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

Anja Ratzmann (anja.ratzmann@uni-greifswald.de)  
Department of Orthodontics, Dental School, University Medicine, Greifswald

Alexander Welk  
Department of Restorative Dentistry, Periodontology, Endodontology, Preventive and Paediatric Dentistry, Dental School, University Medicine Greifswald

Stephanie Hoppe  
Private Dental Office

Jochen Fanghaenel  
Department of Orthodontics, Dental School, University Medicine Regensburg

Christian Schwahn  
Department of Prosthetic Dentistry, Gerontology and Biomaterials, University Medicine Greifswald

Research

Keywords: 2D and 3D, color system, statistical and graphical methods

DOI: https://doi.org/10.21203/rs.3.rs-41190/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License
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 to give some guidance when to use which system and how to interpret color distance measures, including ΔE and d(0M1).

Methods The tooth color of teeth 14 to 24 of 35 patients with a regular bleaching treatment (BT) was visually assessed and electronically measured with the spectrophotometer Shade Inspector™. Tooth color was recorded before BT (T₁/T₂ - Baseline), 14d (T₃/T₄) and 6 months (T₅/T₆) after BT. VITAPAN® Classical (VC) and VITA-3D-Master® (3D) served as reference systems.

Results Intra-rater variability. The 2D system is better than the 3D system both visually and electronically in terms of ΔE and d(0M1) for statistics of agreement and reliability. All four methods show strong patterns of disagreement between repeated measurements in Bland-Altman plots. The 3D system lacks reliability of hue compared with that of lightness and chroma, which is more pronounced visually than electronically. The smallest detectable color difference (SDCD) differs by the four methods used and is most favorable in the electronic 2D system. Inter-rater variability. The agreement between the 2D and 3D system in terms of ΔE is not good according to Byrt’s classification. It is lower within the electronic method than within the visual method. Comparability of the 2D and 3D system is uncertain because confidence intervals of ICCs accounting for systematic error are wide. The systematic error between the 2D and 3D system cannot be neglected. The reliability of the visual and electronic method is substantially the same within the 2D and 3D system; this comparability is fair to good.

Clinical Relevance: According to the results of this study, the 3D system may confuse human raters and even electronic devices. The 2D system is the natural and best choice.

Background

Valid and reliable measurements of tooth color are of major importance in aesthetic and restorative dentistry and in dental technical practice. One of the most important prerequisites is the assessment of tooth color either via a visual comparison with prefabricated color scales or using measuring devices such as a colorimeter, spectrophotometer and digital imaging systems with corresponding software [1]. The most common method in clinical practice still is 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 [2-5]. Concerning the systematic determination, however, an implicit prior belief about the VITA 3D Master® was not checked in 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 human eyes. In fact, dentists and dental technicians believe that the third dimension (hue) is problematic and that the distance between neighbored 3D shades is not large enough.
To quantify color differences, $\Delta E$ has been used in the majority of dental color studies [6-18], although $\Delta E_{00}$ as an modification of $\Delta E$ is preferable [19]. However, numerous studies comparing visual and electronical methods have been published over the past decade [1, 6, 9, 16-18, 20-25].

Tooth color measurements are a complex process. In psychology and statistics, it is well-known that repeated measurements [26, 27] or groups of observations such as patient’s teeth increase reliability [28, 29]. 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 intraclass correlation coefficient (ICC) to estimate measurement variability [30]. These limitations can be overcome by using the distance of each shade from $0M1$ of the 3D color system, denoted by $d(0M1)$ [31]. Because $d(0M1)$ does not distinguish shades of the same radius from $0M1$, $d(0M1)$ and $\Delta E$ are rather complementary than competing. For example, $d(0M1)$ may be favorable in studying bleaching effects towards $0M1$ but is less favorable to compare shades of gender and age groups (or to study whether the gender difference in tooth color increases with age). In general, validity depends on the purpose [32] and is to be redefined for every research question; there is no thing such a universal gold standard [33, 34]. Likewise, choosing methods to assess reproducibility depends on the purpose [35]. Whereas reliability is often related to calibration or comparability of examiners before and while performing 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 asked for in longitudinal studies (at least two measurements per participant; measurement error occured twice or more) [35], when the difference between repeated measurements is in the focus of interest. The smallest detectable difference or herein the smallest detectable color difference (SDCD) describes a statistical property and is different from perceptible or acceptable color difference thresholds. The SDCD of a row of teeth can easily be recalculated from the SDCD of a single tooth [29]. The SDCD may differ from method to method and from study to study; it counterpoints intellectually that the concept of color difference thresholds is universally valid. With 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 differences differ for different colors ($\Delta E$: 1.1 for red and 2.1 for yellow) [36]. Thus, color science is a limited, but rough guidance for color difference thresholds and may be useful for daily tooth color determination in dentistry. Therefore, different aspects are to be considered when comparing the conventionally used 2D with the newer 3D system which seems to be more reasonable, because it is more ordered. Ordering alone, however, is possibly not enough, because the human or electronical rater must have the chance to measure reliably. Whereas directly neighbored 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 ...) [36]. Thus, it can be hypothesized that hue is measured less reliably as lightness or chroma. This can be examined not only for an electronical rater but also for a human rater because within-subject comparisons are justified because the examiner serves as her own control (hue as exposure versus lightness or chroma as reference), similar to n-of-1 trials [37].


The aim of the study was to compare the 2D and the 3D color system concerning a variety of statistical and graphical methods to assess validity and reliability, and to give some guidance, when to use which system and how to trade-off between interpreting ΔE and d(0M1).

Material And Methods

Subjects and clinical procedure

In order to better assess clinically relevant color changes, color measurements were performed in patients with a regular bleaching treatment (BT). The inclusion criteria were good oral hygiene, non-caries, none-endodontically treated and restoration-free permanent teeth. Patients with former BT, periodontal disease, pregnancy, and allergy or hypersensitivity to the bleaching agents were excluded. The study was approved by the ethics committee Aertzekammer Mecklenburg-Vorpommern (Reg. Nr.III UV 15/08). All patients gave informed consent. 35 patients (24 women, 11 men, average age 30 years) from the Dental School at the University of Greifswald participated. The complete clinical procedure was performed by an experienced dentist under standardized conditions according to the standardized clinical protocol for the in-office bleaching process. Bleaching procedure was performed at teeth 15 to 25 and 35 to 45. The supra- and subgingival plaque, stain and calculus were removed, and all teeth polished with non-fluoridated and oil-free pumice before bleaching. The gingiva was protected by a liquid gingiva protection (Dental Dam, Schütz Dental, Rosbach, Germany). Bleach’n Smile, 35% H₂O₂, (Schütz Dental, Rosbach, Germany) was applied three times for ten minutes according to the manufacture’s recommendation. Additionally, a curing light source (Ortholux TM LED Lurnig Light, Fa. 3M Unitek) was used. After bleaching all teeth were fluoridated with Elmex® gelée (CP GABA, Germany).

Visual color and electronic color measurement assessment

The color of labial surfaces of teeth 14 to 24 was visually assessed by an experienced dental technician, who was examined ophthalmologically before this study [38] at diffuse daylight between 11 a.m. and 3 p.m. The time needed for color assessment was not restricted. Electronical measurements were performed with the spectrophotometer Shade Inspector™ (Schuetz-Dental, Rosbach, Germany) by a calibrated dentist [38]. The color systems VITAPAN® Classical (VC; VITA Zahnfabrik, Bad Saeckingen, Germany) and VITA 3D-Master® (3D; VITA Zahnfabrik, Bad Saeckingen, 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) [39]. It serves as 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) [40]. For the measurement procedure, each tooth was categorized into the gingival (S₁), the body (S₂), and the incisal (S₃) segment. The incisal segment S₃ was not included into analysis because of its transparency. Measurements were carried out as described in the previous study.
Time points of visual and electronical measurements were before BT (T1/T2 - Baseline), 14d (T3/T4) and 6-month (T5/T6) after BT (Fig. 1).

### Statistical Methods

\[ \Delta E = ((\Delta L^*)^2 + (\Delta a^*)^2 + (\Delta b^*)^2)^{1/2} \] and \( \Delta E_{00} \) [41] were calculated. The \( \Delta E_{00} \) formula is superior to \( \Delta E \) and yields values that are usually smaller than those of \( \Delta E \) [19]. Herein, we focused on \( \Delta E \) because it is more widespread. The Bland-Altman plot [42] is one of the most frequently cited methods in medicine. Although several adaptations have been discussed [43–47], we present only the classical plot with the mean difference and the limits of agreement for \( d(0M1) \). 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 [48]. Besides the limits of agreement (difference between measurements \( \pm 1.96 \times \) standard deviation of the difference [42] we present the agreement within 2.7 [14] and 3.7 [49] 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 reported also for \( \Delta E \). The standard error of measurement (SEM) is a further agreement statistic and reported in two versions [35] for which the values are very similar herein. The SDCD is defined as \( 1.96 \times \sqrt{2 \times \text{SEM}} \approx 2.77 \times \text{SEM} \) [35]. The SDSC on the level of groups of observations or patient’s teeth is calculated according de Vet et al. 2001 [29]. 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 [32]. The fraction of the total measurement variance due to variance among teeth is estimated by three versions of the intraclass correlation coefficient (ICC) [26]. Whereas the ICC(3,1) ignores systematic differences between the two methods, raters, or measurements of the same rater, the ICC(2,1) includes an additional term of the variance among raters to account for the total measurement variance (denominator) [26, 35]. Thus, the greater the systematic difference between two raters, the smaller the ICC(2,1) compared with the ICC(3,1). The ICC is the most appropriate reliability statistic [35] and recommended besides Bland-Altman plot [30]. To avoid confusing terminology, SEM, SDSC and ICC are presented in the terminology used in Shrout & Fleiss [26]. ICC and kappa, which are closely related [30, 50], are interpreted according to Byrt’s classification [51]. Graphics and statistical analyses were performed using Stata software, release 14.2 (Stata Corporation, College Station, TX, USA).

### Results

#### Intra-rater variability

The agreement within limits of 2.7 \( \Delta E \) is very good for 2D_{elec}, good to very good for 2D_{vis}, good for 3D_{elec} and fair to good for 3D_{vis} (Table 1). The agreement within limits of 3.7 \( \Delta E \) is very good to excellent for 2D_{elec}, very good for 2D_{vis} and 3D_{elec}, and good to very good for 3D_{vis} (Table 1). The agreement in terms of \( \Delta E_{00} \) was better, especially for 3D_{vis}.
Table 1
Agreement of repeated measurements for four methods in terms of ΔE and ΔE₀₀ related to a single tooth.

|                     | Visual 2D | Visual 3D | Electronical 2D | Electronical 3D |
|---------------------|-----------|-----------|-----------------|-----------------|
|                     | Value     | Value     | Value           | Value           |
| Paired observations, number | 840*     | 840*     | 839†            | 840†            |
| Mean ΔE (standard deviation) | 1.12     | 1.99     | 0.97            | 1.55            |
| Agreement within ΔE < 2.7, proportion (95% CI) | 80.1     | 59.4     | 90.9            | 71.7            |
| Agreement within ΔE < 3.7, proportion (95% CI) | 84.6     | 77.9     | 92.8            | 83.3            |
| Mean ΔE₀₀ (standard deviation) | 0.92     | 1.59     | 0.80            | 1.27            |
| Agreement within ΔE₀₀ < 2.7, proportion (95% CI) | 84.2     | 69.5     | 92.1            | 77.4            |
| Agreement within ΔE₀₀ < 3.7, proportion (95% CI) | 91.9     | 88.1     | 96.3            | 86.4            |

* V1 versus V2, V3 versus V4, V5 versus V6 acc. to the flow chart
† E1 versus E2, E3 versus E4, E5 versus E6 acc. to the flow chart

Before presenting the agreement for d(0M1), we show how the difference between two is related to ΔE. For that, a meaningful ΔE is needed, for which the difference between visual and electronical measurements is chosen. This difference in d(0M1) is strongly and substantially symmetrically related to the corresponding ΔE (Fig. 2; R² = 0.69 for 2D and R² = 0.59 for 3D).

The agreement within limits of 2.7 d(0M1) is very good to excellent for 2D_{elec}, very good for 2D_{vis}, and good for 3D_{elec} and for 3D_{vis} (Table 2). The agreement within limits of 3.7 d(0M1) is very good to excellent for all four methods (Table 2). The limits of agreement are narrower for 2D_{elec} than for the remaining three methods (Table 2; Fig. 3). The difference d₂ – d₁, which indicates systematic error, is small for each method (Table 2; Fig. 3). The Bland-Altman plots show clear patterns of disagreement for all methods, which is most pronounced for 2D_{vis}. The largest mean d(0M1) values occurs for 3D_{elec} and
2D_{elec}, the smallest ones for 3D_{elec} and 3D_{vis} (Table 2; Fig. 3). The d(0M1) range is widest for 3D_{elec} (21.6) and narrowest for 2D_{vis} (11.0).
Table 2
Agreement and reliability of repeated measurements for four methods in terms of the distance from 0M1 related to a single tooth.

|                      | Visual 2D | Visual 3D | Electronical 2D | Electronical 3D |
|----------------------|-----------|-----------|------------------|------------------|
|                      | Value     | Value     | Value            | Value            |
| Number of paired observations | 840*      | 840*      | 839†             | 840†             |
| Mean distance (SD) $d_1$ from 0M1 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 0M1 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 $|d(0M1)|$                                 | 83.7 (81.0–86.1) | 70.6 (67.4–73.7) | 93.6 (91.7–95.1) | 77.3 (74.3–80.1) |
| < 2.7, proportion (95% CI)                                         |           |           |                  |                  |
| Agreement within $|d(0M1)|$                                 | 94.0 (92.2–95.6) | 94.0 (92.2–95.6) | 97.0 (95.6–98.1) | 93.1 (91.2–94.7) |
| < 3.7, proportion (95% CI)                                         |           |           |                  |                  |
| 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 $d(0M1)$ value                                      | 22.2       | 20.7       | 24.8             | 24.9             |
| Smallest mean $d(0M1)$ value                                     | 11.2       | 7.3        | 11.2             | 3.3              |
| SEM$_{(2,1)}$                                                    | 1.400      | 1.489      | 1.007            | 1.489            |

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

* V1 versus V2, V3 versus V4, V5 versus V6 acc. to the flow chart

† E1 versus E2, E3 versus E4, E5 versus E6 acc. to the flow chart
|                | Visual 2D | Visual 3D | Electronical 2D | Electronical 3D |
|----------------|----------|----------|-----------------|-----------------|
| SEM_{(3,1)}    | 1.396    | 1.489    | 1.005           | 1.479           |
| SDCD_{(2,1)}   | 3.88     | 4.13     | 2.79            | 4.13            |
| SDCD_{(3,1)}   | 3.87     | 4.13     | 2.79            | 4.10            |
| ICC_{(1,1)} (95% CI) | 0.81 (0.79–0.84) | 0.72 (0.69–0.75) | 0.88 (0.87–0.90) | 0.84 (0.82–0.86) |
| ICC_{(2,1)} (95% CI) | 0.81 (0.79–0.84) | 0.72 (0.69–0.75) | 0.88 (0.87–0.90) | 0.84 (0.82–0.86) |
| ICC_{(3,1)} (95% CI) | 0.82 (0.79–0.84) | 0.72 (0.69–0.75) | 0.88 (0.87–0.90) | 0.84 (0.82–0.86) |

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

* V1 versus V2, V3 versus V4, V5 versus V6 acc. to the flow chart
† E1 versus E2, E3 versus E4, E5 versus E6 acc. to the flow chart

The standard errors of measurement and SDCDs are substantially the same for the four methods except for 2D_{elec}, for which the agreement is better (Table 2). On the level of groups of observations or patient’s teeth, the SDCD of 2D_{elec} is 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} is diminished from 3.9 for a single tooth to 1.9 and 1.4 for four and eight teeth, respectively.

The reliability in terms of the ICC is good (3D_{vis}) to very good (2D_{elec}) (Table 2). Of note, the variability of d(0M1) in terms of the pooled standard deviation is highest for the electronical 3D measurements.

As hypothesized, hue of 3D is less reliably than lightness or chroma of 3D (3D_{elec}: 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); 3D_{vis}: 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).

Inter-methods variability
Concerning the comparability of visual and electronical measurements, the agreement within limits of 2.7 ΔE is fair to good within 2D, and slight to fair within 3D (Table 3). The corresponding agreements within limits of 3.7 ΔE are good.
### 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 | within 3D | within visual | within electronical |
| Paired observations, number | 839* | 840* | 1680† | 1679‡ |
| 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 (59.6–55.5) | 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) |

* V2 versus E1, V3 versus E3, V5 versus E5 acc. to the flow chart
† D2 versus D3 measurements for V1 – V6 acc. to the flow chart
‡ D2 versus D3 measurements for E1 – E6 acc. to the flow chart

Concerning the comparability of 2D and 3D measurements, the agreement within limits of 2.7 $\Delta E$ is *fair* within the visual approach, and poor within the electronical approach (Table 3). The corresponding agreements within limits of 3.7 $\Delta E$ are *fair*.

Concerning the comparability of visual and electronical measurements, the agreement within limits of 2.7 $d(0M1)$ is *good* within 2D, and *fair* within 3D (Table 4). The corresponding agreements within limits of 3.7 $d(0M1)$ are *very good*. 
|                              | Visual versus electronical | 2D versus 3D                  |
|------------------------------|-----------------------------|-------------------------------|
|                              | within                      | within                        |
|                              | 2D                          | 3D                            |
|                              | visual                      | electronical                 |
| Number of paired observations| 839*                        | 840*                         |
|                              | 1680†                       | 1679‡                         |
| Mean distance (SD) $d_1$ from 0M1 for the visual measurement | 15.8 (2.97) | 13.1 (3.69) |
| Mean distance (SD) $d_2$ from 0M1 for the visual measurement | 14.9 (3.28) | 13.4 (2.88) |
| Mean distance (SD) $d_1$ from 0M1 for the 2D measurement | 15.0 (3.25) | 15.9 (2.96) |
| Mean distance (SD) $d_2$ from 0M1 for the 3D measurement | 13.3 (2.82) | 13.3 (3.71) |
| Difference $d_2 – d_1$ (standard deviation) | -0.89 (2.77) | 0.22 (3.05) |
| Agreement within $|d(0M1)| < 2.7$, proportion (95% CI) | 69.1 (65.9–72.2) | 53.3 (49.9–56.7) |
| Agreement within $|d(0M1)| < 3.7$, proportion (95% CI) | 86.3 (83.8–88.5) | 86.3 (83.8–88.6) |
| Limits of agreement | -6.33–4.55 | -5.76–6.19 |
|                              | -5.53–2.25 | -5.90–0.75 |

* V2 versus E1, V3 versus E3, V5 versus E5 acc. to the flow chart
† D2 versus D3 measurements for V1 – V6 acc. to the flow chart
‡ D2 versus D3 measurements for E1 – E6 acc. to the flow chart
** expected number: 30–55
*** expected number: 66–102
|                              | Visual versus electronical | 2D versus 3D |
|------------------------------|----------------------------|--------------|
| Number of observations outside the limits of agreement total (lower; higher) | 58** (33; 25) | 60** (30; 30) |
|                              | 82*** (21; 61) | 49*** (34; 15) |
| ICC(2,1) (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(3,1) (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) |

* V2 versus E1, V3 versus E3, V5 versus E5 acc. to the flow chart
† D2 versus D3 measurements for V1 – V6 acc. to the flow chart
‡ D2 versus D3 measurements for E1 – E6 acc. to the flow chart
** expected number: 30–55
*** expected number: 66–102

Concerning the comparability of 2D and 3D measurements, the agreement within limits of 2.7 d(0M1) is good within the visual approach, and fair within the electronical approach (Table 4). The corresponding agreements within limits of 3.7 d(0M1) are good to very good and very good, respectively.

Concerning the comparability of the visual and electronical measurements, the limits of agreement are wide (Table 4; Fig. 4). The difference $d_2 - d_1$, which indicates systematic error, is moderate within 2D and small within 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 is pronounced within the electronical approach (Table 4; Fig. 4). This difference can be interpreted as constant bias. Assuming proportional bias, the regression line can cautiously be interpreted. The Bland-Altman plots, however, show clear patterns of disagreement for the approaches; the bias between the 2D and 3D system is neither constant nor uniquely proportional.

Concerning the comparability of the visual and electronical measurements, the reliability in terms of the ICC is fair to good.

Concerning the comparability of 2D and 3D measurements, the reliability in terms of the ICC(3,1), which ignores systematic differences, is good to very good. The reliability in terms of the ICC(2,1), which takes into account systematic differences, is poor to very good by interpreting the 95% CIs.
Discussion

The 2D system is better than the 3D system both visually and electronically in terms of \( \Delta E \) and \( d(0M1) \) for statistics of agreement and reliability to assess intra-rater variability. All four methods show strong patterns of disagreement between repeated measurements in Bland-Altman plots. As hypothesized, the 3D system lacks reliability of hue compared with that of lightness and chroma, which is more pronounced visually than electronically. The SDCD differs by the four methods used and is most favorable in the electronic 2D system. The agreement between the 2D and 3D system in terms of \( \Delta E \) is not good. It is lower within the electronic method than within the visual method. The comparability of the 2D and 3D system is uncertain because confidence intervals of ICCs accounting for systematic error are wide. The systematic error between the 2D and 3D system cannot be neglected. The reliability of the visual and electronic method is substantially the same within the 2D and 3D system; this comparability is fair to good.

We discuss following aspects: 2D and 3D, visual and electronic, \( \Delta E \) and \( d(0M1) \), 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 system

The 2D and 3D system differ in the color space assessed [31]. 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; 3D Master shades are more uniformly spaced than that of VC [4]. In contrast, there are spatial gaps of the 3D system, which are filled by the 2D system [31, 39]. In short, both guides are suboptimal and can be improved [12]. The intrarater variability depends on trained skills. For example, the intrarater repeatability of the 3D-Master shade guide is better than that of the VITA Lumin Vacuum in general practitioners but not in specialists (prosthodontics) [52]. Our experienced technician was not only trained, but also calibrated and ophthalmologically examined to ensure an efficacy instead of an effectiveness approach [53]. The variability between raters, which was not investigated herein, may favor the 3D Master shade guide over the VC shade guide [54]. The coverage error favors the 3D system, although it is unclear, whether the difference between the 2D and 3D system is clinically relevant [10, 12, 55–57]. The accuracy of the measurement of tooth shade obtained with an intraoral digital scanner was higher when the color was recorded as 3D Master values rather than VC values, whereas a visually perceptible color difference was found more often for VC values [58]. Repeatability was similar for both values. For some tooth-colored dental materials, it was suggested to convert 3D shades into VC shades (2D) adding a clinically relevant error in comparison with direct shade determination using the VC shade guide [59]. The clear patterns in Bland-Altman plots for \( d(0M1) \) question that this transformation is meaningful.

Visual and electronic method
The aforementioned gaps filled by the 2D system are supported by additional 2D shades to assess quarter points for the second shade designation number [31], 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 herein. Thus, there are reasons to have 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 $d(0M1)$ is better for the visual 2D measurement than that for the electronic 3D measurement. Numerous studies exist comparing instrumental and the conventional visual method [1, 6, 9, 13, 16–18, 20–25, 60]. Several studies found that instrumental methods are more accurate or reliable than visual measurements [9, 17, 21–23, 61–63]. Contrary to these findings, in a recently published study, results of the $\Delta E$ values showed that clinically relevant differences between the visual evaluation and the intraoral scanning device (3Shape) are negligible [18]. According to Li & Wang 2001, the reliability of shade matching can be ensured by neither the instrumental nor the visual approach [60]. Furthermore, studies indicate that the difference in color matching between human-eye assessment and computerized colorimetry dependents on tooth type [16] and shade [6]. The color dimension in with the greatest agreement between operator and spectrophotometer is value (chroma) or lightness [24]. No compatibility between visual and digital methods did exist for MLR and chroma [64]. The compatibility between both methods were determined only for lightness of maxillary central and canine teeth at all regions of labial surfaces [64]. Regarding repeatability, no significant differences were found between three shade guides by visual color assessment, although repeatability was relatively low (33–43%). Agreement with the colorimetric results was also low (8–34%) [65].

$\Delta E$ and $d(0M1)$

$\Delta E$ supports only statistics on agreement; neither Bland-Altman plots nor reliability statistics are feasible. Essentially, $d(0M1)$ enables evaluating patterns of disagreement, further 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 $\Delta E < 2.7$ and $d(0M1)$. The level of agreement within fixed limits, however, is higher for $d(0M1)$. For example, $d(0M1)$ hardly differentiates 3M1 from 2L2.5 ($d(0M1): 15.2$ and $15.3$, respectively) although $\Delta E$ is 8.3. Thus, if lightness is compensated by less chroma or (or chroma by darkness), then $d(0M1)$ will not work well. The systematic errors between 2D and 3D measurements in $d(0M1)$ are plausible, because the 2D and 3D system differ in the color space assessed (see above). Systematic errors between visual and electronic measurements are small, but present within the 2D system, which can be explained by the additional quarter-point shades in the electronic 2D system. It is thus highly plausible that the corresponding systematic error in the 3D system is close to zero – the electronic 3D system does not differ from the visual one.

Bland-Altman plots and statistics – patterns and numbers

According to Bland-Altman plots, bias between the 2D and 3D system is neither constant nor uniquely proportional. Even if these kinds of bias could be adjusted for - as suggested for uniquely proportional
bias [46, 47] - the clear patterns are not appealing for sophisticated statistical methods. Thus, Bland-Altman plots provide important information hardly available in numbers.

Single shade designations of the 3D system and d(0M1)

Although the reliability for the hue component of the visual 3D system is zero, the corresponding d(0M1) 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(0M1), however, supports pseudo-valid measurements, as the range of d(0M1) values differs across the four methods. The bleaching shades added to the electronical 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(0M1) is small, the ICC will be small. As expected, the pooled standard deviation of the electronical 3D system is higher than that of the electronical 2D system. The ICC of the electronical 3D system, however, is lower, which emphasized the problems of the 3D system – independent of human raters.

Smallest detectable color difference, acceptable and perceptible thresholds

An acceptability threshold of 2.7 in ΔE and a perceptibility threshold of 1.2 in ΔE are known [14]. The SDCD in terms of d(0M1) depends on the method and is diminished from 2.8 to 1.0 for a row of eight teeth using electronical 2D measurements. These values are statistical ones and can differ from study to study. However, it is plausible that electronical 2D is the method with the best agreement, including SDCD. For properties of ΔE and d(0M1), electronical 2D is the recommended method for study designs with repeated measurements such as longitudinal studies.

Agreement and reliability (comparability)

Whereas agreement of repeated measurements of the same rater in terms of SEM and SDCD does not differ between visual and electronical 3D measurements, reliability or ICC differ substantially. Thus, a single human rater is not worse than the electronical device for a longitudinal study, when using the 3D system. The comparability of the four methods remains uncertain. Therefore, the same method should be used in multicenter studies, too.

Human and machine

A set of human raters may cause additional problems concerning agreement and reliability. Compared with a set of human raters, a set of devices from the same electronical system should have higher levels of standardization [66], which corresponds to the more favorable ICCs observed. However, n-of-1 trials, as used herein [37] for the single human rater, limit generalizability. It may be further argued that the human rater lacks ability to percept hue. But even if the examiner had lacked this ability, this missing ability would not have been invalidated our conclusions, because we do not make an isolated statement on hue but compare hue with lightness and chroma. These within-human comparisons are supported by the n-of-1 trial design. Moreover, the same within-device comparisons support the hypothesis that hue is not well
reproducible; the electronical 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 electronical measurements is only fair to good. This questions also the validity of visual and electronical 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 electronical values are valid only in the d(0M1) ranges 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 [31], which is mirrored in corresponding Bland-Altman plots. Vice versa, 3D shades 1M1 and 1M2 (both d(0M1) < 11.2 for the minimum of the 2D system) are not well covered by the 2D system [31] and question the validity of neighbored 2D shades, namely A1, B1, and B2. In daily-life 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 human raters and even electronical devices. The 2D system is the natural and best choice.

Abbreviations

2D\text{vis} \\
2D\text{visual} \\
2D\text{elec} \\
2D\text{electronical} \\
3D\text{vis} \\
3D\text{visual} \\
3D\text{elec} \\
3D\text{electronical}

Declarations

Ethical approval: All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional research committee Ärztekammer Mecklenburg-Vorpommern (Reg. Nr.III UV 15/08) and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.
Informed consent: Informed consent was obtained from all individual participants included in the study.

Conflict of Interest: All authors declare that there is no conflict of interest.

Explanation of any issues relating to journal policies: No issues

Declaration of any competing interests: All authors declare that they have no competing interest.

Acknowledgements: Not applicable

Authors’ contributions:
AR: contributed to design, participants recruiting, 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 part of manuscript

All authors gave final approval and agree to be accountable for all aspects of the work.

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.

Availability of data and material: All data are available on request at Department of Orthodontics. Dental school, University Medicine, Walther- Rathenau- Strasse 42, 17475 Greifswald, Germany

Funding: The study was not funded.

Consent of publication: Not applicable.

References

1. 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.
2. Paravina RD, Powers JM, Fay RM. Dental color standards: shade tab arrangement. J Esthet Restor Dent. 2001;13:254–63.
3. Paravina RD, Majkic G, Imai FH, Powers JM. Optimization of tooth color and shade guide design. J Prosthodont. 2007;16:269–76.
4. Paravina RD, Powers JM, Fay RM. Color comparison of two shade guides. Int J Prosthodont. 2002;15:73–8.
5. Paravina RD. Performance assessment of dental shade guides. J Dent. 2009;37(Suppl 1):e15–20.
6. Yap AU, Sim CP, Loh WL, Teo JH. Human-eye versus computerized color matching. Oper Dent. 1999;24:358–63.
7. Sim CP, Yap AU, Teo J. Color perception among different dental personnel. Oper Dent. 2001;26:435–9.
8. 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.
9. Paul S, Peter A, Pietrobon N, Hammerle CH. Visual and spectrophotometric shade analysis of human teeth. J Dent Res. 2002;81:578–82.
10. 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.
11. Hassel AJ, Cevirgen E, Balke Z, Rammelsberg P. Intraexaminer reliability of measurement of tooth color by spectrophotometry. Quintessence Int. 2009;40:421–6.

12. 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.

13. Olms C, Setz JM. The repeatability of digital shade measurement—a clinical study. Clin Oral Investig. 2013;17:1161–6.

14. Paravina RD, Ghinea R, Herrera LJ, Bona AD, Igiel C, Linninger M, Sakai M, Takahashi H, Tashkandi E, Perez Mdel M. Color difference thresholds in dentistry. J Esthet Restor Dent. 2015;27(Suppl 1):1–9.

15. 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.

16. 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.

17. 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.

18. 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.

19. 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.

20. 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.

21. 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.

22. Judeh A, Al-Wahadni A. A comparison between conventional visual and spectrophotometric methods for shade selection. Quintessence Int. 2009;40:e69–79.

23. Browning WD, Chan DC, Blalock JS, Brackett MG. A comparison of human raters and an intra-oral spectrophotometer. Oper Dent. 2009;34:337–43.

24. Gomez-Polo C, Gomez-Polo M, Celemín-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.

25. 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.
26. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. Psychol Bull. 1979;86:420–8.
27. Carroll RJ, Ruppert D, Stefanski LA, Crainiceanu CM. Measurement Error in Nonlinear Models. A Modern Perspective. 2nd ed. Boca Raton: Chapman & Hall/CRC; 2006.
28. Altman DG, Bland JM. Statistics notes - Standard deviations and standard errors. BMJ. 2005;331:903–3.
29. 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.
30. Szkoł M, Nieto FJ. Epidemiology. Beyond the Basics. 3rd ed. Burlington: Jones & Bartlett Learning; 2014.
31. 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.
32. Porta M, Greenland S, Hermán M, dos Santos Silva I, Last JM, Burón A. A Dictionary of Epidemiology. 6th ed. Oxford: Oxford University Press; 2014.
33. Gigerenzer G, Marewski JN. Surrogate Science: The Idol of a Universal Method for Scientific Inference. J Manag. 2015;41:421–40.
34. Greenland S. Invited Commentary: The Need for Cognitive Science in Methodology. Am J Epidemiol. 2017;186:639–45.
35. de Vet HCW, Terwee CB, Knol DL, Bouter LM. When to use agreement versus reliability measures. J Clin Epidemiol. 2006;59:1033–9.
36. Ahn JS, Lee YK. Color distribution of a shade guide in the value, chroma, and hue scale. J Prosthet Dent. 2008;100:18–28.
37. Senn S. Statistical Issues in Drug Development. Chichester:: John Wiley & Sons Ltd.; 2007.
38. Ratzmann A, Klinke T, Schwahn C, Treichel A, Gedrange T. Reproducibility of electronic tooth colour measurements. Biomed Tech (Berl). 2008;53:259–63.
39. Park JH, Lee YK, Lim BS. Influence of illuminants on the color distribution of shade guides. J Prosthet Dent. 2006;96:402–11.
40. Vita Z. Dental Shade Guides. In J Am Dent Assc, vol. 133. pp. 366–367; 2002:366–367.
41. Sharma G, Wu WC, Daa EN. The CIEDE2000 color-difference formula: Implementation notes, supplementary test data, and mathematical observations. Color Research Application. 2005;30:21–30.
42. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet. 1986;1:307–10.
43. Bland JM, Altman DG. Comparing methods of measurement: why plotting difference against standard method is misleading. Lancet. 1995;346:1085–7.
44. 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.
45. Carstensen B. Comparing methods of measurement: Extending the LoA by regression. Stat Med. 2010;29:401–10.
46. Taffe P, Peng M, Stagg V, Williamson T. MethodCompare. An R package to assess bias and precision in method comparison studies. Stat Methods Med Res. 2018;962280218759693.
47. Taffe P, Peng MK, Stagg V, Williamson T. biasplot: A package to effective plots to assess bias and precision in method comparison studies. Stata Journal. 2017;17:208–21.
48. Bland M. Should all my observations lie between the limits of agreement?; 2004.
49. Johnston WM, Kao EC. Assessment of appearance match by visual observation and clinical colorimetry. J Dent Res. 1989;68:819–22.
50. Fleiss JL. Statistical methods for rates and proportions. Ed.2 edn. New York a.o.: Wiley; 1981.
51. Byrt T. How good is that agreement? Epidemiology. 1996;7:561–1.
52. Hammad IA. Intrarater repeatability of shade selections with two shade guides. J Prosthet Dent. 2003;89:50–3.
53. Fletcher RH, Fletcher SW, Fletcher GS. Clinical Epidemiology: The Essentials. fifth edn. Philadelphia (PA): Lippincott Williams & Wilkins; 2014.
54. Oh WS, Koh IW, O’Brien WJ. Estimation of visual shade matching errors with 2 shade guides. Quintessence Int. 2009;40:833–6.
55. 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.
56. 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.
57. 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.
58. Rutkunas V, Dirse J, Bilius V. Accuracy of an intraoral digital scanner in tooth color determination. J Prosthet Dent. 2019.
59. 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.
60. Li Q, Wang YN. Comparison of shade matching by visual observation and an intraoral dental colorimeter. J Oral Rehabil. 2007;34:848–54.
61. 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.
62. Pimentel W, Tiossi R. Comparison between visual and instrumental methods for natural tooth shade matching. The Science of Color. 2014:47–49.
63. Liberato WF, Barreto IC, Costa PP, de Almeida CC, Pimentel W, Tiossi R. A comparison between visual, intraoral scanner, and spectrophotometer shade matching: A clinical study. J Prosthet Dent. 2019;121:271–5.

64. Eryürük SE, Hekimoğlu C, Akçin EF, Çavuşoğlu Y. Comparison of Visual and Digital Color Measurement Methods on Anterior Natural Teeth. Balk J Dent Med. 2018;22:87–92.

65. Klemetti E, Matela AM, Haag P, Kononen M. Shade selection performed by novice dental professionals and colorimeter. J Oral Rehabil. 2006;33:31–5.

66. Todorovic A, Todorovic A, Gostovic AS, Lazic V, Milicic B, Djurisic S. Reliability of conventional shade guides in teeth color determination. Vojnosanit Pregl. 2013;70:929–34.

Tables

**Table 1.** Agreement of repeated measurements for four methods in terms of ΔE and ΔE₀₀ related to a single tooth.

| Method            | ΔE (standard deviation) | ΔE₀₀ (standard deviation) | Agreement within ΔE < 2.7, proportion (95% CI) | Agreement within ΔE₀₀ < 2.7, proportion (95% CI) |
|-------------------|-------------------------|---------------------------|-----------------------------------------------|-----------------------------------------------|
| Paired observations, number | 840* (1.95)            | 0.92 (1.60)               | (77.3 – 82.8)                                 | (81.5 – 86.6)                                 |
| Mean ΔE (4)       | 1.12 (0.95)             | 1.59 (1.58)               | 90.9 (88.8 – 92.8)                            | 92.1 (90.1 – 93.9)                            |
| Mean ΔE₀₀ (4)     | 0.97 (1.41)             | 1.55 (2.11)               | 71.7 (68.5 – 74.7)                            | 77.4 (74.4 – 80.2)                            |
| Agreement within ΔE < 3.7, proportion (95% CI) | 84.2 (82.0 – 87.0) | 88.1 (85.7 – 90.2) | 92.8 (90.9 – 94.5) | 96.3 (94.8 – 97.5) |
| Agreement within ΔE₀₀ < 3.7, proportion (95% CI) | 91.9 (89.8 – 93.7) | 96.3 (94.8 – 97.5) | 83.3 (80.6 – 85.8) | 86.4 (83.9 – 88.7) |

* V1 versus V2, V3 versus V4, V5 versus V6 acc. to the flow chart
† E1 versus E2, E3 versus E4, E5 versus E6 acc. to the flow chart

**Table 2.** Agreement and reliability of repeated measurements for four methods in terms of the distance from 0M1 related to a single tooth.
|                                      | Visual 2D | Visual 3D | Electronical 2D | Electronical 3D |
|--------------------------------------|-----------|-----------|------------------|------------------|
| Number of paired observations        | 840*      | 840*      | 839†             | 840†             |
| Mean distance (SD) \(d_1\) from 0M1 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 0M1 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 \(|d(0M1)|\) < 2.7, proportion (95% CI) | 83.7 (81.0 - 86.1) | 70.6 (67.4 - 73.7) | 93.6 (91.7 - 95.1) | 77.3 (74.3 - 80.1) |
| Agreement within \(|d(0M1)|\) < 3.7, proportion (95% CI) | 94.0 (92.2 - 95.6) | 94.0 (92.2 - 95.6) | 97.0 (95.6 - 98.1) | 93.1 (91.2 - 94.7) |
| 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 \(d(0M1)\) value | 22.2       | 20.7      | 24.8             | 24.9             |
| Smallest mean \(d(0M1)\) value | 11.2       | 7.3       | 11.2             | 3.3              |
| SEM\(_{(2,1)}\)                  | 1.400      | 1.489     | 1.007            | 1.489            |
| SEM\(_{(3,1)}\)                  | 1.396      | 1.489     | 1.005            | 1.479            |
| SDCD\(_{(2,1)}\)                 | 3.88       | 4.13      | 2.79             | 4.13             |
| SDCD\(_{(3,1)}\)                 | 3.87       | 4.13      | 2.79             | 4.10             |
| ICC\(_{(1,1)}\) (95% CI)         | 0.81 (0.79 - 0.84) | 0.72 (0.69 - 0.75) | 0.88 (0.87 - 0.90) | 0.84 (0.82 - 0.86) |
| ICC\(_{(2,1)}\) (95% CI)         | 0.81 (0.79 - 0.84) | 0.72 (0.69 - 0.75) | 0.88 (0.87 - 0.90) | 0.84 (0.82 - 0.86) |
| ICC\(_{(3,1)}\) (95% CI)         | 0.82 (0.79 - 0.84) | 0.72 (0.69 - 0.75) | 0.88 (0.87 - 0.90) | 0.84 (0.82 - 0.86) |

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

* V1 versus V2, V3 versus V4, V5 versus V6 acc. to the flow chart
† E1 versus E2, E3 versus E4, E5 versus E6 acc. to the flow chart

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 | within 3D | within visual | within electronical |
| Value               | Value      | Value      | Value      | Value      |
| Paired observations, number | 839* | 840* | 1680† | 1679‡ |
| Mean ΔE             | 2.53      | 2.99      | 3.46      | 3.91      |
| (standard deviation)| (2.17)    | (2.21)    | (1.66)    | (1.29)    |
| Agreement within ΔE < 2.7, proportion (95% CI) | 59.6 | 40.6 | 45.2 | 18.6 |
| (56.2 – 62.9)       | (37.3 – 44.0) | (42.8 – 47.6) | (16.7 – 20.5) |
| Agreement within ΔE < 3.7, proportion (95% CI) | 67.2 | 68.5 | 52.9 | 46.6 |
| (63.9 – 70.4)       | (65.3 – 71.7) | (50.5 – 55.3) | (44.2 – 49.0) |
| Mean ΔE<sub>00</sub> (standard deviation) | 2.08 | 2.37 | 3.26 | 3.50 |
| (1.80)              | (1.82)    | (1.23)    | (1.00)    |
| Agreement within ΔE<sub>00</sub> < 2.7, proportion (95% CI) | 62.9 | 56.0 | 45.8 | 23.5 |
| (59.6 – 55.5)       | (52.5 – 59.3) | (43.4 – 48.2) | (21.5 – 25.6) |
| Agreement within ΔE<sub>00</sub> < 3.7, proportion (95% CI) | 82.1 | 75.2 | 71.7 | 64.6 |
| (79.4 – 84.7)       | (72.2 – 78.1) | (69.5 – 73.9) | (62.3 – 66.9) |

* V2 versus E1, V3 versus E3, V5 versus E5 acc. to the flow chart
† D2 versus D3 measurements for V1 – V6 acc. to the flow chart
‡ D2 versus D3 measurements for E1 – E6 acc. to the flow chart

Table 4. Comparing methods of measurements of the distance from 0M1 related to a single tooth: 2D versus 3D within visual or electronical measurement; visual versus electronical measurements within 2D and 3D
|                              | Visual versus electronical within 2D | Visual versus electronical within 3D | 2D versus 3D within visual | 2D versus 3D within electronical |
|------------------------------|--------------------------------------|--------------------------------------|---------------------------|----------------------------------|
|                              | Value | Value                        | Value                     | Value                           |
| Number of paired observations | 839*  | 840*                         | 1680†                      | 1679‡                           |
| Mean distance (SD) d₁ from 0M1 for the electronical measurement | 15.8 (2.97) | 13.1 (3.69) |                      |                                 |
| Mean distance (SD) d₂ from 0M1 for the visual measurement | 14.9 (3.28) | 13.4 (2.88) |                      |                                 |
| Mean distance (SD) d₁ from 0M1 for the 2D measurement |                      | 15.0 (3.25) | 15.9 (2.96) |                                 |
| Mean distance (SD) d₂ from 0M1 for the 3D measurement |                      | 13.3 (3.71) | 13.3 (3.71) |                                 |
| Difference d₂ - d₁ (standard deviation) | -0.89 (2.77) | 0.22 (3.05) | -1.64 (1.98) | -2.58 (1.70) |
| Agreement within | | 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 | 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** (33; 25) | 60** (30; 30) | 82*** (21; 61) | 49*** (34; 15) |
| ICC₁ (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) |
| ICC₃ (95% CI) | 0.61 (0.56 - 0.65) | 0.58 (0.53 - 0.62) | 0.79 (0.77 - 0.84) | 0.87 (0.86 - 0.88) |

* V2 versus E1, V3 versus E3, V5 versus E5 acc. to the flow chart
† D2 versus D3 measurements for V1 - V6 acc. to the flow chart
‡ D2 versus D3 measurements for E1 - E6 acc. to the flow chart
** expected number: 30-55
*** expected number: 66-102

Figures
• Anamnesis
• Clinical examination
• First visual color assessment

3 – 7 days

T₁

Color assessment
0
at baseline

V₁

Visual
Electronical

V₂

E₁
E₂

• Informed consent
• Second visual color assessment
• First and second electronical color measurement
• Professional tooth cleaning
• Bleaching treatment

14 days

T₂

• Third visual color assessment
• Third and fourth electronical color measurement

3 – 7 days

T₃

Color assessment
I
14 days after bleaching treatment

V₃

E₃
E₄

• Fourth visual color assessment

6 months

T₄

• Fifth visual color assessment
• Fifth and sixth electronical color measurement

3 – 7 days

T₅

Color assessment
II
6 months after bleaching treatment

V₅

E₅
E₆

• Sixth visual color assessment

T₆

V₆
Figure 1
Consort Flow Diagram.

Figure 2
Scatter plot for the relationship between ΔE of the visual and electronical method and the difference of the distance from 0M1 between the visual and electronical method in 2D and 3D measurements; observations with the same coordinates are jittered to show their number.
Figure 3

Bland-Altman plots for the distance from 0M1 (body surface); observations with the same coordinates are jittered to show their number.
Figure 4

Bland-Altman plots for the distance from 0M1 (body surface); observations with the same coordinates are jittered to show their number.