A measuring system for facial aesthetics in Caucasian adolescents: reproducibility and validity

R. M. A. Kiekens*, J. C. Maltha*, M. A. van ‘t Hof** and A. M. Kuijpers-Jagtman*

Departments of *Orthodontics and Oral Biology and **Preventive and Curative Dentistry, College of Dental Science, Radboud University Nijmegen Medical Centre, The Netherlands

SUMMARY A new measuring system to judge facial aesthetics in young Caucasians is presented. The system uses sets of three photographs (one frontal, one three-quarter smiling, and one lateral) as a stimulus. Scores are performed on a visual analogue scale (VAS) with separate sets of reference photographs for girls and boys. The choice of the reference photographs was based on a panel evaluation of facial aesthetics of 40 boys and 40 girls from the archive of the orthodontic department. Reproducibility of the new measuring system was tested on a series of photographic sets (one frontal, one three-quarter smiling, and one lateral view) of 64 patients, using a panel of 78 adult laymen and 89 professionals. The panel members assessed these sets of photographs on a VAS, in relation to the reference sets.

The system was shown to be reproducible. Although the intra-observer reproducibility was low, the reliability coefficient was excellent (Cronbach’s alpha ≥ 0.98). Validity was tested by comparing the scores on the new scales with those of the three-quarter smiling photographic views on an earlier published scale. The correlation between the ratings on the new measuring system and the earlier published scale was 0.82 for laymen and 0.77 for professionals. The new system is simple and flexible in its use, and reproducible and valid for assessing facial aesthetics in young Caucasians. The system can be used in further investigations on the evaluation of facial aesthetics.

Introduction

Facial aesthetics are an important social concern in current society. Over 70 per cent of parents believe that their child will become more attractive, better liked, and more successful in their future occupational life after orthodontic treatment (Shaw et al., 1979). Children, young adults, and parents rate pleasant aesthetics as an important factor for psychosocial well-being (Shaw, 1981; Shaw et al., 1985; Birkeland et al., 2000). Aesthetic improvement is the most frequently reported subjective reason for seeking orthodontic treatment (Birkeland et al., 1999; Kiyak, 2000; Trulsson et al., 2002). This means that an assessment of dentofacial appearance should be included in the evaluation of orthodontic treatment need and treatment outcome (Tedesco et al., 1983).

Orthodontists often focus on dental aesthetics. This is reflected in the indices or measuring systems accepted within their profession, such as the Index of Orthodontic Treatment Need (IOTN; Evans and Shaw, 1987; Shaw et al., 1995), the Index of Complexity, Outcome, and Need (ICON) (Firestone et al., 2002), the Social Acceptability Scale of Occlusal Conditions (SASOC; Jenny et al., 1980), the Dental-Facial Attractiveness scale (DFA; Tedesco et al., 1983), and the Dental Aesthetic Index (DAI; Cons et al., 1989). These indices and scales all deal with dental malocclusions but not with overall facial aesthetics. However, an orthodontic treatment that is successful in the eyes of the professional does not always improve facial aesthetics (Arnett and Bergman, 1993; Al Yami et al., 1998), or facial balance (Bergman, 1999), and therefore might be considered to be less satisfying in the eyes of the patient. Since dental and facial aesthetics are two different parameters, a scoring system for orthodontic treatment need and treatment outcome should include dental as well as facial aesthetic scales (Phillips et al., 1992b; Al Yami et al., 1998). Such a ‘facial aesthetics’ scoring system preferably has to be simple, applicable in clinical practice, and should lead to quantitative data (Moyers, 1992; Ackerman et al., 1999).

Three components should be considered in developing such a scoring system: the way the subjects are presented, the characteristics of the judges, and the measurement technique.

The use of lateral cephalograms, silhouettes of profiles, or photographs taken in profile, frontal or three-quarter (smiling) orientation as stimuli, has been reported in the literature. All methods have their advantages and disadvantages. Lateral cephalograms and silhouettes have the advantage of reducing or eliminating the influence of confounding variables (Foster, 1973), but they do not represent the whole face and the actual smile cannot be evaluated (Mackley, 1993), a drawback that these stimuli have in common with lateral photographs. Furthermore, other factors may surpass the influence of the profile outline on facial aesthetics (Spyropoulos and Halazonetis, 2001).
Frontal photographs generally are rated more attractive than profile views (Kerr and O'Donnell, 1990), and simultaneous presentation of frontal and profile views probably would be advantageous (Phillips et al., 1992b). Three-quarter (smiling) colour photographs have also been advocated (Howells and Shaw, 1985; Peerlings et al., 1995), and probably the most complete visualization can be achieved by the combined use of frontal, lateral, and three-quarter (smiling) photographs.

In almost all cases, panel assessments have been used to evaluate facial aesthetics. Since the perception of facial aesthetics might be related to regional and/or professional background, age, or gender of the judges, much attention has been paid to the comparison of panels with different composition. Research in this field, however, has led to conflicting results. For example, Peerlings et al. (1995) found no effect of panel composition, while Spyropoulos and Halazonetis (2001) reported professionals to be less critical and Kerr and O'Donnell (1990) found professionals to be more critical than laymen.

Visual analogue scales (VAS) are most often used as a measuring instrument for dental, dentofacial, or facial aesthetics. Most authors have used a VAS without reference photographs (Howells and Shaw, 1985; Phillips et al., 1992a,b), but more recently the use of reference photographs has been advocated (Peerlings et al., 1995; Faure et al., 2002). Reference photographs can help the panel members to use the scale more uniformly, but preferentially their number should be minimized for the sake of simplicity of the scale. Therefore, in this study only one reference set in the mid-point of the scale was chosen.

The aim of this study was to develop a simple and valid measuring system for facial aesthetics in young Caucasian boys and girls, and to test its reproducibility and validity. Reproducibility include the calculation of errors and reliability coefficients.

Materials and methods

The first step was the selection of the reference sets, one for the boys and one for the girls. In a second step the system with the reference sets was evaluated.

Selection of reference photographs

The 1990–2000 files of the Department of Orthodontics and Oral Biology, Radboud University of Nijmegen Medical Centre, The Netherlands, were searched for sets of pre-treatment slides containing standardized frontal, three-quarter smiling and profile photographs of Caucasian children, aged 10 to 16 years, not wearing glasses, and without dental or facial trauma or known congenital defects. These inclusion criteria were met by 366 males and 398 females. From these, 40 males and 40 females were randomly selected to serve as a baseline. From each individual a set of photographs was prepared showing the three views simultaneously. The sets were placed in random order to create a slide show to be projected on a wall screen. A panel of 49 dental students (22 males and 27 females), aged 18 to 26 years was instructed to rate each face on a VAS from 0 to 100 (i.e. very unattractive to very attractive). Each set of pictures was shown for 15 seconds. No additional information about the faces was given.

Means and standard deviations (SD) of the ratings by the panel were calculated for each set of photographs. These mean VAS values ranged from 19 to 71. The median score was 50.1. The SD of the ratings for the individual faces varied from 8.6 to 19.2. One male and one female set of photographs with an average aesthetic score close to the median value and a SD close to 8.6 were then selected to serve as reference sets for the measuring system. The male reference set (age = 14.3 years) had a VAS score of 53.1 (SD = 10.2), and the female set (age = 11.7 years) a VAS score of 56.1 (SD = 9.9) (Figure 1a,b).

The measuring system, its reproducibility and validity

The proposed measuring system for the judgement of facial aesthetics consisted of a set of standardized photographs (frontal, lateral, and three-quarter smiling) of the experimental subjects, together with a set of reference photographs of the same gender. Facial aesthetics of the experimental subjects were scored on a VAS. In order to determine the reproducibility and validity of such a scoring system, it was tested on a series of 64 sets of photographs of orthodontic (pre-treatment) patients, and the judgement was performed by a professional and a lay panel. This resulted in over 8000 individual scores on facial aesthetics. The 64 patients used in the experimental group were randomly selected after stratification to have approximately eight boys and eight girls for each of the four Angle Classes. The determination of the Angle Class was as follows:

Class 1: neutro-occlusion and neutro-relationship of the jaws.
Class II division 1: disto-occlusion and disto-relationship of the jaws with proclined upper incisors.
Class II division 2: disto-occlusion and disto-relationship of the jaws with retroclined upper incisors.
Class III: mesio-occlusion and mesio-relationship of the jaws.

The inclusion criteria were the same as described in the previous section. Existing photographs from each individual were combined into a set of three, showing a frontal, a lateral, and a three-quarter smiling view simultaneously. The sets were randomly placed in a slide show and projected on a wall screen, showing every female face in relation to the female reference set, and every male face in relation to the male reference set. A panel of 78 adult laymen and a panel of 89 professionals (85 orthodontists and 4 postgraduates) were
constructed. The composition of the panels with respect to age and gender is given in Table 1. The members of both panels all had a high social-economic status.

Each set of photographs of one individual, together with the reference set, was shown for 15 seconds and the panel members were asked to assess facial aesthetics in relation to the reference set on which the VAS score was indicated, on a VAS from 0 to 100. In order to evaluate intra-individual reproducibility of the measuring system, six duplicate sets of the individuals were added to the series.

Since no criterion or gold standard exists to define facial aesthetics, other investigations had to be used to validate the measuring system. According to Streiner and Norman (2003) concurrent validity is ‘where a new scale is correlated with the criterion measure, both of which are given at the same time’. The only photographic scales for assessing facial aesthetics in healthy Caucasians found in the current literature were the Peerlings scales (Peerlings et al., 1995). This system uses separate scales for two age groups and for males and females, each with five three-quarter smiling reference photographs. In order to estimate the validity of the measuring system, the photographic three-quarter smiling views of the individuals from 11 to 13 years and from 14 to 16 years ($n = 44$) that were included in the sample, were also evaluated using the Peerlings scales. The ratings

Figure 1 Reference photographs for (a) boys and (b) girls. Indicated score on a visual analogue scale was 53.1 and 56.1, respectively.

Table 1 Composition of the panels.

<table>
<thead>
<tr>
<th></th>
<th>Laymen ($n = 78$)</th>
<th>Professionals ($n = 89$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male/Female</td>
<td>38/40</td>
<td>38/51</td>
</tr>
<tr>
<td>Mean age ± SD (years)</td>
<td>51 ± 10.3</td>
<td>42 ± 8.8</td>
</tr>
<tr>
<td>Median age (years)</td>
<td>50</td>
<td>41</td>
</tr>
<tr>
<td>Age range (years)</td>
<td>28–76</td>
<td>25–65</td>
</tr>
</tbody>
</table>
were undertaken by four orthodontists and four postgraduate students familiar with the Peerlings scales. The ratings on the Peerlings scales given by the small professional panel were compared with the VAS values given by the laymen and the professionals using the new measuring system.

**Statistics**

Means and SD of the ratings for each set of photographs were calculated per panel. The ratings of two laymen were excluded because of missing data. In this situation of sufficient normally distributed ratings, a two-way ANOVA was performed to evaluate the effect of gender and Angle classification on the ratings. Tukey’s test for multiple comparisons was applied in case of significant effects.

In the statistical evaluation of the reproducibility of the measuring system, the random error of the duplicated photographs was calculated according to Dahlberg (1940). The intra-observer consistency was calculated by Pearson’s correlation coefficient and intra-class coefficient (ICC). The reliability coefficient for the final score was calculated by Cronbach’s alpha. The difference between laymen and professionals was studied by a paired t-test.

The reproducibility of the measuring system was also statistically evaluated by Pearson’s correlation between the ratings of the laymen and those of the professionals. Concurrent validity was determined by calculating Pearson’s correlation coefficient between the ratings on the VAS given by the two panels, and the scores on the Peerlings scales given by four orthodontists and four postgraduates.

**Results**

**Reproducibility of the measuring system**

The random error of the duplicate photographs varied over the panel members, from 2 to 17 VAS points in the laymen panel and from 0 to 35 VAS points in the professional panel. The median individual random error was 6.8 VAS points (P25 = 4.7; P75 = 10.5) for the laymen and 10.6 VAS points (P25 = 6.4; P75 = 19.8) for the professionals.

The intra-observer consistency expressed as the Pearson’s correlation coefficient between the first and the second rating of the duplicate photographs was 0.68 (P25 = 0.44; P75 = 0.81) for the laymen and 0.65 (P25 = 0.48; P75 = 0.84) for the professionals. The ICC was 0.56 (P25 = 0.25; P75 = 0.73) for the laymen and 0.53 (P25 = 0.32; P75 = 0.73) for the professionals. The reliability coefficient of the final overall score was excellent: Cronbach’s alpha was 0.98 for the laymen and 0.99 for the professionals. The professionals scored four VAS points lower compared with the laymen (95% CI = 2.6–5.2; P = 0.000). Pearson’s correlation coefficient between the mean VAS scores of the laymen and those of the professionals was 0.92.

**Validity of the measuring system**

Pearson’s correlation between the ratings given on the Peerlings scales by the small panel and the mean VAS values given by the laymen was r = 0.82. Pearson’s correlation with the mean values given by the professionals was r = 0.77. The outcome measures consisted of over 8000 individual assessments of facial aesthetics. VAS means and SDs of the different panels for the two stratification factors, Angle Class and gender, were calculated (Table 2). There was no significant difference in the aesthetic scores for boys for girls (ANOVA, both panels P > 0.25). A significant effect of Angle Class was seen in the ratings of the professionals (P = 0.02). Although the same trends were found, these findings were not significant for the laymen ratings (P = 0.08). Tukey’s test revealed that Angle Class II division 2 patients were considered significantly more attractive than Angle Class III patients (P < 0.05 both for laymen and orthodontists). There was no significant interaction between gender and Angle classification (interaction test, P > 0.30).

**Discussion**

The aim of the present study was to develop a simple and valid measuring system for facial aesthetics. It was decided

<table>
<thead>
<tr>
<th>Angle Class</th>
<th>Boys (n = 32)</th>
<th>Professionals</th>
<th>Girls (n = 32)</th>
<th>Professionals</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Laymen</td>
<td>Professionals</td>
<td>Laymen</td>
<td>Professionals</td>
</tr>
<tr>
<td></td>
<td>n</td>
<td>Mean ± SD</td>
<td>n</td>
<td>Mean ± SD</td>
</tr>
<tr>
<td>Class I</td>
<td>8</td>
<td>52.6 ± 7.8</td>
<td>8</td>
<td>45.2 ± 11.2</td>
</tr>
<tr>
<td>Class II division 1</td>
<td>9</td>
<td>55.8 ± 11.3</td>
<td>9</td>
<td>49.2 ± 11.9</td>
</tr>
<tr>
<td>Class II division 2</td>
<td>8</td>
<td>60.4 ± 4.6</td>
<td>8</td>
<td>58.3 ± 5.6</td>
</tr>
<tr>
<td>Class III</td>
<td>7</td>
<td>51.7 ± 9.4</td>
<td>7</td>
<td>44.3 ± 9.5</td>
</tr>
<tr>
<td>Total</td>
<td>32</td>
<td>55.3 ± 8.9</td>
<td>32</td>
<td>49.4 ± 10.9</td>
</tr>
</tbody>
</table>
to investigate whether this goal could be met, using sets of three colour photographic views of one individual (one frontal, one three-quarter smiling, and one profile) as a stimulus and a similar set as reference pictures. Such sets of photographs do not express a person’s whole facial attractiveness since dynamic characteristics are not taken into account. Howells and Shaw (1985), however, have shown that a close relationship exists between judgements of facial aesthetics on live stimuli and single colour photographs. Simultaneous presentation of multiple views of an individual might even improve this relationship (Phillips et al., 1992b). An important advantage of the use of these sets of facial photographs is that they are usually available in orthodontic offices.

The measuring system used two reference sets of photographs, one for boys and one for girls. These reference sets were chosen because their aesthetic scores were closest to the mean value found for a series of 40 randomly chosen individuals per gender. The use of these reference sets of photographs resulted in a median intra-observer consistency of 0.68 for the laymen and 0.65 for the professionals, and an ICC of 0.56 for the laymen and 0.53 for the professionals. The value of Pearson’s correlation coefficient, and especially the ICC, was low but acceptable. Low correlations are often an issue when measuring subjective commodities such as facial appearance, and therefore a large number of panel members were asked for their opinion. Pearson’s correlation between the VAS scores given by the laymen and those given by the professionals was 0.92, meaning that good agreement was found between both panels. The concurrent validity of this scale with the Peerlings scale was 0.82 for the laymen and 0.77 for the professionals. This indicates that with the reference sets of photographs as used in the present investigation, an adequate measurement of facial aesthetics can be achieved. The measuring system is reproducible and valid and can therefore be used in future inter-centre audit studies.

A point of discussion in the present study might be the composition of the panels. All laymen had a high social-economic status. The choice of panel was such because orthodontic treatment demand in higher socio-economic groups is greater than in lower socio-economic groups, whereas the objective treatment need is independent of socio-economic status (Wheeler et al., 1994). These laymen may be representative of that part of the general public most often dealing with orthodontic treatment demand. Although the professionals gave lower scores (i.e. were more critical) than the laymen, the correlation coefficient between the panels was very high (0.92). This is in agreement with Lundström et al. (1987) and Kerr and O’Donnell (1990). However, it is in contrast with Lines et al. (1978) and Peerlings et al. (1995) who did not find any difference, and with Tedesco et al. (1983), Phillips et al. (1992a,b), Giddon et al. (1996) and Spyropoulos and Halazonetis (2001), who found that dental professionals were less critical than laymen.

In the present study Angle Class II division 2 patients were considered to be more attractive than Class III patients. In most investigations Class I profile patients were considered to be the most attractive (De Smit and Dermaut, 1984; Kerr and O’Donnell, 1990; Michiels and Sather, 1994; Phillips et al., 1995; Bishara and Jakobsen, 1997; Cochrane et al., 1997; 1999). Michiels and Sather (1994) found Class II profile patients the least attractive, but they also stated that marked cheek bones and lower jaw borders were more often mentioned for the most attractive than for the least attractive patients. This finding might be related to the fact that Class II division 2 profile patients have pronounced cheek bones and lower jaw borders. However, although the scores for the Class II division 2 patients were the highest, it cannot be concluded that they are significantly the most attractive of all patients. It can only be concluded that they are significantly more attractive than Class III patients.

The measuring system will be used in future studies, with a panel of laymen only, since the present study has shown that a high correlation exists between the ratings of professionals and laymen. Moreover, as stated by Bowman and Johnston (2001), the opinion of the end-users of orthodontic services may have the most value in determining the appropriateness of aesthetic results. Therefore, the opinion of laymen on facial aesthetics is valued more highly than that of professionals.

Conclusion

This newly developed measuring system for facial aesthetics in young Caucasian boys and girls, in which two separate sets of reference photographs, one for boys and one for girls were used, showing the frontal, three-quarter smiling, and profile face simultaneously, has been shown to be reproducible and valid, and can therefore be used in future research, and especially in inter-centre audit studies.

Address for correspondence

Professor Anne M. Kuijpers-Jagtman
Department of Orthodontics and Oral Biology
College of Dental Science
Radboud University Nijmegen Medical Centre
Postbus 9101
NL-6500 HB Nijmegen
The Netherlands
E-mail: orthodontics@dent.umcn.nl

Acknowledgements

The authors would like to thank L. Wetzels, F. Rangel, O. Van Vlijmen, K. Windels, V. Akkerman, H. Leeuwenburg, H. van der Laan, and K. Alons for producing the computer images and slide show and for their contribution to the data collection.
We would also like to thank the members of the Dutch Association of Orthodontists, the Belgian Orthodontic Society, the members and their partners of the Rotary Club of Oranjewoud, Heerenveen, The Netherlands, and the Professional Women’s Association, Rollarienes of II Roeselare, Belgium, for their kind co-operation as panel members.

References


Foster E J 1973 Profile preferences among diversified groups. The Angle Orthodontist 43: 34–40


