Crowd sourcing word-colour associations

Yun Chen, Boshuo Guo, Dian Li, Stephen Westland[†] and Marjan Vazirian[*]

School of Design, University of Leeds, UK
[*]*AkzoNobel NV, The Netherlands*
[†]*Email: s.westland@leeds.ac.uk*

It is known that colours evoke memories and emotions and that these associations are valuable in design. However, it is difficult to ascertain universal associations because often the word-colour associations are culture- and context-specific. Nevertheless, some important universal colour associations have been shown to exist, at least if the variable of culture is removed. Laboratory studies, however, although effective struggle to collect sufficient information to allow the data to be useful in a practical setting. In this study, a crowd-sourcing internet experiment was carried out to find colours that are associated with different words and the results were compared with those from a previous laboratory study. Approximately 150 participants made a total of 1350 word-colour associations to generate palettes for 30 target words. Comparison with corresponding data from the laboratory study suggests that crowd-sourcing is a promising route and should be explored further.

Introduction

There is a rich history of work that relates colours to language through their colour names. For example, the Berlin and Kay study [1] was concerned with basic colour names and found some similarities between different cultures in the development and use of at least the most basic colour names such as red, yellow and blue. The Berlin and Kay study was interesting in its contribution to arguments about the Sapir-Whorf hypothesis; this much-debated hypothesis states that the language a person speaks will affect their thought (or, in this particular case, their colour perception). This debate still rages on and more recently there has been opposition to the universalist view represented by the original Berlin and Kay study [2]. There have also been studies to map the use of colour names to specific colorimetric boundaries [3]. However, another aspect of the relationship between colour and language is the use of language to represent the emotions or concepts that colours may evoke. It is this aspect of colour and language that this paper is concerned with.

Feelings and memories are evoked by colours [4-5] and even by colour names [6]. Whether these associations are learned associations or shaped by nature or whether they are innate is not clear. The nature of these associations is also confusing in the literature. Some researchers have referred to colour emotions [7-8] and it is certainly the case that colours can elicit an emotional response. On the other hand, it has been argued that the colour associations may be more cognitive [9-10]. No matter what the cause or the precise nature of these associations they are important in many aspects of marketing [11], retail [12-13], art [14] and design [15].

The importance of colour associations is evident in the observation that the internet abounds with information, often presented as infographics, about colour associations. However, often these infographics oversimply colour associations and can be caricatured as stating that 'red means this and blue means that.' The reality, however, is that colour associations are complex and multifaceted. Grieve has argued that colours have no meanings *per se* but, rather, than the meanings they are associated with originate from their use in a particular situation or with a particular object [16]. There is evidence that colour associations vary with culture [17-18] and product category [19]. Nevertheless, it is clear that some universal colour associations exist, at least within cultures [8, 20] even if the extent to which those colour meanings change when applied to a specific product category might be less clear [21]. Consequently a number of studies have been carried out to measure and understand colour associations. Important contributions have been made by Xin *et al.* [7] and by Ou *et al.* [8] in a number of experiments that typically have used semantic differential scaling or categorical scaling as investigative techniques. With these methods participants are shown colours (either physical samples or colours on screen) and are asked to indicate the strength of the association of the colour with regard to two bi-polar terms (such as maleness and femaleness) in the case of semantic differential scaling or the strength of the association with a single term (such as maleness) using a fixed-number discrete scale in the case of categorical scaling. Whilst these methods provide valuable data and have produced some models that allow the associations of any colour to be determined (with regard to the limited number of terms used in the experiments) they are not easy to apply in a design context. Many design processes begin with a brief and it is here where concepts that define the intended design originate; in these situations, the designer seeks colours that can represent or communicate these concepts [22]. In part to address this issue, a recent study presented participants with a word on a computer screen and asked them to select colours that they associate with this word [10]. The term palette in this work is used simply to refer to a collection of colours. This is consistent with its use in colour science [23-24] although we note that in other fields the term has different meanings. We also distinguish the term colour palette from the term colour scheme; the latter implies that the colours have some relationships (for example, colour harmony).

In the previous study [10], colour palettes were generated for each of 30 words based on selections made by two groups of participants (from China and United Kingdom). There was a great deal of similarity between the colour palettes produced from the two groups but some important differences were observed. For example, many UK participants selected red to represent the word 'bad' whereas few Chinese participants used red for this word (see Figure 1).

The primary limitation of that study [10] was that palettes were collected for only 30 words. This may be sufficient to demonstrate and even validate the technique but it is far too few words for any practical application of the study. For example, it has been estimated that an average 20-year-old native speaker of American English knows 42,000 lemmas (a lemma is a word such as 'run' which indexes related words such as run, ran, runs and running) and on average learns a new lemma every two days [25]. However, in practical use less than 10,000 lemma may be needed [26].
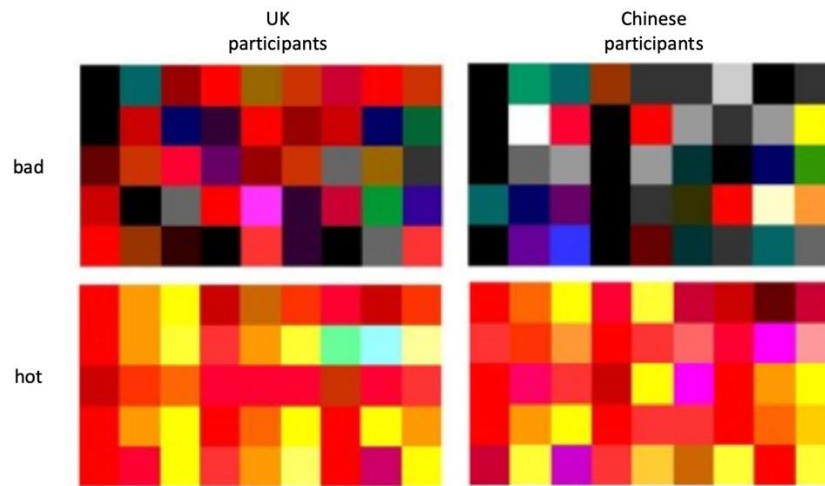
Figure 1: Colour palettes selected by participants in a study by Chen et al. *[10] in which each participant selected three colours to represent a word. Note that Chinese participants used less red to represent the word bad but that UK and Chinese selections were very similar for the word hot.*

Nevertheless, even obtaining palettes for a few thousand words seems challenging and might reasonably require 1440 hours of laboratory time (0.016 hours × 3000 words × 30 participants) or 40 weeks. A radical approach to automatically generate a colour palette for a word was published that used an internet search to collect images that represent words and then used unsupervised machine learning to cluster the colour data from those images [24]. However, although systems that use a machine-learning approach are being developed [27] the relationship between the colour palettes selected by this machine-learning approach and those selected by human participants has yet to be determined. An alternative approach is to use crowd sourcing to collect data over the internet. In one study, participants were asked to name colours that were presented to them over the internet on a browser [28]; although colour names were collected for many colours, each participant provided names for only a few colours so that the experimented was distributed over a large number of participants remotely. A similar colour naming experiment has been developed by Griffin and Mylonas [29].

The purpose of this study is to adopt an online crowd-sourcing approach to word-colour associations and to compare the colour palettes with those that are generated with more traditional laboratory experiments. If crowd-sourcing can be used to obtain reliable estimates of colour palettes that are associated with words then it provides a plausible route to a system that might be able to determine these associations for a large enough number of words to make the system useful in, for example, a design context.

Experimental

An online experiment was constructed using HTML and javascript in which participants were presented with a word on the screen and were asked to select a colour that they associated with this word. A total 30 participants were used in the experiment (see Table 1) and these were the same words that were used in the laboratory study [10] so that a comparison could easily be made. Participants were recruited by email. In the previous laboratory study the experiment was also conducted with samples shown on a screen but the difference was that the screen was calibrated, and the viewing conditions

were highly controlled. This work therefore compares a highly controlled laboratory study with a less controlled study carried out over the internet.

Only UK participants were solicited and the email asked them to undertake the study only if they were from the UK. However, it is impossible with a crowd-sourcing study such as this to exclude the possibility that a small number of participants might have been from a different background.

Active	Dangerous	Good	Married	Poor	Sweet
Bad	Dead	Healthy	Medical	Powerful	Traditional
Clean	Female	Hot	Modern	Religious	Unlucky
Cold	Fresh	Lucky	Natural	Rich	Urban
Cultural	Future	Male	Old	Safe	Young

Table 1: The list of 30 words that were used in the study.

Participants who visited the website were asked to select a colour that was associated with a word (shown at random from the list in Table 1) using a colour picker [30]. The colour picker allowed participants to select a colour using an interface based on a 2-D array of hue and chroma with a lightness slider as shown in Figure 2.
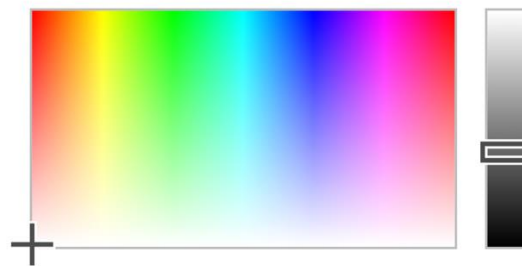

Figure 2: Colour picker [30] used by participants in the study to select colours for each word.

After this, a different word was randomly selected from the list and the procedure repeated. If the same word appeared twice for the same participant they were free to select the same colour as before or a different colour. Participants continued until they decided to leave the page (on average participants each provided one or more colours for about 10 words).

To evaluate similarity between palettes, a method for quantifying the visual difference ΔE_P between two palettes is available [31-32] and was adopted. The algorithm to calculate ΔE_P is according to the following 5 steps given N colours in each palette.

1. For each colour in one palette, the colour differences between this colour and each of the colours in the second palette are calculated. The minimum of these *N* colour differences is recorded.
2. Step 1 is repeated for all the colours in the first palette, for each finding their closest corresponding colours in the second palette, resulting in *N* colour differences.
3. The *N* minimum colour difference values are averaged and the mean value symbolised as m_1.
4. Steps 1-3 are repeated, but this time for each of the colours in the second palette. In other words, for each of these colours the closest corresponding colour in the first palette is found. The mean value of these *N* colour differences is symbolised as m_2.
5. The values of m_1 and m_2 are averaged to obtain the visual colour difference ΔE_P between the two palettes.

Smaller values of ΔE_P are associated with greater similarity between two palettes. The colour differences can be calculated using any standard colour difference equation. In the original algorithm CIELAB colour difference was used [31]; however, in a later study it was found that using the CIEDE2000(2,1,1) equation gave better agreement with the psychophysical data [32]. Therefore in this study the CIEDE2000(2,1,1) equation is used. Note, however, that the algorithm was originally intended for use with pairs of colour palettes that contained the same number of colours. In this study the number of colours in the laboratory experiment is usually not exactly the same as the number in the crowd-sourcing experiment (though these numbers are similar). However, modification of the algorithm to adapt for this situation is trivial.

Results

About 150 participants took part in the experiment. A total number of 1462 responses were obtained and, on average, 48 colours were obtained per word. This number is comparable with the 45-colour palettes obtained in the laboratory experiment [10]. Note, however, that because of the random presentation order of the words in this experiment, and the fact that participants could complete as many words as they wanted to, the size of colour palettes obtained varied from a 32-colour palette (for the word young) to a 69-colour palette (for the word healthy). Table 1 shows the size of the colour palettes that were obtained for each of the words from the crowd-sourcing study. Figure 3 shows representative colour palettes for 6 of the 30 words used in the study.
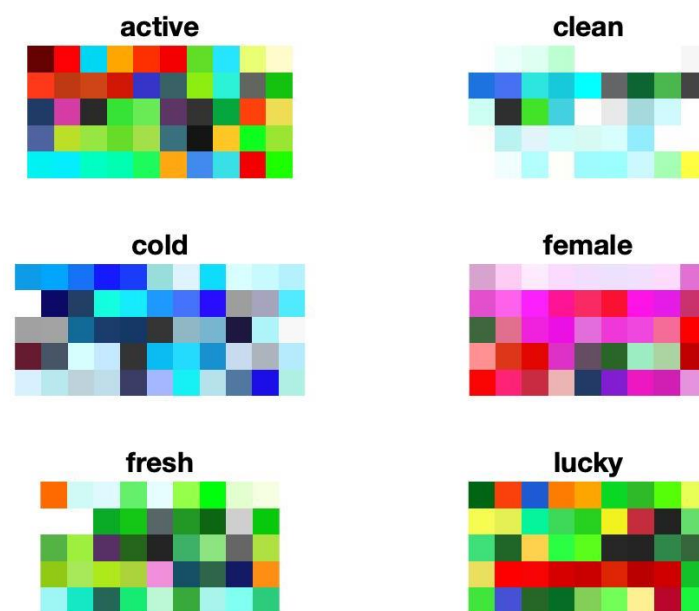

Figure 3: Colour palettes for 6 of the 30 words.

The similarity between the colour palettes generated from the crowd-source experiment and those generated from the previously published laboratory experiment was obtained by calculating the ΔE_P values. The mean value was 6.7 CIEDE2000 units (minimum 1.2; maximum 9.4, standard deviation 1.8). Table 2 shows the ΔE_P score (between the crowd-sourced and laboratory study) for each of the words in the study. Figure 4 shows the two words with the largest differences ('religious' = 9.0 and 'rich'

= 9.1) and the two words with smaller differences ('dead' = 1.2 and 'cold' = 5.0) according to the ΔE_P metric.

Figure 4: Colour palettes for two similar palettes (first row: dead; second row: cold) and two more different palettes (third row: religious; fourth row: rich) according to the ΔE_P metric. The laboratory palettes are on the right and the crowd-sourced palettes are on the left.

More research is needed with the ΔE_P metric. Psychophysical data have been used to validate it in the sense that the larger the ΔE_P the greater the difference between the palettes (and from Figure 4 it is evident that the two palettes with the lowest ΔE_P scores are visually similar). However, what is less clear is how small the ΔE_P value needs to be before we can say that two palettes are similar or indistinguishable. Threshold studies will be needed to determine the critical value in the same way that these studies were used to determined pass/fail values for conventional colour-difference equations. However, experience with other studies suggests that a ΔE_P value of 6 is relatively small. For example, in one study where participants were asked to select a small colour palette to represent fashion runway images, the average ΔE_P score between the palettes generated by the participants (in other words, the inter-observer error) was about 9 and the average ΔE_P score between palettes generated automatically using supervised machine learning and the palettes generated by the participants was about 13 [33]. The ΔE_P between the Chinese palette and the UK palette for hot (see bottom row of Figure 1) was 5.7 [10].

It is evident from Figure 4 that there is quite close agreement in the colours for the words dead and cold between the laboratory and crowd-sourced experiments and less agreement for the words traditional and safe.

A Monte-Carlo analysis was carried out to obtain more insight on whether the ΔE_P scores in Table 2 are smaller than we would expect if there was no relationship between the palettes from the crowd-sourced experiment and the laboratory experiment. To achieve this, for each word, a palette was randomly generated; the CIELAB coordinates for each patch were randomly selected from the ranges, L^* (0 - 100), a^* (-50 - 50) and b^* (-50 – 50).

	Monte-Carlo			
	Number	**DEP**	**Mean**	**Min**
Active	50	5.5	14.3	12.2
Bad	47	5.8	12.9	10.6
Clean	45	5.4	13.8	11.2
Cold	55	5.0	13.8	11.4
Cultural	51	8.8	12.5	10.9
Dangerous	46	4.7	14.8	12.1
Dead	39	1.2	13.7	9.7
Female	50	6.4	13.7	12.0
Fresh	45	9.4	13.2	11.0
Future	42	7.2	11.7	7.1
Good	42	6.4	13.3	10.5
Healthy	69	6.8	14.0	11.9
Hot	43	7.5	18.0	15.1
Lucky	50	7.1	14.7	12.4
Male	55	8.6	12.5	10.9
Married	57	7.5	13.0	11.5
Medical	59	5.6	12.9	10.7
Modern	61	8.0	11.6	10.1
Natural	53	8.1	11.1	9.2
Old	46	7.7	10.4	8.7
Poor	42	7.5	11.2	8.9
Powerful	57	4.6	13.5	11.9
Religious	41	9.0	12.8	11.1
Rich	34	9.1	12.5	7.5
Safe	49	8.8	12.5	11.2
Sweet	49	6.6	12.6	10.8
Traditional	43	7.8	11.4	9.7
Unlucky	61	6.3	11.7	10.5
Urban	59	6.1	11.8	10.2
Young	32	4.7	13.0	8.4

Table 2: The list of 30 words that were used in the study.

Figure 5 shows 10 palettes than have been randomly selected in this way. The ΔE_P between the randomly generated palette and the laboratory palette was calculated. This was repeated 100 times for each word, each times generating a new random palette with the same number of patches as in the crowd sourced experiment (Figure 5 shows 10 of the patches generated for the word, 'dead'). The rightmost two columns in Table 2 show the average ΔE_P between the laboratory experiment and the randomly generated palettes and the minimum ΔE_P. If the ΔE_P from the crowd sourced experiment is lower than the ΔE_P from the randomly generated palettes in > 95% of case then we can be 95% certain that this level of ΔE_P would not occur by chance. For almost all the words, the ΔE_P score from the crowd sourced experiment was lower than in every one of the 100 simulations for that word; this indicates than the low levels of ΔE_P in the crowd sourced experiment are unlikely to have occurred by chance. Note that the minimum ΔE_P score from the random palettes was greater than the crowd sourced score in all cases apart from the words 'future' (where the random palette was closer to the laboratory

experiment in 1% of cases) and 'rich' (where the random palette was closer to the laboratory experiment in 2% of cases).
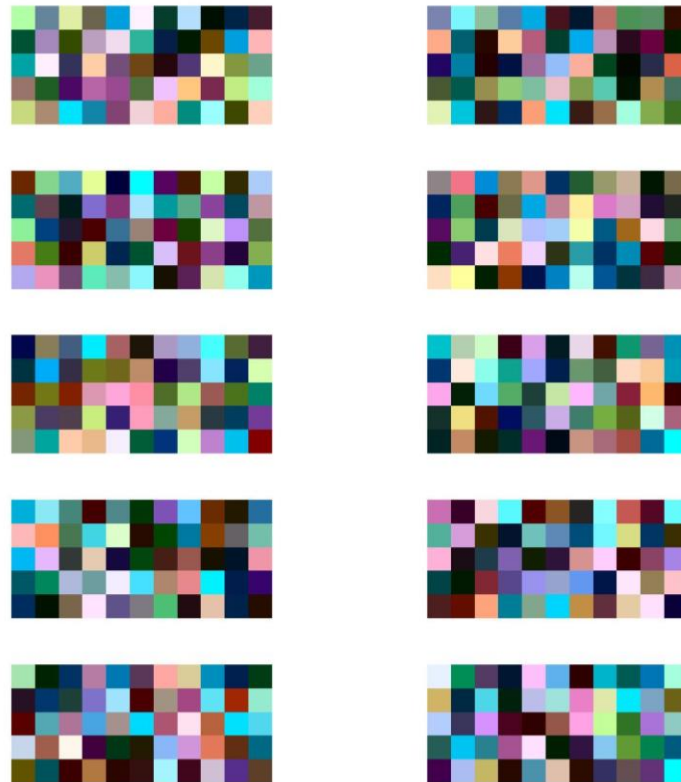

Figure 5: A set of 10 randomly generated 55-colour palettes from the Monte-Carlo simulation.

Discussion

This work is concerned with collecting data on the associations that exist between colours and words. Over at least the last 30 years a number of studies have reported evidence that there are universal word-colour associations. Most studies show that there is more similarity between colour meanings than differences between cultural groups [20, 34]. Nevertheless, when cultural differences do exist they can be significant and it is recommended that care needs to be taken before assuming that a word-colour association found in one cultural group applies to another group. It is also evident that colour meanings can vary greatly with context [35]. For example, the colour red can be associated with many words and concepts. Red can be associated with blood (in the context of a poster about a vampire movie), with danger (in the context of an emergency top button), with strawberries (in the context of a yoghurt pot) and many others. It is no surprise, therefore, that in the colour palettes generated in this study and in the earlier laboratory study [10] it is clear that word-colour associations are far from being one to one relationships. The same colour can appear in palettes that represent many different words and concepts. A tool that could provide users with colour palettes for a target word would be useful even if only as inspiration for further thinking. However, laboratory studies are poorly placed to provide sufficient data to enable such a tool to be constructed. This study has explored the use of crowd-sourcing as a way to generate large amounts of data on word-colour associations which might enable data- or consumer-led design tools. The work in this paper suggests that crowd sourcing can deliver colour palettes that are similar to those found in more rigorous studies but clearly more work (including work that uses more

words and more participants) is required. Although we used a simulation to provide evidence that the palettes from the two experiments were similar, we accept that the evidence is not as strong as it could be since concluding that the two palettes are more similar than randomly generated palettes is a weak test. What is really needed is the ability to be able to use a threshold value of the ΔE_P metric for calculating the visual difference between palettes. Our work has therefore highlighted the need for additional studies in this area.

References

1. Berlin B and Kay P (1999), *Basic Color Terms: Their Universality and Evolution*, CSLI Publications.

2. Wierzbicka A (2006), The semantics of colour: A new paradigm, in *Progress in Colour Studies: Volume I. Language and Culture*, Pitchford NJ and Biggam CP (eds.), 1-24, Philadelphia: John Benjamins Publishing Company.

3. Sivik L and Taft C (1994), Color naming: A mapping in the NCS of common color terms, *Scandinavian Journal of Psychology*, **35**, 144-164.

4. Aaronson BS (1970), Some affective stereotypes of color, *International Journal of Symbology*, **2** (1), 15-27.

5. Madden TJ, Hewett K and Roth MS (2000), Managing images in different cultures: A cross-national study of color meanings and preferences, *Journal of International Marketing*, **8** (4), 90-107.

6. Williams JE and Foley Jr JW (1968), Connotative meanings of color names and color hues, *Perceptual and Motor Skills*, **26** (2), 499-502.

7. Xin JH, Cheng KM, Chong TF, Sato T, Nakamura T, Kajiwara K and Hoshino H (1998), Quantifying colour emotion-what has been achieved, *Research Journal of Textile and Apparel*, **2** (1), 46–54.

8. Ou L, Luo MR, Woodcock A and Wright A (2004), A study of colour emotion and colour preference. I. Colour emotions for single colours, *Color Research and Application*, **29** (3), 232–240.

9. Caivano JL (1998). Color and semiotics: A two-way street, *Color Research and Application*, **23** (6), 390-401.

10. Chen Y, Yang J, Pan Q, Vazirian M and Westland S (2019), A method of exploring word-colour associations, *Color Research and Application*, **45** (1), 85-94.

11. Amsteus M, Al-Shaaban S, Wallin E and Sjöqvist S (2015), Colors in marketing: A study of color associations and context (in) dependence, *International Journal of Business and Social Science*, **6** (3), 32-45.

12. Bellizzi JA, Crowley AE, Hasty RW (1983), The effects of color in store design, *Journal of Retailing*, **59** (1), 21-45.

13. Ares G and Deliza R (2010), Studying the influence of package shape and colour on consumer expectations of milk desserts using word association and conjoint analysis, *Food Quality and Preference*, **21** (8), 930-937.

14. Gage J (1999), *Colour and Meaning: Art, Science and Symbolism*. London: Thames and Hudson.

15. Jiang L, Cheung V, Westland S, Rhodes P, Shen L and Xu L (2020), The impact of color preference on adolescent children's choice of furniture, *Color Research and Application*, **45** (4), 754-767.

16. Grieve KW (1991), Traditional beliefs and colour perception, *Perceptual and Motor Skills*, **72** (3), 1319-1323.

17. Jacobs L, Keown C, Worthley R and Ghymn KI (1991), Cross-cultural colour comparisons: global marketers beware!. *International Marketing Review*, **8** (3), 21-30.

18. Aslam MM (2006), Are you selling the right colour? A cross - cultural review of colour as a marketing cue, *Journal of Marketing Communications*, **12** (1), 15-30.

19. Kauppinen-Räisänen H and Luomala HT (2010), Exploring consumers' product-specific colour meanings, *Qualitative Market Research: An International Journal*, **13** (3), 287-308

20. Gao X-P, Xin JH, Sato T, Hansuebsai A, Scalzo M, Kajiwara K, Guan S-S, Valldeperas J, Lis MJ and Billger M (2007), Analysis of cross-cultural color emotion, *Color Research and Application*, **32** (3), 223-229.

21. Won S and Westland S (2017), Colour meaning in context, *Color Research and Application*, **42** (4), 450-459.

22. Won S & Westland S (2018), Requirements capture for useful colour information for design professionals, *Color Research and Application*, **43** (3), 387-395.

23. Hu YC, Lee MG and Tsai P (2009), Colour palette generation schemes for colour image quantization, *The Imaging Science Journal*, **57** (1), 46-59.

24. Lindner A and Süsstrunk S (2013), Automatic color palette creation from words, *Proceedings of the Twenty-First Color and Imaging Conference*, 69-74, Albuquerque (USA).

25. Brysbaert M, Stevens M, Mandera P and Keuleers E (2016), How many words do we know? Practical estimates of vocabulary size dependent on word definition, the degree of language input and the participant's age, *Frontiers in Psychology*, **7**, 1116, 1-11.

26. Nation I (2006), How large a vocabulary is needed for reading and listening?, *Canadian Modern Language Review*, **63** (1), 59-82.

27. ColourMyIdeas (2020), https://www.colormyideas.com/ – last accessed 2 July 2020.

28. Moroney N (2003), Unconstrained web-based color naming experiment, *Proceedings of the SPIE's Electronic Imaging Conference: Color Imaging VIII: Processing, Hardcopy, and Applications*, **5008**, 36-46, Santa Clara (USA).

29. Griffin LD and Mylonas D (2019), Categorical colour geometry, *PLoS ONE*, **14** (5), e0216296.

30. JScolor (2020), https://jscolor.com/ – last accessed 2 July 2020.

31. Pan QQ and Westland S (2018), Comparative evaluation of color differences between color palettes, *Proceedings of the Twenty-Sixth Color and Imaging Conference*, 110-115, Vancouver (Canada).

32. Yang J, Chen Y, Xiao K and Westland S (2020), Predicting visual similarity between colour palettes, *Color Research and Application*, **45** (3), 401-408.

33. Lai P and Westland S (2020), Machine learning for colour palette extraction from fashion runway images, *International Journal of Fashion Design, Technology and Education*, in press.

34. Lucassen MP, Gevers T and Gijsenij A (2010), Texture affects color emotion, *Color Research and Application*, **36** (6), 426-436.

35. Taft C (1997), Color meaning and context: Comparisons of semantic ratings of colors on samples and objects, *Color Research and Application*, **22** (1), 40-50.