowledge modulating the appearance of afterimages requires the application of priors to predict the appearance of the colour-diagnostic scenes. The hypo-priors account would therefore predict that the interaction of image content (i.e. the colour-diagnosticity of man-made vs. natural scenes) with image orientation should be absent, or at least weaker, in autistic individuals.

6.3 Methods

6.3.1 Participants

Sixteen adults (8 male) diagnosed with an autism spectrum disorder, recruited through two local autism charities, took part. All adults had received an independent clinical diagnosis of autism (n=7) or Asperger's syndrome (n=9) according to ICD-10 (World Health Organization, 1992) or DSM-5 (American Psychiatric Association, 2013) criteria and met threshold for autism on the self-report Social Responsiveness Scale – 2nd edition (SRS-2) (T-score < 60) (Constantino & Gruber, 2012) and the Adult Autism Quotient (AQ; Baron-Cohen et al., 2001) (score < 30). Two of these participants (both male) failed to meet threshold for autism on these questionnaires and were therefore excluded from analysis. Another male participant was excluded due to a low full-scale IQ score (of 72) on the Wechsler Abbreviated Scale of Intelligence – Second Edition (WASI-II; Wechsler & Psychological Corporation, 2011) and another male participant gave responses consistent with a colour vision deficiency on the Ishihara test (Ishihara, 1973) and was excluded, as was another female participant who did not follow the task instructions. A final sample of 11 adults (4 male) formed the autism group.

Sixteen typical adults were recruited from community contacts. Data from two participants were removed as they scored over 60 on the SRS-2. Three further participants were excluded for matching purposes, yielding a final sample of 11 typical adults (4 male). The final groups of typical and autistic participants were well-matched in terms of gender proportion, mean age and mean IQ, but differed significantly on mean SRS-II and mean AQ score (see Table 6.1).

All participants reported normal or corrected-to-normal visual acuity and were assessed as having normal colour vision using Ishihara plates (Ishihara, 1973) and the Lanthony

tritan test (Lanthony, 1998). Participants were paid standard rates of £7.50 per hour. The research protocol was approved by the local university ethics committee and all participants gave written informed consent prior to participation in this study.

Table 6.1. Descriptive statistics and group comparisons for age, IQ, AQ and SRS-2 for autistic and typical adults.

	Group				
	Autistic adults		*Typical adults*		
Measure	*Mean* (SD)	*Range*	*Mean* (SD)	*Range*	*Group difference*
Age (years)	24.8 (4.9)	19-34	23.9 (4.8)	19-33	$t(20) = 0.44$, $p = .668$
IQ[a]	108.6 (12.6)	82-133	109.6 (8.1)	95-121	$t(20) = 0.22$, $p = .827$
AQ[b]	37.7 (5.9)	29-49	15.0 (5.4)	7-23	$t(20) = 9.45$, $p < .001$
SRS-2[c]	78.0 (6.9)	68-90	49.9 (9.4)	26-60	$t(20) = 8.00$, $p < .001$

Table notes: [a]IQ = Intelligence Quotient, as measured by the Wechsler Abbreviated Scale of Intelligence-II (WASI-II; Wechsler & Psychological Corporation, 2011); [b]AQ = Adult Autism Quotient (Baron-Cohen et al., 2001); [c]SRS-2 = Adult Social Responsiveness Scale – 2nd edition (Constantino & Gruber, 2012).

6.3.2 Stimuli

Four photographs depicting books on shelves (“books”), Culzean Castle in Scotland (“castle”), Bells beach, Australia (“beach”) and a row of painted, wooden beach huts in Whitby, England (“huts”) were used as the basis for all stimuli presented in the

experiment. The books (colour non-diagnostic) and castle (colour diagnostic) images were identical to those used by Lupyan while the huts and beach photographs were selected as good examples of images which are colour-diagnostic (beach) or colour non-diagnostic (huts) (figure 6.1).

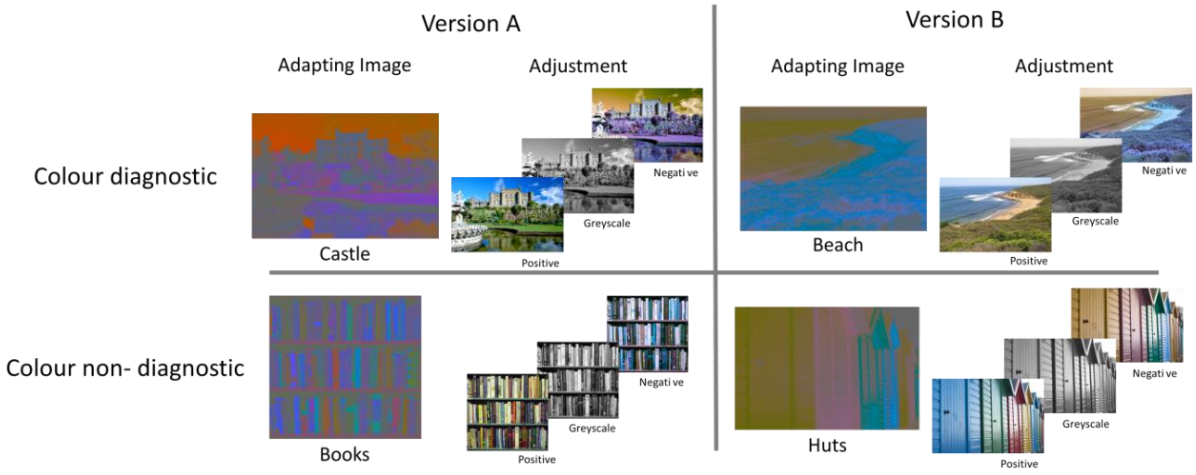

Figure 6.1. Images used in the experiment. At the adjustment phase a continuum of 51 images of gradually changing colour intensity between the positive, greyscale and negative images was available for selection. The dimensions of the images varied slightly due to the proportions of the source images.

The adapting stimuli were generated following the Adobe Photoshop actions described in John Sadowski's online demonstration of the Spanish Castle illusion (Sadowski, 2006). The procedure involves the flattening of the luminance profile of the image, inverting the colours in HSL space and enhancing the colour intensity of the resulting image. These adapting images are not equated in colour saturation, so the image orientation manipulation is used to disentangle the effect of the image content from low-level differences in the adapting stimuli.

Inverted and greyscale versions of each original image were also generated using the hue rotation and desaturate functions of Photoshop in HSL-mode. For each of the four basic images, a smooth continuum consisting of 51 images between the original, "positive" image, through the greyscale image and to the inverted "negative" image were created through linear transformation of the RGB values for each pixel. This continuum was the basis for the adjustment phase of each trial. Throughout the experiment, a background grey with chromaticity equivalent to D65 (xyY (1931) = 0.313, 0.329, 20) was used.

These images (in all coloured and achromatic forms) were always presented centrally on the monitor with a fixation cross in the center. All images subtended 25° of visual angle horizontally and 17° (24° for the books image) of visual angle vertically.

6.3.3 Apparatus

The adaptation tasks were completed on a 22-inch Mitsubishi DiamondPlus 2070SB Diamondtron CRT monitor, with a resolution of 1600 x 1200 pixels, 24-bit colour resolution, and a refresh rate of 100 Hz. Responses were given using the keyboard. A ColourCal colourimeter (Cambridge Research Systems) was used to measure the monitor and calibrate the primary values (RGB) for the background colour regularly during the data collection period. The tasks took place in a blacked-out room, with the monitor as the only source of light. A cardboard viewing tunnel lined with black felt was used to eliminate effect of peripheral objects and colours and a chin rest was used to constrain viewing distance at 57cm, ensuring consistency of the perceived size of the images and overlap of the after-images with adjustment image.

6.3.4 Design

The experiment involved two versions of an adaptation task, both identical in procedure, length and task, but with different images in each. Version A included the books and castle images and version B the huts and beach images. Each version had two images – one colour diagnostic (castle/beach) and one colour non-diagnostic (books/huts). In each version one image would be presented upright (hereafter referred to as "up"), and the other upside-down (hereafter "down"). This diagnosticity-orientation pairing was reversed in the second version (e.g. version A – castle up, books down; version B – huts down, beach up). Thus each participant would adapt to colour-diagnostic scenes in both orientations and colour non-diagnostic scenes in both orientations. The order in which the participant completed the versions (A/B) was also counterbalanced across participants.

For the purpose of analysis the trials were classified according to the class of the image (i.e. books and huts are "non-diagnostic" and castle and beach are "diagnostic") and orientation (upright or upside down), such that, across both versions of the adaptation nulling experiment, each participant provided settings for four conditions – diagnostic upright, diagnostic upside-down, non-diagnostic upright and non-diagnostic upside down. The specific image used in each condition varied between participants (e.g. for some the diagnostic upright image was the castle but for others it was the beach) but was counterbalanced across groups.

6.3.5 Procedure

Participants were seen individually at the university either for a single session lasting approximately two hours or for two shorter sessions. Within a session, the

participant completed one experimental task followed by the two questionnaires tapping autism symptomatology (the AQ and SRS-2) followed by the second experimental task, and finally the WASI-II.

Prior to beginning the task proper, participants were briefed with instruction sheets explaining the trial procedure and the participant's task. These instructions encouraged the participants to try to "make their adjustments as quickly and accurately" as they could and emphasized the importance of maintaining a steady gaze, fixed on the central cross.

To begin, participants completed eight 'norming' trials, in which there was no adaptation phase, only adjustment to greyscale (trials started at a random point on the positive-negative colour continuum). Four norming trials for each stimulus image were completed, with the image only displayed in the orientation it would be presented in for the adaptation trials. This was ostensibly for practice, allowing participants to become accustomed to the adjustment procedure and range of stimuli, but also allowed recording of the pre-adaptation settings for each image and observer (following Lupyan, in press).

Next, participants were presented with version A or B of the experimental task. Each experimental trial required participants to stare at a fixation cross in the center of an adapting (colour) image for 20 seconds, which was immediately replaced by a greyscale version of the same image in the same orientation. A nulling procedure was used to measure the strength of the afterimage in terms of the shift in perception of this greyscale image towards the positive (natural) colours of the photograph. The participant was instructed to adjust the image until it appeared subjectively greyscale. This was done using the up and down arrow keys on the keyboard, which were assigned at random on each trial to either adjust the image towards the positive-colour image or the negative-colour image along the adjustment continuum. When the participant felt the image was

greyscale they pressed the space bar. Participants were given 30s between each trial to allow for the dissipation of the colour aftereffects.

Each version consisted of eight adaptation trials (four for each of two images, of which one image would be upright and one upside down) presented in a pseudo-random order, with the constraint that trials presenting the same image could only appear a maximum of two times in succession. After having completed one version of the task (version A or version B), participants filled in the AQ and SRS-II self-report questionnaires, then completed the second version. The order of versions was counter-balanced across participants.

6.4 Results

Settings were coded initially by an index number that referred to the selected image from the adjustment continuum. However, in order to compare across images it was important to approximate the perceptual equivalence of these indices. The CIE L*u*v* colour space is a perceptually-uniform description of the colours visible to adult human observers. It includes parameters that approximate to saturation or chromatic intensity called “chroma”, as well as hue and lightness. Stronger afterimages require more intense chromatic input to null them. The chroma of each image selected in the experiment to null the aftereffect was subtracted from the chroma of the presented greyscale image that was in the center of the adjustment continuum. This was calculated first for each pixel and then this difference was averaged across the whole image to provide the mean chroma difference from grey for each image. Trials where the selected image was in the opposite direction to that expected given the adaptation image (i.e. towards the positive image) were given a negative chroma score. This allowed us to compare different images using

a perceptually meaningful measure. Higher chroma difference scores indicate that more intensely-coloured images were selected to null the afterimage, indicating that the afterimage was stronger in that trial or condition.

The pre-adaptation settings were first analyzed for group differences (to make sure both groups could set to greyscale) and to confirm that pre-adaptation subjective greyscale points had mean CIE chroma no different from zero. A 2 (image type: diagnostic/non-diagnostic) x 2 (orientation: upright/upside-down) x 2 (group: ASD/typical) mixed ANOVA was run on the pre-adaptation chroma settings. There was a marginal effect of image type ($F(1,20) = 4.13$, $p = .056$), and a marginal interaction of orientation with image type ($F(1,20) = 4.35$, $p = .050$). One-sample *t*-tests indicated that the mean chroma of images selected as appearing subjectively greyscale in the pre-adaptation settings did not deviate significantly from zero in any of the four conditions (natural up, man-made up, natural down, man-made down) (largest $t = 1.75$, smallest $p = .095$). There were no other significant main effects or interactions. There was no effect of autism on the achromatic settings ($F(1,20) = 1.40$, $p = .250$).

Following Lupyan's finding, we predicted that afterimages would be stronger (i.e. require more nulling) for diagnostic than non-diagnostic scenes, but only in the upright condition. We also predicted that that the effect of colour diagnosticity on afterimage strength would be absent or attenuated in autistic adults, compared to typical adults. Finally, if adaptation is generally reduced in autism, afterimages should be generally weaker across all conditions.

Mean chroma differences were submitted to a 2 (image type: diagnostic/non-diagnostic) x 2 (orientation: upright/upside-down) x 2 (group: ASD/typical) mixed ANOVA. The only significant effect was a main effect of image type ($F(1,20) = 29.25$, $p < .001$).

Unexpectedly, there was neither a main effect of orientation ($F(1,20) = 1.95$, $p = .178$), nor a main effect of group ($F(1,20) = 0.08$, $p = .781$). There were also no significant interactions of orientation with image type ($F(1,20) = 1.53$, $p = .231$), orientation with group ($F(1,20) = 0.50$, $p = .489$) image type with group ($F(1,20) = 0.06$, $p = .811$), and no three-way interaction ($F(1,20) = 0.02$, $p = .905$) (figure 6.2).

Bayesian statistics allow inferences to be made about the null, in addition to the alternative, hypothesis. A Bayesian approach was used to assess the strength of the evidence (i.e. the probability of the null hypothesis being true, given the data, $p(H_0|D)$; or the probability of the alternative hypothesis being true, given the data, $p(H_0|D)$) for the crucial three main effects (orientation, image type and group) and the image type x orientation interaction. This was based on calculating the Bayesian information criterion (BIC) following the methods and materials provided by Masson (2011). The effect of image type returned a very high BIC value ($\Delta BIC = -16.74$) indicating very strong support for the alternative hypothesis ($p(H_1|D) > .999$). Support for the null over the alternative was observed for the main effect of orientation ($\Delta BIC = 1.05$, $p(H_0|D) = .628$) and for the interaction between orientation and image type ($\Delta BIC = 2.09$, $p(H_0|D) = .676$). Probabilities below .75 indicate 'weak' support for the null hypothesis in these cases (Raftery, 1999), whereas the main effect of group ($\Delta BIC = 3.00$, $p(H_0|D) = .818$) offers stronger support, with a Bayes factor above 3 (Dienes, 2011) and 'positive' support for the null (Raftery, 1999).

One important factor which affects the magnitude of aftereffects is time since the offset of the adapting stimulus. Afterimages fade with time, so the magnitude recorded in this study is directly related to the time spent adjusting each image. Accordingly, participant mean reaction times (RTs) showed a strong negative correlation with mean afterimage

magnitude ($r(20) = -.53$, $p = .01$), indicating that those who took longer to make their adjustments tended to report weaker afterimages. Participants' reaction times (RTs) were submitted to a 2 (image type: diagnostic/non-diagnostic) x 2 (orientation: upright/upside-down) x 2 (group: ASD/typical) mixed ANOVA (as was carried out for the aftereffect magnitude data). This analysis revealed no significant main effects or interactions (largest F = 2.88, smallest p = .105), indicating that RTs were consistent across conditions and groups (overall mean RT = 6.34 seconds).

A t-test on diagnostic and non-diagnostic images (collapsed across orientation and group) found that nulling settings were significantly more chromatic for diagnostic (M = 12.18, SEM = 1.67) than non-diagnostic images (M = 7.09, SEM = 1.26), $t(21) = 5.53$, $p < .001$).

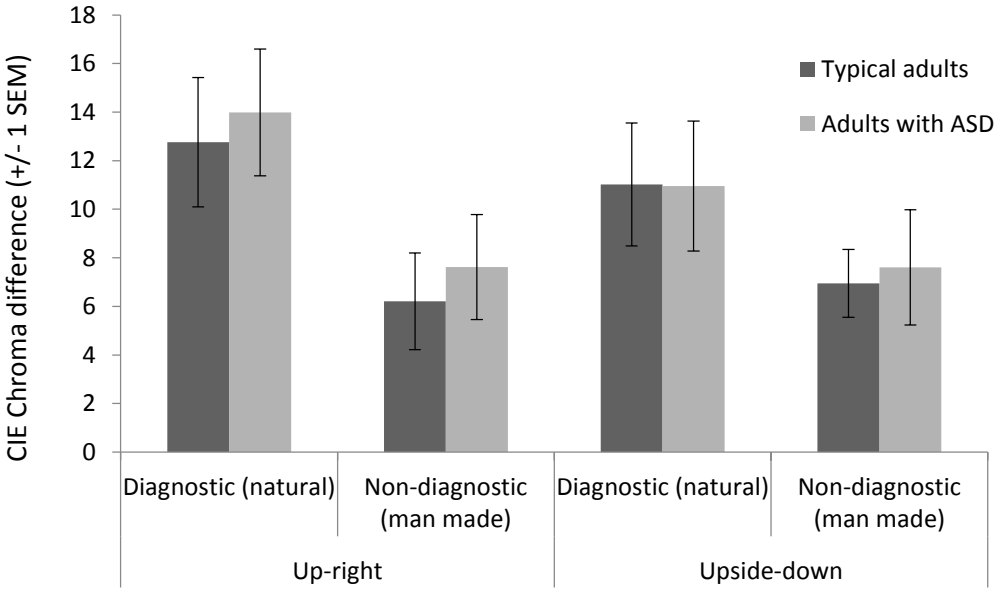

Figure 6.2. Mean whole-image chroma settings for nulling afterimages following adaptation (in CIE chroma units). The diagnostic images were of a castle and beach, while the non-diagnostic were of books and painted huts. Error bars show +/- 1 SEM.

6.5 Discussion

The present study sought to investigate whether evidence for attenuated adaptation to faces (Ewing, Leach, et al., 2013; Ewing, Pellicano, et al., 2013; Fiorentini et al., 2012; Pellicano, Rhodes, & Calder, 2013; Rhodes, Ewing, et al., 2014) and number (Turi et al., 2015) in autism generalize to colour. Furthermore it investigated whether the influence of scene content on the strength of colour afterimages (Lupyan, in press) were different for autistic and typical adults, as the hypo-priors account of autism (Pellicano, 2013; Pellicano & Burr, 2012) would predict. Contrary to expectations, there were no significant group differences observed in the strength of colour afterimages following adaptation to whole-scene images. This finding is in the context of a failure to replicate Lupyan's interaction between adapting image orientation and image content on the strength of the afterimage. The images containing colour diagnostic natural scenes (castle and beach) had stronger afterimages than those containing colour non-diagnostic man-made scenes (books and huts) but, critically, the orientation of the adapting image did not have any significant effect on the strength of the afterimage, regardless of the content (colour diagnostic or not) of the image.

The absence of any group differences with regards to adaptation strength suggests that colour adaptation is intact in autistic adults. The basic phenomenon of colour afterimages is attributable to adaptation very early on in the visual stream, at the level of photoreceptors and ganglion cells in the retina (Zaidi et al., 2012), whereas the effects of luminance contours which help to make the Spanish Castle effect so vivid reflect the contribution of areas of visual cortex modifying the afterimage signal received from the retina to create the perception of filling-in (Anstis et al., 2012; van Lier et al., 2009; Vergeer, Anstis, & van Lier, 2015). In contrast, previous demonstration of attenuated

adaptation in autism has been for faces and number, which are processed at a higher level, in regions of the temporal (i.e., fusiform face area – Kanwisher et al., 1997) and parietal cortex (see Piazza & Izard, 2009 for a review), respectively. In the current study, no general reduction in afterimage strength was observed in the ASD group relative to the typical group. This could suggest that only higher-level adaptation is attenuated, where priors and top-down influence may have more impact on perception, whilst low-level sensory adaptation mechanisms in adults with autism are intact.

It remains possible that colour adaptation is atypical in children on the autism spectrum even if not attenuated in adulthood. Children have a more limited experience of the world than adults and may not yet have developed compensatory strategies for any perceptual atypicality they experience. There is some suggestion that atypical adaptation may be more prevalent in children with autism than adults. For example, whilst children with autism have atypical adaptation aftereffects for facial identity and expression (Ewing, Pellicano, et al., 2013; Rhodes, Ewing, et al., 2014), autistic adults are no different to those of typical adults (Cook et al., 2014). This latter finding is, however, at odds with the finding that the parents of children with autism also exhibit reduced face aftereffects (Fiorentini et al., 2012), suggesting that atypical adaptation of certain stimuli may be part of the broader autism phenotype even in adults. Further studies investigating adaptation in autism at different stages in development for a range of perceptual domains are needed to examine these issues.

Our second hypothesis related to the proposed effect of scene content on colour adaptation. Lupyan proposed that there are top-down effects of knowledge of a scene on the strength of adaptation and that natural scenes therefore have stronger afterimages than man-made scenes, which are not colour diagnostic. While we did observe that natural scenes elicited

stronger afterimages than man-made scenes, and this effect was equally strong across groups, we nevertheless failed to replicate Lupyan's finding of an interaction of this effect with the orientation of the image. Afterimages in our study were of equivalent strength regardless of whether the adapting image was upright or upside down. This lack of an interaction suggests that the difference in adaptation for type of scene may well be due to low-level differences in the images (e.g., chroma) rather than top-down effects of knowledge.

Notably, there were differences in the methods of the current study and Lupyan's, which may explain the discrepancy between the findings. First, the after-effect was measured in different units across studies. Specifically, we quantified the after-effect using the perceptually meaningful measure of chroma rather than units in computational HSL space. However, this difference cannot explain the lack of interaction between image type and orientation in our study[2]. Second, Lupyan's diagnosticity-orientation interaction was derived from comparing participants from different experiments, that is, between participants, while in our study all participants completed all conditions. Seeing both upright and inverted images within one experiment, even if the images are different, could weaken the extent to which the observer views the upright images as a meaningful scene. If so, top-down effects of knowledge may have been weaker in our study than Lupyan's because of our within-subject rather than between-groups manipulation of orientation. However, we feel that the within-subjects design of our study is preferable, since it avoids the possibility of cohort effects, where one group happens to have stronger afterimages than the other, and has greater power by virtue of being based on within-participant factors. The Bayesian analysis confirms that the data provides only 'weak' (Dienes, 2011;

[2] In order to compare directly to Lupyan, analysis in terms of HSL steps was also carried out. Using this measure of afterimage strength we still found no effect of orientation and no orientation-diagnosticity interaction.

Raftery, 1999) support for the null hypothesis relating to main effects of orientation and image type and the interaction between these factors. The lack of a verified top-down effect on colour adaptation in our study means that we are not able to state whether effects of knowledge on adaptation are reduced in autism.

In the present study, we found no relationship between mean response times and the mean measured afterimage magnitudes (i.e. mean response times to all conditions are approximately equal, see results). Lupyan (2015) reported an average adjustment time of just under 10s, while our participants appear to have made their adjustments more rapidly, the average adjustment taking around 6s. With time, the strength of the afterimage fades (Kline & Nestor, 1977), it is therefore likely that Lupyan's (2015) participants were making their achromatic settings for relatively weaker afterimages, compared to the participants in the present study. It is possible that these effects are somewhat sensitive to the time-course of adaptation – perhaps being more likely to be evident when the observer is adjusting to null a less intense afterimage (later) than when the afterimage is stronger (earlier). A weaker afterimage may cause the observer to be less certain of the bottom-up sensory input, and thus more likely to rely on top-down information in making their judgment.

Our data suggest that colour adaptation is intact in autism, even if other types of adaptation are attenuated. This may be due to the lower-level sensory nature of colour adaptation compared to those domains in which adaptation has been shown to be attenuated in autism. Further research should seek to clarify the role of top-down knowledge on colour adaptation in both children and adults with and without autism. Other studies have shown the effect of prior knowledge or expectation on visual perception. An auditory label which is congruent with an image otherwise suppressed

from visual awareness by inter-ocular rivalry can improve detection speed and sensitivity, and is specific to the object (e.g. hearing "pumpkin" speeds detection of pumpkins but not chairs) (Lupyan & Ward, 2013). Similarly, when asked to adjust a familiar and prototypically-coloured object to grayscale, observers appear to overcompensate for the influence of memory colours. For example, observers' grayscale settings for a banana tend to be biased in the blue direction, apparently to counteract the influence of the yellow memory colour (Hansen et al., 2006; Olkkonen, Hansen, & Gegenfurtner, 2008). As such, the brain appears to use prior experience of familiar objects with distinct colours to predictively code their appearance (Bannert & Bartels, 2013). If such effects reflect top-down modulation of the cortical representation of those objects through the application of priors to current percepts, then we should find individual differences in the strength of these effects depending on autism symptomatology, age and experience with the objects in question.

6.6 Conclusion

The present study is the first to demonstrate that colour adaptation is typical in autistic adults, contrary to the evidence for attenuated adaptation in other domains of visual perception. We suggest that this is likely due to the early level at which colour adaptation occurs. The effects of hypo-priors on adaptation in autism appear to be restricted to higher-level visual stimuli and may be strongest in children with autism and reduced or absent in autistic adults.

References

Albrecht, A. R., & Scholl, B. J. (2010). Perceptually Averaging in a Continuous Visual World: Extracting Statistical Summary Representations Over Time. *Psychological Science, 21*(4), 560-567. doi: 10.1177/0956797610363543

Allik, J., Toom, M., Raidvee, A., Averin, K., & Kreegipuu, K. (2013). An almost general theory of mean size perception. *Vision Research, 83*, 25-39. doi: 10.1016/j.visres.2013.02.018

Allik, J., Toom, M., Raidvee, A., Averin, K., & Kreegipuu, K. (2014). Obligatory averaging in mean size perception. *Vision Research, 101*, 34-40. doi: 10.1016/j.visres.2014.05.003

Alvarez, G. A. (2011). Representing multiple objects as an ensemble enhances visual cognition. *Trends in Cognitive Sciences, 15*(3), 122-131. doi: 10.1016/j.tics.2011.01.003

Alvarez, G. A., & Oliva, A. (2008). The representation of simple ensemble visual features outside the focus of attention. *Psychological Science, 19*(4), 392-398. doi: 10.1111/j.1467-9280.2008.02098.x

Alvarez, G. A., & Oliva, A. (2009). Spatial ensemble statistics are efficient codes that can be represented with reduced attention. *Proceedings of the National Academy of Sciences of the United States of America, 106*(18), 7345-7350. doi: 10.1073/pnas.0808981106

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). Washington, DC: Author.

Anstis, S., Vergeer, M., & van Lier, R. (2012). Looking at two paintings at once: Luminance edges can gate colours. *i-Perception, 3*(8), 515-518. doi: 10.1068/i0537sas

Ariely, D. (2001). Seeing sets: Representation by statistical properties. *Psychological Science, 12*(2), 157-162. doi: 10.1111/1467-9280.00327

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. doi: 10.3758/Pp.70.7.1325

Athanasopoulos, P. (2011). Colour and bilingual cognition. *Language and Bilingual Cognition*, 241-261.

Athanasopoulos, P., Dering, B., Wiggett, A., Kuipers, J. R., & Thierry, G. (2010). Perceptual shift in bilingualism: Brain potentials reveal plasticity in pre-attentive colour perception. *Cognition, 116*(3), 437-443. doi: 10.1016/j.cognition.2010.05.016

Attarha, M., Moore, C. M., & Vecera, S. P. (2014). Summary Statistics of Size: Fixed Processing Capacity for Multiple Ensembles but Unlimited Processing Capacity for Single Ensembles. *Journal of Experimental Psychology-Human Perception and Performance, 40*(4), 1440-1449. doi: 10.1037/A0036206

Baijal, S., Nakatani, C., van Leeuwen, C., & Srinivasan, N. (2013). Processing statistics: An examination of focused and distributed attention using event related potentials. *Vision Research, 85*, 20-25. doi: 10.1016/j.visres.2012.09.018

Bakeman, R. (2005). Recommended effect size statistics for repeated measures designs. *Behavior Research Methods, 37*(3), 379-384. doi: 10.3758/Bf03192707

Bannert, M. M., & Bartels, A. (2013). Decoding the yellow of a gray banana. *Current Biology, 23*(22), 2268-2272. doi: 10.1016/j.cub.2013.09.016

Banno, H., & Saiki, J. (2012). Calculation of the mean circle size does not circumvent the bottleneck of crowding. *Journal of Vision, 12*(11). doi: 10.1167/12.11.13

Baron-Cohen, S., Wheelwright, S., Skinner, R., Martin, J., & Clubley, E. (2001). The autism-spectrum quotient (AQ): evidence from Asperger syndrome/high-functioning autism, males and females, scientists and mathematicians. *Journal of Autism and Developmental Disorders, 31*(1), 5-17.

Bauer, B. (2009a). The danger of trial-by-trial knowledge of results in perceptual averaging studies. *Attention Perception & Psychophysics, 71*(3), 655-665. doi: 10.3758/App.71.3.655

Bauer, B. (2009b). Does Stevens's Power Law for Brightness Extend to Perceptual Brightness Averaging? *Psychological Record, 59*(2), 171-185.

Bays, P. M. (2015). Spikes not slots: noise in neural populations limits working memory. *Trends Cogn Sci, 19*(8), 431-438. doi: 10.1016/j.tics.2015.06.004

Behrmann, M., Thomas, C., & Humphreys, K. (2006). Seeing it differently: visual processing in autism. *Trends in Cognitive Sciences, 10*(6), 258-264. doi: 10.1016/j.tics.2006.05.001

Bird, C. M., Berens, S. C., Horner, A. J., & Franklin, A. (2014). Categorical encoding of colour in the brain. *Proceedings of the National Academy of Sciences of the United States of America, 111*(12), 4590-4595. doi: 10.1073/pnas.1315275111

Bolte, S., Holtmann, M., Poustka, F., Scheurich, A., & Schmidt, L. (2007). Gestalt perception and local-global processing in high-functioning autism. *Journal of Autism and Developmental Disorders, 37*(8), 1493-1504. doi: 10.1007/s10803-006-0231-x

Brady, T. F., & Alvarez, G. A. (2011). Hierarchical Encoding in Visual Working Memory: Ensemble Statistics Bias Memory for Individual Items. *Psychological Science, 22*(3), 384-392. doi: 10.1177/0956797610397956

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision, 10*(4), 433-436. doi: 10.1163/156856897x00357

Brainard, D. H., & Hurlbert, A. C. (2015). Colour Vision: Understanding #TheDress. *Current Biology, 25*(13), R551-554. doi: 10.1016/j.cub.2015.05.020

Brophy, A. L. (1986). Alternatives to a Table of Criterion Values in Signal-Detection-Theory. *Behavior Research Methods Instruments & Computers, 18*(3), 285-286. doi: 10.3758/Bf03204400

Brown, R. O., & MacLeod, D. I. A. (1997). Colour appearance depends on the variance of surround colours. *Current Biology, 7*(11), 844-849. doi: 10.1016/S0960-9822(06)00372-1

Buchsbaum, G. (1980). A Spatial Processor Model for Object Colour-Perception. *Journal of the Franklin Institute-Engineering and Applied Mathematics, 310*(1), 1-26. doi: 10.1016/0016-0032(80)90058-7

Burns, B., & Shepp, B. E. (1988). Dimensional interactions and the structure of psychological space: The representation of hue, saturation, and brightness. *Perception & Psychophysics, 43*(5), 494-507.

Calder, A. J., Young, A. W., Perrett, D. I., Etcoff, N. L., & Rowland, D. (1996). Categorical perception of morphed facial expressions. *Visual Cognition, 3*(2), 81-117. doi: 10.1080/713756735

Chong, S. C., Joo, S. J., Emmanouil, T. A., & Treisman, A. (2008). Statistical processing: not so implausible after all. *Percept Psychophys, 70*(7), 1327-1334; discussion 1335-1326. doi: 10.3758/PP.70.7.1327

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

Chong, S. C., & Treisman, A. (2005a). Attentional spread in the statistical processing of visual displays. *Perception & Psychophysics, 67*(1), 1-13. doi: 10.3758/Bf03195009

Chong, S. C., & Treisman, A. (2005b). Statistical processing: computing the average size in perceptual groups. *Vision Research, 45*(7), 891-900. doi: 10.1016/j.visres.2004.10.004

Choo, H., & Franconeri, S. L. (2010). Objects with reduced visibility still contribute to size averaging. *Atten Percept Psychophys, 72*(1), 86-99. doi: 10.3758/APP.72.1.86

Clifford, A., Franklin, A., Holmes, A., Drivonikou, V. G., Ozgen, E., & Davies, I. R. L. (2012). Neural correlates of acquired colour category effects. *Brain and Cognition, 80*(1), 126-143. doi: 10.1016/j.bandc.2012.04.011

Clifford, A., Holmes, A., Davies, I. R. L., & Franklin, A. (2010). Colour categories affect pre-attentive colour perception. *Biological Psychology, 85*(2), 275-282. doi: 10.1016/j.biopsycho.2010.07.014

Clifford, C. W. G. (2012). Visual Perception: Knowing What to Expect. *Current Biology, 22*(7), R223-R225. doi: 10.1016/j.cub.2012.02.019

Collier, G. A. (1976). Further Evidence for Universal Colour Categories. *Language, 52*(4), 884-890. doi: 10.2307/413300

Constantino, J. N., & Gruber, C. P. (2012). *Social Responsiveness Scale*™ (2nd ed.). Torrance, California: Western Psychological Services.

Cook, R., Brewer, R., Shah, P., & Bird, G. (2014). Intact facial adaptation in autistic adults. *Autism Research, 7*(4), 481-490. doi: 10.1002/aur.1381

Corbett, J. E., & Melcher, D. (2014a). Characterizing ensemble statistics: mean size is represented across multiple frames of reference. *Attention Perception & Psychophysics, 76*(3), 746-758. doi: 10.3758/s13414-013-0595-x

Corbett, J. E., & Melcher, D. (2014b). Stable Statistical Representations Facilitate Visual Search. *Journal of Experimental Psychology-Human Perception and Performance, 40*(5), 1915-1925. doi: 10.1037/A0037375

Corbett, J. E., & Oriet, C. (2011). The whole is indeed more than the sum of its parts: Perceptual averaging in the absence of individual item representation. *Acta Psychologica, 138*(2), 289-301. doi: 10.1016/j.actpsy.2011.08.002

Corbett, J. E., & Song, J. H. (2014). Statistical extraction affects visually guided action. *Vis cogn, 22*(7), 881-895. doi: 10.1080/13506285.2014.927044

Corbett, J. E., Wurnitsch, N., Schwartz, A., & Whitney, D. (2012). An aftereffect of adaptation to mean size. *Visual Cognition, 20*(2), 211-231. doi: 10.1080/13506285.2012.657261

Crawford, L. E., Huttenlocher, J., & Engebretson, P. H. (2000). Category effects on estimates of stimuli: Perception or reconstruction? *Psychological Science, 11*(4), 280-284. doi: 10.1111/1467-9280.00256

Dakin, S. (2001). Information limit on the spatial integration of local orientation signals. *J Opt Soc Am A Opt Image Sci Vis, 18*(5), 1016-1026.

Dakin, S. (2012). Seeing statistical regularities. In J. Wagemans (Ed.), *Oxford Handbook of Perceptual Organisation* (pp. 150-166). Oxford: Oxford University Press.

Dakin, S., & Frith, U. (2005). Vagaries of visual perception in autism. *Neuron, 48*(3), 497-507. doi: 10.1016/j.neuron.2005.10.018

Daoutis, C. A., Pilling, M., & Davies, I. R. L. (2006). Categorical effects in visual search for colour. *Visual Cognition, 14*(2), 217-240. doi: 10.1080/13506280600158670

Davidoff, J., Davies, I., & Roberson, D. (1999). Colour categories in a stone-age tribe. *Nature, 398*(6724), 203-204. doi: 10.1038/18335

Daw, N. W. (1962). Why after-images are not seen in normal circumstances. *Nature, 196*, 1143-1145.

De Fockert, J. W., & Marchant, A. P. (2008). Attention modulates set representation by statistical properties. *Perception & Psychophysics, 70*(5), 789-794. doi: 10.3758/Pp.70.5.789

De Fockert, J. W., & Wolfenstein, C. (2009). Rapid extraction of mean identity from sets of faces. *Quarterly Journal of Experimental Psychology, 62*(9), 1716-1722. doi: 10.1080/17470210902811249

de Gardelle, V., & Summerfield, C. (2011). Robust averaging during perceptual judgment. *Proceedings of the National Academy of Sciences of the United States of America, 108*(32), 13341-13346. doi: 10.1073/pnas.1104517108

Demb, J. B., & Brainard, D. H. (2010). Vision: Neurons show their true colours. *Nature, 467*(7316), 670-671. doi: 10.1038/467670b

Demeyere, N., Rzeskiewicz, A., Humphreys, K. A., & Humphreys, G. W. (2008). Automatic statistical processing of visual properties in simultanagnosia. *Neuropsychologia, 46*(11), 2861-2864. doi: 10.1016/j.neuropsychologia.2008.05.014

Derrington, A. M., Krauskopf, J., & Lennie, P. (1984). Chromatic Mechanisms in Lateral Geniculate-Nucleus of Macaque. *Journal of Physiology, 357*, 241-265.

Dienes, Z. (2011). Bayesian Versus Orthodox Statistics: Which Side Are You On? *Perspectives on Psychological Science, 6*(3), 274-290. doi: 10.1177/1745691611406920

Drivonikou, G., Clifford, A., Franklin, A., Ozgen, E., & Davies, I. R. L. (2011). Category training affects colour discrimination but only in the right visual field. In C. P. Biggam, C. A. Hough, C. J. Kay & D. R. Simmons (Eds.), *New Directions in Colour Studies* (pp. 251-264). Amsterdam: John Benjamin.

Emmanouil, T. A., & Treisman, A. (2008). Dividing attention across feature dimensions in statistical processing of perceptual groups. *Percept Psychophys, 70*(6), 946-954.

Ewing, L., Leach, K., Pellicano, E., Jeffery, L., & Rhodes, G. (2013). Reduced face aftereffects in autism are not due to poor attention. *PLoS One, 8*(11), e81353. doi: 10.1371/journal.pone.0081353

Ewing, L., Pellicano, E., & Rhodes, G. (2013). Atypical updating of face representations with experience in children with autism. *Developmental Science, 16*(1), 116-123. doi: 10.1111/desc.12007

Fiorentini, C., Gray, L., Rhodes, G., Jeffery, L., & Pellicano, E. (2012). Reduced face identity aftereffects in relatives of children with autism. *Neuropsychologia, 50*(12), 2926-2932. doi: 10.1016/j.neuropsychologia.2012.08.019

Fletcher, R. (1980). *The City University Colour Vision Test, 2nd Edition.* London: Keeler.

Fouriezos, G., Rubenfeld, S., & Capstick, G. (2008). Visual statistical decisions. *Perception & Psychophysics, 70*(3), 456-464.

Franklin, A., Sowden, P., Burley, R., Notman, L., & Alder, E. (2008). Colour Perception in Children with Autism. *Journal of Autism and Developmental Disorders, 38*(10), 1837-1847. doi: 10.1007/s10803-008-0574-6

Friston, K. J., Lawson, R., & Frith, C. D. (2013). On hyperpriors and hypopriors: comment on Pellicano and Burr. *Trends Cogn Sci, 17*(1), 1. doi: 10.1016/j.tics.2012.11.003

Frith, U., & Happe, F. (1994). Autism – Beyond Theory of Mind. *Cognition, 50*(1-3), 115-132. doi: 10.1016/0010-0277(94)90024-8

Fujita, T., Yamasaki, T., Kamio, Y., Hirose, S., & Tobimatsu, S. (2011). Parvocellular pathway impairment in autism spectrum disorder: Evidence from visual evoked potentials. *Research in Autism Spectrum Disorders, 5*(1), 277-285. doi: 10.1016/j.rasd.2010.04.009

Gegenfurtner, K. R., Bloj, M., & Toscani, M. (2015). The many colours of 'the dress'. *Current Biology, 25*(13), R543-544. doi: 10.1016/j.cub.2015.04.043

Giesel, M., & Gegenfurtner, K. R. (2010). Colour appearance of real objects varying in material, hue, and shape. *Journal of Vision, 10*(9). doi: 10.1167/10.9.10

Gilbert, A. L., Regier, T., Kay, P., & Ivry, R. B. (2006). Whorf hypothesis is supported in the right visual field but not the left. *Proceedings of the National Academy of Sciences of the United States of America, 103*(2), 489-494. doi: 10.1073/pnas.0509868103

Gliga, T., Bedford, R., Charman, T., Johnson, M. H., & Team, Basis. (2015). Enhanced Visual Search in Infancy Predicts Emerging Autism Symptoms. *Current Biology, 25*(13), 1727-1730. doi: 10.1016/j.cub.2015.05.011

Haberman, J., & Whitney, D. (2007). Rapid extraction of mean emotion and gender from sets of faces. *Current Biology, 17*(17), R751-753. doi: 10.1016/j.cub.2007.06.039

Haberman, J., & Whitney, D. (2009). Seeing the Mean: Ensemble Coding for Sets of Faces. *Journal of Experimental Psychology-Human Perception and Performance, 35*(3), 718-734. doi: 10.1037/A0013899

Haberman, J., & Whitney, D. (2010). The visual system discounts emotional deviants when extracting average expression. *Attention Perception & Psychophysics, 72*(7), 1825-1838. doi: 10.3758/App.72.7.1825

Haberman, J., & Whitney, D. (2011). Efficient summary statistical representation when change localization fails. *Psychon Bull Rev, 18*(5), 855-859. doi: 10.3758/s13423-011-0125-6

Haberman, J., & Whitney, D. (2012). Ensemble Perception: Summarizing the scene and broadening the limits of visual processing. In J. Wolfe & L. Robertson (Eds.), From Perception to Consciousness: Searching with Anne Treisman (pp. 339-349). Oxford: Oxford University Press.

Hansen, T., Giesel, M., & Gegenfurtner, K. R. (2008). Chromatic discrimination of natural objects. *Journal of Vision, 8*(1), 2 1-19. doi: 10.1167/8.1.2

Hansen, T., Olkkonen, M., Walter, S., & Gegenfurtner, K. R. (2006). Memory modulates colour appearance. *Nature Neuroscience, 9*(11), 1367-1368. doi: 10.1038/nn1794

Happe, F., & Frith, U. (2006). The weak coherence account: Detail-focused cognitive style in autism spectrum disorders. *Journal of Autism and Developmental Disorders, 36*(1), 5-25. doi: 10.1007/s10803-005-0039-0

He, X., Witzel, C., Forder, L., Clifford, A., & Franklin, A. (2014). Colour categories only affect post-perceptual processes when same- and different-category colours are equally discriminable. *Journal of the Optical Society of America: A, 31*(4), A322-331. doi: 10.1364/JOSAA.31.00A322

Heaton, P., Ludlow, A., & Roberson, D. (2008). When less is more: Poor discrimination but good colour memory in autism. *Research in Autism Spectrum Disorders, 2*(1), 147-156. doi: 10.1016/j.rasd.2007.04.004

Heider, E. R. (1972). Universals in Colour Naming and Memory. *Journal of Experimental Psychology, 93*(1), 10-&. doi: 10.1037/H0032606

Huang, L., Treisman, A., & Pashler, H. (2007). Characterizing the limits of human visual awareness. *Science, 317*(5839), 823-825. doi: 10.1126/science.1143515

Huttenlocher, J., Hedges, L. V., & Vevea, J. L. (2000). Why do categories affect stimulus judgment? *Journal of Experimental Psychology-General, 129*(2), 220-241. doi: 10.1037//0096-3445.129.2.220

Im, H. Y., & Chong, S. C. (2009). Computation of mean size is based on perceived size. *Attention Perception & Psychophysics, 71*(2), 375-384. doi: 10.3758/APP.71.2.375

Im, H. Y., & Chong, S. C. (2014). Mean size as a unit of visual working memory. *Perception, 43*(7), 663-676. doi: 10.1068/P7719

Im, H. Y., & Halberda, J. (2013). The effects of sampling and internal noise on the representation of ensemble average size. *Attention Perception & Psychophysics, 75*(2), 278-286. doi: 10.3758/s13414-012-0399-4

Im, H. Y., Park, Woon Ju, & Chong, Sang Chul. (2014). Ensemble statistics as units of selection. *Journal of Cognitive Psychology, 27*(1), 114-127. doi: 10.1080/20445911.2014.985301

Ishihara, S. (1973). *Ishihara's test chart for colour deficiency*. Tokyo: Kanehara Trading INC.

Juricevic, I., & Webster, M. A. (2009). Variations in normal colour vision. V. Simulations of adaptation to natural colour environments. *Visual Neuroscience, 26*(1), 133-145. doi: 10.1017/S0952523808080942

Kanwisher, N., McDermott, J., & Chun, M. M. (1997). The fusiform face area: a module in human extrastriate cortex specialized for face perception. *Journal of Neuroscience, 17*(11), 4302-4311.

Kay, P., Berlin, B., Maffi, L., Merrifield, W. R., & Cook, R. (2011). *The World Colour Survey*. Stanford: Center for the Study of Language & Information Publications.

Kay, P., & Regier, T. (2003). Resolving the question of colour naming universals. *Proceedings of the National Academy of Sciences of the United States of America, 100*(15), 9085-9089. doi: 10.1073/pnas.1532837100

Kersten, D., Mamassian, P., & Yuille, A. (2004). Object perception as Bayesian inference. *Annual Review of Psychology, 55*, 271-304. doi: 10.1146/annurev.psych.55.090902.142005

Kleiner, M., Brainard, D., & Pelli, D. (2007). What's new in Psychtoolbox-3? *Perception, 36*, 14-14.

Kline, D. W., & Nestor, S. (1977). Persistence of complementary afterimages as a function of adult age and exposure duration. *Experimental Aging Research, 3*(3), 191-201. doi: 10.1080/03610737708257102

Knill, D. C., & Pouget, A. (2004). The Bayesian brain: the role of uncertainty in neural coding and computation. *Trends Neurosci, 27*(12), 712-719. doi: 10.1016/j.tins.2004.10.007

Koh, H. C., Milne, E., & Dobkins, K. (2010). Contrast sensitivity for motion detection and direction discrimination in adolescents with autism spectrum disorders and their siblings. *Neuropsychologia, 48*(14), 4046-4056. doi: 10.1016/j.neuropsychologia.2010.10.008

Kohn, A. (2007). Visual adaptation: Physiology, mechanisms, and functional benefits. *Journal of Neurophysiology, 97*(5), 3155-3164. doi: 10.1152/jn.00086.2007

Komarova, N. L., & Jameson, K. A. (2013). A Quantitative Theory of Human Colour Choices. *Plos One, 8*(2). doi: 10.1371/journal.pone.0055986

Kraft, J. M., & Brainard, D. H. (1999). Mechanisms of colour constancy under nearly natural viewing. *Proceedings of the National Academy of Sciences of the United States of America, 96*(1), 307-312. doi: 10.1073/pnas.96.1.307

Krauskopf, J., Williams, D. R., & Heeley, D. W. (1982). Cardinal Directions of Colour Space. *Vision Research, 22*(9), 1123-1131. doi: 10.1016/0042-6989(82)90077-3

Kuehni, R. G. (2014). Unique Hues and Their Stimuli-State of the Art. *Colour Research and Application, 39*(3), 279-287. doi: 10.1002/Col.21793

Kuriki, I. (2004). Testing the possibility of average-colour perception from multi-coloured patterns. *Optical Review, 11*(4), 249-257. doi: 10.1007/s10043-004-0249-2

Lafer-Sousa, R., Hermann, K. L., & Conway, B. R. (2015). Striking individual differences in colour perception uncovered by 'the dress' photograph. *Current Biology, 25*(13), R545-546. doi: 10.1016/j.cub.2015.04.053

Lanthony, P. (1998). *Album Tritan* (2nd ed.). Paris: Laboratoire de la Vision des Couleurs.

Lanzoni, L., Melcher, D., Miceli, G., & Corbett, J. E. (2014). Global statistical regularities modulate the speed of visual search in patients with focal attentional deficits. *Frontiers in Psychology, 5*(514), 1-12. doi: 10.3389/fpsyg.2014.00514

Lawson, R. P., Rees, G., & Friston, K. J. (2014). An aberrant precision account of autism. *Frontiers in Human Neuroscience, 8*. doi: 10.3389/Fnhum.2014.00302

Leib, A. Y., Fischer, J., Liu, Y., Qiu, S., Robertson, L., & Whitney, D. (2014). Ensemble crowd perception: A viewpoint-invariant mechanism to represent average crowd identity. *Journal of Vision, 14*(8). doi: 10.1167/14.8.26

Leib, A. Y., Fischer, J., Liu, Y., Whitney, D., & Robertson, L. (2013). Ensemble Crowd Perception: A Viewpoint Invariant Mechanism to Represent Average Crowd Identity. *Journal of Vision, 13*(9), 424.

Leib, A. Y., Landau, A. N., Baek, Y., Chong, S. C., & Robertson, L. (2012). Extracting the mean size across the visual field in patients with mild, chronic unilateral neglect. *Frontiers in Human Neuroscience, 6*. doi: 10.3389/Fnhum.2012.00267

Leib, A. Y., Puri, A. M., Fischer, J., Bentin, S., Whitney, D., & Robertson, L. (2012). Crowd perception in prosopagnosia. *Neuropsychologia, 50*(7), 1698-1707. doi: 10.1016/j.neuropsychologia.2012.03.026

Li, A., & Lennie, P. (1997). Mechanisms underlying segmentation of coloured textures. *Vision Research, 37*(1), 83-97.

Lindsey, D. T., & Brown, A. M. (2009). World Colour Survey colour naming reveals universal motifs and their within-language diversity. *Proceedings of the National Academy of Sciences of the United States of America, 106*(47), 19785-19790. doi: 10.1073/pnas.0910981106

Ludlow, A. K., Heaton, P., Hill, E., & Franklin, A. (2014). Colour obsessions and phobias in autism spectrum disorders: The case of JG. *Neurocase, 20*(3), 296-306. doi: 10.1080/13554794.2013.770880

Lupyan, G. (2013). Semantic effects on colour afterimages. *Journal of Vision, 13*(9), 466-466. doi: 10.1167/13.9.466

Lupyan, G. (in press). Object knowledge changes visual appearance: Semantic effects on colour afterimages. *Acta Psychologica*. http://sapir.psych.wisc.edu/publications/

Lupyan, G., & Ward, E. J. (2013). Language can boost otherwise unseen objects into visual awareness. *Proceedings of the National Academy of Sciences, 110*(35), 14196-14201. Doi: 10.1073/pnas.1303312110

Marchant, A. P., & De Fockert, J. W. (2009). Priming by the mean representation of a set. *Quarterly Journal of Experimental Psychology, 62*(10), 1889-1895. Doi: 10.1080/17470210902871045

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 Psychologica, 142*(2), 245-250. Doi: 10.1016/j.actpsy.2012.11.002

Masson, M. E. J. (2011). A tutorial on a practical Bayesian alternative to null-hypothesis significance testing. *Behavior Research Methods, 43*(3), 679-690. Doi: 10.3758/s13428-010-0049-5

Maule, J., & Franklin, A. (2015). Effects of ensemble complexity and perceptual similarity on rapid averaging of hue. *Journal of Vision, 15*(4), 1-18. Doi: 10.1167/15.4.6

Maule, J., & Franklin, A. (under review). Accurate rapid averaging of multi-hue ensembles is due to a limited capacity sub-sampling mechanism.

Maule, J., Witzel, C., & Franklin, A. (2014). Getting the gist of multiple hues: metric and categorical effects on ensemble perception of hue. *Journal of the Optical Society of America: A, 31*(4), A93-A102. Doi: 10.1364/Josaa.31.000a93

McCollough, C. (1965). Colour Adaptation of Edge-Detectors in Human Visual System. *Science, 149*(3688), 1115-1116. Doi: 10.1126/science.149.3688.1115

McDermott, K. C., & Webster, M. A. (2012). Uniform colour spaces and natural image statistics. *Journal of the Optical Society of America: A, 29*(2), A182-A187.

Michael, E., de Gardelle, V., & Summerfield, C. (2014). Priming by the variability of visual information. *Proceedings of the National Academy of Sciences of the United States of America, 111*(21), 7873-7878. Doi: 10.1073/pnas.1308674111

Milojevic, Z., Ennis, R. J., & Gegenfurtner, K. R. (2014). Colour classification of leaves. *Perception, 43* (ECVP Abstract Supplement), 146.

Mottron, L., Burack, J. A., Iarocci, G., Belleville, S., & Enns, J. T. (2003). Locally oriented perception with intact global processing among adolescents with high-functioning autism: evidence from multiple paradigms. *Journal of Child Psychology and Psychiatry, 44*(6), 904-913.

Mottron, L., Dawson, M., Soulieres, I., Hubert, B., & Burack, J. (2006). Enhanced perceptual functioning in autism: an update, and eight principles of autistic perception. *Journal of Autism and Developmental Disorders, 36*(1), 27-43. Doi: 10.1007/s10803-005-0040-7

Myczek, K., & Simons, D. J. (2008). Better than average: Alternatives to statistical summary representations for rapid judgments of average size. *Perception & Psychophysics, 70*(5), 772-788. Doi: 10.3758/Pp.70.5.772

Olkkonen, M., & Allred, S. R. (2014). Short-term memory affects colour perception in context. *PloS One, 9*(1), e86488. Doi: 10.1371/journal.pone.0086488

Olkkonen, M., Hansen, T., & Gegenfurtner, K. R. (2008). Colour appearance of familiar objects: effects of object shape, texture, and illumination changes. *Journal of Vision, 8*(5), 13 11-16. Doi: 10.1167/8.5.13

Olkkonen, M., Hansen, T., & Gegenfurtner, K. R. (2009). Categorical colour constancy for simulated surfaces. *Journal of Vision, 9*(12). Doi: 10.1167/9.12.6

Olkkonen, M., McCarthy, P. F., & Allred, S. R. (2014). The central tendency bias in colour perception: effects of internal and external noise. *Journal of Vision, 14*(11). Doi: 10.1167/14.11.5

Olkkonen, M., Witzel, C., Hansen, T., & Gegenfurtner, K. R. (2010). Categorical colour constancy for real surfaces. *Journal of Vision, 10*(9). Doi: 10.1167/10.9.16

Oriet, C., & Brand, J. (2013). Size averaging of irrelevant stimuli cannot be prevented. *Vision Research, 79*, 8-16. Doi: 10.1016/j.visres.2012.12.004

Parkes, L., Lund, J., Angelucci, A., Solomon, J. A., & Morgan, M. (2001). Compulsory averaging of crowded orientation signals in human vision. *Nature Neuroscience, 4*(7), 739-744. Doi: 10.1038/89532

Pavlovskaya, M., Soroker, N., Bonneh, Y. S., & Hochstein, S. (2015). Computing an average when part of the population is not perceived. *Journal of Cognitive Neuroscience, 27*(7), 1397-1411. Doi: 10.1162/jocn_a_00791

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision, 10*(4), 437-442. Doi: 10.1163/156856897x00366

Pellicano, E. (2013). Sensory Symptoms in Autism: A Blooming, Buzzing Confusion? *Child Development Perspectives, 7*(3), 143-148. Doi: 10.1111/Cdep.12031

Pellicano, E., & Burr, D. (2012). When the world becomes 'too real': a Bayesian explanation of autistic perception. *Trends in Cognitive Sciences, 16*(10), 504-510. Doi: 10.1016/j.tics.2012.08.009

Pellicano, E., Jeffery, L., Burr, D., & Rhodes, G. (2007). Abnormal adaptive face-coding mechanisms in children with autism spectrum disorder. *Current Biology, 17*(17), 1508-1512. Doi: 10.1016/j.cub.2007.07.065

Pellicano, E., Rhodes, G., & Calder, A. J. (2013). Reduced gaze aftereffects are related to difficulties categorising gaze direction in children with autism. *Neuropsychologia, 51*(8), 1504-1509. Doi: 10.1016/j.neuropsychologia.2013.03.021

Piazza, M., & Izard, V. (2009). How humans count: numerosity and the parietal cortex. *Neuroscientist, 15*(3), 261-273. Doi: 10.1177/1073858409333073

Pilling, M., Wiggett, A., Ozgen, E., & Davies, I. R. L. (2003). Is colour "categorical perception" really perceptual? *Memory & Cognition, 31*(4), 538-551. Doi: 10.3758/Bf03196095

Powell, G., Bompas, A., & Sumner, P. (2012). Making the incredible credible: afterimages are modulated by contextual edges more than real stimuli. *Journal of Vision, 12*(10). Doi: 10.1167/12.10.17

Raftery, A. E. (1999). Bayes factors and BIC – Comment on "A critique of the Bayesian information criterion for model selection". *Sociological Methods & Research, 27*(3), 411-427. Doi: 10.1177/0049124199027003005

Ratnasingam, S., & Anderson, B. L. (2015). The role of chromatic variance in modulating colour appearance. *Journal of Vision, 15*(5). Doi: 10.1167/15.5.19

Regier, T., Kay, P., & Cook, R. S. (2005). Focal colours are universal after all. *Proceedings of the National Academy of Sciences of the United States of America, 102*(23), 8386-8391. Doi: 10.1073/pnas.0503281102

Rhodes, G., Ewing, L., Jeffery, L., Avard, E., & Taylor, L. (2014). Reduced adaptability, but no fundamental disruption, of norm-based face-coding mechanisms in cognitively able children and adolescents with autism. *Neuropsychologia, 62*, 262-268. Doi: 10.1016/j.neuropsychologia.2014.07.030

Rhodes, G., Neumann, M. F., Ewing, L., & Palermo, R. (2014). Reduced set averaging of face identity in children and adolescents with autism. *Quarterly Journal of Experimental Psychology*, 1-13. Doi: 10.1080/17470218.2014.981554

Rhodes, G., Pellicano, L., Jeffery, L., & Burr, D. (2007). Abnormal adaptive face-coding mechanisms in autism. *Perception, 36*, 152-152.

Roberson, D., & Davidoff, J. (2000). The categorical perception of colours and facial expressions: The effect of verbal interference. *Memory & Cognition, 28*(6), 977-986. Doi: 10.3758/Bf03209345

Roberson, D., Davies, I., & Davidoff, J. (2000). Colour categories are not universal: Replications and new evidence from a stone-age culture. *Journal of Experimental Psychology-General, 129*(3), 369-398. Doi: 10.1037//0096-3445.129.3.369

Roberson, D., Pak, H., & Hanley, J. R. (2008). Categorical perception of colour in the left and right visual field is verbally mediated: Evidence from Korean. *Cognition, 107*(2), 752-762. Doi: 10.1016/j.cognition.2007.09.001

Robitaille, N., & Harris, I. M. (2011). When more is less: Extraction of summary statistics benefits from larger sets. *Journal of Vision, 11*(12). Doi: 10.1167/11.12.18

Romero, J., Hita, E., & Delbarco, L. J. (1986). A Comparative-Study of Successive and Simultaneous Methods in Colour Discrimination. *Vision Research, 26*(3), 471-476. Doi: 10.1016/0042-6989(86)90189-6

Ropar, D., & Mitchell, P. (2002). Shape constancy in autism: the role of prior knowledge and perspective cues. *Journal of Child Psychology and Psychiatry, 43*(5), 647-653.

Rutherford, M. D., Troubridge, E. K., & Walsh, J. (2012). Visual Afterimages of Emotional Faces in High Functioning Autism. *Journal of Autism and Developmental Disorders, 42*(2), 221-229. Doi: 10.1007/s10803-011-1233-x

Sadowski, J. (2006). Big Spanish Castle. Retrieved 6th June 2014, from http://www.johnsadowski.com/big_spanish_castle.php

Shevell, S. K., & Kingdom, F. A. A. (2008). Colour in complex scenes. *Annual Review of Psychology, 59*, 143-166. Doi: 10.1146/annurev.psych.59.103006.093619

Simmons, D. R., Robertson, A. E., McKay, L. S., Toal, E., McAleer, P., & Pollick, F. E. (2009). Vision in autism spectrum disorders. *Vision Research, 49*(22), 2705-2739. Doi: 10.1016/j.visres.2009.08.005

Simons, D. J., & Myczek, K. (2008). Average size perception and the allure of a new mechanism. *Perception & Psychophysics, 70*(7), 1335-1336. Doi: 10.3758/Pp.70.7.1335

Sinha, P., Kjelgaard, M. M., Gandhi, T. K., Tsourides, K., Cardinaux, A. L., Pantazis, D., . . . Held, R. M. (2014). Autism as a disorder of prediction. *Proceedings of the*

National Academy of Sciences of the United States of America, 111(42), 15220-15225. Doi: 10.1073/pnas.1416797111

Siok, W. T., Kay, P., Wang, W. S. Y., Chan, A. H. D., Chen, L., Luke, K. K., & Tan, L. H. (2009). Language regions of brain are operative in colour perception. *Proceedings of the National Academy of Sciences of the United States of America, 106*(20), 8140-8145. Doi: 10.1073/pnas.0903627106

Smithson, H. E. (2005). Sensory, computational and cognitive components of human colour constancy. *Philosophical Transactions of the Royal Society B: Biologicial Sciences, 360*(1458), 1329-1346. Doi: 10.1098/rstb.2005.1633

Sunaga, S., & Yamashita, Y. (2007). Global colour impressions of multicoloured textured patterns with equal unique hue elements. *Colour Research and Application, 32*(4), 267-277. Doi: 10.1002/Col.20330

Sweeny, T. D., Haroz, S., & Whitney, D. (2012). Reference repulsion in the categorical perception of biological motion. *Vision Research, 64*, 26-34. Doi: 10.1016/j.visres.2012.05.008

Sweeny, T. D., & Whitney, D. (2014). Perceiving Crowd Attention: Ensemble Perception of a Crowd's Gaze. *Psychological Science, 25*(10), 1903-1913. Doi: 10.1177/0956797614544510

Sweeny, T. D., Wurnitsch, N., Gopnik, A., & Whitney, D. (2014). Ensemble perception of size in 4-5-year-old children. *Developmental Science, 18*(4), 556-568. Doi: 10.1111/desc.12239

te Pas, S. F., & Koenderink, J. J. (2004). Visual discrimination of spectral distributions. *Perception, 33*(12), 1483-1497.

Thierry, G., Athanasopoulos, P., Wiggett, A., Dering, B., & Kuipers, J. R. (2009). Unconscious effects of language-specific terminology on preattentive colour perception. *Proceedings of the National Academy of Sciences of the United States of America, 106*(11), 4567-4570. Doi: 10.1073/pnas.0811155106

Treisman, A. (2006). How the deployment of attention determines what we see. *Visual Cognition, 14*(4-8), 411-443. Doi: 10.1080/13506280500195250

Tsao, D. Y., & Freiwald, W. A. (2006). What's so special about the average face? *Trends in Cognitive Sciences, 10*(9), 391-393. Doi: 10.1016/j.tics.2006.07.009

Turi, M., Burr, D. C., Igliozzi, R., Aagten-Murphy, D., Muratori, F., & Pellicano, E. (2015). Children with autism spectrum disorder show reduced adaptation to

number. *Proceedings of the National Academy of Sciences of the United States of America, 112*(25), 7868-7872. Doi: 10.1073/pnas.1504099112

Uchikawa, K., & Ikeda, M. (1981). Temporal Deterioration of Wavelength Discrimination with Successive Comparison Method. *Vision Research, 21*(4), 591-595. Doi: 10.1016/0042-6989(81)90106-1

Utochkin, I. S., & Tiurina, N. A. (2014). Parallel averaging of size is possible but range-limited: A reply to Marchant, Simons, and De Fockert. *Acta Psychologica, 146*, 7-18. Doi: 10.1016/j.actpsy.2013.11.012

van Boxtel, J. J. A., & Lu, H. J. (2013). A predictive coding perspective on autism spectrum disorders. *Frontiers in Psychology, 4*. Doi: 10.3389/Fpsyg.2013.00019

Van de Cruys, S., de-Wit, L., Evers, K., Boets, B., & Wagemans, J. (2013). Weak priors versus overfitting of predictions in autism: Reply to Pellicano and Burr (TICS, 2012). *I-Perception, 4*(2), 95-97. Doi: 10.1068/i0580ic

van Lier, R., Vergeer, M., & Anstis, S. (2009). Filling-in afterimage colours between the lines. *Current Biology, 19*(8), R323-R324. Doi: 10.1016/j.cub.2009.03.010

Vergeer, M., Anstis, S., & van Lier, R. (2015). Flexible colour perception depending on the shape and positioning of achromatic contours. *Frontiers in Psychology, 6*, 620. Doi: 10.3389/fpsyg.2015.00620

Vul, E., & MacLeod, D. I. (2006). Contingent aftereffects distinguish conscious and preconscious colour processing. *Nature Neuroscience, 9*(7), 873-874. Doi: 10.1038/nn1723

Watamaniuk, S. N. J., & Duchon, A. (1992). The Human Visual-System Averages Speed Information. *Vision Research, 32*(5), 931-941. Doi: 10.1016/0042-6989(92)90036-I

Watamaniuk, S. N. J., & McKee, S. P. (1998). Simultaneous encoding of direction at a local and global scale. *Perception & Psychophysics, 60*(2), 191-200.

Watamaniuk, S. N. J., Sekuler, R., & Williams, D. W. (1989). Direction Perception in Complex Dynamic Displays – the Integration of Direction Information. *Vision Research, 29*(1), 47-59. Doi: 10.1016/0042-6989(89)90173-9

Webster, J., Kay, P., & Webster, M. A. (2014). Perceiving the average hue of colour arrays. *Journal of the Optical Society of America: A, 31*(4), A283-A292. Doi: 10.1364/Josaa.31.00a283

Webster, M. A. (2011). Adaptation and visual coding. *Journal of Vision, 11*(5). Doi: 10.1167/11.5.3

Webster, M. A., & Leonard, D. (2008). Adaptation and perceptual norms in colour vision. *Journal of the Optical Society of America: A, 25*(11), 2817-2825.

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. Doi: 10.1098/rstb.2010.0360

Webster, M. A., Mizokami, Y., & Webster, S. M. (2007). Seasonal variations in the colour statistics of natural images. *Network-Computation in Neural Systems, 18*(3), 213-233. Doi: 10.1080/09548980701654405

Webster, M. A., & Mollon, J. D. (1995). Colour Constancy Influenced by Contrast Adaptation. *Nature, 373*(6516), 694-698. Doi: 10.1038/373694a0

Webster, M. A., & Mollon, J. D. (1997). Adaptation and the colour statistics of natural images. *Vision Research, 37*(23), 3283-3298. Doi: 10.1016/S0042-6989(97)00125-9

Wechsler, D., & Psychological Corporation. (2011). *WASI –II: Wechsler Abbreviated Scale of Intelligence* (2nd ed.). San Antonio, Texas: Psychological Corporation.

Wheatstone, Charles. (1838). Contributions to the Physiology of Vision. Part the First. On Some Remarkable, and Hitherto Unobserved, Phenomena of Binocular Vision. *Philosophical Transactions of the Royal Society of London B: Biological Sciences, 128*, 371-394. Doi: 10.1098/rstl.1838.0019

Whiting, B. F., & Oriet, C. (2011). Rapid averaging? Not so fast! *Psychonomic Bulletin & Review, 18*(3), 484-489. Doi: 10.3758/s13423-011-0071-3

Winawer, J., Witthoft, N., Frank, M. C., Wu, L., Wade, A. R., & Boroditsky, L. (2007). Russian blues reveal effects of language on colour discrimination. *Proceedings of the National Academy of Sciences of the United States of America, 104*(19), 7780-7785. Doi: 10.1073/pnas.0701644104

Winkler, A. D., Spillmann, L., Werner, J. S., & Webster, M. A. (2015). Asymmetries in blue-yellow colour perception and in the colour of 'the dress'. *Current Biology, 25*(13), R547-548. Doi: 10.1016/j.cub.2015.05.004

Witzel, C., & Gegenfurtner, K. R. (2013). Categorical sensitivity to colour differences. *Journal of Vision, 13*(7). Doi: 10.1167/13.7.1

World Health Organization. (1992). *The ICD-10 classification of mental and behavioural disorders: Clinical descriptions and diagnostic guidelines*. Geneva: World Health Organization.

Wright, O. (2011). Effects of stimulus range on colour categorization. In C. P. Biggam, C. A. Hough, C. J. Kay & D. R. Simmons (Eds.), *New Directions in Colour Studies* (pp. 265-276). Amsterdam: John Benjamin.

Younger, B. A. (1985). The segregation of items into categories by ten-month-old infants. *Child Development, 56*(6), 1574-1583.

Younger, B. A., & Cohen, L. B. (1986). Developmental change in infants' perception of correlations among attributes. *Child Development, 57*(3), 803-815.

Zaidi, Q., Ennis, R., Cao, D. C., & Lee, B. (2012). Neural Locus of Colour Afterimages. *Current Biology, 22*(3), 220-224. Doi: 10.1016/j.cub.2011.12.021

Zaidi, Q., Spehar, B., & DeBonet, J. (1998). Adaptation to textured chromatic fields. *Journal of the Optical Society of America: A, 15*(1), 23-32.