

RESEARCH

Open Access

# New insights in the reproducibility of visual and electronic tooth color assessment for dental practice



Anja Ratzmann<sup>1\*</sup>, Alexander Welk<sup>2</sup>, Stephanie Hoppe<sup>3</sup>, Jochen Fanghaenel<sup>4</sup> and Christian Schwahn<sup>5</sup>

## Abstract

**Background:** The aim of the study was to compare a 2D and 3D color system concerning a variety of statistical and graphical methods to assess validity and reliability of color measurements, and provide guidance on when to use which system and how to interpret color distance measures, including  $\Delta E$  and d(OM1).

**Methods:** The color of teeth 14 to 24 of 35 patients undergoing regular bleaching treatment was visually assessed and electronically measured with the spectrophotometer Shade Inspector™. Tooth color was recorded before bleaching treatment, after 14 days, and again after 6 months. VITAPAN® Classical (2D) and VITA-3D-Master® (3D) served as reference systems.

**Results:** Concerning repeated measurements, the 2D system was superior to the 3D system, both visually and electronically in terms of  $\Delta E$  and d(OM1), for statistics of agreement and reliability. All four methods showed strong patterns in Bland-Altman plots. In the 3D system, *hue* was less reliable than *lightness* and *chroma*, which was more pronounced visually than electronically. The smallest detectable color difference varied among the four methods used, and was most favorable in the electronic 2D system. Comparing the methods, the agreement between the 2D and 3D system in terms of  $\Delta E$  was not *good*. The reliability of the visual and electronic method was essentially the same in the 2D and 3D systems; this comparability is *fair to good*.

**Clinical relevance:** The 3D system may confuse human raters and even electronic devices. The 2D system is the simple and best choice.

## Background

Valid and reliable measurements of tooth color are of major importance in esthetic and restorative dentistry as well as in dental technical practice. Tooth color is usually described based on the Munsell color space in terms of hue, value, and chroma [1, 2]. Hue measures the basic color, value indicates the lightness of a color, and chroma measures the saturation or intensity of a color. Value is determined first, followed by chroma, yielding

hue as the third dimension. One of the most important prerequisites is the assessment of tooth color either via visual comparison with prefabricated color scales or using measuring devices such as a colorimeter, spectrophotometer or digital imaging systems with corresponding software [3]. The most common method in clinical practice is still the visual method using VITAPAN® Classical shade guide, which is a 2D system. In 1998, the VITA 3D-Master® shade guide was launched on the dental market. It was developed to systematize color determination, thereby enhancing the likelihood of valid and reliable color measurements [4–7]. Concerning the systematic determination, however, an implicit prior belief about the VITA 3D Master® was not checked in

\* Correspondence: [anja.ratzmann@uni-greifswald.de](mailto:anja.ratzmann@uni-greifswald.de)

<sup>1</sup>Department of Orthodontics and Department of Dental Propaedeutics/Community Dentistry, Dental School, University Medicine, Walther-Rathenau-Strasse 42, 17475 Greifswald, Germany  
Full list of author information is available at the end of the article



© The Author(s). 2020 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

developing this color guide: namely, that any two 3D shades within the same dimension at given constant shade values of the other two dimensions can be well differentiated by the human eye. In fact, dentists and dental technicians believe that the third dimension (*hue*) is problematic and that the distance between adjacent 3D shades is not large enough in this dimension. To quantify color differences,  $\Delta E$  as the Euclidean distance between two points in the color space of the three dimensions (value, chroma, and hue) has been used in the majority of dental color studies [8–20], although a modification of  $\Delta E$  is preferable [21]. However, numerous studies comparing visual and electronic methods have been published over the past decade [3, 8, 11, 18–20, 22–27].

Taking tooth color measurements is a complex process. In psychology and statistics, it is well known that repeated measurements [28, 29] or groups of observations such as on patients' teeth increase reliability [30, 31]. Moreover, the favored  $\Delta E$  to measure color differences cannot be applied to important graphical and statistical methods for the assessment of validity and reliability, including Bland-Altman plots to examine patterns of disagreement and the intraclass correlation coefficient (ICC) to estimate measurement variability [32]. These limitations can be overcome by using the distance of each shade from 0M1 of the 3D color system, denoted by  $d(0M1)$  [33]. Because  $d(0M1)$  does not distinguish shades of the same radius from M1,  $d(0M1)$  and  $\Delta E$  are complementary rather than competing. For example, in studying bleaching effects,  $d(0M1)$  may be favorable for 0M1 but less favorable for comparing shades by gender and age groups (or to study whether the gender difference in tooth color increases with age). In general, validity depends on the purpose [34] and is to be redefined for every research question; there is no such thing as a universal gold standard [35, 36]. Likewise, choosing methods to assess reproducibility depends on the purpose [37]. Whereas reliability is often related to calibration or comparability of examiners before and during performance of large cross-sectional or multicenter studies (only one measurement per participant in the full-scale investigation), the smallest detectable difference or the smallest detectable change is sought in longitudinal studies (at least two measurements per participant; measurement error occurred twice or more) [37], when the difference between repeated measurements is in the focus of interest. The smallest detectable difference or, in the present context, the smallest detectable color difference (SDCD), describes a statistical property and is different from perceptible or acceptable color difference thresholds. The SDCC of a row of teeth can easily be recalculated from the SDCC of a single tooth [31]. The SDCC may differ from method to method and from

study to study; it contradicts the idea that color difference thresholds are universally valid. In other words, the concept of a universal color difference threshold is scientifically misleading because it confuses validity and reliability. Moreover, color metrics are arbitrary, color perception is subjective, and acceptable color shade differences vary among different colors ( $\Delta E$ : 1.1 among red shades and 2.1 among yellow shades) [38]. Despite these limitations of color science, it can serve as a rough guide for color difference thresholds and may be useful in daily tooth color determination in dentistry. Therefore, different aspects must be considered when comparing the conventional 2D system with the newer 3D system. This seems more reasonable, because it is more ordered. Ordering alone, however, may not be enough, because the human or electronic rater must have the chance to measure reliably. Whereas directly adjacent shades of the 3D system have mean  $\Delta E$  values of about 3.8 for *lightness* (1M1–2M1–3M1) and 4.4 for *chroma* (2M1–2M2–2M3), the mean  $\Delta E$  value is only about 1.5 for the six direct neighbors of *hue* (2L1.5–2R1.5; 2L2.5–2R2.5) [38].

Thus, it can be hypothesized that *hue* is measured less reliably than *lightness* or *chroma*. This can be examined not only for an electronic rater but also for a human rater; within-subject comparisons are justified because the examiner serves as her/his own control (*hue* as exposure versus *lightness* or *chroma* as reference), similar to n-of-1 trials [39].

The aim of this study was to compare the 2D and the 3D color systems concerning a variety of statistical and graphical methods to assess validity and reliability, as well as to provide guidance on when to use which system and how to interpret  $\Delta E$  and  $d(0M1)$ .

## Materials and methods

### Subjects and clinical procedure

In order to better assess clinically relevant color changes, color measurements were performed in patients receiving a regular in-office bleaching treatment (BT). The tooth-inclusion criteria for performing BT were no caries, endodontic treatment or restorations. Patients with insufficient oral hygiene, previous BT, periodontal disease, pregnancy, and allergy or hypersensitivity to the bleaching agents were excluded. The study was approved by the ethics committee of the Medical Association (Ärztzammer) of Mecklenburg-Vorpommern (Reg. Nr.III UV 15/08). All patients gave informed consent. Thirty-five patients (24 women, 11 men, average age 30 years) from the Dental Clinic at the University of Greifswald participated. The complete clinical procedure was performed under standardized conditions according to the standardized clinical protocol for in-office bleaching under the supervision of an experienced dentist (AW).

The bleaching procedure was performed on teeth 15 to 25 and 35 to 45. Supra- and subgingival plaque, stains and calculus were removed, and all teeth were polished with non-fluoridated, oil-free pumice before bleaching.

The gingiva was protected by a liquid gingiva protectant (Dental Dam, Schütz Dental, Rosbach, Germany) activated by a light-curing unit (Ortholux TM LED Lurnig Light, Fa. 3 M Unitek). Bleach'n Smile, 35% H<sub>2</sub>O<sub>2</sub>, (Schütz Dental, Rosbach, Germany) was applied three times for 10 minutes according to the manufacturer's recommendation."

After bleaching, all teeth were fluoridated with Elmex® gelée (CP GABA, Germany).

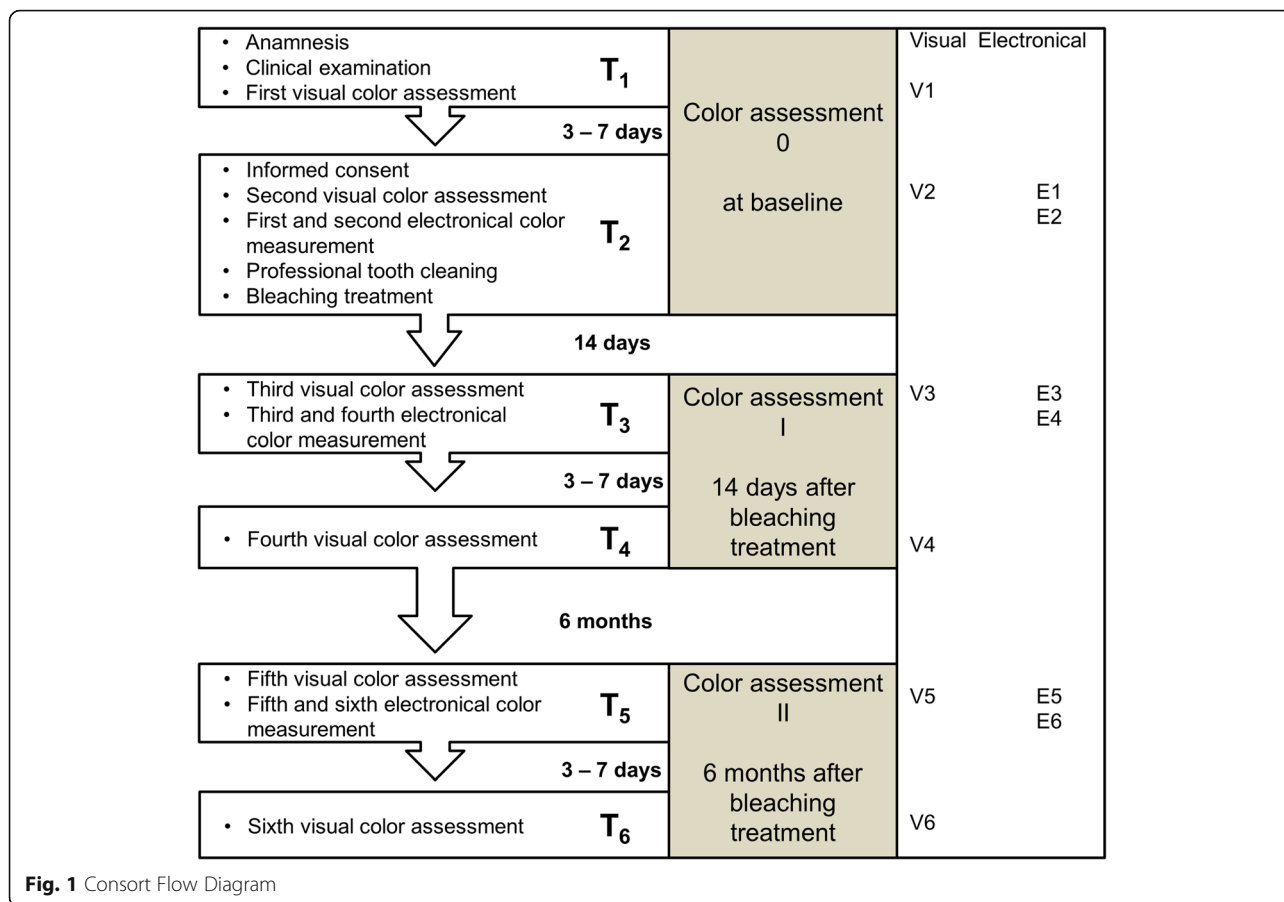
### Visual and electronic color assessment

The color of labial surfaces of teeth 14 to 24 was visually assessed by an experienced dental technician, who was ophthalmologically examined before this study [40], under diffuse daylight between 11 a.m. and 3 p.m. The time needed for color assessment was not restricted. Electronic measurements were performed with the spectrophotometer Shade Inspector™ (Schütz-Dental, Rosbach, Germany) by a dentist calibrated prior to this study [40]. The color systems VITAPAN® Classical (2D-VC; VITA Zahnfabrik, Bad Säckingen, Germany) and VITA 3D-Master® (3D; VITA Zahnfabrik, Bad Säckingen, Germany) served as reference systems. The VC color system has a two-dimensional structure that enables the description of *hue* (category A to D) and *lightness* including *chroma* (group 1 to 4) [41]. It serves as the standard shade guide for visual color assessment in dental practice. The 3D color system has a three-dimensional structure that enables the separate description of *lightness* (1 to 5 and 0 for bleaching), *chroma* (1 to 3, including half points), and *hue* (M, L, R) [42]. For the measurement procedure, each tooth was categorized into the gingival (S<sub>1</sub>), the body (S<sub>2</sub>), and the incisal (S<sub>3</sub>) segment. The incisal segment S<sub>3</sub> was not included in the analysis due to its transparency. Measurements were carried out as described in the previous study [33]. Time points of visual and electronic measurements were before BT (T<sub>1</sub>/T<sub>2</sub>- Baseline), 14 days (T<sub>3</sub>/T<sub>4</sub>) and 6 months (T<sub>5</sub>/T<sub>6</sub>) after BT (Fig. 1).

### Statistical methods

$\Delta E = ((\Delta L^*)^2 + (\Delta a^*)^2 + (\Delta b^*)^2)^{1/2}$  and  $\Delta E_{00}$  [43] were calculated.  $\Delta E_{00}$  is superior to  $\Delta E$ , but its calculation is quite sophisticated. Irregularities in the color space are corrected as follows: 1. the differences in the individual dimensions are calculated; 2. weighting is carried out; 3. finally a term for the interaction between the chroma differences and the hue differences is added; the calculation includes 22 lines of formulae [43].  $\Delta E_{00}$  values are usually smaller than those of  $\Delta E$

[21]. Here, we focused on  $\Delta E$  because it is more commonly used. The Bland-Altman plot [44] is one of the most frequently cited methods in medicine. Although several adaptations have been discussed [45–49], we present only the classical plot with the mean difference and the limits of agreement for d(0M1), which is  $\Delta E$  of each shade from 0M1. For method comparisons, but not for intra-rater comparisons, the regression line was added. Out of 840 paired observations, a total of 30–55 observations can be expected to be outside the limits of agreement according to M. Bland [50]. Besides the limits of agreement (difference between measurements  $\pm 1.96 \cdot$  standard deviation of the difference [44]), we present the agreement within 2.7 [16] and 3.7 [51] units of d(0M1) and  $\Delta E$ . These agreement statistics and the difference between the pairs of observations (denoted by  $d_2 - d_1$  for d(0M1), including standard deviation, are the only measurement error statistics also reported for  $\Delta E$ . The standard error of measurement (SEM) is a further agreement statistic and reported in two versions [37], for which the values are very similar herein. The SDCCD is defined as  $1.96 \cdot \sqrt{2} \cdot \text{SEM} \approx 2.77 \cdot \text{SEM}$  [37]. The SDSC on the level of groups of observations or patient's teeth is calculated according de Vet et al. 2001 [31]. In addition to agreement statistics, which are related to differences of repeated measurements, we present reliability statistics, which are related to calibration or comparability of raters or methods [34]. The fraction of the total measurement variance due to variance among teeth is estimated by three versions of the intraclass correlation coefficient (ICC) [28]. Whereas the ICC<sub>(3,1)</sub> ignores systematic differences between the two methods, raters, or measurements of the same rater, the ICC<sub>(2,1)</sub> includes an additional term of the variance among raters to account for the total measurement variance (denominator) [28, 37]. Thus, the greater the systematic difference between two raters, the smaller the ICC<sub>(2,1)</sub> compared with the ICC<sub>(3,1)</sub>. The ICC is the most appropriate reliability statistic [37] and recommended besides the Bland-Altman plot [32]. To avoid confusing terminology, SEM, SDSC and ICC are presented in the terminology used in Shrout & Fleiss [28]. ICC and kappa, which are closely related [32, 52], are interpreted according to Byrt's classification [53]. Graphics and statistical analyses were performed using Stata software, release 14.2 (Stata Corporation, College Station, TX, USA). As the American Statistical Association took a stand against Null Hypothesis Significance Testing [54, 55], we present confidence intervals as recommended [56]. Because accuracy requires a large sample size [44], we looked for at least 200 observations as recommended [57].



**Results**

**Intra-rater variability**

The agreement within the limits of  $\Delta E < 2.7$  was better for 2D than for 3D, both visually and electronically (Table 1). Figure 2 shows how the difference between two values of  $d(0M1)$  is related to  $\Delta E$ , for which the difference between visual and electronic measurements was chosen. This difference in  $d(0M1)$  was strongly and substantially symmetrically related to  $\Delta E$  (Fig. 2;  $R^2 = 0.69$  for 2D and  $R^2 = 0.59$  for 3D). The agreement within the limits of  $|d(0M1)| < 2.7$  was also better for 2D than for 3D, both visually and electronically (Table 2). The limits of agreement were narrower for  $2D_{elec}$  than for the remaining three methods (Table 2; Fig. 3). The Bland-Altman plots show clear patterns of disagreement for all methods, which is most pronounced for  $2D_{vis}$  (Fig. 3). The  $d(0M1)$  range is narrowest for  $2D_{vis}$  (11.0) and widest for  $3D_{elec}$  (21.6) (Fig. 3); the variability of  $d(0M1)$  in terms of the pooled standard deviation is highest for  $3D_{elec}$ . The reliability in terms of the ICC is good to very good for  $d(0M1)$  (Table 2).

As hypothesized for the three single dimensions, *hue* is less reliable than *lightness* or *chroma*, both electronically (Kappa value for *hue* = 0.45, 95% CI: 0.40–0.50;

$ICC_{(1,1)}$  for *lightness* = 0.76, 95% CI: 0.74–0.79;  $ICC_{(1,1)}$  for *chroma* = 0.67, 95% CI: 0.63–0.70) and visually (Kappa value for *hue* = 0.01, 95% CI: -0.05 – 0.06;  $ICC_{(1,1)}$  for *lightness* = 0.52, 95% CI: 0.47–0.57;  $ICC_{(1,1)}$  for *chroma* = 0.66, 95% CI: 0.62–0.69).

The standard errors of measurement and SDCDs were essentially the same for the four methods, except for  $2D_{elec}$  which was better (Table 2). On the level of groups of observations or patient’s teeth, the SDCD of  $2D_{elec}$  diminished from 2.8 for a single tooth to 1.4 and 1.0 for four and eight teeth, respectively. The SDCD of  $2D_{vis}$  decreased from 3.9 for a single tooth to 1.9 and 1.4 for four and eight teeth, respectively.

**Inter-method variability**

The comparability of visual and electronic measurements was fair to good in 2D and slight to fair in 3D for the agreement within the limits of  $\Delta E < 2.7$  (Table 3). The corresponding agreement of 2D and 3D measurements was fair in the visual approach, and poor to slight in the electronic approach (Table 3).

The comparability of visual and electronic measurements was good in 2D and fair in 3D for the agreement within the limits of  $|d(0M1)| < 2.7$  (Table 4). The

**Table 1** Agreement of repeated measurements for four methods in terms of  $\Delta E$  and  $\Delta E_{00}$  related to a single tooth

	Visual 2D Value	Visual 3D Value	Electronical 2D Value	Electronical 3D Value
Paired observations, number	840 <sup>a</sup>	840 <sup>a</sup>	839 <sup>b</sup>	840 <sup>b</sup>
Mean $\Delta E$ (standard deviation)	1.12 (1.95)	1.99 (1.95)	0.97 (1.41)	1.55 (2.11)
Agreement within $\Delta E < 2.7$ , proportion (95% CI)	80.1 (77.3–82.8)****	59.4 (56.0–62.7)***	90.9 (88.8–92.8)***,†	71.7 (68.5–74.7)**
Agreement within $\Delta E < 3.7$ , proportion (95% CI)	84.6 (82.0–87.0)***	77.9 (74.9–80.6)***	92.8 (90.9–94.5)***,†	83.3 (80.6–85.8)***
Mean $\Delta E_{00}$ (standard deviation)	0.92 (1.60)	1.59 (1.58)	0.80 (1.19)	1.27 (1.74)
Agreement within $\Delta E_{00} < 2.7$ , proportion (95% CI)	84.2 (81.5–86.6)***	69.5 (66.3–72.6)**	92.1 (90.1–93.9)***,†	77.4 (74.4–80.2)****
Agreement within $\Delta E_{00} < 3.7$ , proportion (95% CI)	91.9 (89.8–93.7)***,†	88.1 (85.7–90.2)***	96.3 (94.8–97.5)†	86.4 (83.9–88.7)***

<sup>a</sup> V1 versus V2, V3 versus V4, V5 versus V6 acc. to the flow chart

<sup>b</sup> E1 versus E2, E3 versus E4, E5 versus E6 acc. to the flow chart

Classifications for the interpretation of agreement

\* fair [40–60]; \*\* good [60–80]; \*\*\* very good [80–92]; † excellent [92–100]

corresponding agreement of 2D and 3D measurements was good in the visual approach, and fair in the electronic approach (Table 4).

Concerning the comparability of the visual and electronic measurements, the difference  $d_2 - d_1$ , which indicates systematic error, was moderate in 2D and small in 3D (Table 4; Fig. 4). The Bland-Altman plots show marked patterns of disagreement for the approaches.

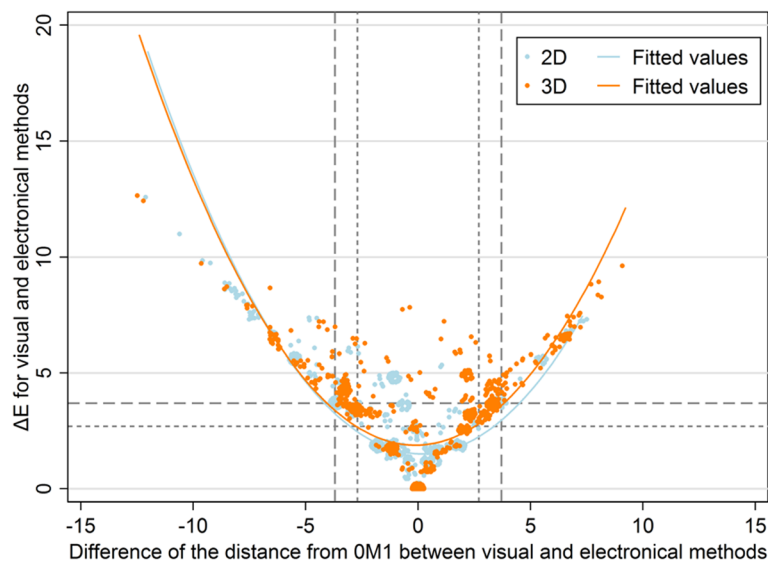
Concerning the comparability of 2D and 3D measurements, the difference  $d_2 - d_1$  indicates systematic error, which was pronounced in the electronic approach (Table 4; Fig. 4). This difference can be interpreted as constant bias. Assuming proportional bias, the regression line can be cautiously interpreted. The Bland-Altman plots, however, showed clear patterns of disagreement for the

approaches; the bias between the 2D and 3D system is neither constant nor uniquely proportional.

The reliability in terms of the ICC was fair to good for visual and electronic measurements. The reliability in terms of the  $ICC_{(3,1)}$ , which ignores systematic differences, was good to very good for 2D and 3D measurements. The reliability in terms of the  $ICC_{(2,1)}$ , which takes into account systematic differences, was poor to very good.

### Discussion

The 2D system proved superior to the 3D system both visually and electronically in terms of  $\Delta E$  and  $d(0M1)$  for statistics of agreement and reliability to assess intrarater variability. All four methods showed strong



**Fig. 2** Scatter plot for the relationship between  $\Delta E$  of the visual and electronic method and the difference of the distance from 0M1 between the visual and electronic method in 2D and 3D measurements; observations with the same coordinates are jittered to show their number

**Table 2** Agreement and reliability of repeated measurements for four methods in terms of the distance from OM1 related to a single tooth

	Visual 2D Value	Visual 3D Value	Electronical 2D Value	Electronical 3D Value
Number of paired observations	840 <sup>a</sup>	840 <sup>a</sup>	839 <sup>b</sup>	840 <sup>b</sup>
Mean distance (SD) $d_1$ from OM1 for the 1st measurement	15.0 (3.28)	13.4 (2.89)	15.8 (2.97)	13.1 (3.69)
Mean distance (SD) $d_2$ from OM1 for the 2nd measurement	14.9 (3.23)	13.3 (2.76)	15.9 (2.94)	13.4 (3.73)
Pooled SD of the 1st and 2nd measurement	3.25	2.83	2.96	3.71
Difference $d_2 - d_1$ (standard deviation)	-0.17 (1.98)	-0.08 (2.11)	0.09 (1.42)	0.26 (2.09)
Agreement within $ d(OM1)  < 2.7$ , proportion (95% CI)	83.7 (81.0–86.1) <sup>***</sup>	70.6 (67.4–73.7) <sup>**</sup>	93.6 (91.7–95.1) <sup>***,†</sup>	77.3 (74.3–80.1) <sup>***</sup>
Agreement within $ d(OM1)  < 3.7$ , proportion (95% CI)	94.0 (92.2–95.6) <sup>†</sup>	94.0 (92.2–95.6) <sup>†</sup>	97.0 (95.6–98.1) <sup>†</sup>	93.1 (91.2–94.7) <sup>***,†</sup>
Limits of agreement	-4.04 – 3.70	-4.21 – 4.06	-2.70 – 2.88	-3.84 – 4.36
Number of observations outside the limits of agreement total (lower; higher); expected: 30–55	50 (38; 12)	38 (13; 25)	53 (26; 27)	52 (20; 32)
Largest mean $d(OM1)$ value	22.2	20.7	24.8	24.9
Smallest mean $d(OM1)$ value	11.2	7.3	11.2	3.3
SEM <sub>(2,1)</sub>	1.400	1.489	1.007	1.489
SEM <sub>(3,1)</sub>	1.396	1.489	1.005	1.479
SDCD <sub>(2,1)</sub>	3.88	4.13	2.79	4.13
SDCD <sub>(3,1)</sub>	3.87	4.13	2.79	4.10
ICC <sub>(1,1)</sub> (95% CI)	0.81 (0.79–0.84) <sup>****</sup>	0.72 (0.69–0.75) <sup>**</sup>	0.88 (0.87–0.90) <sup>***</sup>	0.84 (0.82–0.86) <sup>***</sup>
ICC <sub>(2,1)</sub> (95% CI)	0.81 (0.79–0.84) <sup>****</sup>	0.72 (0.69–0.75) <sup>**</sup>	0.88 (0.87–0.90) <sup>***</sup>	0.84 (0.82–0.86) <sup>***</sup>
ICC <sub>(3,1)</sub> (95% CI)	0.82 (0.79–0.84) <sup>****</sup>	0.72 (0.69–0.75) <sup>**</sup>	0.88 (0.87–0.90) <sup>***</sup>	0.84 (0.82–0.86) <sup>***</sup>

SD denotes standard deviation, CI denotes confidence interval, SEM denotes standard error of measurement, SDCD denotes smallest detectable color difference, ICC denotes intraclass correlation coefficient

<sup>a</sup> V1 versus V2, V3 versus V4, V5 versus V6 acc. to the flow chart

<sup>b</sup> E1 versus E2, E3 versus E4, E5 versus E6 acc. to the flow chart

Classifications for the interpretation of agreement

\*\* good [60–80]; \*\*\* very good [80–92]; † excellent [92–100]

Classifications for the interpretation of reliability in terms of ICC

\* fair [0.4–0.6]; \*\* good [0.6–0.8]; \*\*\* very good [0.8–0.92]; † excellent [0.92–1.0]

patterns of disagreement between repeated measurements in Bland-Altman plots. As hypothesized, the 3D system is less reliable for *hue* than for *lightness* and *chroma*, a phenomenon which was more pronounced visually than electronically. The SDCD differs by the four methods used and was most favorable in the electronic 2D system. The agreement between the 2D and 3D systems in terms of  $\Delta E$  was not *good*. It was lower in the electronic than in the visual method. The comparability of the 2D and 3D systems was uncertain, because confidence intervals of ICCs accounting for systematic error were wide. The systematic error between the 2D and 3D systems cannot be neglected. The reliability of the visual and electronic method was substantially the same in the 2D and 3D systems; this comparability was *fair* to *good*.

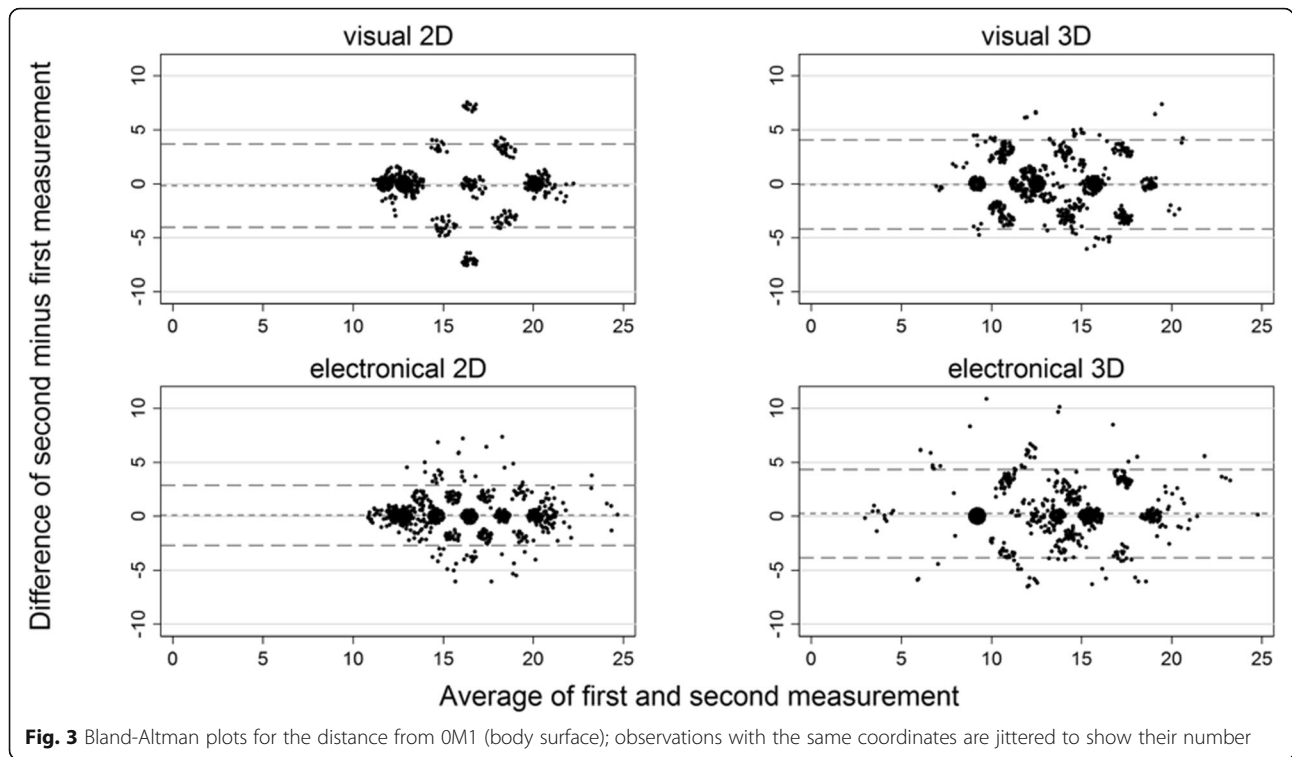
Below, the following aspects are discussed: 2D and 3D, visual and electronic,  $\Delta E$  and  $d(OM1)$ , Bland-Altman plots and statistics (patterns and numbers), single shade designations of the 3D system, validity and reliability, statistical SDCD and known thresholds, agreement and

reliability (comparability), human and machine, and intra- and inter-method variability.

## 2D and 3D systems

The 2D and 3D systems differ in the color space assessed [33]. Some 3D shades that are lighter (*lightness*) or stronger (*chroma*) are not well covered by the 2D system, which is especially pronounced for the additional bleaching shades available only in the 3D system. Compared to VC, *hue* ranges of 3D Master are extended toward yellow-red, and 3D Master shades are more uniformly spaced than that of VC [6]. In contrast, there are spatial gaps in the 3D system which are filled in the 2D system [33, 41]. In short, both guides are suboptimal and can be improved [14].

The variability between raters may favor the 3D Master shade guide over the VC shade guide [58]. The coverage error favors the 3D system, although it is unclear whether the difference between the 2D and 3D systems is clinically relevant [12, 14, 59–61]. However, the clear patterns in Bland-Altman plots for  $d(OM1)$  cast



doubt on the meaningfulness of converting 3D shades into VC shades (2D) as suggested elsewhere [62].

**Visual and electronical method**

The gaps mentioned above that are filled by the 2D system are supported by additional 2D shades to assess quarter-points for the second shade designation number [33], which is an important difference between the visual and electronic method. A further important difference is

the extension of the second shade designation number from the visual four-point scale to the electronic five-point scale. Similarly, the electronic 3D system includes bleaching shades not used by the visual 3D system evaluated here. Thus, it could have been expected that a human rater is inferior to the electronic rater, especially for the 2D system. It is of note that the agreement of intra-rater variability in terms of  $\Delta E$  and  $d(OM1)$  is better for the visual 2D measurement than that for the electronic 3D measurement.

**Table 3** Comparing methods of measurements in terms of  $\Delta E$  and  $\Delta E_{00}$ : 2D versus 3D within visual or electronical measurement; visual versus electronical measurements within 2D and 3D

	Visual versus electronical		2D versus 3D	
	within 2D Value	within 3D Value	within visual Value	within electronical Value
Paired observations, number	839 <sup>a</sup>	840 <sup>a</sup>	1680 <sup>b</sup>	1679 <sup>c</sup>
Mean $\Delta E$ (standard deviation)	2.53 (2.17)	2.99 (2.21)	3.46 (1.66)	3.91 (1.29)
Agreement within $\Delta E < 2.7$ , proportion (95% CI)	59.6 (56.2–62.9)***	40.6 (37.3–44.0)*	45.2 (42.8–47.6)*	18.6 (16.7–20.5)---
Agreement within $\Delta E < 3.7$ , proportion (95% CI)	67.2 (63.9–70.4)**	68.5 (65.3–71.7)**	52.9 (50.5–55.3)*	46.6 (44.2–49.0)*
Mean $\Delta E_{00}$ (standard deviation)	2.08 (1.80)	2.37 (1.82)	3.26 (1.23)	3.50 (1.00)
Agreement within $\Delta E_{00} < 2.7$ , proportion (95% CI)	62.9 (59.6–65.5)***	56.0 (52.5–59.3)*	45.8 (43.4–48.2)*	23.5 (21.5–25.6)---
Agreement within $\Delta E_{00} < 3.7$ , proportion (95% CI)	82.1 (79.4–84.7)****	75.2 (72.2–78.1)**	71.7 (69.5–73.9)**	64.6 (62.3–66.9)**

<sup>a</sup> V2 versus E1, V3 versus E3, V5 versus E5 acc. to the flow chart

<sup>b</sup> D2 versus D3 measurements for V1 – V6 acc. to the flow chart

<sup>c</sup> D2 versus D3 measurements for E1 – E6 acc. to the flow chart

Classifications for the interpretation of agreement

--- poor < 20; - slight [20–40]; \* fair [40–60]; \*\* good [60–80]; \*\*\* very good [80–92]

**Table 4** Comparing methods of measurements of the distance from OM1 related to a single tooth: 2D versus 3D within visual or electronical measurement; visual versus electronical measurements within 2D and 3D

	Visual versus electronical		2D versus 3D	
	within 2D	within 3D	within visual	within electronical
	Value	Value	Value	Value
Number of paired observations	839 <sup>a</sup>	840 <sup>a</sup>	1680 <sup>b</sup>	1679 <sup>c</sup>
Mean distance (SD) d <sub>1</sub> from OM1 for the electronical measurement	15.8 (2.97)	13.1 (3.69)		
Mean distance (SD) d <sub>2</sub> from OM1 for the visual measurement	14.9 (3.28)	13.4 (2.88)		
Mean distance (SD) d <sub>1</sub> from OM1 for the 2D measurement			15.0 (3.25)	15.9 (2.96)
Mean distance (SD) d <sub>2</sub> from OM1 for the 3D measurement			13.3 (2.82)	13.3 (3.71)
Difference d <sub>2</sub> – d <sub>1</sub> (standard deviation)	–0.89 (2.77)	0.22 (3.05)	–1.64 (1.98)	–2.58 (1.70)
Agreement within  d(OM1)  < 2.7, proportion (95% CI)	69.1 (65.9–72.2)**	53.3 (49.9–56.7)*	66.5 (64.2–68.8)**	47.1 (44.6–49.5)*
Agreement within  d(OM1)  < 3.7, proportion (95% CI)	86.3 (83.8–88.5)***	86.3 (83.8–88.6)***	80.9 (78.9–82.7)***	84.0 (82.2–85.8)***
Limits of agreement	–6.33 – 4.55	–5.76 – 6.19	–5.53 – 2.25	–5.90 – 0.75
Number of observations outside the limits of agreement total (lower; higher)	58 <sup>d</sup> (33; 25)	60 <sup>d</sup> (30; 30)	82 <sup>e</sup> (21; 61)	49 <sup>e</sup> (34; 15)
ICC <sub>(2,1)</sub> (95% CI)	0.58 (0.50–0.65)***	0.58 (0.53–0.62)***	0.69 (0.27–0.84)***	0.67 (–0.06–0.88)***
ICC <sub>(3,1)</sub> (95% CI)	0.61 (0.56–0.65)***	0.58 (0.53–0.62)***	0.79 (0.77–0.81)***	0.87 (0.86–0.88)***

<sup>a</sup> V2 versus E1, V3 versus E3, V5 versus E5 acc. to the flow chart

<sup>b</sup> D2 versus D3 measurements for V1 – V6 acc. to the flow chart

<sup>c</sup> D2 versus D3 measurements for E1 – E6 acc. to the flow chart

<sup>d</sup> expected number: 30–55

<sup>e</sup> expected number: 66–102

Classifications for the interpretation of agreement

— poor < 20; – slight [20–40]; \* fair [40–60]; \*\* good [60–80]; \*\*\* very good [80–92]

Classifications for the interpretation of reliability in terms of ICC

— poor < 0.2; – slight [0.2–0.4]; \* fair [0.4–0.6]; \*\* good [0.6–0.8]; \*\*\* very good [0.8–0.92]

Several studies have found that instrumental methods are more accurate or reliable than visual measurements [11, 19, 23–25, 63–65]. A recent study, however, has shown that clinically relevant differences between the visual evaluation and the intraoral scanning device (3Shape) are negligible [20]. According to Li & Wang, the reliability of shade matching can be ensured neither by the instrumental nor by the visual approach [66]. Furthermore, the difference in color matching between human-eye assessment and computerized colorimetry depends on tooth type [18] and shade [8].

#### ΔE and d(OM1)

ΔE supports only statistics on agreement; neither Bland-Altman plots nor reliability statistics are feasible. Essentially, d(OM1) enables evaluating patterns of disagreement, other agreement statistics such as SDCD, and reliability statistics including versions of ICC accounting for systematic errors. Regarding agreement of repeated measurements of the same rater, the differences among the four methods are substantially the same for ΔE < 2.7 and d(OM1). The level of agreement within fixed limits, however, is higher for d(OM1). For example, d(OM1) hardly differentiates 3M1 from 2L2.5 (d(OM1): 15.2 and 15.3, respectively) although ΔE is 8.3. Thus, if *lightness* is compensated by less *chroma* (or *chroma* by darkness),

then d(OM1) will not work well. The systematic errors between 2D and 3D measurements in d(OM1) are plausible, because the 2D and 3D systems differ in the color space assessed (see above). Within the 2D system, systematic errors between visual and electronic measurements are small, which can be explained by the additional quarter-point shades in the electronic 2D system.

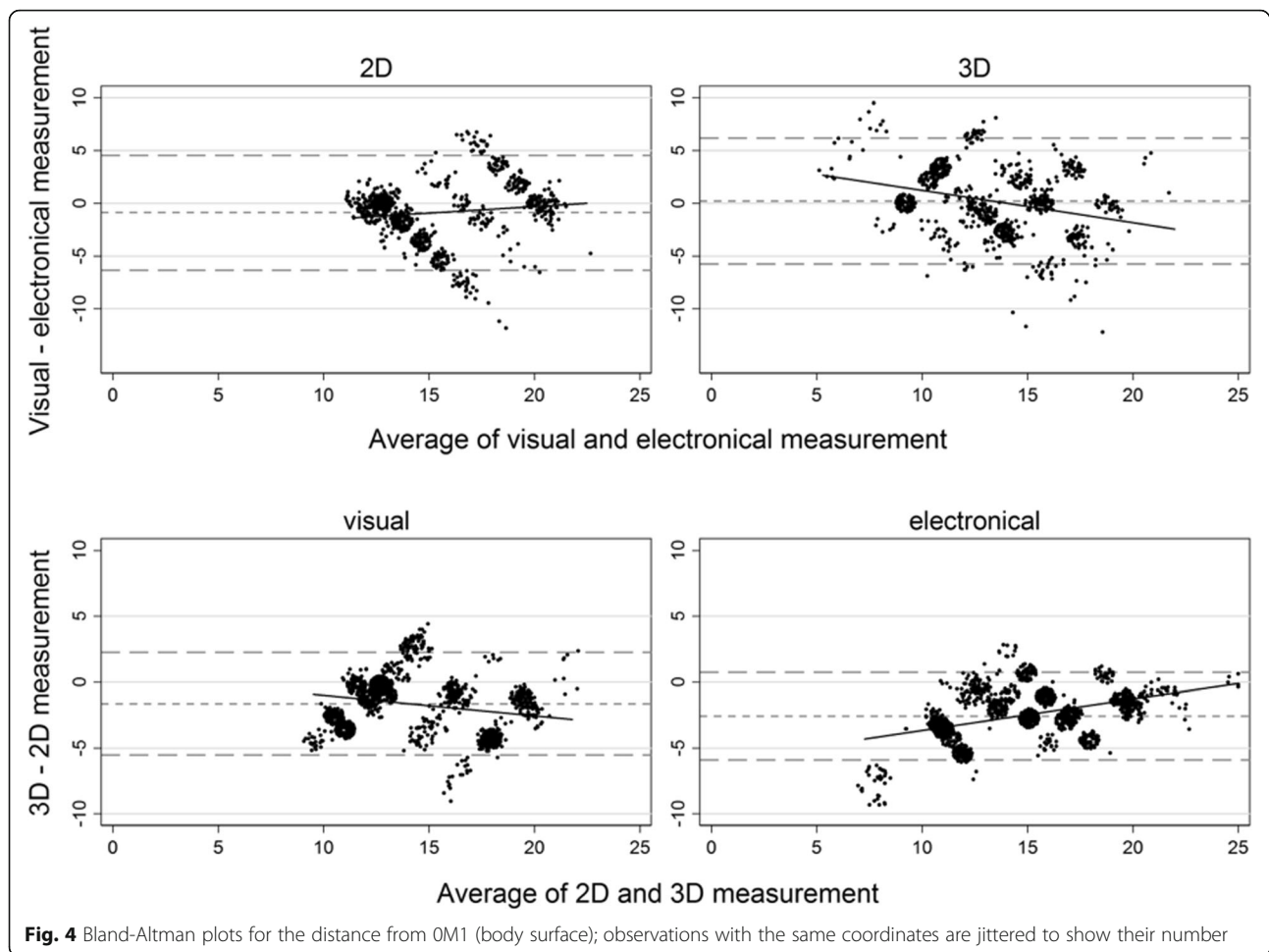
#### Bland-Altman plots and statistics – patterns and numbers

According to Bland-Altman plots, bias between the 2D and 3D systems is neither constant nor uniquely proportional. Even if these kinds of bias could be adjusted for – as suggested for uniquely proportional bias [48, 49] – the clear patterns are not appropriate for sophisticated statistical methods. Thus, Bland-Altman plots provide important information hardly available in numbers.

#### Single shade designations of the 3D system and d(OM1)

Although the reliability for the *hue* component of the visual 3D system is zero, the corresponding d(OM1) indicates good reliability. Likewise, the reliabilities are *fair* versus *very good* for the electronical 3D system, respectively. Thus, reliabilities of single shade designations can be misleading, especially for *hue*, for which ΔE values





are only about 1.5 (see above). Nevertheless, the *hue* component of the 3D system is problematic, because its reliability is lower than those of *lightness* and *chroma*.

#### Validity and reliability

Colorimetry does not facilitate valid measurements. The value of  $d(OM1)$ , however, supports pseudo-valid measurements, as the range of  $d(OM1)$  values differs across the four methods. The bleaching shades added to the electronic 3D system (not to the visual 3D system) make the difference: this range (21.6) is twice as high compared to visual 2D (11.0). Reliability in terms of the ICC depends on this range – if the variability of  $d(OM1)$  is small, the ICC will be small. As expected, the pooled standard deviation of the electronic 3D system is higher than that of the electronic 2D system. The ICC of the electronic 3D system, however, is lower, which emphasizes the problems with the 3D system – independent of human raters.

#### Smallest detectable color difference, acceptable and perceptible thresholds

An acceptability threshold of 2.7 in  $\Delta E$  and a perceptibility threshold of 1.2 in  $\Delta E$  are known [16]. The SDCD in terms of  $d(OM1)$  depends on the method and decreases from 2.8 to 1.0 for a row of eight teeth using electronic 2D measurements. These are statistical values and can differ from study to study. However, it is plausible that electronic 2D is the method with the best agreement, including SDCD. For properties of  $\Delta E$  and  $d(OM1)$ , electronic 2D is the recommended method for study designs with repeated measurements, such as longitudinal studies.

#### Agreement and reliability (comparability)

Whereas intra-rater agreement of repeated measurements in terms of SEM and SDCD does not differ between visual and electronic 3D measurements, the reliability or ICC differ substantially. Thus, a single human rater is not worse than the electronic device for a longitudinal study when using the 3D system. The

comparability of the four methods remains uncertain. Therefore, the same method should also be used in multicenter studies.

### Human and machine

Compared with a set of human raters, a set of devices from the same electronic system should have higher levels of standardization [67], which corresponds to the more favorable ICCs observed. However, n-of-1 trials, as used herein for the single human rater, limit generalizability. It may be further argued that the human rater lacks the ability to perceive *hue* [39]. But even if the examiner had lacked this ability, this would not have invalidated our conclusions, because we did not make an isolated statement on *hue*, but rather compared *hue* with *lightness* and *chroma*. These intra-human comparisons are supported by the n-of-1 trial design. Moreover, the same intra-device comparisons support the hypothesis that *hue* is not well reproducible; the electronic reliability of *hue* is merely *fair*. In addition to our findings, background knowledge further supports that 3D *hue* cannot be well assessed (see Introduction).

### Intra- and inter-method variability – validity revisited

Whereas the reliability within each of the four methods is *good* to *very good*, comparability of the visual and electronic measurements is only *fair* to *good*. This also questions the validity of visual and electronic measurements. In turn, this question also refers to the difference between the 2D and 3D system. In fact, Bland-Altman plots using the 2D system suggest that both visual and electronic values are valid only for d(OM1) values of about 12 (A1 – A2, B1 – B2) and greater than 20 (A4, B3 – B4, C3 – C4, D4). The shades B1 and A2 are not well covered by the 3D system [33], which is mirrored in the corresponding Bland-Altman plots. Vice versa, 3D shades 1M1 and 1M2 (both d(OM1) < 11.2 for the minimum of the 2D system) are not well covered by the 2D system [33] and question the validity of adjacent 2D shades, namely A1, B1, and B2. In daily practice, the 3D system may be useful for shades not available in the 2D system. Nevertheless, switching between methods cannot be recommended in scientific studies. The 3D system, however, can be favorable in bleaching studies owing to the added bleaching shades.

### Conclusion

The 3D system may confuse both human raters and electronic devices. The 2D system is the simple and best choice.

### Abbreviations

2D<sub>vis</sub>: 2D<sub>visual</sub>; 2D<sub>elec</sub>: 2D<sub>electronic</sub>; 3D<sub>vis</sub>: 3D<sub>visual</sub>; 3D<sub>elec</sub>: 3D<sub>electronic</sub>

### Acknowledgements

Not applicable.

### Informed consent

Informed consent was obtained individually from all participants included in the study.

### Explanation of any issues relating to journal policies

No issues.

### Confirmation that all authors have read and approved the manuscript

All authors have read and approved the manuscript.

### Confirmation that the content of the manuscript has not been published or submitted for publication elsewhere

The authors confirm that the content of the manuscript has not been published or submitted for publication elsewhere.

### Authors' contributions

AR: contributed to design, recruiting participants, analysis and interpretation, writing of manuscript, AW: contributed to design, supervision clinical treatment, analysis and interpretation, revising the manuscript, JF: critically revised the manuscript, StH: contributed to data acquisition. CS: contributed to statistical analysis and interpretation, writing statistical section of manuscript. All authors gave final approval and agreed to be accountable for all aspects of the work.

### Funding

The study was not funded. Open Access funding enabled and organized by Projekt DEAL.

### Availability of data and materials

All data are available on request at the Department of Orthodontics, Dental School, University Medicine, Walther-Rathenau-Strasse 42, 17475 Greifswald, Germany.

### Ethics approval

All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional research committee of the Ärztekammer Mecklenburg-Vorpommern (Reg. Nr.III UV 15/08) and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

### Consent for publication

Not applicable.

### Competing interests

All authors declare that they have no competing interests.

### Author details

<sup>1</sup>Department of Orthodontics and Department of Dental Propaedeutics/Community Dentistry, Dental School, University Medicine, Walther-Rathenau-Strasse 42, 17475 Greifswald, Germany. <sup>2</sup>Department of Restorative Dentistry, Periodontology, Endodontology, Preventive and Paediatric Dentistry, Dental School, University Medicine, Walther-Rathenau-Strasse 42, 17475 Greifswald, Germany. <sup>3</sup>Private Dental Office, Promenadestrasse, 296047 Bamberg, Germany. <sup>4</sup>Department of Orthodontics, Dental School, University Medicine, Franz-Josef-Strauß-Allee 11, 93053 Regensburg, Germany. <sup>5</sup>Department of Prosthetic Dentistry, Gerontology and Biomaterials, University Medicine, Walther-Rathenau-Strasse 42, 17475 Greifswald, Germany.

Received: 8 July 2020 Accepted: 18 November 2020

Published online: 16 December 2020

### References

1. Terry DA, Geller W, Tric O, Anderson MJ, Tourville M, Kobashigawa A. Anatomical form defines color: function, form, and aesthetics. *Pract Proced Aesthet Dent.* 2002;14:59–67 quiz 68.
2. Sikri VK. Color: implications in dentistry. *J Conserv Dent.* 2010;13:249–55.

3. Chu SJ, Trushkowsky RD, Paravina RD. Dental color matching instruments and systems. Review of clinical and research aspects. *J Dent*. 2010;38(Suppl 2):e2–16.
4. Paravina RD, Powers JM, Fay RM. Dental color standards: shade tab arrangement. *J Esthet Restor Dent*. 2001;13:254–63.
5. Paravina RD, Majkic G, Imai FH, Powers JM. Optimization of tooth color and shade guide design. *J Prosthodont*. 2007;16:269–76.
6. Paravina RD, Powers JM, Fay RM. Color comparison of two shade guides. *Int J Prosthodont*. 2002;15:73–8.
7. Paravina RD. Performance assessment of dental shade guides. *J Dent*. 2009;37(Suppl 1):e15–20.
8. Yap AU, Sim CP, Loh WL, Teo JH. Human-eye versus computerized color matching. *Oper Dent*. 1999;24:358–63.
9. Sim CP, Yap AU, Teo J. Color perception among different dental personnel. *Oper Dent*. 2001;26:435–9.
10. Wee AG, Monaghan P, Johnston WM. Variation in color between intended matched shade and fabricated shade of dental porcelain. *J Prosthet Dent*. 2002;87:657–66.
11. Paul S, Peter A, Pietrobon N, Hammerle CH. Visual and spectrophotometric shade analysis of human teeth. *J Dent Res*. 2002;81:578–82.
12. Li Q, Yu H, Wang YN. In vivo spectroradiometric evaluation of colour matching errors among five shade guides. *J Oral Rehabil*. 2009;36:65–70.
13. Hassel AJ, Cevirgen E, Balke Z, Rammelsberg P. Intraexaminer reliability of measurement of tooth color by spectrophotometry. *Quintessence Int*. 2009;40:421–6.
14. Cocking C, Cevirgen E, Helling S, Oswald M, Corcodel N, Rammelsberg P, Reinelt G, Hassel AJ. Colour compatibility between teeth and dental shade guides in Quinquagenarians and septuagenarians. *J Oral Rehabil*. 2009;36:848–55.
15. Olms C, Setz JM. The repeatability of digital shade measurement—a clinical study. *Clin Oral Investig*. 2013;17:1161–6.
16. Paravina RD, Ghinea R, Herrera LJ, Bona AD, Igiel C, Linninger M, Sakai M, Takahashi H, Tashkandi E, Perez MM. Color difference thresholds in dentistry. *J Esthet Restor Dent*. 2015;27(Suppl 1):S1–9.
17. Knezovic D, Zlataric D, Illes IZ, Alajbeg M, Zagar. In vivo evaluations of inter-observer reliability using VITA Easyshade(R) advance 4.0 dental shade-matching device. *Acta Stomatol Croat*. 2016;50:34–9.
18. Igiel C, Weyhrauch M, Wentaschek S, Scheller H, Lehmann KM. Dental color matching: a comparison between visual and instrumental methods. *Dent Mater J*. 2016;35:63–9.
19. Lehmann K, Devigus A, Wentaschek S, Igiel C, Scheller H, Paravina R. Comparison of visual shade matching and electronic color measurement device. *Int J Esthet Dent*. 2017;12:396–404.
20. Mehl A, Bosch G, Fischer C, Ender A. In vivo tooth-color measurement with a new 3D intraoral scanning system in comparison to conventional digital and visual color determination methods. *Int J Comput Dent*. 2017;20:343–61.
21. Gomez-Polo C, Portillo Munoz M, Lorenzo Luengo MC, Vicente P, Galindo P, Martin Casado AM. Comparison of two color-difference formulas using the Bland-Altman approach based on natural tooth color space. *J Prosthet Dent*. 2016;115:482–8.
22. van der Burgt TP, ten Bosch JJ, Borsboom PC, Kortsmit WJ. A comparison of new and conventional methods for quantification of tooth color. *J Prosthet Dent*. 1990;63:155–62.
23. Fani G, Vichi A, Davidson CL. Spectrophotometric and visual shade measurements of human teeth using three shade guides. *Am J Dent*. 2007;20:142–6.
24. Judeh A, Al-Wahadni A. A comparison between conventional visual and spectrophotometric methods for shade selection. *Quintessence Int*. 2009;40:e69–79.
25. Browning WD, Chan DC, Blalock JS, Brackett MG. A comparison of human raters and an intra-oral spectrophotometer. *Oper Dent*. 2009;34:337–43.
26. Gomez-Polo C, Gomez-Polo M, Celemin-Vinuela A, Martinez Vazquez De Parga JA. Differences between the human eye and the spectrophotometer in the shade matching of tooth colour. *J Dent*. 2014;42:742–5.
27. Parameswaran V, Anilkumar S, Lylajam S, Rajesh C, Narayan V. Comparison of accuracies of an intraoral spectrophotometer and conventional visual method for shade matching using two shade guide systems. *J Indian Prosthodont Soc*. 2016;16:352–8.
28. Shrout PE, Fleiss JL. Intra-class correlations: uses in assessing rater reliability. *Psychol Bull*. 1979;86:420–8.
29. Carroll RJ, Ruppert D, Stefanski LA, Crainiceanu CM. Measurement error in nonlinear models. A modern perspective. 2nd ed. Boca Raton: Chapman & Hall/CRC; 2006.
30. Altman DG, Bland JM. Statistics notes - standard deviations and standard errors. *Br Med J*. 2005;331:903.
31. de Vet HCW, Bouter LM, Bezemer PD, Beurskens AJHM. Reproducibility and responsiveness of evaluative outcome measures - theoretical considerations illustrated by an empirical example. *Int J Technol Assess Health Care*. 2001;17:479–87.
32. Szklo M, Nieto FJ. Epidemiology. Beyond the basics. 3rd edn. Burlington: Jones & Bartlett Learning; 2014.
33. Ratzmann A, Schwahn C, Treichel A, Faltermeier A, Welk A. Assessing the effect of multibracket appliance treatment on tooth color by using electronic measurement. *Head Face Med*. 2018;14:22.
34. Porta M, Greenland S, Hernán M, dos Santos SJ, Last JM, Burón A. A dictionary of epidemiology. 6th ed. Oxford: Oxford University Press; 2014.
35. Gigerenzer G, Marewski JN. Surrogate science: the idol of a universal method for scientific inference. *J Manag*. 2015;41:421–40.
36. Greenland S. Invited commentary: the need for cognitive science in methodology. *Am J Epidemiol*. 2017;186:639–45.
37. de Vet HCW, Terwee CB, Knol DL, Bouter LM. When to use agreement versus reliability measures. *J Clin Epidemiol*. 2006;59:1033–9.
38. Ahn JS, Lee YK. Color distribution of a shade guide in the value, chroma, and hue scale. *J Prosthet Dent*. 2008;100:18–28.
39. Senn S. Statistical issues in drug development. Chichester: Wiley; 2007.
40. Ratzmann A, Klinke T, Schwahn C, Treichel A, Gedrange T. Reproducibility of electronic tooth colour measurements. *Biomed Tech (Berl)*. 2008;53:259–63.
41. Park JH, Lee YK, Lim BS. Influence of illuminants on the color distribution of shade guides. *J Prosthet Dent*. 2006;96:402–11.
42. Vita Z. Dental Shade Guides. In *J Am Dent Assoc*. 2002;133:366–7.
43. Sharma G, Wu WC, Daa EN. The CIEDE2000 color-difference formula: implementation notes, supplementary test data, and mathematical observations. *Color Res Appl*. 2005;30:21–30.
44. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet*. 1986;1:307–10.
45. Bland JM, Altman DG. Comparing methods of measurement: why plotting difference against standard method is misleading. *Lancet*. 1995;346:1085–7.
46. Krouwer JS. Why Bland-Altman plots should use X, not (Y+X)/2 when X is a reference method. *Stat Med*. 2008;27:778–80.
47. Carstensen B. Comparing methods of measurement: extending the LoA by regression. *Stat Med*. 2010;29:401–10.
48. Taffé P, Peng M, Stagg V, Williamson T. MethodCompare: an R package to assess bias and precision in method comparison studies. *Stat Methods Med Res*. 2019;28(8):2557–65. <https://doi.org/10.1177/0962280218759693>.
49. Taffé P, Peng MK, Stagg V, Williamson T. Biasplot: a package to effective plots to assess bias and precision in method comparison studies. *Stata J*. 2017;17:208–21.
50. Bland M. Should all my observations lie between the limits of agreement? 2004.
51. Johnston WM, Kao EC. Assessment of appearance match by visual observation and clinical colorimetry. *J Dent Res*. 1989;68:819–22.
52. Fleiss JL. Statistical methods for rates and proportions. 2nd ed. New York: Wiley; 1981.
53. Byrt T. How good is that agreement? *Epidemiology*. 1996;7:561.
54. Wasserstein RL, Assoc AS. ASA statement on statistical significance and *P*-values. *Am Stat*. 2016;70:131–3.
55. Wasserstein RL, Lazar NA. The ASA's statement on *p*-values: context, process, and purpose. *Am Stat*. 2016;70:129–31.
56. Greenland S, Senn SJ, Rothman KJ, Carlin JB, Poole C, Goodman SN, Altman DG. Statistical tests, *P* values, confidence intervals, and power: a guide to misinterpretations. *Eur J Epidemiol*. 2016;31:337–50.
57. How can I decide the sample size for a study of agreement between two methods of measurement? 2004. Last accessed October 5, 2020. <https://wwwusers.york.ac.uk/~mb55/meas/sizemeth.htm>.
58. Oh WS, Koh IW, O'Brien WJ. Estimation of visual shade matching errors with 2 shade guides. *Quintessence Int*. 2009;40:833–6.
59. Hassel AJ, Koke U, Schmitter M, Beck J, Rammelsberg P. Clinical effect of different shade guide systems on the tooth shades of ceramic-veneered restorations. *Int J Prosthodont*. 2005;18:422–6.

60. Bayindir F, Kuo S, Johnston WM, Wee AG. Coverage error of three conceptually different shade guide systems to vital unrestored dentition. *J Prosthet Dent.* 2007;98:175–85.
61. Ongul D, Sermet B, Balkaya MC. Visual and instrumental evaluation of color match ability of 2 shade guides on a ceramic system. *J Prosthet Dent.* 2012; 108:9–14.
62. Zenthofer A, Wiesberg S, Hildenbrandt A, Reinelt G, Rammelsberg P, Hassel AJ. Selecting VITA classical shades with the VITA 3D-master shade guide. *Int J Prosthodont.* 2014;27:376–82.
63. Paul SJ, Peter A, Rodoni L, Pietrobon N. Conventional visual vs spectrophotometric shade taking for porcelain-fused-to-metal crowns: a clinical comparison. *Int J Periodontics Restorative Dent.* 2004;24:222–31.
64. Pimentel W, Tiozzi R. Comparison between visual and instrumental methods for natural tooth shade matching. *Sci Color.* 2014:47–9.
65. Liberato WF, Barreto IC, Costa PP, de Almeida CC, Pimentel W, Tiozzi R. A comparison between visual, intraoral scanner, and spectrophotometer shade matching: a clinical study. *J Prosthet Dent.* 2019;121:271–5.
66. Li Q, Wang YN. Comparison of shade matching by visual observation and an intraoral dental colorimeter. *J Oral Rehabil.* 2007;34:848–54.
67. Todorovic A, Todorovic A, Gostovic AS, Lazic V, Milicic B, Djuricic S. Reliability of conventional shade guides in teeth color determination. *Vojnosanit Pregl.* 2013;70:929–34.

### Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Ready to submit your research? Choose BMC and benefit from:**

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

**At BMC, research is always in progress.**

Learn more [biomedcentral.com/submissions](https://biomedcentral.com/submissions)

