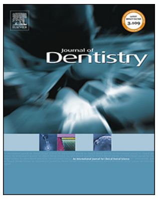
Journal of Dentistry 76 (2018) 132–136

Contents lists available at ScienceDirect

Journal of Dentistry

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



Evaluation of colorimetric indices for the assessment of tooth whiteness

Qianqian Pan[a], Stephen Westland[a,*], Roger Ellwood[b]

[a] School of Design, University of Leeds, Leeds, LS2 9JT, United Kingdom
[b] Colgate Palmolive Dental Health Unit, Williams House, Manchester Science Park, Manchester, M15 6SE, United Kingdom

ARTICLE INFO

Keywords:
Tooth Whiteness
Whiteness index
Visual assessments
Color space

ABSTRACT

Objectives: To evaluate the performance of existing equations that measure perceptual whiteness of teeth.
Methods: Three new psychophysical experiments were conducted and combined with two previously published experiments to form a large set of data to test performance of whiteness indices. Three whiteness indices (WIC, WIO, WI_D,) were compared with regard to their ability to measure perceived whiteness. Coefficient of determination (r^2) and '% wrong decisions' were used as measures of performance. One of the new experiments involved 500 participants across five different countries to explore the effect of gender, age and culture on whiteness perception.
Results: Equations (WIO and WI_D) that have been optimized for use with tooth whiteness better correlated with visual perceptions of changes in tooth whiteness than the more general CIE whiteness index (WIC). The best performance was given by WIO (in terms of both r^2 and % wrong decisions). No effect of age, gender or culture was found on whiteness perception.
Conclusions: WIO is a robust method for assessing whiteness of human teeth.

1. Introduction

The assessment of tooth whiteness is important in dentistry both in terms of communication of the benefits of products to consumers and the evaluation and comparison of different oral care products in clinical trials. Tooth whitening may be beneficial to patients because it can lead to better oral-care routines and higher self-esteem [1]. The development of equations that can predict perceptual whiteness is important to quantify the performance of whitening treatments and hence to optimize their efficacy.

Traditionally shade guides have been used to visually assess tooth color with the Vitapan Classical shade guide (which consists of 16 shade guide tabs) and the Vita Bleachedguide 3D-Master (which has 15 tabs that consist of the odd numbers, 1, 3, 5, etc., from a 29-point scale) being used extensively (VITA Zahnfabrik, Bad Sackingen, Germany). When using instrumental methods of assessing tooth color such as spectrophotometers, colorimeters or cameras, color is typically communicated using color systems that use three numbers for the complete identification of a color. There are many such color space systems available such as the CIE XYZ tristimulus values or the CIELAB system [2].

It is complex to relate three-dimensional changes in CIE XYZ or CIELAB values to changes in perceptual whiteness or yellowness, particularly with respect to the weighting of relative changes for the individual components. Therefore industrial applications of color science for materials such as paper and paint have traditionally assessed whiteness using a univariant metric known as a whiteness index [3]. A number of such whiteness indices have been developed for various industries, most notably for paper and textiles, and the CIE whiteness index WIC [4,5] is widely used. Thus

$$\mathrm{WIC} = Y + 800(x_n - x) + 1700(y_n - y) \tag{1}$$

where Y, x and y are the colorimetric properties (luminance and chromaticity values) of the sample to be assessed and x_n, y_n are the chromaticity values of the reference white (usually the light source used to view the samples). The CIE whiteness formula was modified and optimized for use with dentistry and this modified form is known as the WIO whiteness formula [6,7]. Thus

$$\mathrm{WIO} = Y + 1075.012(x_n - x) + 145.516(y_n - y) \tag{2}$$

The WIO formula was developed to predict perceptual whiteness of teeth under daylight (specifically the D65 illuminant) and therefore the values of x_n and y_n are 0.3138 and 0.3310 respectively. The WIO equation has been shown to be effective [8] and has been used in a number of clinical studies [8,9]. However, one limitation, as with other whiteness indices, is that it requires the user to understand the XYZ

[*] Corresponding author.
E-mail addresses: q.pan@leeds.ac.uk (Q. Pan), s.westland@leeds.ac.uk (S. Westland), roger.ellwood@manchester.ac.uk (R. Ellwood).

https://doi.org/10.1016/j.jdent.2018.07.004
Received 15 February 2018; Received in revised form 11 June 2018; Accepted 6 July 2018

color space and the associated chromaticity values xy when in dentistry the most widely used color space is CIELAB [10]. Recently, a new equation has been developed for use in dentistry that is based upon the CIELAB color space and is referred to as WI_D [8]. Thus,

$$WI_D = 0.511L^* - 2.24a^* - 1.100b^* \quad (3)$$

It is interesting to note that this equation weights changes of L* as being less significant in terms of whiteness than changes in b* (yellow-blue) which in turn is weighted less than a* (green- red). This seems slightly counter-intuitive since increases in whiteness are often particularly associated with changes in L* and b* [11].

The performance of the WI_D equation was compared with that of several other whiteness indices (including WIO and WIC) using data from four psychophysical experiments. The WI_D equation was shown to perform better than any other previously published whiteness equation based on CIELAB and performance was comparable to WIO (in one of the psychophysical experiments WI_D performed slightly better and in another WIO performed slightly better) [8].

It is far from clear, however, just how universal the construct of perceived whiteness of teeth is. For example, would observers in China rank a set of teeth samples for whiteness in the same order as observers in UK? In addition it is not known whether males and females share identical concepts of tooth whiteness. For this reason a new large-scale experiment was conducted in which 500 panelists (100 from each of 5 geographical regions) each ranked shade guide tabs in terms of whiteness. Each group of 100 subjects were balanced so that each included 50 male and 50 female and 50 younger adults (aged 18–30) and 50 older (aged > 30). The purpose of this experiment was to assess the validity of WIO and WI_D using a much larger set of data than has previously been used and to also explore whether a single equation is suitable for people of different age and from different cultures.

One limitation of most of the samples (usually shade guide tabs) that have been used in previous studies of this nature is that there are correlations between the colorimetric values. For example, generally we find that L* decreases as b* increases; and a* decreases as b* decreases. This is because the samples represent stages along the locus in color space that is naturally described by bleaching and ageing (that is the locus of tooth color). An example is shown in Fig. 1. The problem with correlations in the data set is that this allows multiple equations to fit the data with similar performance. However, these equations may perform very differently when presented with data that are not correlated. If the only processes that were used to change tooth color were bleaching and the natural ageing process then these correlations would probably always exist. However, recent years have seen the introduction of novel methods of achieving tooth whitening where these correlations may not exist (or where a different set of correlations may exist) and it is not clear whether the previously published equations will perform well in such cases. Therefore an additional set of data was generated in this study using stimuli that were not correlated in color space. Because of the difficulty in creating such samples as physical shade guide tabs, the stimuli for this experiment were digitally simulated on a color-calibrated display.

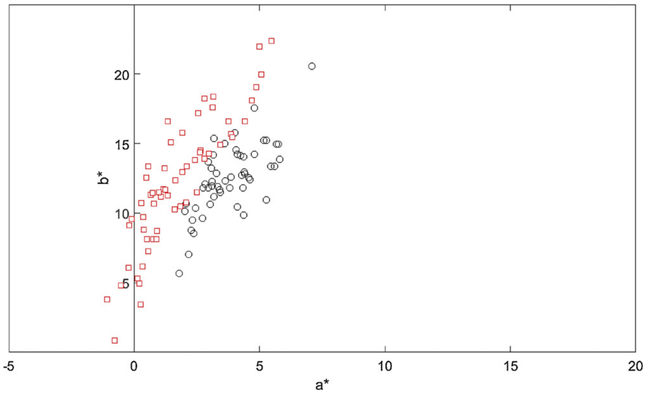


Fig. 1. CIE a*–b* values of 29 Vita 3D Master Shade Guide and 29 Vita Extended Bleachedguide tabs used in PE3 (red squares) and custom-made disks used in PE4 (black circles) (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article).

Table 1
Summary of experimental details of the five data sets.

	N of observers	N of samples	Nature of the samples
PE1	9	26	Vita 3D Master shade guide
PE2	9	16	Vitapan Classical shade guide
PE3	500	58	Vita 3D-Master shade guide (29 tabs) Vita 3D Master Extended Bleachedguide (29 custom made tabs)
PE4	80	52	Custom-made ceramic disks (see Fig. 2)
PE5	53	45	Digital simulations on screen

2. Methodology

2.1. Psychophysical experiments

The data from five separate psychophysical experiments (PE1 – PE5) were used in this study. Two of these experiments (PE1 and PE2) were carried out by Luo et al. and have previously been published [7]. The other three experiments were carried out for this study; PE3 and PE4 were carried out with shade guide tabs (or ceramic disks) and PE5 was carried out using digital simulations of tabs presented on a computer display. The materials and methods for the five experiments are summarized in Table 1.

In PE1-PE4, participants were asked to rank the samples in order of perceptual whiteness when viewed under D65 daylight in a viewing cabinet. In PE5 a paired-comparison method was used where participants were asked to select which of a pair of digitally simulated samples was whiter when viewed on a color-calibrated display. The methods to calculate interval scale perceptual whiteness values from paired comparison or from ranking are widely published and understood [12–14]. For all the data sets (PE1-PE5) the raw data were used to calculate interval scale values that represent the relative perceptual whiteness of the samples involved. Two methods (r^2 and %WD) were used to quantify the agreement between these perceptual whiteness values and the values of the whiteness indices that were being studied.

2.2. PE1 and PE2

In PE1 and PE2 9 observers ranked the 26 Vita 3D Master shade guide tabs and the 16 Vitapan Classical shade guide tabs respectively. Observations were carried out in a viewing cabinet with D65 illumination and a grey interior. Although the number of participants was small, these data are included in this study because they are the original data on which the WIO formula was developed (the equation was originally validated using a separate experiment with 88 observers). These data were previously published [7] but were available for use in this study.

2.3. PE3

PE3 was carried out at 5 distinct geographical locations (UK, India, Brazil, USA and China). In each location 100 participants were recruited in a balanced design to allow the effect of age, gender and culture to be assessed (25 young males, 25 young females, 25 old males and 25 old females). The young group was aged 18–30 and the old group were aged 30-60. Each participant was asked to rank each of 58 shade guide tabs in order of whiteness in a viewing cabinet under D65 illumination. The set of shade guide tabs consisted of 29 Vita Toothguide 3D-Master tabs and 29 custom-made Vita Extended Bleachedguide 3D-Master tabs. The CIELAB a*-b* values of these tabs are illustrated in Fig. 1. The L* ranged from 56.5 to 77.8. A neutral grey card was provided for each geographical center to cover the interior base of the viewing cabinet so that the same background would be used in each study. The spectral reflectance factors of each tab were measured using a Konica-Minolta CM-2600d reflectance spectrophotometer and subsequently converted to CIE XYZ values for D65 illuminant (1964 CIE observer).

2.4. PE4

PE4 was carried out at two geographical locations (UK and China). In each location 40 participants were recruited (an equal number of males and females). Each participant was asked to rank each of 52 custom-made circular shade guide tabs in order of whiteness in a viewing cabinet under D65 illumination. A neutral grey card was provided for each center for the base of the viewing cabinet so that the same background would be used in each study. The samples were custom-made disks of 7.5 mm diameter with an overall thickness of 3 mm (enamel thickness 0.7 mm, dentine thickness 1.7 mm, base 0.6 mm) designed to cover the gamut of tooth color. They were fabricated out of dental porcelain (Vita VMK Master, Vita Zahnfabrik, Bad Sackingen, Germany).

The samples were custom-made disks with an overall thickness of 3 mm (enamel thickness 0.7 mm, dentine thickness 1.7 mm, base 0.6 mm) designed to cover the gamut of tooth color. The CIELAB a*-b* values of these disks are illustrated in Fig. 1. Fig. 2 shows a selection of the disks that were used (PE4). The L* ranged from 60.0 to 74.8. The spectral reflectance factors of each disk were measured using a Konica-Minolta CM-2600d reflectance spectrophotometer and subsequently converted to CIE XYZ values for D65 illuminant (1964 CIE observer).

2.5. PE5

The data for PE5 were grouped into 5 sets of 9 stimuli. Each set of 9 stimuli were derived from a color center. The target CIELAB values for the five color centers are displayed in Table 2.

For each color center 8 additional samples were generated using ΔL*, Δa* and Δb* values of ± 2.

As shown in Fig. 3, around each color center there are four samples that are 2 CIELAB units lighter and four that are 2 CIELAB units darker. There are also four that have a Δa* of 2 and four that have a Δa* of -2 compared to the center (similarly for b*). Each set of 9 samples are completely decorrelated with each other in color space. However, if we consider all 45 samples together they would show correlations because of the fact that there are correlations in Table 2 (which are necessary to ensure that the samples fall in the gamut of tooth color). For this reason, each participant viewed each of the five data sets independently. A two-alternative forced-choice paired comparison method was used. So for each set of 9 samples, each participant would be presented with 36 paired comparisons in random order. However, samples in one set would never be compared with those from another set. In total 53 participants took part and made a total of 9540 paired comparisons (53 participants × 36 comparisons × 5 sets). Data set PE5 therefore needs to be treated as five separate data sets (PE5a, PE5b, PE5c, PE5d and PE5e); interval scale whiteness values were calculated for the 9 samples in each set separately.

Table 2
: Target CIELAB values for the five color centers for PE5.

	L*	a*	b*
C1	78.9	−0.8	0.8
C2	70.9	0.2	4.0
C3	66.4	0.8	8.1
C4	64.8	1.2	13.2
C5	61.0	2.9	17.4

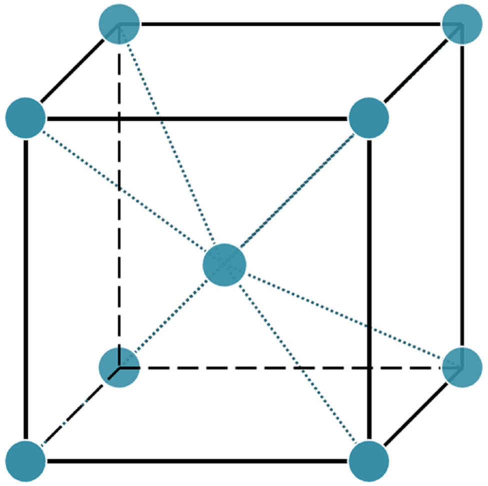


Fig. 3. Distribution of 9 samples for one color center in CIELAB space.

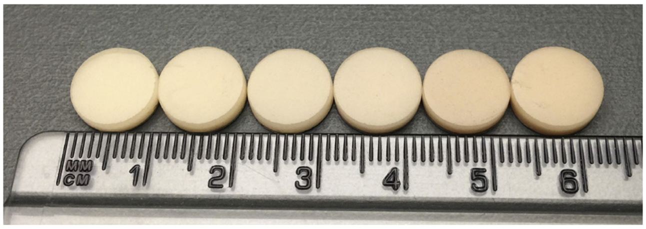


Fig. 2. A selection of the 52 custom-made circular disks used in PE4.

2.6. Assessment of whiteness metrics

For all five psychophysical experiments the rank data (or, in one case, the paired comparison data) were converted to interval scale values for the perception of whiteness using standard methods [12]. It is important to recognize that the perceptual scale data that are derived are interval data and not ratio data. The significance of this is that they cannot be pooled together into a single set of data; rather each set of data needs to be considered individually (we cannot even pool the five separate experiments that constitute PE5). However, we use methods to compare the outputs of whiteness indices with the visual data. Two such methods are used (correlation r^2 and % wrong-decisions) and these have been used in other studies [7,8].

The r^2 value between two sets of data is the coefficient of determination and this is the square of the Pearson correlation coefficient r. Values closer to one indicate a high correlation between the two sets of data. The % wrong-decisions criterion is obtaining by comparing each sample in a data set with each other and calculating the number of times that the whiteness metric would disagree about which one of a pair the whitest compared with the average visual decision of the whole group of observers (denoted by the visual scale values). A whiteness metric agrees with the visual data if the % wrong decisions value is low.

3. Results

3.1. Performance of Whiteness metrics

Table 3 shows the coefficient of determination r^2 between the values of the whiteness indices and the perceptual whiteness scale values for each of the data sets.

For data sets PE1 – PE4 the two equations that have been optimized

Table 3
Coefficient of determination for each of the whiteness metrics and data sets.

	PE3	PE4	PE1	PE2	PE5				
					C1	C2	C3	C4	C5
WIO	0.97	0.87	0.93	0.85	0.69	0.72	0.78	0.79	0.73
WI$_D$	0.96	0.88	0.91	0.83	0.25	0.45	0.60	0.51	0.49
WIC	0.96	0.80	0.87	0.79	0.43	0.39	0.38	0.46	0.36

Table 4
% wrong decisions for each of the whiteness metrics and data sets.

	PE3	PE4	PE1	PE2	PE5				
					C1	C2	C3	C4	C5
WIO	4.8	8.5	7.4	10.8	11.1	13.9	19.4	13.9	19.4
WI$_D$	5.4	11.9	8.0	10.0	30.6	27.8	22.2	25.0	25.0
WIC	5.6	14.2	12.9	12.5	19.4	30.6	27.8	19.4	36.1

for tooth whiteness (WIO and WI$_D$) perform better than the general CIE whiteness equation with the best performance being found for the WIO equation. The difference between the three equations is much better for data set PE5 where WIO > > WI$_D$ > > WIC. Recall that the difference between PE1 – PE4 and PE5 is that the samples in PE5 were designed so that changes in the various dimensions in color space were not correlated.

Table 4 shows the % wrong decisions made by the whiteness indices for each of the data sets.

The criterion exhibited in Table 4 shows a similar trend to that exhibited in Table 3. The lowest % wrong decisions are found for WIO and the highest % wrong decisions are generally found for WIC with the largest differences being seen for the PE5 data set.

3.2. *Effect of culture and gender*

For experiment PE3 the data were pooled across culture and gender in the previous analysis. However, in this section separate whiteness scale values are calculated from the data collected in the five geographical locations, UK, USA, India, China and Brasil. Fig. 4 shows a typical correlation between the scale values calculated from the 100 participants in China and the 100 participants in UK. The r^2 value between the Chinese and UK whiteness scale values was 0.9912. Table 5 shows the correlations between all of the five regions.

For the various subgroups of participants within PE3 the % wrong decisions were calculated (by calculating the number of wrong decisions that each observer makes when compared with the average performance of each subgroup). These are shown in Table 6.

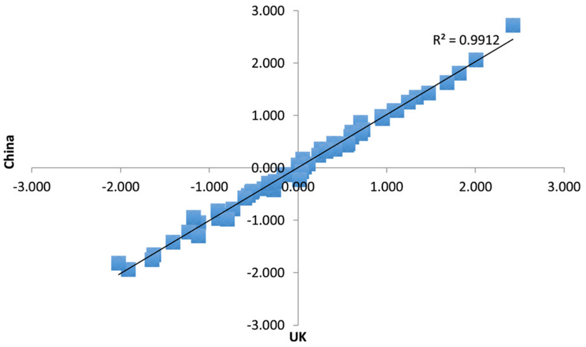


Fig. 4. Correlation between the perceptual whiteness scale values calculated from the 58 samples in PE3 between the UK and Chinese participant sets.

Table 5
Coefficient of determination between the whiteness values calculated in PE3 between the different regions.

	UK	China	USA	Brazil	India
UK		0.9912	0.9918	0.9913	0.9967
China	X		0.9877	0.9939	0.9926
USA	X	X		0.9917	0.9918
Brazil	X	X	X		0.9917
India	X	X	X	X	

Table 6
% wrong decisions for the subgroups of participants (PE3 data).

Subgroup	% wrong decision
All pooled (N = 500)	5.3
UK (N = 100)	5.6
China (N = 100)	4.8
Brazil (N = 100)	4.8
USA (N = 100)	4.9
India (N = 100)	5.6
Females (N = 250)	5.1
Males (N = 250)	5.4
Young (N = 250)	5.1
Old (N = 250)	5.4

4. Discussion

Overall, it is clear that the optimized whiteness equations perform better than the older CIE whiteness equation. If we consider the ability of the two optimized equations to predict perceptual whiteness when observers rank the commercially available shade guide tabs, the performance of the two equations is rather similar but with some evidence that WIO performs the best. There is evidence that WIO performs substantially better than WI$_D$ for the PE5 data set and therefore we can conclude that WIO may be more robust as a predictor of changes in tooth whiteness.

Both optimized dental equations (WIO and WI$_D$) performed better than the CIE whiteness index (WIC). The best performance was given by WIO (in terms of both r^2 and % wrong decisions) and this advantage was most marked when uncorrelated changes in the three dimensions of color space occur (data set PE5). No effect of age, gender or culture was found on whiteness perception. WIO is a robust method for assessing whiteness of human teeth. It should be noted that the data sets PE1 and PE2 were derived more than a decade ago. Although the current study finds no effect of geographical location on the whiteness judgements whether such judgements change over time remains an open question.

5. Conclusion

This study evaluated the performance of existing equations that measure perceptual whiteness of teeth. Three new psychophysical experiments were conducted and two previously published studies were used to evaluate three whiteness indices (WIC, WIO, WI$_{D,}$) in terms of their ability to measure perceived whiteness. No effect of age, gender or culture was found on whiteness judgements and best performance was given by WIO. WIO is a robust method for assessing whiteness of human teeth.

Conflict of interest statement

Part of this work was carried through financial support from the Colgate Palmolive Company. The work also used previously published data that was also carried out with the financial support of the Colgate Palmolive Company.

Acknowledgements

This project was possible because of financial support received from Colgate-Palmolive Company.

References

[1] P.W. Kihn, Vital tooth whitening, Dent. Clin. North Am. 51 (2) (2007) 319–331.

[2] W.S. Wyszecki, G. Stiles, Color Science: Concepts and Methods, Quantitative Data and Formulae, Wiley-Interscience, 2000.

[3] S. Westland, CIE Whiteness, in: M.R. Luo (Ed.), Encyclopedia of Color Science and Technology, Springer, 2015, pp. 1–5, , https://doi.org/10.1007/978-3-642-27851-8_5-1.

[4] E. Ganz, Whiteness formulas: a selection, Appl. Opt. 18 (7) (1979) 1073–1078.

[5] ASTM, Designation E313-73: Standard Test Method for Indices of Whiteness and Yellowness of Near-White, Opaque Materials, (1993).

[6] W. Luo, S. Westland, P. Brunton, R. Ellwood, I.A. Pretty, N. Mohan, Comparison of the ability of different colour indices to assess changes in tooth whiteness, J. Dent. 35 (2) (2007) 109–116.

[7] W. Luo, S. Westland, R. Ellwood, I. Pretty, V. Cheung, Development of a whiteness index for dentistry, J Dent. 37 (Suppl 1) (2009) e21–e26, https://doi.org/10.1016/j.jdent.2009.05.011.

[8] M. del Mar Pérez, R. Ghinea, M.J. Rivas, A. Yebra, A.M. Ionescu, R.D. Paravina, L.J. Herrera, Development of a customized whiteness index for dentistry based on CIELAB color space, Dent. Mater. 32 (3) (2016) 461–467.

[9] M. Oliveira, E. Fernández, J. Bortolatto, O. Oliveira Junior, M. Bandeca, S. Khajotia, F. Florez, Optical dental whitening efficacy of Blue covarine toothpaste in teeth stained by different colors, J. Esthet. Restor. Dent. 28 (S1) (2016) S68–S77.

[10] S. Westland, C. Ripamonti, V. Cheung, Computational Colour Science: Using MATLAB, 2nd edition, John Wiley, 2012 ISBN-10: 0470665696.

[11] A. Joiner, I. Hopkinson, Y. Deng, S. Westland, A review of tooth colour and whiteness, J Dent. 36 (Suppl 1) (2008) 2–7, https://doi.org/10.1016/j.jdent.2008.02.001.

[12] S. Westland, Y. Li, V. Cheung, Monte-carlo analysis of incomplete paired-comparison experiments, J. Imaging Sci. Technol. 58 (5) (2014) 050506-1.

[13] W.S. Togerson, Theory and Methods of Scaling, Wiley, New York, 1958.

[14] P.G. Engeldrum, Psychometric Scaling; A Toolkit Imaging Systems Development, Imcotek Press, Winchester, 2000.