Q-sort assessment vs visual analog scale in the evaluation of smile esthetics

Brian J. Schabel,[a] James A. McNamara, Jr,[b] Lorenzo Franchi,[c] and Tiziano Baccetti[e]
Santa Cruz, Calif, Ann Arbor, Mich, and Florence, Italy

Introduction: This study was designed to compare the reliability of the Q-sort and visual analog scale (VAS) methods for the assessment of smile esthetics. Furthermore, agreement between orthodontists and parents of orthodontic patients, and between male and female raters, was assessed in terms of subjective evaluation of the smile. **Methods:** Clinical photographs and digital video captures of 48 orthodontically treated patients were rated by 2 panels: 25 experienced orthodontists (15 men, 10 women) and 20 parents of the patients (8 men, 12 women). Interrater reliability of the Q-sort and VAS methods was evaluated by using single-measure and average-measure intraclass correlation (ICC). Kappa agreement and the McNemar test were used to evaluate agreement between orthodontists and parents, and between men and women, for "attractive" and "unattractive" images of smiles captured with clinical photography. **Results:** The single-measure ICC coefficients showed fair to good reliability of the Q-sort and poor reliability of the VAS for measuring esthetic preferences of an individual orthodontist or parent. Both rating groups agreed significantly (*P* >0.05) on the total percentage of "attractive" images of smiles captured with clinical photography. Men and women, however, significantly disagreed on the total percentages of "attractive" and "unattractive" smiles. Women rated higher percentages of both image groups as "attractive" than did their male counterparts. **Conclusions:** The Q-sort was more reliable than the VAS for measuring smile esthetics. Orthodontists and parents of orthodontic patients agreed with respect to "attractive" and "unattractive" smiles. Men and women agreed poorly with respect to "attractive" and "unattractive" smiles. (Am J Orthod Dentofacial Orthop 2009;135:S61-71)

It has been stated that beauty defies measurement. As orthodontists and craniofacial biologists indulge in scientific investigations attempting to define or redefine dentofacial esthetics (the smile in particular), some fundamental questions arise that pertain to the subjective measurement and the reliability of the smile under evaluation. Is this dilemma because a rating method capable of measuring perceptions of esthetics in a valid way has not been used (or assessed) yet, or is beauty of the face or the smile truly immeasurable?

Recently, in response to the notion that the clinician's attitudes and judgments are still the driving force of the decision-making process, research in esthetics has been geared toward evaluating esthetic judgments and values of the general public against professional opinions.[1] These judgments often are contradictory, thus causing a learned professional to question whether these differences are due to variability in study design or the inherent subjectivity in measuring esthetic preference. For example, in consecutive articles published in 2005, Moore et al[2] reported that broader smiles and reduced buccal corridor spaces are preferred by laypersons, whereas Roden-Johnson et al[3] found that laypersons, dentists, and orthodontists showed no preference in buccal corridor spaces, and laypersons had no preference between wide or narrow arch forms. The main reason that these studies cannot be compared directly is that they used different scales of measurement.

Various rating methods have been used[2-25] to assess esthetic preferences applied to dentofacial structures

[a] Graduate orthodontic program, University of Michigan, Ann Arbor; private practice, Santa Cruz, Calif.
[b] Thomas M. and Doris Graber Endowed Professor of Dentistry, Department of Orthodontics and Pediatric Dentistry, School of Dentistry; professor of Cell and Developmental Biology, School of Medicine; and research professor, Center for Human Growth and Development, University of Michigan; private practice, Ann Arbor, Mich.
[c] Assistant professor, Department of Orthodontics, University of Florence, Florence, Italy; Thomas M. Graber Visiting Scholar, Department of Orthodontics and Pediatric Dentistry, School of Dentistry, University of Michigan, Ann Arbor.

The authors report no commercial, financial, or proprietary interest in the products or companies described in this article.

Funded in part by the LeGro Fund and sources made available through the Thomas M. and Doris Graber Endowed Professorship of the University of Michigan.

Reprint requests to: Tiziano Baccetti, Department of Orthodontics, Università degli Studi di Firenze, Via del Ponte di Mezzo, 46-48, 50127, Firenze, Italy; e-mail, t.baccetti@odonto.unifi.it.

Submitted, May 2007; revised and accepted, August 2007.
0889-5406/$34.00

doi:10.1016/j.ajodo.2007.08.019

Table I. Scales used to measure perception of various aspects of dentofacial esthetics

Author, year	*Specific scale used*
Profile esthetics	
Q-sort	
Cox and van der Linden, 1970[4]	29-subject Q-sort
Rating scale	
Riedel, 1950[5]	3-category ordinal scale
De Smit and Dermaut, 1984[6]	9-point ordinal ranking
Barber and Ghafari, 1985[7]	2-category nominal scale
Spyropoulous and Halazonetis, 2001[8]	10-point ordinal ranking
VAS	
Phillips et al, 1992[9]	100 mm; unattractive/attractive
O'Neill et al, 2000[10]	100 mm; % more attractive
Hall et al, 2000[11]	100 mm; unacceptable/acceptable
Altered dental esthetics	
Rating scale	
Romani et al, 1993[12]	2-category nominal scale
Hunt et al, 2002[13]	10-category ordinal scale
Faure et al, 2002[14]	24-point ordinal ranking
Moore et al, 2005[2]	5-category ordinal scale
Soh et al, 2005[15]	7-point ordinal ranking
VAS	
Kokich et al, 1999[16]	50 mm; least to most attractive
Faure et al, 2002[14]	100 mm; relative to reference photo
Roden-Johnson et al, 2005[3]	100 mm; unattractive/attractive
Schlosser et al, 2005[17]	100 mm; least to most attractive
Components of attractive smiles	
Rating Scale	
Hulsey, 1970[18]	5-category ordinal scale
Mackley, 1993[19]	5-point ordinal ranking
Dunn et al, 1996[20]	8-point ordinal ranking
VAS	
Ritter et al, 2006[21]	100 mm; least to most attractive
Parekh et al, 2006[22]	100 mm; % more attractive
Influence of extractions on smile esthetics	
Rating scale	
Johnson and Smith, 1995[23]	5-category ordinal scale
Boley et al, 1998[24]	2-category nominal scale
Kim and Gianelly, 2003[25]	10-category ordinal scale

(Table I), each with its own inherent strengths, limitations, and reliability. These methods can be categorized in 2 groups. The first group of methods comprises rating and rank-order scales, in which outcomes of judgments are not separated by equal intervals. The rating scales delineate differences that are relative (not absolute) and organized ordinally, thus having limitations in terms of statistical evaluation. The second group of methods includes the visual analog scale (VAS) and the Q-sort method, which are intended to create a normal or quasi-normal distribution of the outcomes of judgments, respectively. Statistics for data organized in a continuous scale of values can be applied to the results of VAS and Q-sort. Because of their clinical and statistical applicability, these methods have attracted the interest of researchers who evaluated dentofacial esthetics.

The VAS has been used extensively to evaluate opinions regarding various aspects of dentofacial appearance: profile esthetics,[9-11] preferences to altered dental esthetics,[3,14,16,17] and smile esthetics in adolescents seeking orthodontic treatment.[26] The VAS also has been used to appraise facial esthetic preferences of alternate photographic views of the same subject.[9,22,27]

A VAS is a convenient, simple, economical, and rapid method of obtaining value judgments.[9] Ease of use, however, does not mean that these rating scales have no weaknesses or limitations. Raters tend to spread their responses over the entire scale and avoid the ends at the anchor points, independently of the actual preferences.[28-30] Furthermore, raters might be incapable of making equally discriminative judgments at each level of a scale.[29]

The Q-sort method, originally proposed in 1953 by Stephenson,[31] is an alternative approach that generates a valid 9-category ordinal ranking of large samples based on various subjective criteria. Although originally intended for a sample size of 96, the Q-sort technique has been applied to samples of 48 to rank items according to complex, subjective criteria.[32] The Q-sort method uses a progressive forced-choice winnowing of the sample to create a quasi-normal distribution to rate subjects on an esthetic scale from "least pleasing" to "most pleasing."

VAS and Q-sort analyses have methodological differences. VAS scores are absolute, not relative, with raters scoring each object relatively independently from the others. In contrast, in the Q-sort technique all objects are assessed together, and the evaluation is strictly related to that sample of objects.

Reports of the validity of methods to rate esthetic variables in the dentofacial region are lacking; no information is available in the literature with regard to the reliability of subjective measurements of the esthetics of the smile. Therefore, our aim in this study was to compare the reliability of the VAS and the Q-sort methods used to quantify subjective opinions regarding smile esthetics. The more reliable scale served then as the gold-standard, and it was used to quantify and compare the esthetic preferences of orthodontists with those of parents of orthodontic patients, and the preferences of female raters with those of male raters.

MATERIAL AND METHODS

The subjects were recruited from the University of Michigan Graduate Orthodontic Clinic during routine posttreatment appointments (final records or retainer checks). Potential subjects were given a brief introduction to the study and asked whether they would be willing to participate. No subject received compensation for participation.

Each adult subject (≥18 years of age) reviewed and signed a consent form created in accordance with the rules and regulations of the University of Michigan Health Sciences Institutional Review Board. Each subject under the age of 18 reviewed and signed a child's assent form, and a legal guardian reviewed and signed a consent form in accordance with the Institutional Review Board. Each subject also reviewed and completed a consent form created by the University of Michigan in accordance with the Health Insurance Portability and Accountability Act (HIPAA) for the use and disclosure of protected health information.

To be included in the study, subjects needed the following characteristics: (1) age from 12 to 20 years, (2) white ancestry, (3) orthodontic treatment completed within the last 6 months, (4) no missing or malformed teeth, and (5) a complete set of diagnostic posttreatment records including intraoral and extraoral photographs and a good-quality video clip of the smile. The protocol proposed for the study required 48 subjects to satisfy the design of the Q-sort. A test was performed to determine the power of this sample size with respect to correlation tests (type I error = 0.05). For a bivariate normal distribution and a sample size of 48, a test of H_0: $P = 0$ (ie, the correlation coefficient under the null hypothesis) was found to have a power of 0.80 to detect a linear correlation of r = 0.38. Thus, the default sample size for the Q-sort procedure was deemed adequate for testing for correlation.

The extraoral photographic series included photographs of the subject in repose, during smiling, and in profile. For this study, only the smiling photograph was used. All photographs were taken by 1 of 2 dental school staff photographers. A 35-mm SLR camera (EF, Canon, Lake Success, NY) was mounted to a frame set at a fixed distance of 36 inches between the lens and the subject. The camera was connected to a 2-strobe lighting source that illuminated the subject indirectly from a flash that reflected from a photographic umbrella.

Before taking the smiling image, the photographer instructed the subject to smile. The reproducibility of the posed smile derived from a static photograph was shown by Ackerman et al.[33] Each image was captured on Kodak EV-100 slide film (Eastman Kodak, Rochester, NY). The film was developed, and the 2 × 2-in slides were used in this study. The slides were scanned with the Super Coolscan 4000 ED (Nikon, Melville, NY) and imported directly into a commercially available image-editing software program (version 7.0, Adobe Photoshop, Adobe, San Jose, Calif). Each slide was scanned at maximum dpi to enhance the image quality.

A 3 × 5-in template was created to standardize the size and location of each image. The images were opened in Photoshop, and the template was superimposed on the image. The smile images were enlarged until the outer commissures of the lips matched the vertical tick marks inset .75 inch from the border of the template. The smiling images then were positioned so that the maxillary incisal edges coincided with the horizontal line of the template (Fig 1).

After enlarging and positioning the images correctly, the portion of the image outside the template was cropped. The resulting images were edited further in Photoshop by using the healing brush tool to remove blemishes, skin irregularities, or other extraneous marks that could influence the rater when evaluating the image. The images were labeled with a 4-digit number, unique to each subject, obtained from a random number generator. After the number, the photos obtained from still photography were denoted with "p," and photos from digital video clips were denoted with "v." Once the editing was complete, each image was compressed to approximately 150 KB and saved as a JPEG file.

A digital video camera was used to record the dynamic range of each subject's smile, with slight modifications to the protocol of Ackerman and Ackerman.[34] To standardize the technique, a digital video camera (PV-GS200, Panasonic, Knoxville, Tenn) was used in the same location under standard fluorescent lighting; each video clip was obtained by the senior author (B.J.S.). The camcorder was mounted on an adjustable microphone stand and set at a fixed distance of 60 inches from the subject. Each subject was seated and had his or her head positioned so that an imaginary line between the top of the ear and the midpoint between the upper eyelash and eyebrow paralleled the floor. The video camera was adjusted vertically to be directly in line with the subject's mouth, and the zoom feature was used to focus only on the mouth and adjacent soft tissues to protect the subject's anonymity.

Before the video clip was recorded, subjects were given the following instructions: (1) you will be asked to smile and then relax 3 times; (2) when you relax, please touch your lips lightly together; and (3) when you are asked to smile, please smile until you are told to relax again. When the instructions were understood, the recording began. The duration of each video clip was approximately 10 to 15 seconds.

The raw digital video clips of each subject were transferred to a computer by using a commercially available video-editing software package (version 6.0, Premiere, Adobe). This program allowed the streaming video to be converted into individual photographic frames at the rate of approximately 30 frames per second. Thus, a 10-second video resulted in about 300 frames. The frame representing the subject's posed unstrained social smile was selected, as advocated by Ackerman et al[33] and Ackerman and Ackerman.[34] This frame, identified as the "held smile" by the examiner, was 1 of 15 consecutive frames in which the smile did not change. This unedited image was saved as a JPEG file.

Esthetic judgments were given by orthodontists and parents of orthodontic patients. Each panel member was given a background information questionnaire before rating any photographs. The orthodontic panel comprised 25 full-time and part-time orthodontic faculty members from the University of Michigan Graduate Orthodontic Program. This panel consisted of 15 men and 10 women whose clinical experience ranged from 1 to 35 years, with an average of 17 years. Their ages were 29 to 64 years, with an average age of 46 years. Sixteen panelists were certified by the American Board of Orthodontics. Of the 25 orthodontic raters, 13 were graduates of the University of Michigan, and 2 had received their orthodontic specialty training internationally.

The nonorthodontic panel comprised 20 parents whose children were undergoing orthodontic treatment at the University of Michigan Graduate Orthodontic Clinic. This panel consisted of 12 women and 8 men whose ages ranged from 36 to 52 years, with an average of 43 years. Of the 20 parents, 9 had undergone orthodontic treatment themselves. The time since their child had begun orthodontic treatment was 1 to 13 years, with an average of 5 years.

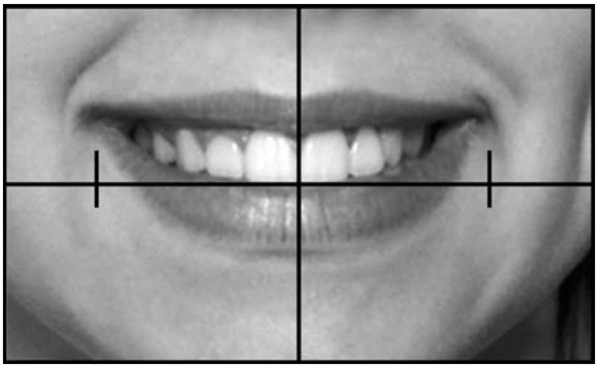

Fig 1. A standardized smile image using the 3 × 5-in template.

At 2 separate rating sessions (T1 and T2), the panelists were asked to use a VAS and the Q-sort technique to rate the attractiveness of 48 of the following photographs: at T1, smile images captured with clinical photography; and, at T2, smile images obtained from digital video clips.

The rating sessions for the parents coincided with consecutively scheduled orthodontic appointments; the rating sessions for the orthodontists occurred sporadically over a 2-month period.

The panelists were asked to rate the attractiveness of each image in the 2 series (smiles taken with clinical photography and smiles obtained from digital video clips) on a 100-mm horizontal VAS anchored on the extremes by "least attractive" and "most attractive." During T1 and T2, the panelists were given a 48-page packet. Each page consisted of a single image of each subject and its corresponding VAS. The pages were arranged in random order, and this sequence was consistent for each rater and the 2 rating sessions.

The panelists were given the following verbal instructions at T1: (1) view each of the 48 images in the packet before rating them; (2) rate the attractiveness of the smile, from the least to the most attractive smile imaginable, using your own standards of attractiveness, by marking a line that intersects the VAS anywhere along the scale; and (3) keeping the packet in order, rate the images in sequence without looking forward or backward. After the verbal instructions, the panelists were given written instructions to review before beginning.

At T2, the same verbal instructions were given as for T1. The panelists were given no additional instructions. They were not told that these were the same subjects evaluated at T1 or that the smiles were obtained from video clips.

The VAS anchors represented the "least attractive" (ie, ugly) at 0 mm and the "most attractive" (ie, beautiful) at 100 mm. The distance between the panelist's mark and the zero anchor point was measured in millimeters and served as an estimate of the esthetic value placed on the image. These distances were measured by using a Dentaguage 2 digital slide caliper (Erkinedental, Marina Del Rey, Calif), and the measurements were recorded to the nearest hundredth of a millimeter.

Stephenson[31] developed the Q-sort as a valid 9-point ordinal ranking technique that allows ranking many objects on the basis of various subjective criteria. Specifically, the panelists were asked to rate the attractiveness of each of 3 series of images using the Q-sort method. The images were the same as those rated with the VAS. During each rating session (T1 and T2), the panelists were given a stack of 48 images in random order.

At T1, the panelists were given the following verbal instructions: (1) from the 48 images, select the 2 least and the 2 most attractive smiles and set them aside; (2) from the remaining 44 photographs, choose the 4 least and the 4 most attractive smile photographs and set them aside; (3) continue this process and set aside 5 and then 8 photographs from each extreme; (4) the remaining 10 photos should represent smiles you consider to have neutral attractiveness; (5) once the Q-sort is completed, survey the distribution and draw a line (cut point) between the 2 columns separating "unattractive" from "attractive" smiles; (6) after drawing your line, leave the Q-sort intact so that it can be scored later.

After the verbal instructions, the panelists were given written instructions to review before beginning. The cut point between "unattractive" and "attractive" smiles was marked on the distribution located on the written instruction form.

At T2, the same verbal instructions were given as for T1. The panelists were given another written instruction form so that they could mark the cut point on the distribution separating the unattractive from attractive smiles. The raters were not told that these were the same subjects evaluated at T1 or that the smiles were obtained from video clips.

With the panelists' Q-sort distribution intact, the 4-digit number for each subject was transferred to a grid so that the photographs could be reused. Each of the 9 groups of images was given a score ranging from 0, for the 2 images with the least perceived attractiveness, to 8, for the 2 images with the greatest perceived attractiveness. The scores each subject received from the various judges were averaged to generate each subject's overall Q-sort score. Each subject received 3 Q-sort scores: smile captured with clinical photography, smile from a digital video clip, and lips in repose. The assignment of scores to each Q-sort distribution is shown in Figure 2. The x-axis represents the score assigned to each photograph in a group; the y-axis lists the number of subjects (photos).

The cut point separating the "unattractive" and "attractive" images also was given a numeric value. For example, a line drawn between columns 2 and 3 on the Q-sort distribution was given a score of 2.5 (Fig 3), which represented the esthetic boundary for that panel member. The cut points from the various judges were averaged to generate the overall demarcation between "attractive" and "unattractive" images.

Statistical analysis

Descriptive statistics (means, standard deviations, and ranges) were calculated for age, VAS scores, Q-sort scores, and esthetic boundary cut-point scores. The Shapiro-Wilks test for normality was used; it showed that these variables were distributed normally. Parametric statistics, therefore, were used for all inferential tests.

Interrater reliability of the Q-sort and VAS methods of assessing esthetic preference was evaluated by using single-measure and average-measure intraclass correlation (ICC). Single-measure ICC is a measurement of the reliability of an individual Q-sort or VAS score; it takes into account the proportion of total variance between raters accounted for by within-rater variation of the individual scores. Simply stated, the single-measure ICC was used to evaluate the likelihood that a rater would generate the same Q-sort or VAS score with repeated measures of the same variables. The average-measure ICC measures the reliability of the average Q-sort or VAS score and takes into account the proportion of total variance between raters accounted for by within-rater variation of the average scores. In short, the average-measure ICC was used to evaluate the likelihood that a group of raters would generate the same average Q-sort or VAS scores with repeated measures of the same variables.

The relationships between average Q-sort and VAS scores for the smile images captured with clinical photography and those from digital video clips were examined by means of the Pearson product-moment correlation, which also was used to assess the relationship among average Q-sort (and VAS) scores of the various raters for images of smiles captured with clinical photography and smiles obtained from digital video clips. Kappa agreement was

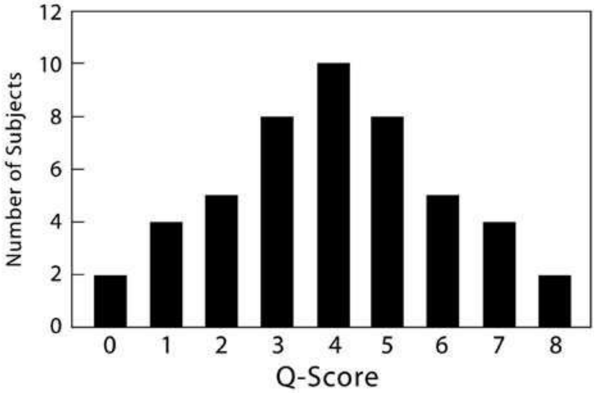

Fig 2. Assignment of scores to the Q-sort distribution.

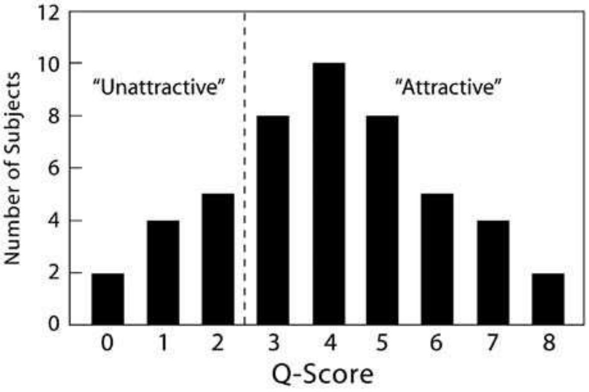

Fig 3. Assignment of scores to the cut-off point used to separate "attractive" from "unattractive" smiles in the Q-sort distribution. A line between columns 2 and 3 was given a score of 2.5.

used to evaluate whether raters agreed as to which images were "attractive" or "unattractive," correcting for agreement that could occur by chance alone. Kappa values had a possible range of 0 to +1. Fleiss[35] characterized different ranges of values for kappa with respect to the level of agreement they suggest; kappa values greater than 0.75 or so represent excellent agreement beyond chance, values below 0.40 represent poor agreement beyond chance, and values between 0.40 and 0.75 represent fair to good agreement beyond chance.

The McNemar test was used to evaluate whether the raters agreed on the combined frequency (ie, percentage) of "attractive" and "unattractive" images. A statistically significant result (P <0.05) indicated that raters disagreed on the percentage of "attractive" images.

The level of significance for all statistical tests was set at 0.05. All statistical tests were performed with a software program (SPSS version 12.0, Statistical Package for the Social Science for Windows, Chicago, Ill).

RESULTS

The subjects ranged in age from 12 years 1 month to 19 years 6 months, with an average age of 15 years. Of the 48 participants, 24 were male and 24 were female.

Standard descriptive statistics were calculated for the average VAS and Q-sort assessments of the various images and the esthetic boundary scores (Table II). Average VAS and Q-sort scores were calculated by combining the results of the respective raters for each subject. Esthetic boundary scores represent the average cut point between "unattractive" and "attractive" photographs during the Q-sort assessment.

Because of the ordinal nature and the normal distribution of the Q-sort, the mean Q-sort score was 4 when scores of all subjects were combined. The VAS scores ranged from 0 to 100, and the Q-sort scores and esthetic boundary scores ranged from 0 to 8. For each image, the mean VAS scores of the orthodontic raters resided near the midpoint of the scale (ie, 50: neutral attractiveness), whereas the mean VAS rating of the parents was closer to the unattractive extreme (ie, the mean VAS scores were about 40).

Interrater reliability of the Q-sort and VAS assessments was determined by using single-measure and average-measure ICCs, obtained from the orthodontic and parent raters with respect to the various images; they are summarized in Table III. The single-measure ICC coefficients demonstrated fair to good reliability of the Q-sort and poor reliability of the VAS for measuring esthetic preferences of an individual orthodontist or parent. The average-measure ICC coefficients had excellent reliability of both the Q-sort and VAS; however, the Q-sort was more reliable for measuring esthetic preferences of a group of orthodontists or parents. The Q-sort and VAS scores of smiles captured with clinical photography were highly variable among raters, reflecting the subjectivity of esthetic assessment.

Pearson product-moment coefficients of linear correlation, r, were used to characterize the relationship between average Q-sort and VAS scores of images of smiles captured with clinical photography and smiles obtained from digital video clips. Correlation coefficients vary from -1 to +1. The strength of the relationship is indicated by the size of the coefficient, whereas its direction is indicated by the sign. The strength of these relationships (r = 0.96 for both assessments) suggested that average Q-sort and VAS scores were strongly related.

Pearson correlation coefficients were calculated to characterize the relationship between average

Table II. Descriptive statistics for average Q-sort, VAS, and esthetic boundary scores of the various raters for each image type

			Range	
Variable and rater	*Mean*	*SD*	*Min*	*Max*
Smile photo				
Average VAS score (range, 0-100)				
Orthodontists	52.2	11.1	24.4	76.8
Parents	41.3	11.7	16.0	66.5
Average Q-sort score (range, 0-8)				
Orthodontists	4*	1.4	0.5	6.7
Parents	4*	1.4	0.7	7.1
Average esthetic boundary (range, 0-8)				
Orthodontists	3.2	1.3	0.5	6.5
Parents	3.7	1.7	1.5	6.5
Smile video				
Average VAS score (range, 0-100)				
Orthodontists	51.3	12.5	19.7	75.1
Parents	42.1	11.1	15.3	64.2
Average Q-sort score (range, 0-8)				
Orthodontists	4*	1.5	0.3	6.6
Parents	4*	1.4	0.5	6.5
Average esthetic boundary (range, 0-8)				
Orthodontists	3.1	1.3	0.5	6.5
Parents	3.3	1.7	1.5	6.5

*The normal distribution of the Q-sort results in a mean of 4 when subjects are combined.
Min, Minimum; *Max*, maximum.

Q-sort and VAS scores of the various rating groups (male orthodontists, female orthodontists, male parents, and female parents) for smiles captured with clinical photography (Table IV). Stronger positive correlations were found among average Q-sort scores (r values, 0.78-0.92) than among average VAS scores (r values, 0.69-0.86) for each rating group. Because both the single-measure and average-measure ICCs suggested that the Q-sort was more reliable than the VAS, and that the Pearson correlation coefficients were consistently higher among average Q-sort than VAS scores, average Q-sort scores were used for the rest of the study to represent the esthetic values of the various images.

Kappa agreement and the McNemar test were used to evaluate the level of agreement between orthodontists and parents for "attractive" and "unattractive" images of smiles captured with clinical photography. The images were categorized as "attractive" or "unattractive" depending on whether the average Q-sort scores of the respective raters were above or below their average cut point, respectively. A statistically significant result (P <0.05) of the McNemar test indicated that the raters disagreed on the percentage of "attractive" images.

Orthodontists and parents showed fair to good agreement beyond chance when evaluating smiles captured with clinical photography (kappa value = 0.56; P <0.001). Both rating groups agreed significantly (P >0.05) on the total percentage of "attractive" images of smiles captured with clinical photography (P values from the McNemar test = 0.11). Interestingly, orthodontists rated a higher percentage of both groups of images as "attractive" than did the parent raters.

Kappa agreement and the McNemar test also were used to evaluate agreement between the sexes for "attractive" and "unattractive" images of smiles captured with clinical photography. Men and women had relatively poor agreement beyond chance between "attractive" and "unattractive" smiles captured with clinical photography (kappa value = 0.43; P <0.001). Men and women also significantly disagreed on the total percentages of "attractive" and "unattractive" smiles captured with clinical photography (P value from the McNemar test = 0.003). Interestingly, women rated a higher percentage of both image groups as "attractive" than did their male counterparts.

DISCUSSION

Our aim in this study was to evaluate the reliability of and the relationship between the Q-sort and VAS methods used to measure esthetic preferences. A collateral objective was to evaluate agreement between orthodontists and parents of orthodontic patients and between the sexes with respect to the attractiveness ratings of the smile.

The smile images obtained from clinical photographs and the digital video clips were rated near the

Table III. Single-measure and average-measure ICC: reliability of the Q-sort and VAS methods of assessing esthetic preference

	Single-measure ICC		*Average-measure ICC*	
Rater and image	*Q-sort*	*VAS*	*Q-sort*	*VAS*
Orthodontists (n = 25)				
Smile from clinical photograph	0.50	0.29	0.96	0.91
Smile from digital video clip	0.57	0.32	0.97	0.92
Parents (n = 20)				
Smile from clinical photograph	0.49	0.30	0.95	0.91
Smile from digital video clip	0.51	0.28	0.95	0.89

midpoint of the VAS scale by orthodontists (mean VAS scores of 52 and 51, respectively) and closer to the "least attractive" anchor by parents (mean VAS scores of 41 and 42, respectively).

Howells and Shaw[37] evaluated the relationship between dental and facial attractiveness of photographic images and live patients rated with a VAS. For ratings of dental attractiveness, the mean VAS score of the photographic images was 50, and the mean rating after live viewing was 49. These findings, along with our results, support the claims that raters avoid the extremes of the VAS scale and that mean VAS scores tend to cluster around the midpoint, independently of actual preferences.[28-30]

Consistent with our study, other investigators have reported that orthodontists tend to rate the esthetics of various dentofacial attributes higher than do laypeople.[3,9,26] Differences in the mean VAS scores between orthodontists and parents of orthodontic patients suggest that the general public is more critical about the esthetics of the smile. These perceptions might be derived from the "Hollywood" smile in the media. Peck and Peck[38] reported that the layman's perception of esthetics is purely driven by external observations and that the mass media is influential in unifying the general public's perceptions of beauty. Orthodontists' perceptions of smiles, on the contrary, might be driven by previous experience—what is a reasonable orthodontic result rather than what is possible with other forms of cosmetic dentistry.

Smile preferences of orthodontists and laypersons in subjects seeking orthodontic treatment were evaluated by McNamara et al[26] using a VAS. Interestingly, the mean, minimum, and maximum VAS scores generated from the pretreatment smiles were similar to those of the posttreatment smiles in our study.

Raters used Q-sort and VAS to specify their esthetic preferences with various images (smiles from clinical photography and digital video clips) of all 48 subjects. Because the raters evaluated 96 images (2 images of each of the 48 subjects) using 2 scales, it did not seem reasonable to measure intrarater reliability, since it would have required the raters to repeat their measurements. Interrater reliability was examined by using single-measure and average-measure ICC. The single-measure ICC coefficient is an index of the reliability of a Q-sort or VAS score of 1 rater. The results of the single-measure ICC suggest that the panelists were more likely to rate the esthetics of individual images similarly using the Q-sort (single-measure ICC coefficients, 0.49-0.57) than the VAS (single-measure ICC coefficients, 0.24-0.32) if they had repeated their measurements.

The average-measure ICC coefficient is an index of the reliability of mean ratings of a group of raters. The strength of the average-measure ICC coefficients suggests that mean Q-sort and VAS scores would most likely be similar if the 45 panelists repeated their measurements using these scales. Although excellent reliability was found for both scales, the Q-sort was shown to be more reliable than the VAS (ICC coefficients, 0.95-0.97 and 0.86-0.92, respectively). The variability of a measurement decreases when the results of many raters are combined, resulting in a higher average-measure ICC coefficient.

The relationship between the average Q-sort and VAS scores of the various images (smiles from clinical photography and digital video clips) were examined with the Pearson product-moment correlation. The strength of these relationships (r = 0.96) suggested that average scores of Q-sort and VAS were in basic agreement with respect to evaluating esthetic preference.

The strength of the correlation between the average Q-sort and VAS scores could be due in part to the large number of raters (n = 45). It has been reported that correlation coefficients increase simply by adding more raters. For example, Howells and Shaw[37] used a VAS to measure preferences of facial esthetics; they reported that correlation coefficients of average VAS scores increased from 0.67 to 0.86 when the number of panelists increased from 2 to 22.

Our results are difficult to compare with previous investigations because no other study has evaluated

Table IV. Correlation among average Q-sort and VAS scores of the various rating groups for smile images captured with clinical photography

	Orthodontists		*Parents*	
Smiles from clinical photography	*Female*	*Male*	*Female*	*Male*
Average Q-sort score				
Female orthodontists				
Correlation (r)	1	0.92*	0.78*	0.79*
Male orthodontists				
Correlation (r)	0.92*	1	0.83*	0.87*
Female Parents				
Correlation (r)	0.78*	0.83*	1	0.90*
Male Parents				
Correlation (r)	0.79*	0.87*	0.90*	1
Average VAS score				
Female orthodontists				
Correlation (r)	1	0.86*	0.69*	0.70*
Male orthodontists				
Correlation (r)	0.86*	1	0.81*	0.83*
Female Parents				
Correlation (r)	0.69*	0.81*	1	0.84*
Male Parents				
Correlation (r)	0.70*	0.83*	0.84*	1

*P <0.001; mean r for Q-sort = 0.85; mean r for VAS = 0.79.

the relationship between the VAS and Q-sort methods. Faure et al[14] evaluated the reliability of 3 scales used to measure facial esthetics: a 100-mm VAS, a ratio scale in which photographs were scored in reference to a photograph deemed to have a mean score of 60 of a possible 100, and a ranking scale used to organize photographs from least to most attractive. They reported that panelists rated the photos "equally" with the 3 scales, although no correlation values were given to support this.

High variability was found among Q-sort and VAS scores in this study. This observation could be expected because of the subjectiveness of esthetics and the images of 48 subjects. Many studies, however, used a VAS to evaluate esthetic preferences of smiles obtained from a few subjects that were altered digitally in small increments.[2,3,14,16,17,22] For these studies, variability of the scores would likely be small; therefore, a more reliable scale (eg, the Q-sort) might have been better suited for differentiating between the esthetic scores of similar images.

It has been reported that the VAS is a rapid, convenient, and reliable method of rating dental and facial appearance.[37] Phillips et al[9] suggested that ranking procedures (Q-sort or other categorical rating scales) rather than the VAS should be used to evaluate preferences because the VAS can introduce a level of precision beyond the discriminatory ability of the judges. The use of ranks also eliminates the problem that the distribution of scores in a continuous scale (eg, VAS) would not be the same for all judges; this can occur when certain portions of the scale are neglected or intervals in the rating scale are not viewed as having equal value.[39]

Statistically significant chance corrected agreement (kappa P value <0.001) was found between "attractive" and "unattractive" images of smiles captured with clinical photography, as perceived by orthodontists and parents. The agreement between orthodontists and parents coincides with the findings of previous investigations that examined agreement between orthodontists and laypersons in regard to various aspects of dentofacial appearance: smile esthetics,[3,20-22] specific dental attributes,[40] facial attractiveness,[38] and profile esthetics.[4,11,15] Others, however, concluded that the perceptions of orthodontists do not correspond specifically to those of laypeople.[12,16,27] To support this contention, it was suggested that orthodontists might be biased by their formal training in esthetics, which could lead to differences of opinion between them and laypeople.[3]

The McNemar test showed significant agreement between orthodontists and parents with respect to the combined frequency (percentage) of "attractive" images of smiles captured with clinical photography (P >0.05). Orthodontists rated a higher percentage of both groups of images as "attractive" than did the parents, although the differences were not statistically significant. Similarly, other investigators reported that orthodontists were more likely than laypersons to rate the esthetics of various dentofacial attributes as "attractive."[9,41] Contrary to our findings, however, Shaw et al[42] reported that orthodontists are more critical than laypeople regarding the esthetics of photographs of the dentition. This report, however, did not specify whether these differences were statistically significant.

Male and female raters were found to have significant chance corrected agreement (kappa *P* values <0.001) between "attractive" and "unattractive" images of smiles captured with clinical photography (kappa values = 0.43). The kappa values calculated from the male and female ratings were strikingly similar to those of orthodontists and parents, suggesting that opinions of smile esthetics were consistent among all raters. Agreement between the sexes with respect to smile esthetics in this study is consistent with the findings of previous investigators who have examined agreement between men and women for various aspects of dentofacial appearance: smile esthetics[2,18] and profile esthetics.[6,10] Men and women did not agree on the percentages of attractive and unattractive images of smiles captured with clinical photography; this difference was significant (*P* <0.01). Women were more likely than men to judge the images of smiles captured with clinical photography as "attractive." This finding is in line with other investigators who reported that female raters were more likely to judge various dentofacial attributes more attractive than the male raters.[7,37]

CONCLUSIONS

The findings of this investigation are as follows.

1. The Q-sort was more reliable than the VAS for measuring the esthetics of the smile.
2. Orthodontists and parents of orthodontic patients agreed with respect to "attractive" and "unattractive" smiles. Orthodontists rated more smiles "attractive," but this difference was not significant.
3. Men and women agreed poorly with respect to "attractive" and "unattractive" smiles. Women rated significantly more smiles as "attractive" than males.

REFERENCES

1. Hayes C. The use of patient based outcome measures in clinical decision making. Community Dent Health 1997;15:19-21.
2. Moore T, Southard KA, Casko JS, Qian F, Southard TE. Buccal corridors and smile esthetics. Am J Orthod Dentofacial Orthop 2005;127:208-13.
3. Roden-Johnson D, Gallerano R, English J. The effects of buccal corridor spaces and arch form on smile esthetics. Am J Orthod Dentofacial Orthop 2005;127:343-50.
4. Cox NH, van der Linden FPGM. Facial harmony. Am J Orthod 1971;60:175-83.
5. Riedel RA. The relation of maxillary structures to cranium in malocclusion and normal occlusion. Angle Orthod 1952;22:142-5.
6. De Smit A, Dermaut L. Soft-tissue profile preferences. Am J Orthod 1984;86:67-73.
7. Barber JG, Ghafari J. Silhouette profiles in the assessment of facial esthetics: a comparison of treated cases with various orthodontic appliances. Am J Orthod 1985;87:385-91.
8. Spyropoulous MN, Halazonetis DJ. Significance of the soft tissue profile on facial esthetics. Am J Orthod Dentofacial Orthop 2001;119:464-71.
9. Phillips C, Tulloch C, Dann C. Rating of facial attractiveness. Community Dent Oral Epidemiol 1992;20:214-20.
10. O'Neill K, Harkness M, Knight R. Ratings of profile attractiveness after functional appliance treatment. Am J Orthod Dentofacial Orthop 2000;118:371-6.
11. Hall D, Taylor RW, Jacobson A, Sadowsky PL, Bartolucci A. The perception of optimal profile in African Americans versus white Americans as assessed by orthodontists and the lay public. Am J Orthod Dentofacial Orthop 2000;118:514-25.
12. Romani KL, Agahi F, Nanda R, Zernik JH. Evaluation of horizontal and vertical differences in facial profiles by orthodontists and lay people. Angle Orthod 1993;63:175-82.
13. Hunt O, Johnston C, Hepper P, Burden D, Stevenson M. The influence of maxillary gingival exposure on dental attractiveness ratings. Eur J Orthod 2002;24:199-204.
14. Faure JC, Rieffe C, Maltha JC. The influence of different facial components of facial aesthetics. Eur J Orthod 2002;24:1-7.
15. Soh J, Chew MT, Wong HB. Professional assessment of facial profile attractiveness. Am J Orthod Dentofacial Orthop 2005;128:201-5.
16. Kokich VO, Kiyak HA, Shapiro PA. Comparing the perception of dentists and lay people to altered dental esthetics. J Esthet Dent 1999;11:311-24.
17. Schlosser JB, Preston CB, Lampasso J. The effects of computer-aided anteroposterior maxillary incisor movement on ratings of facial attractiveness. Am J Orthod Dentofacial Orthop 2005;127:17-24.
18. Hulsey CM. An esthetic evaluation of lip-teeth relationships present in the smile. Am J Orthod 1970;57:132-44.
19. Mackley RJ. An evaluation of smiles before and after orthodontic treatment. Angle Orthod 1993;63:183-9.
20. Dunn WJ, Murchison DF, Broome JC. Esthetics: patients' perceptions of dental attractiveness. J Prosthodont 1996;5:166-71.
21. Ritter DE, Gandini LG, Pinto Ados S, Locks A. Esthetic influence of negative space in the buccal corridor during smiling. Angle Orthod 2006;76:198-203.
22. Parekh SM, Fields HW, Beck M, Rosenstiel S. Attractiveness of variations in the smile arc and buccal corridor space as judged by orthodontists and laymen. Angle Orthod 2006l;76:557-63.
23. Johnson DK, Smith RJ. Smile esthetics after orthodontic treatment with and without extraction of four first premolars. Am J Orthod Dentofacial Orthop 1995;108:162-7.
24. Boley JC, Pontier JP, Smith S, Fulbright M. Facial changes in extraction and nonextraction patients. Angle Orthod 1998;68:539-46.
25. Kim E, Gianelly AA. Extraction vs. nonextraction: arch widths and smile esthetics. Angle Orthod 2003;73:354-8.
26. McNamara L, McNamara JA Jr, Ackerman MB, Baccetti T. Hard and soft tissue contributions to the esthetics of the posed smile in growing patients seeking orthodontic treatment. Am J Orthod Dentofacial Orthop 2008;133:491-9.
27. Flores-Mir C, Silva E, Barriga MI, Lagravere MO, Major PW. Lay person's perception of smile aesthetics in dental and facial views. J Orthod 2004;31:204-9.
28. Weinstein MC, Fineberg HV. Clinical decision analysis. Philadelphia: Saunders; 1980.
29. Kaplan RM, Ernst JA. Do category rating scales produce biased preference weights for a health index? Med Care 1983;21:193-207.
30. Bleichrodt H, Johannesson M. An experimental test of the theoretical foundation for the rating-scale valuations. Med Decis Making 1997;17:208-16.
31. Stephenson W. The study of behavior. Chicago: University of Chicago Press; 1953.

32. Coy K, Speltz ML, Jones K. Facial appearance and attachment in infants with orofacial clefts: a replication. Cleft Palate Craniofac J 2002;39:66-72.
33. Ackerman JL, Ackerman MB, Brensinger CM, Landis JR. A morphometric analysis of the posed smile. Clin Orthod Res 1998;1:2-11.
34. Ackerman JL, Ackerman MB. Digital video as a clinical tool in orthodontics: dynamic smile analysis and design in diagnosis and treatment planning. In: McNamara JA Jr, editor. Information technology and orthodontic treatment. Monograph 40. Craniofacial Growth Series. Ann Arbor: Center for Human Growth and Development; University of Michigan; 2003. p. 195-203.
35. Fleiss JL. Statistical methods for rates and proportions. 2nd ed. New York: John Wiley & Sons; 1981.
36. Fleiss JL. The design and analysis of clinical experiments. New York: John Wiley & Sons; 1986.
37. Howells DJ, Shaw WC. The validity and reliability of ratings of dental and facial attractiveness for epidemiologic use. Am J Orthod 1985;88:402-8.
38. Peck H, Peck S. A concept of facial esthetics. Angle Orthod 1970;40:284-318.
39. Conover WJ, Iman RL. Rank transformation as a bridge between parametric and nonparametric statistics. Am Statistician 1981;35:124-9.
40. Brisman AS. Esthetics: a comparison of dentists' and patients' concepts. J Am Dent Assoc 1980;100:345-52.
41. Bell R, Kiyak HA, Joondeph DR, McNeill RW, Wallen TR. Perceptions on facial profile and their influence on the decision to undergo orthognathic surgery. Am J Orthod 1985;88:323-32.
42. Shaw WC, Richmond S, O'Brien KD, Brook PH. Quality control in orthodontics: indices of orthodontic treatment need and treatment standards. Br Dent J 1991;170:107-12.