THE UNIVERSITY of EDINBURGH

Edinburgh Research Explorer

Acquisition of colour categories through learning

Citation for published version:
Martinovic, J 2024, 'Acquisition of colour categories through learning: Differences between hue and lightness', *Cognition*, vol. 242, 105657. https://doi.org/10.1016/j.cognition.2023.105657

Digital Object Identifier (DOI):
10.1016/j.cognition.2023.105657

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Publisher's PDF, also known as Version of record

Published In:
Cognition

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.

OPEN ACCESS

Download date: 13. Dec. 2023

Cognition 242 (2024) 105657

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/cognit

Acquisition of colour categories through learning: Differences between hue and lightness

Jasna Martinovic [*]

Department of Psychology, School of Philosophy, Psychology and Linguistics, University of Edinburgh, 7 George Square, EH8 9JZ Edinburgh, Scotland, UK

ARTICLE INFO

Keywords:
Colour perception
Colour categorisation
Perceptual learning
Relational learning
Acquisition of colour categories
Labelling
Colour terms
Attention

ABSTRACT

Colour categories are acquired through learning, but the nature of this process is not fully understood. Some category distinctions are defined by hue (e.g. red/purple) but other by lightness (red/pink). The aim of this study was to investigate if the acquisition of key information for making accurate cross-boundary discriminations poses different challenges for hue-defined as opposed to lightness-defined boundaries. To answer this question, hue- and lightness-learners were trained on a novel category boundary within the GREEN region of colour space. After training, hue- and lightness-learners as well as untrained controls performed delayed same-different discrimination for lightness and hue pairs. In addition to discrimination data, errors during learning and category-labelling strategies were examined. Errors during learning distributed non-uniformly and in accordance with the Bezold-Brücke effect, which accounts for darker colours at the green-blue boundary appearing greener and lighter colours appearing bluer. Only hue-learners showed discrimination improvements due to category boundary acquisition. Thus, acquisition is more efficient for hue-category compared to lightness-category boundaries. Almost all learners reported using category-labelling strategies, with hue-learners almost exclusively using 'green'/'blue' and lightness learners using a wider range of labels, most often 'light'/'dark'. Thus, labels play an important role in colour category learning and such labelling does not conform to everyday naming: here, the label 'blue' is used for exemplars that would normally be named 'green'. In conclusion, labelling serves the purpose of highlighting key information that differentiates exemplars across the category boundary, and basic colour terms may be particularly effective in facilitating such attentional guidance.

1. Introduction

What is the nature and origin of colour categorisation? This is a key question that has led to a large, interdisciplinary body of research, intersecting perception, cognition and linguistics (Bornstein, Kessen, & Weiskopf, 1976; Bornstein & Korda, 1984; Roberson, 2005; Roberson, Davies, & Davidoff, 2000; for recent reviews see Lindsey & Brown, 2021; Witzel, 2019; Witzel & Gegenfurtner, 2018). Originating in experimental phonetics, categorical perception is the term initially used to describe the phenomenon that two stimuli belonging to different categories are easy to discriminate while two equally-spaced stimuli belonging to the same category are difficult to discriminate (Liberman, Harris, Hoffman, & Griffith, 1957; for a review see Schouten, Gerrits, & van Hessen, 2003). Similarly to continuous changes in speech signals which map on to categorically different phonemes, we perceive certain parts of the colour continuum as instances of distinct categories.

To probe how colour categorisation influences perception, researchers commonly focus on reaction times (RTs), discrimination accuracy or event-related potential (ERPs) latencies or amplitudes for stimuli that cross linguistic category boundaries. Faster RTs, higher accuracies and faster and/or larger ERP components are taken as a signature of categorical perception (for recent reviews, see Siuda-Krzywicka, Boros, Bartolomeo, & Witzel, 2019; Witzel, 2019; Witzel & Gegenfurtner, 2018). Traditionally, the phenomenon of categorical perception was thus equated with the behavioural or neural signatures derived from methods used to study it (Schouten et al., 2003). But rather than being interested in quantifying the degree of categorical perception that can be observed under various conditions and inferring whether its mechanisms are pre-attentive or malleable to experience (e.g. Athanasopoulos, Dering, Wiggett, Kuipers, & Thierry, 2010; Thierry, Athanasopoulos, Wiggett, Dering, & Kuipers, 2009), recent models of categorisation in speech perception are more concerned with defining the mechanisms that partition the auditory space into categories (McMurray & Jongman, 2011). Similarly, two contemporary models of

[*] Corresponding author.
E-mail address: j.martinovic@ed.ac.uk.

https://doi.org/10.1016/j.cognition.2023.105657
Received 29 May 2023; Received in revised form 20 September 2023; Accepted 30 October 2023
Available online 17 November 2023
0010-0277/© 2023 The Author(s). Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

colour categorisation attempt to explain the mechanisms through which cross-category advantages occur. Label-feedback hypothesis proposes that on-line co-activation of linguistic category labels produces the warping of perceptual representations, wherein within-category exemplars become drawn together and cross-category exemplars pushed further apart (Lupyan, 2012). On the other hand, categorical facilitation hypothesis proposes that top-down effects of linguistic categories are more likely to stem from attention being directed towards salient information at the category boundary, rather than an actual 'warping' of the perceptual space (Witzel, 2019). While Lupyan (2017) acknowledges the possibility of attention as a mediator of cognitive influences on colour perception, the label-feedback hypothesis posits that the main mechanism is language and that the link between language and attention stems from the fact that linguistic labels act as highly effective cues in guiding attention (Forder & Lupyan, 2019).

Experimental evidence in favour of categorical colour perception consists overwhelmingly of findings of behavioural or ERP advantages at lexically established category boundaries (for a review see Lindsey & Brown, 2021). However, interpreting these results as a product of cognition influencing perception is complicated by the fact that a universally valid metric for colour difference does not exist (Luo, 2016). Also, it is impossible to eliminate the possibility that categorical boundaries coincide with discrimination minima originating in low-level and/or higher-order perceptual mechanisms. For example, just-noticeable differences (JNDs) are non-linear along the transition from blue to green and reaction times (RTs) are considerably slower for blues as opposed to greens (Brown, Lindsey, & Guckes, 2011; Witzel & Gegenfurtner, 2015). Also, discrimination is enhanced around the region of colour space coinciding with unique blue and yellow (Danilova & Mollon, 2012).

Evaluating the nature of categorical colour perception along familiar, linguistically-represented boundaries thus seems less than optimal. A more controlled way would be to investigate categorical effects in a between-subjects design, with groups of participants that differ in whether they possess a certain lexical distinction. Indeed, language-specific basic colour categories *sinij* 'dark blue' and *goluboj* 'light blue' have a somewhat sharper categorical transition in Russian as opposed to English speakers, although this does not translate into a cross-category RT advantage (Martinovic, Paramei, & MacInnes, 2020). To circumvent the problems inherent with cross-colour or cross-cultural comparisons, it is also possible to teach a group of participants a novel category boundary in an area of colour space that would otherwise constitute a single category (e.g., 'green'). Within experimental psychology, such category-learning studies would then compare discrimination in participants who have successfully learned the novel boundary with those who failed to learn it (e.g., Perez-Gay Juarez et al., 2017) or with discrimination in untrained controls (e.g., Clifford et al., 2012; Grandison et al., 2016; Özgen & Davies, 2002), for whom that same area of stimulus space still represents a single category.

Using a control-group category-learning approach, Özgen and Davies (2002) provided evidence that categorical perception of colour may be an outcome of perceptual learning for both hue and lightness-defined categories. In their fourth and final experiment, category-learners were trained on either a novel hue or a novel lightness boundary in the GREEN area of colour space, with the hue boundary lying close to the green prototype. A same/different delayed discrimination task was then used to evaluate if learners would exhibit the classical boundary advantage after learning. While hue learners had improved cross-category relative to within-category sensitivity on both hue and lightness dimensions, lightness learners improved only for lightness. The emergence of categorical perception effects through perceptual learning was taken to imply that similar mechanisms might be involved when children acquire colour terms. As perceptual learning should only occur for the attended dimension, the fact that hue categorisers learned both hue and lightness could be due to an asymmetrical relationship of integrality (Garner & Felfoldy, 1970), where it may not be possible to extract hue information independently of lightness (Burns & Shepp, 1988). Meanwhile, the possibility of acquiring both hue and lightness-based categories through such learning was taken as evidence in favour of genuine basicness of multiple lightness-based colour terms for 'blue' that exist in several languages (for an overview, see Paramei, 2005, 2007).

Recently, it was reported that lightness-based "Russian blues" *sinij* and *goluboj* are likely to be less structured than hue-based categories (Martinovic et al., 2020). On the other hand, in Özgen and Davies (2002) study lightness categories seemed to be more universally learned, leading to enhanced discrimination in both hue and lightness learners. It may appear that those findings are inconsistent with each other: if lightness-based colour categories are less firmly demarcated, they should be learnt less efficiently than hue-based categories. To reconcile these findings and gain more insight into potential mechanisms that drive the differences in the acquisition of lightness-based and hue-based categories, we conducted two experiments that used a similar novel category learning approach to the original work by Özgen and Davies (2002). We expected to replicate findings of enhanced distinctiveness across the newly learned boundary, but were interested in underlying reasons as to why learning might differ between hue and lightness. To gain further insight into how perceptual attributes might affect category learning, we manipulated the location of the hue boundary. It either coincided with the green prototype, demarcating a narrow range of colours between 'green-blue' and 'green-yellow' (as in Özgen & Davies, 2002) or was shifted towards 'green-yellow', demarcating a more extended range of colours between 'green-yellow' and prototype-green. We collected data on errors made during such training, expecting that errors might turn out to be non-uniform for hue learners, consistent with the suggestion made in the original paper that hue cannot be processed independently of its lightness value (Burns & Shepp, 1988). We also expected more errors when learning from the extended set in which the boundary was shifted away from the green prototype and towards 'green-yellow'. The final set of predictions concerned the ease with which the lightness boundary would be acquired. If hues of different lightness levels pose different categorisation challenges, this might lead to an obligatory learning of lightness information and could potentially make lightness-based categories easier to acquire. Conversely, if lightness-based categories are less well demarcated, with less salient reference information at the lightness boundary, they should prove harder to acquire.

To be able to monitor error rates driven by different colour samples, we opted for a somewhat different learning protocol. While Özgen and Davies (2002) trained their participants on colour samples randomly chosen from each novel category's area, our participants learned from the actual 16 stimulus colours. This type of learning is likely to have more ecological validity, as random-exemplar learning does not resemble how children acquire categorical knowledge in everyday life. Children would learn from sets that include both highly familiar objects and novel objects, rather than experiencing random sets of objects. More importantly, rather than relying on perceptual mechanisms alone, children would also use language, making systematic hypotheses about colour word meanings and inferring what they denote in relation to other colour terms to determine where one category ends and another begins (for an overview, see Wagner, Tillman, & Barner, 2016). Such label-guided perceptual learning at the boundary is likely to be the process through which early categories become narrower, and start to align with the colour term boundaries imposed by language (Istomina, 1960a, 1960b; Wagner, Dobkins, & Barner, 2013; Wagner, Jergens, & Barner, 2018). Because of this important role of language in colour category acquisition, we asked our participants to report any category-labelling strategies that they might have used during the experiment. We expected that 'light'/'dark' or 'yellow-green'/'blue-green' might be the dominant labels, depending on the probed boundary, as participants rely on terms that are useful in differentiating salient differences that demarcate the two novel categories (Zettersten & Lupyan, 2020).

Foreshadowing our results, novel category acquisition produced discrimination benefits only at hue boundaries. Hue learners labelled the two categories as ‘green’ and ‘blue’ irrespective of the boundary location, while lightness learners used a mixture of strategies and hue, lightness and saturation labels. Reliance on ‘blue’/’green’ labels and patterns of learning errors aligned with the Bezold-Brücke effect (at the green/blue boundary, lighter colours appear bluer and darker appear greener; Lillo, Aguado, Moreira, & Davies, 2004) both support the argument that labels are likely to act as cues that help to differentiate category-diagnostic information.

2. Methods

2.1. Participants

43 participants (12 male, 31 female; age ranging from 17 to 53 years, mean = 22) completed Experiment 1. Controls and hue learner groups had 15 participants each, while there were 13 lightness learners as 2 participants from that group failed to complete the study according to protocol. A separate set of 43 participants completed Experiment 2 (18 male, 23 female and two of undisclosed gender, age ranging from 17 to 55 years, mean = 23). Here, lightness learners and controls had 15 participants per group, but hue learners ended up with 13 participants due to failure to complete the training protocol in two participants.

All our participants had normal colour vision, as evaluated by the Cambridge Colour Test (Regan, Reffin, & Mollon, 1994) and normal or corrected-to-normal visual acuity. Participants were recruited from the University of Aberdeen student population and completed the study for class credit or reimbursement. They were naïve to the purpose of the study. Participants gave written informed consent prior to taking part. The study was approved by the Psychology Ethics Committee of the University of Aberdeen and was in line with the Declaration of Helsinki (1964).

With 10 participants per group, Özgen and Davies’ (2002) experiment was sensitive to a within-between repeated measures ANOVA interaction (e.g., group by categorisation type) of effect size f = 0.290 at 80% power, assuming a correlation of 0.618 between the repeated measures (estimated from our own data). In that study, the selective improvement observed because of hue category acquisition equalled f = 0.398 when hue categorisers were compared to lightness categorisers and f = 0.236 when hue categorisers were compared to controls. Meanwhile, the general improvement in lightness categorisation brought about by category acquisition had an effect size of f = 0.251 (f = 0.244 for lightness learners vs. controls; for more detail on effect size calculations, see Supplementary Materials). Thus, it appears that the Özgen and Davies’ study had borderline sensitivity towards its effects of interest. To extend our sensitivity towards medium effect sizes (f = 0.25, according to Cohen, 2013) we planned for a sample size of 15 participants per learner group. This would enable us to detect a between-within ANOVA interaction of f = 0.232 at 80% power. Combining our two experiments would provide a sample of ~30 per learner group, providing sensitivity to an interaction of f = 0.161 at 80% power (calculated with G*Power 3.1.9.7; Faul, Erdfelder, Lang, & Buchner, 2007). This would allow us to capture the improvements brought about by category acquisition even if they were overestimated by Özgen and Davies (2002) and were small to medium-sized. Perception research utilises precise measurement techniques and focuses on effects that are assumed to be large and stable across participants (for a discussion, see Baker, Lygo, Meese, & Georgeson, 2018). If no effect is observed, this would indicate that the influence of category learning on discrimination is at best small-to-medium (f < 0.161 for combined data) and potentially less interesting, as such small effects would be an unlikely outcome of more general and lawful relationships between cognition and perception. Larger follow up studies would be needed to affirm the presence or absence of such smaller effects (Lakens, 2022; Schäfer & Schwarz, 2019).

2.2. Apparatus

Stimuli were presented on a 21” Viewsonic P227f CRT display, driven by a CRS (Cambridge Research Systems Ltd., Rochester, UK) ViSaGe system, giving 14-bit resolution per RGB channel. Monitor output was calibrated prior to testing using a ColorCal2 (CRS, UK). The monitor was switched on at least 30 min before the start of the experiment. Observers viewed the display from a distance of 96 cm. Participants gave responses via a Cedrus-530 button-box (Cedrus, San Pedro, CA). Measurements of monitor phosphors by a SpectroCAL (CRS, UK) were used in combination with CIE 1931 colour matching functions to ensure accurate colour representation. CRS Toolbox and CRS Colour Toolbox (Westland, Ripamonti, & Cheung, 2012) for Matlab (The Mathworks Inc., Natick, MA) were used to run the experiment.

2.3. Stimuli

The stimulus arrays consisted of 16 colour samples and were situated within the GREEN region of CIE LAB space (see Fig. 1a and Table 1). This is a relatively large area of the perceptual colour space, and thus allows multiple colour exemplars to be drawn and used as stimulus materials. Colour samples were superimposed over a background metameric with D65 and set to 50 cd/m^2 and were darker than the background, as in Özgen and Davies (2002). Stimuli stretched over an area of GREEN nested between yellow-green and green-blue (Bosten & Lawrance-Owen, 2014). The boundaries introduced by training divided the 16 samples in two equally sized sections, either with regard to hue (between the middle rows: Fig. 1, hue 2 and 3) or lightness (columns: Fig. 1, lightness 2 and 3).

Each colour sample was presented as a square that subtended 2.6° visual angle.

We selected our stimulus colours for Experiment 1 from CIE LAB

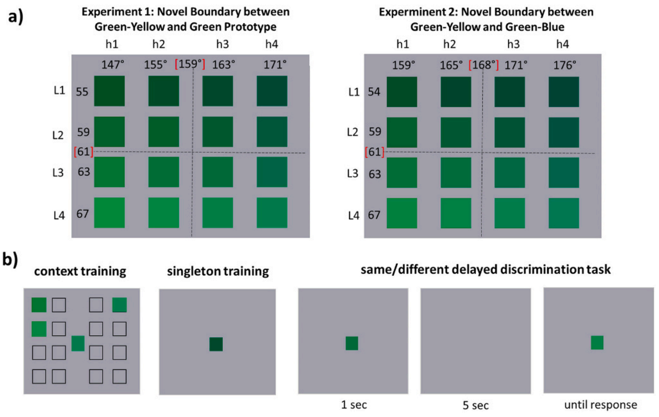
a)
Experiment 1: Novel Boundary between Green-Yellow and Green Prototype
h1
h2
h3
h4
147°
155°
[159°]
163°
171°
L1
55
L2
59
[61]
L3
63
L4
67
Experminent 2: Novel Boundary between Green-Yellow and Green-Blue
h1
h2
h3
h4
159°
165°
[168°]
171°
176°
L1
54
L2
59
[61]
L3
63
L4
67
b)
context training
singleton training
same/different delayed discrimination task
1 sec
5 sec
until response

Fig. 1. Stimuli and procedure. a) Stimulus samples, with CIE LCh hue and lightness values for each experiment listed to the left/above. For Experiment 2, the values are approximate (see Table 1) as the Munsell colour space does not equate to a 3-dimensional colour appearance model in which distances between hues hold across value levels and vice versa. h1-h4 labels the four hues, while L1-L4 labels the four lightness values of the stimuli. b) On the left, training phases 1 and 2 are depicted, with the first phase including context during learning and second phase involving viewing a singleton sample to be categorised. On the right, the timeline of the same/different delayed discrimination task is depicted. The first sample is shown for 1 s, followed by a 5 s delay. After this, the second sample is shown until participant responds on whether it is same or different as the first colour sample. Note: colours in the figure have been adjusted to appear as similar as possible to true stimulus colours (as displayed on a calibrated device) using the export image function in the CRS toolbox. While this is still only an approximation, the image successfully captures some of the shift in yellowness/greenness/blueness across the stimulus space (the reader is referred to the web version of this article for a coloured version of the figure).

Table 1
CIE LAB and LCh coordinates and luminance values for colour samples used in the study. L1 – L4 are the four lightness levels, while h1-h4 are the four hue levels used in the study (see Fig. 1).

		Experiment 1: Novel Boundary between Green-yellow and Green Prototype				Experiment 2: Novel Boundary between Green-yellow and Green-blue			
		h1	h2	h3	h4	h1	h2	h3	h4
	L	55	55	55	55	54.38	54.38	54.38	54.38
L1	a	−36.90	−39.88	−42.08	−43.46	−36.77	−38.34	−39.59	−40.57
	b	23.96	18.60	12.86	6.88	13.56	9.72	6.02	2.49
	C	44	44	44	44	39.19	39.56	40.04	40.64
	H	147	155	163	171	159.76	165.78	171.35	176.49
	Lum			22.93 cd/m^2				22.34 cd/m^2	
	L	59	59	59	59	58.56	58.56	58.56	58.56
L2	a	−36.90	−39.88	−42.08	−43.46	−37.54	−39.10	−40.37	−41.45
	b	23.96	18.60	12.86	6.88	14.20	10.34	6.67	3.17
	C	44	44	44	44	40.13	40.45	40.92	41.57
	H	147	155	163	171	159.28	165.19	170.61	175.63
	Lum			27.03 cd/m^2				26.56 cd/m^2	
	L	63	63	63	63	62.68	62.68	62.68	62.68
L3	a	−36.90	−39.88	−42.08	−43.46	−38.20	−39.77	−41.06	−42.18
	b	23.96	18.60	12.86	6.88	14.79	10.91	7.27	3.80
	C	44	44	44	44	40.96	41.24	41.70	42.35
	H	147	155	163	171	158.83	164.66	169.96	174.84
	Lum			31.59 cd/m^2				31.21 cd/m^2	
	L	67	67	67	67	66.74	66.74	66.74	66.74
L4	a	−36.90	−39.88	−42.08	−43.46	−38.78	−40.37	−41.68	−42.80
	b	23.96	18.60	12.86	6.88	15.32	11.44	7.81	4.39
	C	44	44	44	44	41.69	41.96	42.41	43.02
	H	147	155	163	171	158.44	164.18	169.38	174.14
	Lum			36.63 cd/m^2				36.29 cd/m^2	

space rather than the Munsell system as the former is better at controlling saturation between hues (for more detail, see Fairchild, 2013; Schiller, Valsecchi, & Gegenfurtner, 2018). Samples covered a highly similar area of colour space as in Özgen and Davies (2002). To maintain the same distance in colour space between stimulus samples (ΔE = 6 for hue and ΔE = 4 for lightness), hues were shifted towards green-yellow, covering the area of 147°-171° with boundary at 159° (between green-yellow and the green prototype; Fig. 1a). In Experiment 2, we used the same hues as Özgen and Davies (2002: Munsell 4.48 G – 0.9 BG, at 7.9 Chroma, ~159°-174° in CIE LCh, with a boundary at 7.5G or ~167-8°, co-incidental with the green prototype). R package *colorspace* was used to transform colours from CIE LUV coordinates provided in their paper (Zeileis et al., 2020). The most significant difference in category-belongingness between the two experiments is for hues of ~163–165°. They are proximal to the boundary in both experiments, but shifted away from the green prototype category and into the 'yellow-green' category in Experiment 1 while in Experiment 2 they are placed in the 'green-blue' category and away from the 'green-yellow' category (see Table 1 and Fig. 1a). Another difference is in chroma and saturation (i.e. ratio of chroma and lightness, C/L), with samples in Experiment 2 being slightly less colourful and saturated.

2.4. Procedure

Both experiments were designed to follow the same procedure. Participants were allocated to one of three groups: hue learners, lightness learners or controls. Category learner participants were asked to come to the lab on two successive days. Hue learners performed supervised learning on the hue boundary, while lightness learners were trained on the lightness boundary. On day 1, they would complete the two training phases. On day 2, they completed a briefer, "top-up" training session prior to completing a *same/different* delayed judgment task.

In the first, context-training phase, participants learned the categorisation rule while being able to see all correctly assigned colours. The context was provided by a grid of 16 empty slots on the screen - 8 on each side - to be filled in by correctly assigned stimulus colours. Colours appeared in the centre of the screen in random order, one at a time (see Fig. 1b). The first colour could be assigned to either the left or the right of the screen, using the left and right buttons on the button box. No specific guidance as to how the colours should be assigned was given, other than that the colours must be sorted into two categories. Since the first colour can be allocated to either side, this determined which side of the screen each category would end up being on. If a correct response was made, the colour would appear in the next slot on the side to which the participant assigned it. If an error was made, participants were given auditory feedback, and the colour sample disappeared rather than being allocated to a slot, to be replaced by a new sample. The incorrectly allocated samples were appended to the end of the stimulus queue, so participants would have to categorise them again, once they had allocated all the previous samples. In this way, participants continued to allocate the colours until all slots were filled. Context-training phase had to be performed for at least 20 min and the completion criterion was that at least 10 trials were done with at least 3 trials error free. This took around 20 mins on average.

In the second, singleton-training phase, participants performed the same task of assigning colours to one of two categories, but this time they were not provided with the context of colours that were already assigned. In other words, there was no grid of colour slots on the screen. Randomly selected colour singletons would appear in the centre of the screen and participants had to allocate them to one of two categories. After the colour has been allocated, it disappeared. If a colour was categorised incorrectly, there would be a beep and the word "incorrect" would appear on the screen, prior to the presentation of the next sample. As in the context-training phase, the first sample could be allocated to either the 'left' or 'right' category, and subsequent colours were to be allocated accordingly. Participants had to complete at least 50 trials, with 25 in a row being correct. This took around 5 min on average.

On day 2, participants completed both phases of training again, but with less stringent criteria. Context training had to last for at least 10 min, and the criterion for completion was to have done at least 5 trials and at least 2 of them error free, which took about 10 mins. Singleton training remained the same and took roughly the same time of 5 min.

The top-up training was followed by the discrimination task. A delayed *same/different* judgment task was used, with stimuli blocked by hue or lightness. A single coloured square appeared in the centre of the screen for 1 s and then disappeared. After a 5-s interval, a second coloured square appeared and stayed on the screen until response. The square would either be physically identical or would be adjacent to the previous sample in the stimulus grid, either in lightness or hue (depending on the block; see Fig. 1). Once the participant had responded whether the two colours were same or different by pressing the left or right button, a beep would sound signalling the beginning of the next trial. No feedback was given to the accuracy of responses.

The discrimination task contained 96 trials. There were 16 same-category pairs, shown 3 times each (48 trials) and 12 different-category pairs (8 within and 4 across category boundary), shown 4 times each (48 trials) in random order. This was preceded by 20 practice trials. Participants were given a short, self-paced break in the middle of the block. Twenty-four different-category pairs and 24 same-category pairs (16 of each colour, and another 8 randomly selected) appeared in each sub-section of the discrimination block. Thus, there were two blocks of 96 trials for hue and lightness discriminations. The order of hue and lightness discriminations was counterbalanced across participants. At the end of the experiment, participants were asked to report any labels that they used for the two categories. Meanwhile, the control group only took part in one session, which consisted of a brief presentation of the stimulus colours, arranged in a 4 × 4 grid, followed by the two discrimination tasks. For category-learners, the second session lasted ~50 min, while for controls this took ca. 35 min.

There were a few departures from the procedure used by Özgen and Davies (2002). First, as mentioned in the Introduction, their training included randomly generated colours from the GREEN area on each side of the category boundary. We opted, instead, to train with stimulus colours themselves. The rationale was as follows: while Özgen and Davies deployed training to manipulate category acquisition but focused their analyses on discrimination data only, we wanted to evaluate error patterns specific to each employed colour sample during learning. Our aims were also very different: while Özgen and Davies (2002) aimed to assess if colour category acquisition could be driven purely by perceptual learning, or at the very least by category labelling, we also aimed to investigate potential asymmetries in acquiring hue and lightness dimensions from errors made during training and reports of labelling strategies. Learning from randomly selected samples is indeed a valid approach in eliminating the likelihood that participants will come up with labels for each individual colour sample, even if adoption of individual labels for a set of 16 highly similar exemplars covering a narrow colour space area is not highly likely. ΔE is the colour difference metric in CIE LAB space and 2.3ΔE is an estimate of a just noticeable difference (JND) for adjacent colours (Fairchild, 2013). With Özgen and Davies' (2002) stimuli spanning a section of 'green' that is ~18ΔE across, there could be at most three just distinguishable within-category hues on each side of the boundary. Thus, it is highly unlikely that individual labelling is a useful strategy for participants, even once they start to repeatedly encounter the same samples (Özgen & Davies, 2002).

As mentioned above, the use of the stimulus grid samples for training is likely to have accelerated learning, since the training colours covered the hue/lightness space evenly, providing a more stable training context for relational learning (for a discussion, see Doumas, Hummel, & Sandhofer, 2008; Sandhofer & Doumas, 2008). In line with this, in Özgen and Davies' study, the context phase lasted about 30 min, while for almost all our participants it only took the required 20 min. The singleton phase, in comparison, took about 5 min in both studies. Second, our participants performed the two phases of training on day 1 only once, while Özgen and Davies' participants performed the two phases twice. Based on piloting, we deemed that one repetition was sufficient, due to low number of errors in the latter stages of learning (see Supplementary Materials for more detail). As a last change, we adjusted the number of discrimination pairs. In Özgen and Davies (2002) each of the 12 different-category pairs was repeated four times, while each of the 16 same-exemplar pairs was shown twice. This resulted in 80 (48 different and 32 same) trials per dimension. To avoid the potential response bias introduced by having an unequal number of *same* and *different* trials (Macmillan & Creelman, 2004), we added 8 further same pairs to each 48-trial sub-section of the block.

2.5. Data analysis

The analyses were performed in R version 3.6.1 (R_Core_Team, 2016), using packages *ggplot2* (Wickham, 2016), *lme4* (Bates, Maechler, Bolker, & Walker, 2015), *emmeans* (Lenth, Singmann, Love, Buerkner, & Herve, 2019), *effectsize* (Ben-Shachar, Lüdecke, & Makowski, 2020) and *performance* (Lüdecke, Ben-Shachar, Patil, Waggoner, & Makowski, 2021).

To evaluate differences in errors between hue and lightness learners, we compared their error counts using generalised linear mixed-effect models (GLMMs). We first fitted a model with experiment, training group, hue and lightness of colour samples as a fixed effect and participant-level and observation-level intercepts as random effects. Observation-level intercepts were added to the model to account for a frequent problem with highly non-normally distributed count data: overdispersion, when the variance of data is higher than the mean (Harrison, 2014). Fitting of the model relied on a Poisson distribution with a log link, using maximum likelihood with a Laplace approximation. Contrasts were set to simple coding: each level of the factor was compared to the reference level, with the intercept set to grand mean. When fitting GLMMs, we applied the maximal random effect structure that was possible while maintaining goodness of fit (Barr, Levy, Scheepers, & Tily, 2013). We then evaluated whether we could remove the four-way interaction between all the fixed effects from the model without significantly reducing its goodness of fit. In Supplementary Materials we report all estimates of the best-fitting, final model – which in all instances turned out to be the full model, with the four-way interaction. Post-hoc tests on this final model were performed using omnibus paired *t*-tests, corrected for multiple comparisons ($p < .05$) with the 'mvt' method from *emmeans* package (Lenth et al., 2019). This method relies on the multivariate *t* distribution with the same covariance structure as the estimates to determine the *p*-value adjustment.

When it came to discrimination data, we followed the approach of Özgen and Davies (2002) to perform repeated measures ANOVA on the non-parametric measure of sensitivity A' for *same/different* judgments, in order to be able to directly compare our outcomes with theirs. A' is calculated from hit and false alarm rates according to the above chance (h ≥ f) and below chance (h < f) formulae below (formulas *1* and *2*; h – hits; f – false alarms). When performance is high, A′ is consistent with a threshold model assuming (roughly) rectangular distributions, as the ROCs are approximately linear. However, as performance lowers, the shape of the A′ ROC curve increasingly mimics that of a signal detection model assuming logistic evidence distributions (for more detail, see Rhodes, Cowan, Parra, & Logie, 2019).

$$A' = 0.5 + \frac{(h-f)(1+h-f)}{4h\,(1-f)}, h \geq f \tag{1}$$

$$A' = 0.5 - \frac{(f-h)(1+f-h)}{4f\,(1-h)}, h < f \tag{2}$$

To avoid problems of interpreting ANOVA interactions from such a non-uniformly behaving variable as A', we also fit GLMMs to our binomial single-trial hit/miss data and evaluate and report them in the same way as for error counts. Compared to extraction of A's from single participants, GLMMs have the additional advantage that they do not discard information about trial repetitions and subject-specific variability (Moscatelli, Mezzetti, & Lacquaniti, 2012). Furthermore, subsuming same-exemplar judgments under the measure of A' is also problematic from a theoretical point of view, since the *same/different* dichotomy represents two opposite poles in relational processing, which is thought to be the key mechanism in acquiring concepts such as colour categories (for a review, see Hespos, Gentner, Anderson, & Shivaram, 2021).

2.6. *Open Science*

Pre-registration, data, R analysis scripts, detailed stimulus coordinates and Matlab scripts for running the experiments are available in an online repository (https://osf.io/vn5t2/).

3. Results

This section will first present labelling patterns, and then show error counts during learning and discrimination performance. Even though participants reported their category labels at the end of the experiment, it makes sense to present these results first as they offer important context for understanding and interpreting the remaining data.

3.1. *Category labels across both experiments*

Participants were asked if they used any labels to refer to the colours while they performed the tasks. In Experiment 1, hue learners reported a relatively narrow range of labels: 14 participants used green/blue and 1 used green/turquoise. Lightness learners reported a much wider variety of category labels. Non-hue related labels were reported by 9 participants: 6 dark/light, 1 low/high saturation, 1 matte/bright, 1 dark/bright. Hue-related labels were reported by 3 participants: 1 green/aqua, 1 green/blue and 1 yellow-green/turquoise. One participant reported using no labels.

In Experiment 2, all hue learners used green/blue labels. Lightness learners, again, reported a mixture of non-hue and hue-related labels. Dark/light was reported by 5, while 4 reported using both lightness and hue-related labels: 3 used both light/dark and green/blue, while 1 used light/dark, hard/soft and green/blue. Finally, 3 participants reported using no labels, with one of them making a further comment that they relied on judging the contrast of the stimulus edge against the background.

As control participants did not perform categorisation training, it initially appeared pointless to ask them for labels. However, they still experienced all the colour samples, having performed the same discrimination tasks as the learners. Therefore, we asked the controls from Experiment 2 to report any labels for the colour samples viewed during this task. No labels were reported by 4, while the remaining 11 reported multiple non-basic colour terms. These were either green hyponyms, modified, compounded, or 'fancy' terms (2 dark green, 4 forest green, 6 grass green, 3 jade, 3 light green, 4 yellow-green) or non-basic terms straddling the blue-green category boundary (2 aqua, 4 sea green, 3 teal, 3 turquoise). While turquoise is generally considered non-basic, it has been argued to be an emerging basic colour term in English (Mylonas & MacDonald, 2016). Interestingly, not a single control participant used the basic term 'blue', but rather relied on the aforementioned non-basic 'blue-/green' terms.

3.2. *Error patterns during category acquisition*

3.2.1. *Context training phase*

Fig. 2 shows error counts in the first, context-training phase of

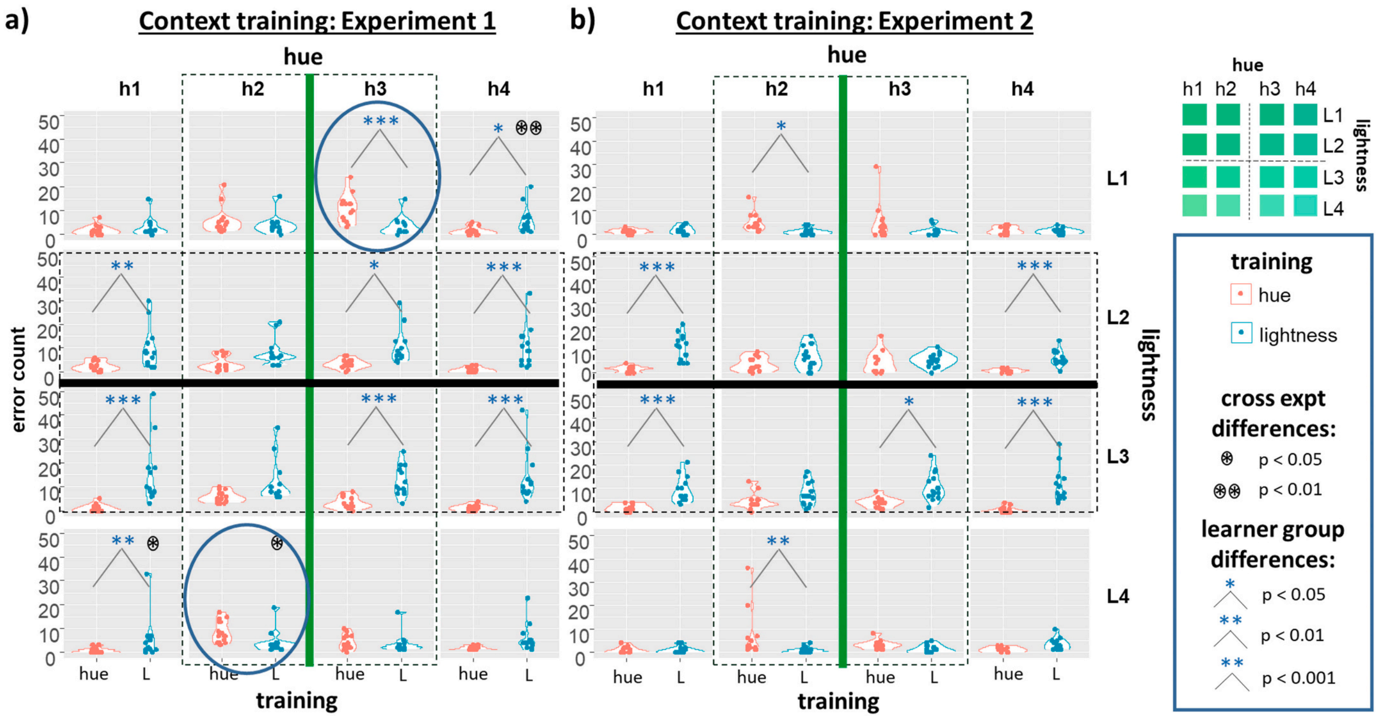

Fig. 2. Error count distributions from the contextual category-learning phase, for hue (red) and lightness (turquoise) learners in Experiments 1 (a) and 2 (b). h1-h4 denote the four hue values while L1-L4 denote the four lightness values, as in Fig. 1 and Table 1 (see also inset in the top-right corner of the figure). The novel hue boundary is depicted by the green vertical line between h2 and h3 while the novel lightness boundary is depicted by the horizontal black line between L2 and L3. Dashed rectangles outline the colour samples that surround the boundaries. In each subplot, hue learners are depicted on the left (in red) while lightness learners are depicted on the right (in turquoise). Violin plots are used to show the error distributions, with individual data points superimposed onto a kernel density estimation. Significant differences in between-subject post-hoc tests (between experiments and between training groups) are indicated by different symbols (see legend on the left). Participants in Experiment 1, where the hue boundary is shifted away from the green prototype and the range is extended, make more errors on lightness judgments. In this experiment, we also observe systematic within-participant differences driven by the task-irrelevant colour dimension in hue learners alone: errors are elevated for h3 at L1 and h2 at L4, in line with the Bezhold-Brucke effect (i.e. darker shades of 'turquoise' appear greener and lighter shades bluer; relevant subplots are circled to highlight this difference). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

learning. Participants produce different patterns of errors across the colour samples and training groups in the two experiments, with a significant four-way interaction ($\chi^2(9) = 29.043$, $p < .001$; for full statistical details of the best fitting model, see Supplementary Materials). In Experiment 1, hue categorisers make 57 ± 22 errors, while lightness categorisers make 128 ± 90 errors (mean +/− SD). In Experiment 2, hue learners make 48 ± 33 errors, while lightness learners make 80 ± 30 errors. Thus, lightness learners consistently make more errors during training (z = 4.902, p < .001) and there are fewer errors in Experiment 2 compared to Experiment 1 (z = −3.037, p = .002). To decompose the four-way interaction between our factors, we follow it up with omnibus pairwise comparisons.

The main difference between experiments is that Experiment 1 gave participants an extended hue range that was shifted away from green-blue towards green-yellow, moving the boundary away from the green prototype. Meanwhile, the lightness range is constant across the two experiments. In this light, it is interesting that the only significant between-experiment differences occur in lightness learners, who make more errors for L1 h4 (z = 4.139, p = .007), L4 h1 (z = 3.982, p = .015) and L4 h2 (z = 3.725, p 0.042) in the first experiment. More errors for these peripheral colour samples (i.e., not proximal to the L2/L3 lightness boundary) could indicate that in the presence of an extended and somewhat shifted hue range, there is more interference from task-irrelevant hue information leading to more widespread errors. All other across-experiment differences are not significant (hue: z > 1.907, p > .961; lightness: z > 2.990, p > .21).

It is also of interest to identify colour samples that attract higher error counts at certain levels of the task-irrelevant dimension for both hue and lightness learners, implying increased interference. In Experiment 1, hue learners are affected by the lightness of boundary hues in an asymmetric way: more errors are made for near-boundary sample h3 L1 (i.e. the lowest level of lightness) than for any other level (L1 vs. L2: z = 5.305, p < .001; L1 vs. L3: z = 5.617, p < .001; L1 vs. L4: z = 4.612, p < .001). For h2 on the opposite side of the boundary, however, these learners exhibit the opposite influence, with more errors at the highest lightness level (L4 vs. L2: z = 4.381, p = .003). Thus, at low lightness levels there is a tendency for the ‘green’ labelled category to expand into its neighbouring sample, while at high lightness levels the opposite occurs, with the ‘blue’ category expanding onto its neighbour. Similar asymmetries in firmness of the lightness boundary across hue levels are absent for lightness learners (z > 3.104, p > .302). On the contrary, while hue learners in Experiment 2 do not exhibit asymmetries in errors across lightness levels at their newly acquired hue boundary (z > 1.726, p > .999), lightness learners show a dissimilar asymmetry, with more errors for h4 compared to h2 at the highest level of lightness L4 (z = 3.894, p = .023). This statistically more modest asymmetry is away from the boundary (L2/L3) and could again be indicative of a reduced salience of the lightness boundary, therefore leading to more widespread errors.

3.2.2. Singleton-training phase

Fig. 3 shows error counts from the singleton, no-context phase of learning. It is notable that very few errors appear in the singleton phases of both experiments. This indicates that participants had successfully acquired the two categories. In Experiment 1, there is a total of 7 ± 4 errors for hue learners and 19 ± 12 for lightness learners. In Experiment 2, hue learners make an average of 8 ± 4 errors, whereas lightness learners make an average of 7 ± 7 errors. Again, there is a four-way

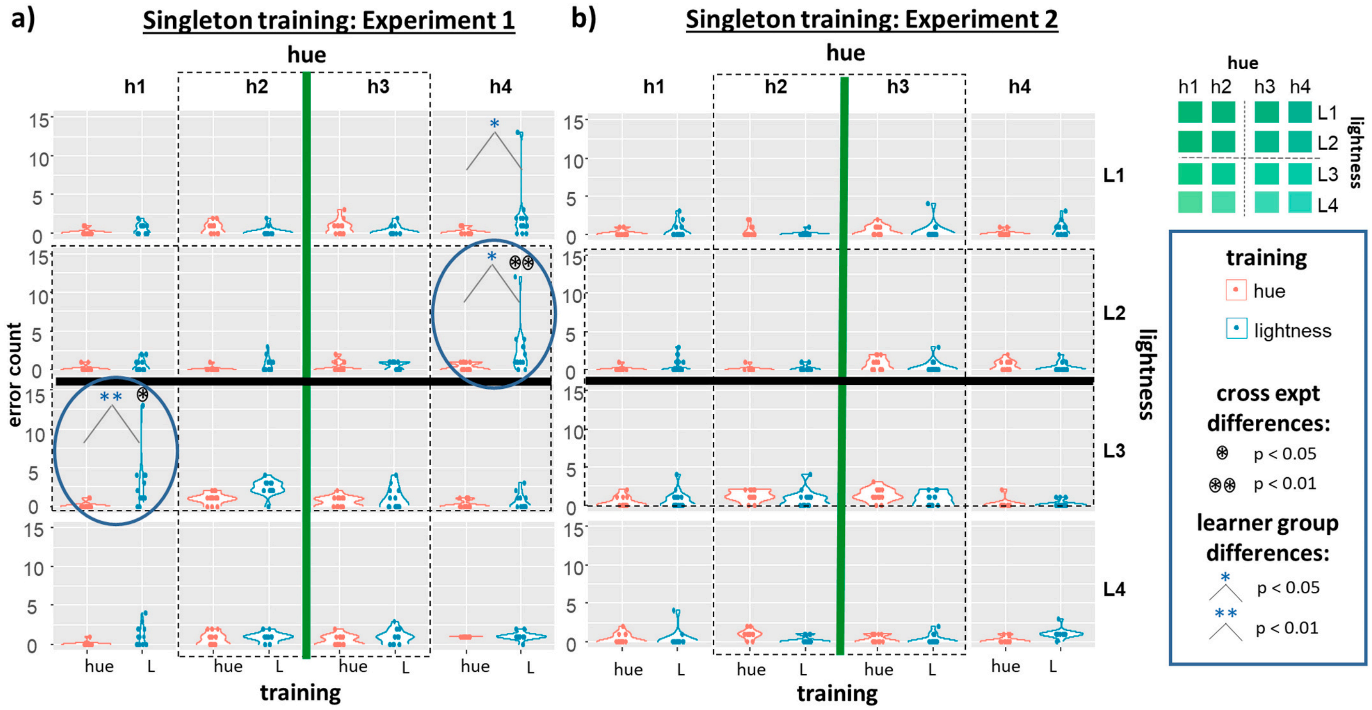

Fig. 3. Error count distributions from the singleton category-learning phase, for hue (red) and lightness (turquoise) learners in Experiments 1 (a) and 2 (b). As in Fig. 2, h1-h4 denote the four hue values while L1-L4 denote the four lightness values (for a visualisation, see inset in the top-right corner of the figure). The novel hue boundary is depicted by the green vertical line between h2 and h3 while the novel lightness boundary is depicted by the horizontal black line between L2 and L3. Dashed lines outline the colour samples that surround the boundaries. In each subplot, hue learners are depicted on the left (in red) while lightness learners are depicted on the right (in turquoise). Violin plots are used to show the error distributions, with individual data points superimposed onto a kernel density estimation. Significant between-subject differences are indicated by different types of symbols (see legend). Data from Experiment 1, where the hue boundary is shifted away from the green prototype and the range is extended, again exhibits higher errors in lightness learners and within-participant differences driven by the task-irrelevant dimension, but this time in lightness learners alone: errors are elevated for h1 at L3 and h4 at L2, which are again in line with the Bezhold-Brucke effect (i.e. near the boundary, greener shades of turquoise appear darker and bluer shades lighter; the relevant parts of the graph are circled to highlight this difference). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

interaction between the fixed effects ($\chi^2(9)$ =19.704, p = .020; for full statistical details of the best fitting model, see Supplementary Materials), meaning that errors accumulate on different colour samples across experiments and training groups. The interaction is largely driven by increased errors in lightness learners for h1 L3 (z = 3.690, p = .044 and h4 L2 (z = 4.183, p = .007) in the first experiment. This asymmetry across the boundary is like the one observed in hue learners during context training: here, the hue on the 'green' side (h1) tends to be erroneously grouped with darker samples and the hue on the 'blue' side (h4) exhibits the opposite tendency and is more likely to be categorised into the 'light' category. Any other differences between experiments are not statistically robust (z > 3.039, p > .332). This indicates that less efficient learning in phase 1 of Experiment 1, wherein lightness learners make many errors across the stimulus grid, seems to extend into the singleton training phase – but now only certain exemplars (h1 L3 and h4 L2) pose a challenge. Meanwhile, other participants (all hue learners and lightness learners in Experiment 2) seem to have acquired the boundary with a similar degree of efficiency by the start of phase 2, producing only sporadic errors that do not follow any specific pattern.

3.2.3. *Interim discussion: Error count patterns*

It appears that small changes to the range of hues encompassed within the to-be-categorised GREEN area and to the location of the boundary can have a sizeable impact on error patterns. This indicates that independent processing of lightness and hue for the purpose of colour categorisation may not be possible, irrespective of which of the two is the attended dimension. Such pattern of results confirms the integrality of hue and lightness (Garner & Felfoldy, 1970) when categorising colours. In line with this, patterns of errors showed an interactive effect in both experiments and both phases of learning (context and singleton phase). Experiments 1 and 2 use stimuli from a very similar GREEN area, with the main categorical difference being a shift in the position of the boundary, effectively putting the 163–165° hues in the same category with green prototype samples in Experiment 1 but with green-yellow samples in Experiment 2. Özgen and Davies (2002) chose to position their boundary at the green prototype. As revealed by our experiments, this boundary is indeed much easier to acquire: a small shift in the position of the hue boundary and the range of hues selectively disrupts the efficiency of lightness category acquisition. In our Experiment 1, the darkest 163° (h3,L1) colour is frequently (erroneously) co-categorised with the samples (h1,h2) labelled as 'green'. Meanwhile, the lightest 155° (h2, L4) colour is more often erroneously co-categorised with the samples labelled as 'blue'. Similarly, darker-than-boundary samples on the side of the green prototype tend to be categorised as lighter (h4 L2), while lighter-than-boundary samples on the side of green-yellow tend to be categorised as darker (h1L3). Experiment 2 relies on the same stimuli as Özgen and Davies (2002) and inspection of Fig. 2b hints that similar miscategorisation tendencies may be present in a few less efficient (i.e., outlier) hue learners yet are greatly subdued across the whole sample. Observed asymmetries are consistent with the Bezold-Brücke effect: near the boundary between green and blue, the label "blue" is dominant for lighter and "green" for darker samples (Lillo et al., 2004). Indeed, the turquoise-labelled area is known to increase and expand towards focal green as stimuli get lighter, making their 'blueness' more salient and increasing the nameability of otherwise more ambiguous colour exemplars between green and blue prototypes (see Figs. 4-6 in Guest & Van Laar, 2000). This implies that it is the greenness/blueness that represents the key attribute along which the newly acquired categories are divided. Indeed, the labelling data indicate that our participants used 'green' and 'blue' labels consistently across experiments, which means that category acquisition is guided by existing labels and that their employment is influenced by the 'greenness' and 'blueness' content of samples even when our stimulus range is moved away from nominally green-blue shades. Not only does the efficiency of acquisition of the lightness boundary become pronouncedly poorer in this case, but boundaries also exhibit asymmetries in their firmness which can only be explained by a failure of pre-existing labels to guide efficient category acquisition when the 'blue' category encompasses 163° - 171° rather than 171°-176° hue samples.

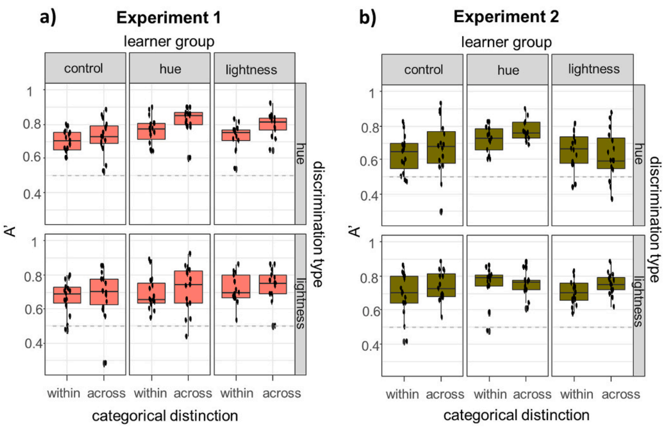

Fig. 4. Box plots of A's from the two experiments. a) Experiment 1. b) Experiment 2. Dashed lines indicate an A' of 0.5, which is equivalent to chance. Dots indicate individual data points, the center-line is the median and lines demarcate the range between the 25th and 75th percentile. In Experiment 1, there is increased sensitivity for hue, as well as for cross-boundary categorisation, without any interactions. In Experiment 2, on the other hand, the key three-way interaction between discrimination type, learner group, and categorical distinction is significant, driven by a cross-boundary advantage specific to novel hue category acquisition.

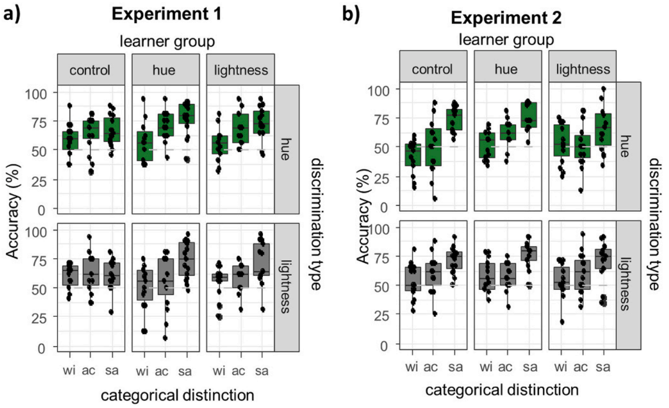

Fig. 5. Box plots of same/different judgment accuracies. a) Experiment 1. b) Experiment 2. The graph makes it clear that above-chance sensitivity (see Fig. 4) is mainly driven by an increased tendency to respond that colour samples are the same, which is most prominent in hue learners (see also interaction plot in Fig. 6). Dashed lines indicate 50% accuracy, which is equivalent to chance, while dots indicate individual data points. Categorical distinctions of judgments are abbreviated: 'wi' for within, 'ac' for across, and 'sa' for same.

3.3. *Same/different delayed discrimination*

As outlined in the Methods section, A' is a non-parametric measure of sensitivity. Since this measure was used in the original study by Özgen and Davies (2002), we report it to facilitate comparisons between the two sets of results. As explained in the Data Analysis section, computing A's from *same/different* judgments reduces some important parameters of participant performance. Moreover, A's are harder to intuitively interpret when compared to proportions of correct accuracy judgments for within-category, across-category and same-exemplar performance. Therefore, to cast further light on the effects of learning novel hue and lightness-based categories, we further analyse the discrimination data

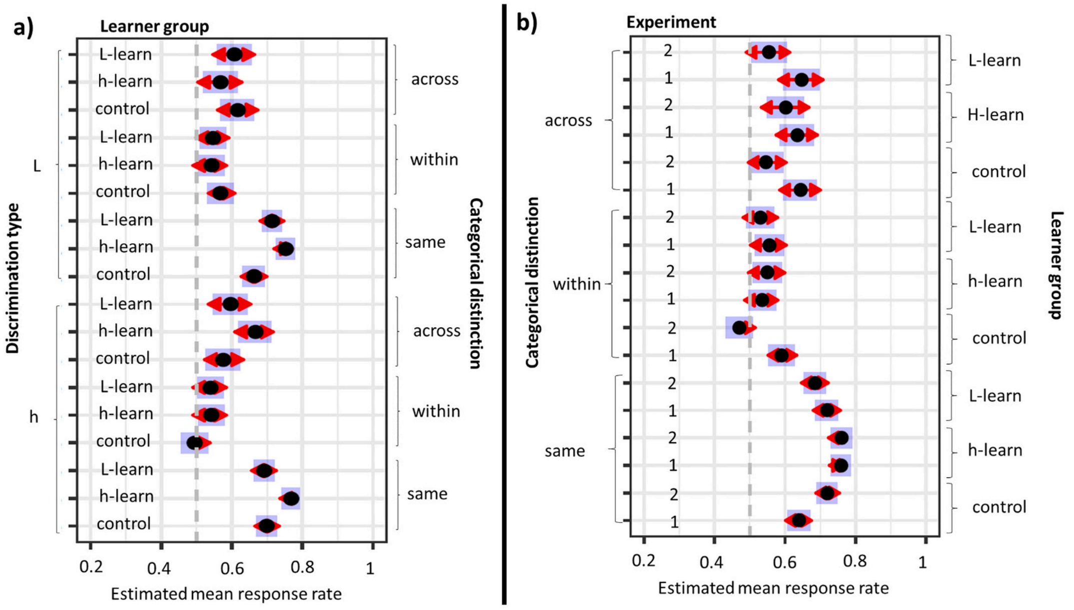

Fig. 6. Interaction plots for estimated response rates derived from GLMMs. A) Learner group by Discrimination type by Categorical distinction. B) Experiment by Learner group by Categorical distinction. Dashed line indicates chance performance. Shaded blue areas indicate 95% confidence intervals of the estimate, black dots indicate mean estimates, while red arrows are used to demarcate statistically significant differences – those conditions for whom the red errors overlap are not statistically different from each other, as evaluated by mvt-corrected omnibus t-tests reported in the main text. Abbreviations used in the plot are as follows – Discrimination Type: L – lightness, h – hue, Learner Group: L-learn – Lightness learners, h-learn – hue learners, control – control group, Experiment: 1 – Experiment 1, 2 – Experiment 2, Categorical distinction: same – same-exemplar, within – within-category, across- cross-category judgment. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

by fitting GLMMs to binary (correct, incorrect) judgments in the experiments.

3.3.1. *A' measure of sensitivity*

3.3.1.1. Experiment 1: Novel boundary between green-yellow and green. Fig. 4a presents A's for Experiment 1, which are analysed with a mixed-model ANOVA with within-subjects factors of discrimination type (hue or lightness) and categorisation type (within- or across- the novel category boundary), and the between-subjects factor of learner group (hue, lightness or control). We find a main effect of discrimination type (F (1, 40) = 10.153, p = .003, $\eta_p^2 = 0.202$), with higher sensitivity for hue. There is also a main effect of categorisation type (F(1, 40) = 21.054, $p <$ .001, $\eta_p^2 = 0.345$), with higher sensitivity across the boundary. There is no interaction between the two within-subjects factors (F(1, 40) = 1.678, p = .203, $\eta_p^2 = 0.04$), nor any interactions with the learner groups (discrimination type: F(2, 40) = 1.321, p = .278, $\eta_p^2 = 0.062$; categorisation type: F(2, 40) = 0.508, p = .605, $\eta_p^2 = 0.025$; full interaction: F (2, 40) = 0.244, p = .785, $\eta_p^2 = 0.012$).

3.3.1.2. Experiment 2: Novel boundary between green-yellow and green-blue. Fig. 4b shows A's for Experiment 2. These are analysed using the same mixed-model ANOVA as above. Again, there is a main effect of discrimination type (F(1, 40) = 13.919, p = .001, $\eta_p^2 = 0.258$), but this time sensitivity is higher for lightness. This is in line with Özgen and Davies (2002) who also observed a similar main effect ($\eta_p^2 = 0.29$). There is also a main effect of categorisation type (F(1, 40) = 4.939, p = .032, η_p^2 = 0.110), with higher sensitivity across the boundary. There is no interaction between the two within-subjects factors (F(1, 40) = 0.029, p = .866, $\eta_p^2 = 0.001$). The main effects do not interact with learner groups (discrimination type: F(2, 40) = 2.635, p = .084, $\eta_p^2 = 0.116$; categorisation type: F(2, 40) = 0.594, p = .557, $\eta_p^2 = 0.029$); however, the key interaction between all three factors shows significance (F(2, 40) = 3.268, p = .048, $\eta_p^2 = 0.14$). To follow this up, we conduct three paired t-tests, contrasting sensitivity for within- and across-category boundary comparisons for each group. There are no robust differences in controls (t(14) = 0.966, p = .350) or lightness learners (t(14) = 1.104, p = .288), but hue learners are significantly better for cross-category judgments (t (12) = 3.209, p = .008, d = 0.890).

3.3.1.3. Interim discussion: Comparison of A's across experiments. Neither of the two experiments fully replicate the findings of Özgen and Davies (2002): specifically, in Experiment 1, we find no effect of training, while in Experiment 2 we find an improvement only for hue learners, and only on their trained dimension. Both studies use the same lightness range while Experiment 2 also uses the same hue range as Özgen and Davies. In Experiment 2, we replicate the selective improvement of cross-category judgments in hue learners, with a similar effect size. In Experiment 1, we find facilitated hue discrimination in all groups, rather than just in hue learners, indicating that no facilitation of discrimination has taken place. However, we find no improvement for lightness judgments in either of the experiments, for either hue learners or lightness learners. For more detailed comparisons of our mean A' data with those of Özgen and Davies, see Supplementary Materials.

One could argue that longer training would be necessary for hue learners to acquire lightness categories. However, this is inconsistent with the fact that fully trained lightness categorisers also fail to show any improvement. It may be possible that the slightly lower saturation of colours in our stimulus set, necessitated by the technical limitations of the display, selectively compromised the acquisition of lightness categories. It is, however, unclear why this would happen, as differences are small and the chroma of our colour samples remains very high. The effect size for discrimination improvements due to lightness category acquisition is not large and Özgen and Davies (2002) did not have enough sensitivity to capture this effect reliably. To evaluate if such an effect can be captured with increased power, in the subsequent analysis we combine data from both experiments to achieve an even larger sample (~30 per group). We then perform a comprehensive re-analysis of these data using GLMMs on binary accuracy outcomes (correct vs. incorrect), which also allows us to look at *same* and *different* judgments separately.

3.3.2. GLMM analysis of accuracies across both experiments

Acccuracies are visualised in Fig. 5. The maximal model that could be fit to the data included the categorical distinction (same exemplar, within-category exemplar or across-category exemplar), discrimination type (hue or lightness), learner group (controls, hue learners and lightness learners) and the experiment (1 or 2) as fixed effects, and by-participant intercepts as a random effect.

The best-fitting final model (Table S3, Supplementary Results) includes three-way interactions between learner group, discrimination type and categorical distinction ($\chi^2(4) = 15.76$, $p = .003$) and the experiment, learner group and categorical distinction ($\chi^2(4) = 36.551$, $p < .001$). These two interactions are crucial: the first one concerns the predicted improvements on cross-category distinctions in the learned dimensions, while the second one concerns between-experiment differences in category learning effects.

The main hypothesis on cross-category learning effects concerns the interaction between learner group, discrimination type and categorical distinction. We decompose it using omnibus *t*-tests corrected for multiple comparisons. As can be seen in Fig. 6a, key differences in performance concern hue learners, who show an advantage for across-category compared to within-category hue discrimination pairs ($z = 4.352$, $p = .002$), unlike either controls ($z = 2.873$, $p = .251$) or lightness learners ($z = 1.952$, $p = .879$). However, hue learners also show a tendency to respond more accurately when presented with same samples compared to lightness learners ($z = 3.876$, $p = .012$) and, not statistically robustly, when compared with controls ($z = 3.451$, $p = .0516$). In the absence of any between-group differences for within-category pairs (all ps > 0.9451), and with non-robust between-group differences in across-category pairs (hue learners vs. controls: $z = 2.645$, $p = .401$; lightness learners vs. controls: $z = 0.588$, $p = 1.00$; hue learners vs. lightness learners: $z = 2.056$, $p = .826$), it appears that the tendency of hue learners to increase their 'same' responses is another outcome specific to hue training.

The other significant interaction indicated by the GLMM analysis involves between-experiment differences between learner groups on same/within−/across-category judgments. As shown in Fig. 6b, the largest differences between the two experiments concern the 'same'-response rates. In the control group, these are lower in Experiment 1 as opposed to Experiment 2 ($z = 4.363$, $p = .002$). This is logical, as in Experiment 1 colour samples are more saturated and somewhat more spread out in terms of hue. This would have made the distinctions between colours more salient and thus potentially reduced the bias towards responding that samples are identical. However, these between-experiment differences only affect the controls - there are no differences in 'same'-response rates for either hue learners ($z = 0.056$, $p = 1.00$) or lightness learners ($z = 1.507$, $p = .988$), who would have had more extensive exposure to the colour samples during their training.

4. Discussion

Our study confirms that the acquisition of novel hue-based categories leads to discrimination advantages, but also shows how this is highly dependent on the employed colour set. With colours constrained to a narrower GREEN region and the boundary close to the prototype, we replicate, for hue learners, the advantage for cross-category compared to within-category discriminations (Özgen & Davies, 2002) and demonstrate it is accompanied by a tendency towards increased response rates for 'same'-category judgments. We fail, however, to replicate the learning advantage for lightness-based novel categories. Lightness-boundary learners make more errors than hue-boundary learners during category acquisition. Less efficient acquisition of lightness-based categories is in line with previous work, which shows that lightness-based categories are less well demarcated than hue-based categories (Martinovic et al., 2020). Labelling patterns provide further context for interpreting these findings: while lightness learners use category labels that are mainly lightness-based but sometimes also involve hue and saturation signifiers, hue learners consistently use 'green' and 'blue' labels. Error patterns during category acquisition increase for cross-boundary neighbours in accordance with the Bezold-Brücke effect, with darkest neighbours tending to be classified as 'green' and lightest neighbours tending to be classified as 'blue'. The use of simple, basic-term labels 'green' and 'blue' to facilitate categorisation is in line with the label-feedback hypothesis (Lupyan, 2012). While the 'blue' label would not be used in everyday situations to name any of the displayed colour samples, it would be remarkably useful in guiding attention to the most salient change in information across the hue-boundary in our stimulus set, i.e. that of the change of tinge between greenness and blueness. Consistent with this interpretation, an ERP study of the neural correlates of acquired colour category effects found that the only component that differs between learners and controls is the P300 (Clifford et al., 2012). Enhancements in P300 amplitude are firmly related to increased attentional processing (for an overview, see Polich, 2007). This suggests that labels may drive attention to category-diagnostic chromatic content and facilitate categorical modulations of discrimination, with the efficiency of label use dependent on the ease with which these labels can be applied. The importance of attention in guiding the effect of labels is in line with the categorical facilitation model (Witzel, 2019).

With a colour set that covers an identical area of 'green' as in Özgen and Davies (2002), we replicate the key finding that hue categorisers perform better on cross-category than within-category hue pairs. Still, neither hue learners nor lightness learners improve their performance on across-category pairs varying in lightness. Lightness discrimination was not any easier than in the original study, so this cannot explain the lack of the effect. Lightness discrimination is slightly higher for across-category than within-category pairs (Figs. 4-6), but this difference is not statistically robust. Thus, lightness categories may be more difficult to acquire, warranting longer training, and may produce weaker categorical effects on delayed discrimination than hue categories. Lightness category effects are at best small-to-medium sized ($f < 0.16$), as they fail to be captured by the joint analysis of data from our two experiments and are likely to have been overestimated in the original study, which relied on a small sample size. Either way, our findings resolve the apparent inconsistency between Özgen and Davies' (2002) and Martinovic et al. (2020) findings on lightness-based categories. It appears that lightness-based categories are indeed less structured and therefore harder to acquire compared to hue-based categories. Even the labels used for novel lightness categories are more variegated: along with achromatic modifiers, they include the elaborated non-lightness labels, which indicates verbal interference from hue and saturation dimensions. The lower structuredness of lightness-based categories and the less consistent labelling of these categories does not mean, however, that they are in any way less fundamental than hue-based categories. In fact, the less structured internal relations within the primary colour categories 'cool' and 'warm' could have acted as a key driver to further structuring of the colour space into additional basic categories.

Accuracy for within-category discriminations is low in both Experiments (see Fig. 5), often remaining close to chance level. With such poor within-category accuracy, it is not surprising that the data manifest acquired across-category distinctiveness, rather than within-category similarity. To be able to test effects of categorical perception on within- and across-category pairs in the same experiment, it would be useful to obtain a baseline accuracy of about 75% for both types of pairs, allowing equal room for effects of both across-category distinctiveness and within-category similarity to emerge. In our controls, the accuracy is highest for judgments on same exemplars (~70%). More distinctive colour samples that would spread over a bigger region of the colour space could allow better baseline accuracy, but it would be difficult to create an array of such samples while remaining within a single colour category. Although the GREEN region occupies the biggest area of the colour space, we already observe prominent intrusions of the ‘blue’ category in our labelling data. Therefore, it would be tricky to further extend the stimulus range whilst avoiding intrusions from ‘yellow’ and ‘blue’. The limited options in drawing truly ‘category-neutral’ stimulus samples represent a significant challenge for studying mechanisms that guide colour categorisation through novel category training.

Low overall accuracy in the discrimination task is also the inevitable consequence of delayed discrimination. It introduces a working memory component, which is likely to compromise discrimination judgments and introduce a higher contribution from memory colours and their related labels (Bae, Olkkonen, Allred, & Flombaum, 2015; see also Uchikawa & Shinoda, 1996). In fact, it would be important to investigate in detail such *just memorable differences (JMDs)* for colour, as they involve both perceptual and working memory contributions. Memory contributions would act by attracting representations towards focal colours (Bae et al., 2015) and become stronger with extension of the delay between exemplars (e.g. Ajda & Bračko, 2019). On the contrary, fewer categorical contributions would be in play when measuring JNDs for simultaneously viewed colours. This corresponds to the conclusions drawn by Webster and Kay (2012), who found weak and inconsistent categorical effects in a perceptual grouping task that did not require explicit naming, but stronger categorical influence in a task that required explicit judgment of hue. In that sense, conceivably the effects we observe are partly driven by memory and thus cannot be exclusively interpreted as categorical *perception* (for a very similar debate in relation to categorical perception of speech signals, see Gerrits & Schouten, 2004; Schouten et al., 2003).

Our findings also provide evidence that existing colour labels play an important role in perceptual learning of novel colour categories and that this labelling does not necessarily have to conform to everyday naming (e.g., ‘blue’ was used by most learners for stimuli that would in everyday circumstances be named ‘green’ or ‘blue-green’). This complicates the interpretation of perceptual learning as the sole/main mechanism for acquisition of colour categories. Language and perceptual learning are more likely to interact during this process (Goldstein, Davidoff, & Roberson, 2009; Wagner et al., 2013). Future research in this field should further scrutinise learning performance, contrasting successful learners with unsuccessful learners, as in Perez-Gay Juarez et al. (2017). This is preferable to the control group approach, as the control group has very limited experience of the stimulus range but mere exposure also affects category learning (Folstein, Gauthier, & Palmeri, 2010). Controlling for exposure in this way would enable stronger generalisations from patterns in the discrimination data, such as higher ‘same’-response rates in hue-learners observed in this study. These could be due to response biases introduced by mere exposure, or, more interestingly, may be a consequence of enhanced similarity in appearance due to category learning (Goldstone & Hendrickson, 2010; Livingston, Andrews, & Harnad, 1998). Labelling differences between controls and learners, with controls using a more diverse set of labels, hints towards such learning effects. Different labelling strategies may also account for at least some of the large individual differences in the tendency to align response patterns with categories that were observed by Webster and Kay (2012).

Moreover, the unusual use of “blue” as a label for ‘green’ or ‘blue-green’ exemplars is consistent with the model in which relative colour differences subserve attentional guidance (e.g., bluer vs. greener; Becker, 2010; Becker, Folk, & Remington, 2010; Harris, Remington, & Becker, 2013). It is inevitable that relational learning would benefit from relational attention. Likewise, contemporary computational models of speech perception reframe categorical perception as expectation-relative encoding of features and cues, where categories act to set expectations relative to which perceptual information is interpreted (Apfelbaum & McMurray, 2015; McMurray & Jongman, 2011). Attention also shapes the encoding of visual features (Dube, Emrich, & Al-Aidroos, 2017) and has recently been proposed as a potential mechanism for categorical facilitation of colour perception (Witzel, 2019), where improvements only occur in those participants whose attention is drawn to the categorical distinction between different hues. Our findings suggest that attention may indeed be the key mechanism behind categorical colour perception, and that linguistic labels are likely to mediate this effect by facilitating attentional guidance (Lupyan, 2017; Zettersten & Lupyan, 2020) to the most salient information that differentiates exemplars across the category-boundary. In this way, our findings reconcile the label-feedback hypothesis and the categorical facilitation hypothesis within a single framework that can parsimoniously explain both the observed categorical perception phenomena and their underlying mechanisms.

Open practices

Pre-registration of the study design and analyses, data, R analysis scripts, detailed stimulus coordinates and Matlab scripts for running the experiments are available at https://osf.io/vn5t2/?view_only=9a6570677cff4e66bab1ef05311c86d

Credit statement

JM designed the study, collected and analysed the data, and wrote and revised the manuscript.

Acknowledgments

This project was partly supported by a German Academic Exchange Service (Deutscher Akademischer Austauschdienst; DAAD) re-invitation programme award (application no. 57378442). The author would like to thank Galina Paramei for her valuable comments on an earlier draft of the manuscript and Sophia Thompson, Meredith Brown and Joshua Bugg for help with data collection. Pre-registration, data, R analysis scripts, detailed stimulus coordinates and Matlab scripts for running the experiments are available at https://osf.io/vn5t2/

Appendix B. Supplementary data

Additional methodological details, further data visualisations and full summaries of the best-fitting General Linear Mixed-effect Models (GLMMs) are provided. Supplementary data to this article can be found online at https://doi.org/10.1016/j.cognition.2023.105657.

References

Ajda, C., & Bračko, S. (2019). Influence of basic colour parameters on colour memory. *TEKSTILEC, 62*, 232–241. https://doi.org/10.14502/Tekstilec2019.62.232-241

Apfelbaum, K. S., & McMurray, B. (2015). Relative cue encoding in the context of sophisticated models of categorization: Separating information from categorization. *Psychonomic Bulletin & Review, 22*(4), 916–943. https://doi.org/10.3758/s13423-014-0783-2

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. https://doi.org/10.1016/j.cognition.2010.05.016

Bae, G. Y., Olkkonen, M., Allred, S. R., & Flombaum, J. I. (2015). Why some colors appear more memorable than others: A model combining categories and particulars in color working memory [article]. *Journal of Experimental Psychology. General, 144* (4), 744–763. https://doi.org/10.1037/xge0000076

Baker, D. H., Lygo, F. A., Meese, T. S., & Georgeson, M. A. (2018). Binocular summation revisited: Beyond $\sqrt{2}$. *Psychological Bulletin, 144*(11), 1186–1199. https://doi.org/10.1037/bul0000163

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68*(3). https://doi.org/10.1016/j.jml.2012.1011.1001

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Becker, S. I. (2010). The role of target–distractor relationships in guiding attention and the eyes in visual search. *Journal of Experimental Psychology: General, 139*(2), 247–265. https://doi.org/10.1037/a0018808

Becker, S. I., Folk, C. L., & Remington, R. W. (2010). The role of relational information in contingent capture [article]. *Journal of Experimental Psychology. Human Perception and Performance, 36*(6), 1460–1476. https://doi.org/10.1037/a0020370

Ben-Shachar, M., Lüdecke, D., & Makowski, D. (2020). Effectsize: Estimation of effect size indices and standardized parameters. *Journal of Open Source Software, 5*(56), 2815. https://doi.org/10.21105/joss.02815

Bornstein, M. H., Kessen, W., & Weiskopf, S. (1976). Color vision and hue categorization in young human infants. *Journal of Experimental Psychology: Human Perception and Performance, 2*, 115–129. https://doi.org/10.1037/0096-1523.2.1.115

Bornstein, M. H., & Korda, N. O. (1984). Discrimination and matching within and between hues measured by reaction-times - some implications for categorical perception and levels of information-processing [article]. *Psychological Research Psychologische Forschung, 46*(3), 207–222. https://doi.org/10.1007/bf00308884

Bosten, J., & Lawrance-Owen, A. J. (2014). No difference in variability of unique hue selections and binary hue selections. *Journal of the Optical Society of America. A, Optics, Image Science, and Vision, 31*(4), A357–A364.

Brown, A. M., Lindsey, D. T., & Guckes, K. M. (2011). Color names, color categories, and color-cued visual search: Sometimes, color perception is not categorical [article]. *Journal of Vision, 11*(12), 21. article 2 https://doi.org/10.1167/11.12.2.

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. https://doi.org/10.3758/BF03207885

Clifford, A., Franklin, A., Holmes, A., Drivonikou, V. G., Özgen, E., & Davies, I. R. L. (2012). Neural correlates of acquired color category effects. *Brain and Cognition, 80* (1), 126–143. https://brain.unboundmedicine.com/medline/citation/22722021/Neural_correlates_of_acquired_color_category_effects_https://linkinghub.elsevier.com/retrieve/pii/S0278-2626(12)00073-5.

Cohen, J. (2013). *Statistical power analysis for the behavioral sciences*. Routledge.

Danilova, M. V., & Mollon, J. D. (2012). Foveal color perception: Minimal thresholds at a boundary between perceptual categories. *Vision Research, 62*, 162–172. https://doi.org/10.1016/j.visres.2012.04.006

Doumas, L. A. A., Hummel, J. E., & Sandhofer, C. M. (2008). A theory of the discovery and predication of relational concepts. *Psychological Review, 115*(1), 1–43. https://doi.org/10.1037/0033-295X.115.1.1

Dube, B., Emrich, S. M., & Al-Aidroos, N. (2017). More than a filter: Feature-based attention regulates the distribution of visual working memory resources. *Journal of Experimental Psychology. Human Perception and Performance, 43*(10), 1843–1854. https://doi.org/10.1037/xhp0000428

Fairchild, M. D. (2013). *Colour appearance models* (3rd revised ed.). Wiley-Blackwell.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*, 175–191. https://doi.org/10.3758/BF03193146

Folstein, J., Gauthier, I., & Palmeri, T. (2010). Mere exposure alters category learning of novel objects [original research]. *Frontiers in Psychology, 1*(40). https://doi.org/10.3389/fpsyg.2010.00040

Forder, L., & Lupyan, G. (2019). Hearing words changes color perception: Facilitation of color discrimination by verbal and visual cues. *Journal of Experimental Psychology. General, 148*(7), 1105–1123. https://doi.org/10.1037/xge0000560

Garner, W. R., & Felfoldy, G. L. (1970). Integrality of stimulus dimensions in various types of information processing. *Cognitive Psychology, 1*(3), 225–241. https://doi.org/10.1016/0010-0285(70)90016-2

Gerrits, E., & Schouten, M. E. H. (2004). Categorical perception depends on the discrimination task. *Perception & Psychophysics, 66*(3), 363–376. https://doi.org/10.3758/BF03194885

Goldstein, J., Davidoff, J., & Roberson, D. (2009). Knowing color terms enhances recognition: Further evidence from English and Himba. *Journal of Experimental Child Psychology, 102*(2), 219–238. https://doi.org/10.1016/j.jecp.2008.06.002

Goldstone, R. L., & Hendrickson, A. T. (2010). Categorical perception. *WIREs Cognitive Science, 1*(1), 69–78. https://doi.org/10.1002/wcs.26

Grandison, A., Sowden, P. T., Drivonikou, V. G., Notman, L. A., Alexander, I., & Davies, I. R. L. (2016). Chromatic perceptual learning but no category effects without linguistic input [original research]. *Frontiers in Psychology, 7*. https://doi.org/10.3389/fpsyg.2016.00731

Guest, S., & Van Laar, D. (2000). The structure of colour naming space. *Vision Research, 40*, 723–734. https://doi.org/10.1016/S0042-6989(99)00221-7

Harris, A. M., Remington, R. W., & Becker, S. I. (2013). Feature specificity in attentional capture by size and color [article]. *Journal of Vision, 13*(3). ://WOS:000319810800012.

Harrison, X. A. (2014). Using observation-level random effects to model overdispersion in count data in ecology and evolution. *PeerJ, 2*, Article e616. https://doi.org/10.7717/peerj.616

Hespos, S., Gentner, D., Anderson, E., & Shivaram, A. (2021). The origins of same/different discrimination in human infants. *Current Opinion in Behavioral Sciences, 37*, 69–74. https://doi.org/10.1016/j.cobeha.2020.10.013

Istomina, Z. M. (1960a). O vzaimootnošenii vosprijatija i nazyvanija cveta u detej doškol'nogo vozrasta (eksperimentalnoe issledovanie). *Isvestija Akademii pedagogičeskix nauk RSFSR, 113*, 76–102.

Istomina, Z. M. (1960b). Vosprijatie i nazyvanie cveta v rannem vozraste. In *, 113. Izvestija Akademii pegagogičeskih nauk RSFSR vyp*.

Lakens, D. (2022). Sample size justification. *Collabra: Psychology, 8*(1), 33267. https://doi.org/10.1525/collabra.33267

Lenth, R. V., Singmann, H., Love, J., Buerkner, P., & Herve, M. (2019). emmeans: Estimated Marginal Means, aka Least Squares Means (Version 1.4.3.01). https://CRAN.R-project.org/package=emmeans.

Liberman, A. M., Harris, K. S., Hoffman, H. S., & Griffith, B. C. (1957). The discrimination of speech sounds within and across phoneme boundaries. *Journal of Experimental Psychology, 54*, 358–368. https://doi.org/10.1037/h0044417

Lillo, J., Aguado, L., Moreira, H., & Davies, I. (2004). Lightness and hue perception: The Bezold-Brucke effect and colour basic categories. *Psicologica, 25*(1), 23–43.

Lindsey, D. T., & Brown, A. M. (2021). Lexical Color Categories. *Annual Review of Vision Science, 7*(1), 605–631. https://doi.org/10.1146/annurev-vision-093019-112420

Livingston, K. R., Andrews, J. K., & Harnad, S. (1998). Categorical perception effects induced by category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24*(3), 732–753. https://doi.org/10.1037/0278-7393.24.3.732

Lüdecke, D., Ben-Shachar, M. S., Patil, I., Waggoner, P., & Makowski, D. (2021). Performance: an R package for assessment, comparison and testing of statistical models. *Journal of Open Source Software, 6*(60), 3139. https://doi.org/10.21105/joss.03139

Luo, M. R. (2016). CIELAB. In R. Luo (Ed.), *Encyclopedia of color science and technology* (pp. 207–212). Berlin Heidelberg: Springer. https://doi.org/10.1007/978-3-642-27851-8_11-1.

Lupyan, G. (2012). Linguistically modulated perception label-feedback hypothesis [article]. *Frontiers in Psychology, 3, 13*, Article 54. https://doi.org/10.3389/fpsyg.2012.00054

Lupyan, G. (2017). Changing what you see by changing what you know: The role of attention [hypothesis and theory]. *Frontiers in Psychology, 8*(553). https://doi.org/10.3389/fpsyg.2017.00553

Macmillan, N. A., & Creelman, C. D. (2004). *Detection theory: A user's guide* (Second edition ed.). Psychology Press.

Martinovic, J., Paramei, G. V., & MacInnes, W. J. (2020). Russian blues reveal the limits of language influencing colour discrimination. *Cognition, 201*, Article 104281. https://doi.org/10.1016/j.cognition.2020.104281

McMurray, B., & Jongman, A. (2011). What information is necessary for speech categorization? Harnessing variability in the speech signal by integrating cues computed relative to expectations. *Psychological Review, 118*, 219–246. https://doi.org/10.1037/a0022325

Moscatelli, A., Mezzetti, M., & Lacquaniti, F. (2012). Modeling psychophysical data at the population-level: The generalized linear mixed model. *Journal of Vision, 12*(11), 26. https://doi.org/10.1167/12.11.26

Mylonas, D., & MacDonald, L. (2016). Augmenting basic colour terms in english. *Color Research & Application, 41*(1), 32-42. https://doi.org/10.1002/col.21944

Özgen, E., & Davies, I. R. L. (2002). Acquisition of categorical color perception: A perceptual learning approach to the linguistic relativity hypothesis. *Journal of Experimental Psychology: General, 131*(4), 477–493. https://doi.org/10.1037/0096-3445.131.4.477

Paramei, G. V. (2005). Singing the Russian blues: An argument for culturally basic color terms [article; proceedings paper]. *Cross-Cultural Research, 39*(1), 10–38. https://doi.org/10.1177/1069397104267888

Paramei, G. V. (2007). Russian "blues": Controversies of basicness. In R. E. MacLaury, G. V. Paramei, & D. Dedrick (Eds.), *Anthropology of color: Interdisciplinary multilevel modeling* (pp. 75–106). John Benjamins.

Perez-Gay Juarez, F., Theriault, C., Gregory, M., Sabri, H., Rivas, D., & Harnad, S. (2017). How and why does category learning cause categorical perception? *International Journal of Comparative Psychology, 30*.

Polich, J. (2007). Updating p300: An integrative theory of P3a and P3b [review]. *Clinical Neurophysiology, 118*(10), 2128–2148. https://doi.org/10.1016/j.clinph.2007.04.019

R_Core_Team. (2016). *R: A language and environment for statisticall computing*. R Foundation for Statistical Computing. https://www.R-project.org/.

Regan, B. C., Reffin, J. P., & Mollon, J. D. (1994). Luminance noise and the rapid determination of discrimination ellipses in color deficiency [article]. *Vision Research, 34*(10), 1279–1299. ://A1994NZ95400006.

Rhodes, S., Cowan, N., Parra, M. A., & Logie, R. H. (2019). Interaction effects on common measures of sensitivity: Choice of measure, type I error, and power. *Behavior Research Methods, 51*(5), 2209–2227. https://doi.org/10.3758/s13428-018-1081-0

Roberson, D. (2005). Color categories are culturally diverse in cognition as well as in language [article]. *Cross-Cultural Research, 39*(1), 56–71. https://doi.org/10.1177/1069397104267890

Roberson, D., Davies, I., & Davidoff, J. (2000). Color categories are not universal: Replications and new evidence from a stone-age culture. *Journal of Experimental Psychology. General, 129*(3), 369–398. <Go to ISI>://000089151400006.

Sandhofer, C. M., & Doumas, L. A. A. (2008). Order of presentation effects in learning color categories. *Journal of Cognition and Development, 9*(2), 194–221. https://doi.org/10.1080/15248370802022639

Schäfer, T., & Schwarz, M. A. (2019). The meaningfulness of effect sizes in psychological research: Differences between sub-disciplines and the impact of potential biases [original research]. *Frontiers in Psychology, 10*. https://doi.org/10.3389/fpsyg.2019.00813

Schiller, F., Valsecchi, M., & Gegenfurtner, K. R. (2018). An evaluation of different measures of color saturation. *Vision Research, 151*, 117–134. https://doi.org/10.1016/j.visres.2017.04.012

Schouten, B., Gerrits, E., & van Hessen, A. (2003). The end of categorical perception as we know it. *Speech Communication, 41*, 71–80. https://doi.org/10.1016/S0167-6393(02)00094-8

Siuda-Krzywicka, K., Boros, M., Bartolomeo, P., & Witzel, C. (2019). The biological bases of colour categorisation: From goldfish to the human brain. *Cortex: A Journal Devoted to the Study of the Nervous System and Behavior, 118*, 82–106. https://doi.org/10.1016/j.cortex.2019.04.010

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

Uchikawa, K., & Shinoda, H. (1996). Influence of basic color categories on color memory discrimination. *Color Research & Application, 21*(6), 430–439. https://doi.org/10.1002/(SICI)1520-6378(199612)21:6<430::AID-COL5>3.0.CO;2-X

Wagner, K., Dobkins, K., & Barner, D. (2013). Slow mapping: Color word learning as a gradual inductive process. *Cognition, 127*(3), 307–317. https://doi.org/10.1016/j.cognition.2013.01.010

Wagner, K., Jergens, J., & Barner, D. (2018). Partial color word comprehension precedes production. *Language Learning and Development, 14*(4), 241–261. https://doi.org/10.1080/15475441.2018.1445531

Wagner, K., Tillman, K., & Barner, D. (2016). Inferring number, time and color concepts from core knowledge and linguistic structure. In D. Barner, & A. S. Baron (Eds.), *Core knowledge and conceptual change*. Oxford University Press.

Webster, M. A., & Kay, P. (2012). Color categories and color appearance [article]. *Cognition, 122*(3), 375–392. https://doi.org/10.1016/j.cognition.2011.11.008

Westland, S., Ripamonti, C., & Cheung, V. (2012). *Computational colour science using MATLAB* (2nd ed.). John Wiley & Sons.

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag.

Witzel, C. (2019). Misconceptions About Colour Categories [journal article]. *Review of Philosophy and Psychology, 10*, 499–540. https://doi.org/10.1007/s13164-018-0404-5

Witzel, C., & Gegenfurtner, K. R. (2015). Categorical facilitation with equally discriminable colors [article]. *Journal of Vision, 15*(8), 33. Article 22 https://doi.org/10.1167/15.8.22.

Witzel, C., & Gegenfurtner, K. R. (2018). Color perception: objects, constancy, and categories. In J. A. Movshon, & B. A. Wandell (Eds.)*, 4*. *Annual review of vision science* (pp. 475–499). Annual Reviews https://doi.org/10.1146/annurev-vision-091517-034231.

Zeileis, A., Fisher, J. C., Hornik, K., Ihaka, R., McWhite, C. D., Murrell, P., … Wilke, C. O. (2020). Colorspace: A toolbox for manipulating and assessing colors and palettes. *Journal of Statistical Software, 96*(1), 1–49.

Zettersten, M., & Lupyan, G. (2020). Finding categories through words: More nameable features improve category learning. *Cognition, 196*, Article 104135.