©Operative Dentistry, 2009, **34-3,** 337-343

A Comparison of Human Raters and an Intra-oral Spectrophotometer

WD Browning • DC Chan
JS Blalock • MG Brackett

Clinical Relevance

Using the Vita 3D-Master shade guide, the accuracy of three experienced clinicians was compared to shade choices using an intraoral spectrophotometer. Compared to the clinicians, the shades chosen by the Easyshade guide were more frequently an exact match. This study indicates shade matching may be improved by using an electronic device.

SUMMARY

Consistently choosing an accurate shade match is far more difficult than it appears. Recently, several electronic shade-matching devices have been marketed. One device is an intraoral spectrophotometer, Easyshade. The current study compared the accuracy and consistency of the Easyshade (ES) device to three clinicians experienced in tooth whitening trials and trained in the use of the Vitapan 3D Master shade.

The maxillary anteriors of 16 participants were matched on three separate occasions one month apart. At each appointment, the three clinicians (R1, R2 & R3) and ES independently chose a single 3D Master tab. A trained research assistant used the Easyshade device to record CIE L*, C* and H* and a shade tab. In addition, color differences between shade tabs were calculated using the Delta E 2000 (Δ e 00) formula. The CIE L*C*H* data were also used to establish standards for the five lightness groups of the 3D Master. An intra-rater agreement was evaluated using an intra-class correlation statistic, and an inter-rater agreement was evaluated using a weighted Kappa statistic.

The percentages of exact matches were: ES = 41%; R1 = 27%; R2 = 22% and R3 = 17%. Matches within a half-shade were also calculated. This represents a mismatch that is perceptible but acceptable. The percentages of matches within a half-tab were: ES = 91%; R1 = 69%; R2 = 85% and R3 = 79%. In terms of lightness, the intra-rater

*William D Browning, DDS, MS, professor and Indiana Dental Association endowed chair, Restorative Dentistry, Indiana University School of Dentistry, Indianapolis, IN, USA

Daniel C Chan, DDS, MS, associate dean, clinical services, University of Washington, School of Dentistry, Seattle, WA, USA

John S Blalock, DMD, associate professor, Department of Oral Rehabilitation, School of Dentistry, Medical College of Georgia, Augusta, GA, USA

Martha G Brackett, DDS, MSD, assistant professor, Department of Oral Rehabilitation, School of Dentistry, Medical College of Georgia, Augusta, GA, USA

*Reprint request: 1121 W Michigan St, Room S317, Indianapolis, IN 46202, USA; e-mail: wbrownin@iupui.edu

DOI: 10.2341/08-106

agreement was considered to be very good for ES and R2 and good for R1 and R3. For chroma, agreement for ES was considered good, and for the three clinicians, it was considered moderate.

The mean color difference for the L*, C*, H* data recorded at each evaluation was 1.5, or only slightly greater than the color difference between the same tab on different guides (1.2). The $\Delta_{e\,00}$ data were the most accurate data collected, and they were used to establish a standard to which the tab choices of the four raters were compared. A weighted Kappa statistic was performed and, in terms of lightness, agreement was found to be good for all raters. For chroma, agreement was very good for ES and it was good for the clinicians.

In terms of the number of exact matches and matches within a half-shade, the performance of ES was at least comparable to, if not better than, the dentists. Statistically, the same was true in terms of consistency and accuracy when making repeated matches of lightness and chroma using the 3D Master shade guide.

INTRODUCTION

The color and appearance of teeth is a complex phenomenon. Lighting conditions, translucency, opacity, gloss and the human eye influence the overall perception of color.[1] Visual color determination by comparing the teeth to a shade guide has been the most often used method in dentistry. Research has shown that shade guides do not fully represent the color of natural teeth, and the effectiveness of color determination using shade guides is contradictory. Some authors[2] concluded that the human eye is efficient in detecting even small differences, while other authors have commented that the human evaluation of tooth shade is unreliable.[3]

As a result, more technical and objective methods of color measurement have been sought. These methods include tristimulus colorimeters, spectrophotometers and computer analysis of digital images. This terminology is sometimes confusing. Since both colorimeters and spectrophotometers measure color, both could be referred to as colorimeters. A tristimulus colorimeter typically contains filters that are correlated to analyze light from only one of the several standardized light sources acknowledged by the Commission Internationale de l'Eclairage (CIE). These devices only measure red, green and blue points. Spectrophotometers measure the full spectrum. In opposition to tristimulus colorimeters, spectrophotometers can calculate color space coordinates for any of the standard illuminants, rather than just one. All three of the technical methods listed above can provide readings from the CIE L*a*b* color space.

The ultimate objective is to relate the color differences observed to clinical practice, with the human eye being the final arbiter. Accordingly, a color difference detectable by an electronic device, but imperceptible to the human eye, is of no clinical importance. Two thresholds are relevant: The first is perceptability or the minimum color difference perceived by the human eye. The second is acceptability or the greatest color difference that still represents an acceptable match.

A number of computer-based instruments have entered the market. Studies have been conducted that evaluated the reliability of computer-based instruments.[4-6] However, few studies have compared these instruments with human color perception under clinical conditions. The current study compared the intrarater agreement for three human evaluators and an intraoral spectrophotometer using the Vitapan 3D-Master system. In addition, inter-rater agreement was determined.

METHODS AND MATERIALS

General Description

Sixteen participants were informed about the project and gave written consent to participate. The project and the informed consent process were approved by the local Institutional Review Board. Fifteen participants had six suitable teeth and one participant had only five. Thus, 95 teeth were available for shade determination. Each participant was evaluated on three separate occasions one month apart. At each evaluation, three operators independently determined the best shade match. In addition, an intraoral spectrophotometer was used.

Human Evaluators

Three experienced, calibrated raters, R1, R2 & R3, independently matched the shade. These three dentists ranged in age from 38 to 57 years and had between eight and 31 years of experience in dentistry. All had participated extensively in tooth-whitening trials. Each dentist successfully completed three separate calibration exercises at the 85% level with the Vitapan 3D-Master (Vident, Brea, CA, USA). The operators matched the shade of the middle-third of the study teeth.

Vitapan 3D-Master

The manufacturer's suggested use of the Vitapan 3D-Master system (Figure 1) is to first determine the lightness (L) group, then chroma (C), and finally the hue (H). This order better matches the capabilities of the human eye and corresponds to the importance of the three color elements in obtaining an accurate shade match.

The Vitapan 3D-Master was chosen, because the tabs were more uniformly spaced[7] and it has been shown to more closely match the population of the teeth studied.[8]

Figure 1: *Vita 3-D Shade Guide.*

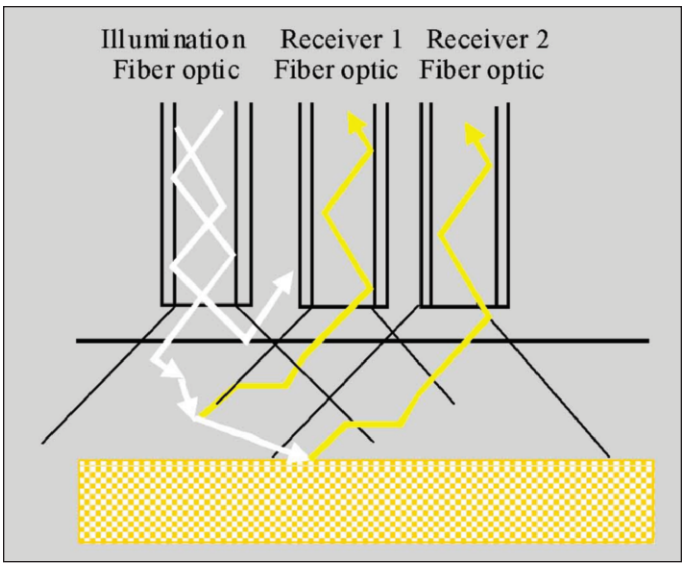

Figure 2: *Choice and placement of fiber optics assure spectral light is not included in color measurements.*

In addition, the 3D system is set up in a more logical arrangement.[7] There are five distinct lightness groups, and the difference between each pair of adjacent groups is equivalent. Within each lightness group, the change in chroma for adjacent pairs of tabs is equivalent. Three hue groups, L, M and R, are represented. The M represents the median hue. The L is more yellow and the R is more red.

The even spacing included in the 3D-Master arrangement affords the operator a choice of intermediate levels of lightness and chroma. For example, if a tooth was darker than the lightness in group 2 but lighter than the lightness in group 3, the operator could request the ceramist to create a restoration whose lightness matched 2.5. These intermediate shades are created not as a result of the ceramist's artistic ability, but as a function of the 3D system. The use of these intermediate choices adds the equivalent of an additional 46 shade tabs to the existing 26. For each match, the operator was required to note the lightness, chroma and hue of the tooth.

VITA Easyshade (ES)—General Description

The VITA EasyShade (Vident) is an intraoral reflectance spectrophotometer consisting of a base unit and a handpiece with a 5-mm wide probe tip. The unit provides a full spectrum light source, fiber-optic bundles that prevent the unit from making a color reading if motion or improper angulation is detected and two spectrometers. One unit measures light reflected from shallower surfaces and the other measures from deeper surfaces. When the unit is activated, light is projected into the deeper structures of the tooth, the reflected light is captured by a spectrometer, the spectral distribution of the light is analyzed and, finally, the CIE L*C*H* values are determined.

The CIE has established guidelines for relating physical measurements of color to a system representing color as perceived by the human eye.[9] This includes formulas for calculating color differences between two colors. In the CIE system, L* is a measure of darkness and lightness, with zero representing the darkest color and 100 the lightest color. C* represents color saturation and is measured on a 0 to 40 scale, with 0 representing the least color saturation and 40 the highest color saturation. H* is the hue measured on a scale of 0° to 360°. Zero represents red and 90° represents yellow.

Since spectral light or the light reflected off the enamel surface contains no color information, it is desirable to exclude it from the color determination process. This is possible, because light reflects from the enamel surface at such an angle that it is not captured by either spectrometer #1 or #2. This is a result of the unit's design. Fiber optic bundles are carefully chosen with specific numerical aperture ratings (NA), and their position is manipulated in terms of critical height (Ch) or distance from the fiber optic bundle to the object to be measured. The NA and Ch of spectrometer #2 are arranged to analyze light reflected only from the dentin, while spectrometer #1 captures light from surfaces shallower than dentin (Figure 2). Spectrometer #2 is active when the unit is in tooth measurement mode. Spectrometer #1 is active only when the Easyshade spectrophotometer is set in the verify mode. The verify mode is used to confirm that ceramic restorations received by the clinician accurately reflect the shade chosen and it measures shade tabs.

VITA Easyshade (ES)—Measurements

All shade measurements made with Easyshade were performed by a trained research assistant. The probe

of the Easyshade spectrophotometer was placed perpendicular to the tooth surface and the unit was activated. Measurements were made only in the middle-third of the tooth. At each evaluation and for each tooth, CIE L* (lightness), C* (chroma) and H* (hue) measurements were recorded. The Easyshade spectrophotometer compares CIE L*C*H* measurements for the tooth to known color standards for the 3D-Master tabs stored in its memory chip and reports which tab represents the closest match. This interpolated shade tab was also recorded. The standards for each tab and the internal algorithm used to determine the closest 3D-Master shade tab are proprietary.

Intra-rater Agreement

For each evaluator, shade choices for each tooth and each evaluation session were compared. Shade choices made by the same evaluator at two different evaluation sessions were considered to be an exact match only if the evaluator chose exactly the same tab at both sessions.

Since the interpolated Vita 3D-Master shade uses the CIE L*C*H* data to determine the closest tab match, it is subject to rounding error. An example may help to clarify. Assume that the L* value for the lightness group 2.0 is 78.0, and for group 3.0, it is 73.0. The interpolated 2.5 lightness group then is 75.5. The closest match for a tooth that has an L* value of 74.45 would be 2.5, while for an L* value of 74.44, it would be 3.0. Minor operator errors that lead to differences of 0.01 in CIE L* are unavoidable and, in terms of Δ_e, they are inconsequential. So, the fact that, at two different sessions, the same tooth might be interpreted as being in the 3.0 and 2.5 lightness groups, is a rounding error. Similar examples can also be given for chroma and hue. Accordingly, matches within a half-step of one another were calculated as an additional observation. In addition, the intra-class correlation (ICC) statistic was computed using SAS 9.1.3. For each evaluator, all possible pairs of ratings were evaluated.

Inter-rater Agreement

While it was possible to determine the number of matches among all possible pairs of raters, the authors believed this information would be of little value. More typically, in rater agreement studies, raters are compared to a gold standard to determine whether the level of agreement between each rater and the gold standard is significant. This approach is more relevant than noting that two evaluators, both of whom could have been wrong, agreed. However, due to the subjective nature of color and shade matching, establishing a gold standard is difficult. From past work, the authors of the current study anticipated that the CIE L*C*H* data provided by the Easyshade spectrophotometer would be reliable and precise, and the authors used this data to establish a standard to which the raters could be compared. Rather than considering this a gold standard, the authors consider it to be a "reasonable standard."

Three 3D-Master shade guides were used exclusively in the current study. Using the Easyshade spectrophotometer in the appropriate mode, each tab in these three guides was measured three times. The nine CIE L* readings were then averaged to determine the L* of each lightness group. This information was then used to establish an L* range for all lightness groups, including intermediate groups, such as 2.5, 3.5, etc. For each tooth and each evaluation session, the spectrometer data for L* was compared to these ranges, and the tab that was the best match in terms of lightness was recorded as the current study's reasonable standard. In statistical testing, the tab choice of each evaluator at each evaluation was compared to this reasonable standard. In the same manner, the ranges for C* were established and the C* data for each tooth and each evaluator were compared to these ranges to establish the reasonable standard for C*, and this was used for statistical purposes.

Rater agreement statistics calculate agreement above that which would occur by chance alone. Given the fact that there are only three levels for hue chance, agreement was fairly likely. Considering also that the hue groups chosen as the best tab match, such as lightness and chroma, were subject to rounding errors, the authors of the current study did not attempt to create a "reasonable standard" for hue. Accordingly, only CIE L* and C* spectrometer data were used and each was considered independently. Using a weighted Kappa statistic (SAS 9.1.3), the authors of the current study compared each of the four raters (ES, R1, R2 and R3) to the reasonable standard. For this comparison, a Kappa of less than 0.2 is considered poor. A Kappa of 0.2 to 0.4 represents fair agreement, 0.4 to 0.6 is considered moderate, 0.6 to 0.8 is considered good and 0.8 to 1.0 is considered very good.

Delta E 2000

The CIE L*C*H* data captured using the Easyshade guide was used to calculate the color difference or Delta E. The CIE color difference formula Delta E 2000 ($\Delta_{e\ 00}$) was used. From theoretical considerations and through experimentation under optimal viewing conditions, color scientists estimate that the threshold of perceptability equates to a Δ_e of 1.[10] Under ideal viewing conditions, 50% of the observers will be able to perceive a color difference this small and the other half will not. The threshold of acceptability is represented by a Δ_e of 2.7.[11] Color differences in this range are considered by 50% of the people to be an acceptable match.

Table 1: *Intra-rater Matches*

	ES	Rater #1	Rater #2	Rater #3
Exact Match	41.4%	27.0%	21.8%	16.5%
Match ± 0.5 tabs	91.1%	69.0%	84.6%	79.4%

Table 2: *Intra-class Correlation*

Rater	Intra-rater ICC	
	Lightness	Chroma
ES	0.87	0.80
Rater #1	0.79	0.56
Rater #2	0.87	0.45
Rater #3	0.73	0.52

Table 3: *Weighted Kappa Statistics*

	Lightness	*Chroma*
Rater	Overall Weighted Kappa	Overall Weighted Kappa
ES	0.69	0.92
Rater #1	0.68	0.84
Rater #2	0.63	0.82
Rater #3	0.61	0.82

RESULTS

Intra-rater Agreement

Each evaluator made 285 shade matches over the course of the study. Considering the 26 actual and 46 interpolated shade tabs, the evaluators chose from 72 tabs. Accordingly, choosing the same tab for a tooth at two separate evaluations would occur by chance alone in only 0.02% of the matches. By contrast, the percentage of exact matches ranged between 17% and 41%. The percentage of exact matches and matches within 0.5 tabs of one another are listed in Table 1. The intra-class correlation for each evaluator in terms of judging lightness and chroma is listed in Table 2. In terms of consistency in judging lightness, the intra-rater agreement for each evaluator was: ES–very good; R1–good; R2–very good and R3–good. For chroma, the level of intra-rater agreement for ES would be considered good and for the other three raters, it would be considered moderate.

Delta E 2000

The mean size of the color difference between tabs in the five lightness groups 5M1 to 4M1, etc, was 4.7. Repeated measurements of the same shade tab in the same 3D-Master guide resulted in a mean $\Delta_{e\,00}$ of 0.22. The mean $\Delta_{e\,00}$ for differences between the same tab in each of the three guides was 1.2.

The spectrometer data for each tooth at each evaluation was used to calculate the color differences $\Delta_{e\,00}$. All possible pairs of evaluation sessions were included. The mean $\Delta_{e\,00}$ for the spectrometer data was 1.5. The color difference between repeated measures of the same tooth was slightly higher than that between the same tab on each of the three shade guides. At a Δ_e of 1.5 and 1.2, respectively, both would be a minor, but perceptible, color difference. Both were well within the threshold for acceptability.

Inter-rater Agreement

The weighted Kappa statistics for lightness and chroma are listed in Table 3. In terms of judging lightness, the agreement between the evaluators and the reasonable standard established using the spectrometer data was good for all evaluators. For chroma, ES was rated very good and for the other evaluators, it was rated good.

DISCUSSION

In terms of reliability of a measurement system, there are two keys issues. The first is whether the measurements are accurate, or how often does the measurement system get the right answer. The second is precision or, when measuring the same object at different times, does the measurement system provide the same answer each time. In terms of target shooting, accuracy would be represented by the number of hits that are dead center. Precision would be represented by how closely repeated attempts were clustered together. Both are important.

Intra-rater evaluations reflect the precision of the measurement system. Inter-rater evaluations that compare raters to a standard reflect the accuracy of a measurement system. Since evaluating the agreement of the four raters to one another would have provided only a slightly different assessment of precision, the authors thought it important to create a reasonable standard.

Research has demonstrated that shade guides do not fully represent the wide variety of tooth colors present in the population.[12-13] As a result, most teeth are not an exact match to a specific shade tab. The tooth lies between tabs, forcing the evaluator to choose the tab that is the closest match. Research has also found that tab color varies from shade guide to shade guide among the same manufacturer,[14] and the current results regarding color differences of ostensibly the same tab in the three different shade guides confirm this.

The percentage of times that the four raters matched a tooth to exactly the same shade tab on two separate occasions was lower than the three human raters expected to achieve. Given the concerns with shade

guides and the differences in how people perceive color, the use of shade guides to measure color is subjective. Errors in measuring color difference with a shade guide are documented in the dental literature.[15]

In terms of evaluating lightness, the intra-agreement levels of raters ES and R2 were very good, while those for raters R1 and R3 were good. The strength of this intra-class correlation statistic may seem at odds with the percentage of exact matches achieved. It demonstrates that, where teeth had higher L* values, the raters chose lighter shade tabs and, where the teeth had a lower L*, they chose darker tabs. While the accuracy of the raters may have had room for improvement, the data supports a fairly high level of precision in terms of judgment of lightness. The fact that the percentage of matches increased dramatically when matches of plus or minus a half-tab were included also reflects raters' precision.

Comparing the intra-rater statistics for the four evaluators, rater ES was as precise as rater R2 and more precise than raters R1 and R3 for lightness and more precise than all three of the other raters for chroma. In terms of accuracy and precision, the results can be summarized as follows: For lightness, all raters were more precise than they were accurate. For Chroma, the human raters were more accurate than they were precise, and rater ES was both more accurate and more precise than the human raters.

The Weighted Kappa statistic reflected the accuracy of the raters when compared to the reasonable standard. Since only 50% of observers can perceive a Δ_e of one, the mean $\Delta_{e\,00}$ of 1.5 for the CIE L*C*H* data reflects a color difference that would still be difficult to perceive. By contrast, the mean size of the color difference between tabs 1M1, 2M1, etc, was estimated as 4.7. For all four raters, the accuracy of their determinations of the lightness or darkness of the teeth would be rated good, and for chroma, they would be rated very good. In both categories, rater ES was as accurate as or more accurate than the other raters.

The use of the CIE L*C*H* data to establish a reasonable standard causes some concerns. Since Easyshade uses the CIE L*C*H* data captured by the spectrometer to generate its interpolated 3D-Master tab choice, the reasonable standard established and the 3D tab chosen by rater ES are clearly not independent of one another. For both the intra-class correlation and the weighted Kappa statistics, the performance of raters R1, R2 and R3 were very comparable to those of rater ES. The authors of the current study believe that, if the reasonable standard and the ES 3D-Master tab chosen by ES were as highly correlated as it might appear that they would be, the human raters would not have been as comparable. Since the CIE L*C*H* data were shown to be highly precise, it seems reasonable to give the choice to create a reasonable standard some weight; however, one must be mindful that the tab choices for rater ES were not fully independent. However, the tab choices for the other three operators were made independent of any knowledge of the spectrophotometry data, and they were not correlated to the L* and C* data. Accordingly, comparison of the three clinicians to the standard created is useful.

CONCLUSIONS

The CIE L*C*H* data captured by the Easyshade guide were the most accurate. The 3D-Master interpolated shades provided by the Easyshade guide were at least as accurate and precise as those chosen by the three clinicians who were highly experienced at shade matching.

(Received 6 August 2008)

References

1. Joiner A (2004) Tooth colour: A review of the literature *Journal of Dentistry* **32(Supplement 1)** 3-12.
2. Paul S, Peter A, Pietrobon N & Hammerle CH (2002) Visual and spectrophotometric shade analysis of human teeth *Journal of Dental Research* **81(8)** 578-582.
3. Horn DJ, Bulan-Brady J & Hicks ML (1998) Sphere spectrophotometer versus human evaluation of tooth shade *Journal of Endodontics* **24(12)** 786-790.
4. Baltzer A & Kaufmann-Jinoian V (2005) Shading of ceramic crowns using digital tooth shade matching devices *International Journal of Computerized Dentistry* **8(2)** 129-152.
5. Okubo SR, Kanawati A, Richards MW & Childress S (1998) Evaluation of visual and instrument shade matching *The Journal of Prosthetic Dentistry* **80(6)** 642-648.
6. Klemetti E, Matela AM, Haag P & Kononen M (2006) Shade selection performed by novice dental professionals and colorimeter *Journal of Oral Rehabilitation* **33(1)** 31-35.
7. Paravina RD, Powers JM & Fay RM (2002) Color comparison of two shade guides *International Journal of Prosthodontics* **15(1)** 73-78.
8. Bayindir F, Kuo S, Johnston WM & Wee AG (2007) Coverage error of three conceptually different shade guide systems to vital unrestored dentition *The Journal of Prosthetic Dentistry* **98(3)** 175-185.
9. Commission Internationale de l'Eclairage (1978) Recommendations on Uniform Colour Spaces, Colour Terms, Publication 15, Supplement 2 Bureau Central de la Commission Internationale de l'Eclairage.
10. Kuehni RG & Marcus RT (1979) An experiment in visual scaling of small color differences *Color Research and Application* **4** 83-91.
11. Wee AG, Monaghan P & Johnston M (2002) Variation in color between intended matched shade and fabricated shade of dental porcelain *The Journal of Prosthetic Dentistry* **87(6)** 657-666.

12. Miller L (1993) Shade matching *Journal of Esthetic Dentistry* **5(4)** 143-153.
13. Miller L (1987) Organizing color in dentistry *Journal of the American Dental Association* **(Special Issue; Dec)** 26-E–40-E.
14. Preston JD (1985) Current status of shade selection and color matching *Quintessence International* **16(1)** 47-58.
15. Browning WD (2003) Use of shade guides for color measurement in tooth-bleaching studies *Journal of Esthetic & Restorative Dentistry* **15(Supplement)** 1 S13-20.