Development and Psychometric Evaluation of the FACE-Q Satisfaction with Appearance Scale

A New Patient-Reported Outcome Instrument for Facial Aesthetics Patients

Andrea L. Pusic, MD, MD, MHS, FRCSC[a,*],
Anne F. Klassen, DPhil, BA[b], Amie M. Scott, MPH[a],
Stefan J. Cano, PhD[c]

KEYWORDS

• Facial cosmetic surgery • Aesthetic surgery • Outcomes • Quality of life • Patient satisfaction • Psychometrics • Questionnaire • Rasch measurement

KEY POINTS

- Accurate and reliable measurement of patient-centered outcomes is critical to ongoing practice improvement and clinical research in facial aesthetics.
- Modern psychometric methods overcome the limitations of traditional psychometric methods by providing clinically meaningful interval-level data.
- The FACE-Q Satisfaction with Facial Appearance scale is a new-generation condition-specific patient-reported outcome instrument, capable of providing clinically meaningful and scientifically sound data reflecting patient perceptions of outcome.

BACKGROUND

Facial aesthetics procedures are an important area of continued growth in plastic surgery; 13.8 million cosmetic procedures were performed in the United States in 2011, an increase of 5% from 2010.[1] Rhinoplasty (n = 244,000) and blepharoplasty (n = 196,000) were second and third to breast augmentation (n = 307,000) in popularity. Botulinum toxin type A (n = 5.7 million), soft tissue fillers (n = 1.9 million) and chemical peels (n = 1.1 million) were the top three cosmetic minimally invasive procedures.[1]

Specially designed questionnaires known as patient-reported outcome (PRO) instruments, developed to measure a range of outcomes (eg, symptoms, satisfaction, body image, and quality

Disclosure: The FACE-Q © is owned by Memorial Sloan-Kettering Cancer Center (MSKCC). Drs Cano, Klassen, and Pusic are codevelopers of the FACE-Q © and, as such, receive a share of any license revenues based on MSKCC's inventor sharing policy.

This study was funded by grants from the Plastic Surgery Education Foundation.

[a] Memorial Sloan-Kettering Cancer Center, Room MRI-1007, 1275 York Avenue, New York, NY 10065, USA; [b] McMaster University, Department of Pediatrics, 3N27, 1280 Main Street W, Hamilton, ON L8S 4K1, Canada; [c] Plymouth University Peninsula Schools of Medicine and Dentistry, Clinical Neurology Research Group, Tamar Science Park, Plymouth PL6 8BX, UK

[*] Corresponding author. Clinical Neurology Research Group, Plastic and Reconstructive Surgery, Memorial Sloan-Kettering Cancer Center, Tamar Science Park, Room MRI-1007, 1275 York Avenue, New York, NY 10065, USA.
E-mail address: pusica@mskcc.org

Clin Plastic Surg 40 (2013) 249–260
http://dx.doi.org/10.1016/j.cps.2012.12.001

of life), have become a mainstay of clinical research in all areas of medicine and surgery.[2–4] To provide meaningful measurement, such PRO instruments must be shown to be reliable, valid, and responsive (**Table 1**).[2,5] Although understanding the patient's perspective is especially important in facial aesthetics, a systematic review performed by our team identified that there is a lack of reliable and valid PRO instruments available for measuring the range of issues important to facial aesthetic patients.[6] We therefore set out to develop a new PRO instrument following the methodology we previously used to develop other plastic surgery–specific PRO instruments.[7–9] This new PRO instrument is called the FACE-Q and includes a range of separate scales that measure important outcomes for patients having any type of facial cosmetic surgery, minimally invasive cosmetic procedure, or facial injectable.

This article describes the development and psychometric evaluation of the core FACE-Q scale, called the Satisfaction with Facial Appearance scale.

QUALITATIVE AND QUANTITATIVE METHODS

We obtained local institutional ethics review board approval before commencing our study. The content for the Satisfaction with Facial Appearance scale was developed as part of a larger suite of scales that cover a range of concepts important to facial aesthetics patients.[10] These scales were constructed with strict adherence to recommended guidelines for PRO instrument development.[11–15] The guidelines outline three phases required to develop a scientifically credible and clinically meaningful tool.

In the first phase, a conceptual framework is formally defined, and a pool of items is generated. These items are developed from the following three sources: review of the literature, qualitative patient interviews, and expert opinion. The item pool is developed into a series of scales that are pilot tested in the target participant sample to clarify ambiguities in item wording, confirm appropriateness, and determine acceptability and completion time. This phase of our research is described in a separate publication[10] and is summarized later in this paper. In the second phase (the main focus of this article), the scales undergo psychometric evaluation in a large sample of target subjects. Questions representing the best indicators of outcome are retained based on their performance against a standardized set of psychometric criteria. In the third phase, further psychometric evaluation is performed by administering the item-reduced scales to a large sample of participants to further examine their scientific soundness.[16,17]

Phase 1: Qualitative Phase

Qualitative interviews were conducted with 50 patients recruited from 7 offices of plastic surgeons and dermatologists practicing in New York (United States) and Vancouver (Canada) between January 2008 and February 2009. Participants ranged in age from 20 to 79 years (mean age 51 years) and had undergone 1 or more of the following facial procedures: botulinum toxin (n = 20), resurfacing (n = 15), filler (n = 15), blepharoplasty (n = 25), facelift (n = 22), rhinoplasty (n = 9), neck lift (n = 8), brow lift (n = 4), and chin implant (n = 2).

Patients were interviewed using open-ended questions. Interviews were digitally recorded and transcribed verbatim and coded within NVivo8 software[18] using a line-by-line coding approach. Data collection and analysis took place concurrently to gather data to refine emerging codes and categories. Data analyses led to the development of a conceptual framework that depicts important concepts for facial aesthetic patients (**Fig. 1**).

To develop scales with items covering the concepts in **Fig. 1**, we examined codes (ie, key phrases expressed by patients) and linked these to specific patient characteristics (eg, type of procedure, age, and gender). Attaching key patient characteristics to each code provided the information needed to develop core items (common to all patients), and unique items (specific to a subgroup). To develop a set of scales, we then iteratively and interactively examined the item lists developed from the coded material to identify a set of items that mapped out a continuum for each major concept. For each item we examined Flesch-Kincaid grade level scores[19] and adjusted as necessary to ensure the lowest possible grade level for reading. Scale instructions and appropriate response options were then developed for each scale.

The scales were then presented to 26 experts (15 plastic surgeons, 4 dermatologists, 3 psychologists, 4 office staff) to further appraise and refine. In addition, 35 facial aesthetic patients participated in one-on-one cognitive debriefing interviews to identify any ambiguous wording and confirm appropriateness, acceptability, and completion time of the preliminary scales. The process resulted in the development of a set of independently functioning scales that measure the concepts forming the conceptual framework (**Table 2**).

Table 1
Glossary of terms

Term	Definition
Ad hoc questionnaire	A PRO instrument that has not been developed and/or validated using acknowledged guidelines.[6,17,49–51] Such PRO instruments may pose clinically reasonable questions, but one cannot be confident about their reliability (ie, ability to produce consistent and reproducible scores) or validity (ie, ability to measure what is intended to be measured)
Conceptual framework	The expected relationships of items within a domain and between domains within a PRO concept. The validation process confirms the conceptual framework
Domain	A domain is a collective word for a group of related concepts. All the items in a single domain contribute to the measurement of the domain concept
Generic questionnaires	PRO instruments that can be used in any patient group regardless of their health condition, and allow direct comparisons across disease groups and/or healthy populations. An example of a generic questionnaire is the Short Form 36-Item Health Survey, which is the most widely used generic measure in the world[52]
Health-related quality of life	In quality-of-life measurement, the terms quality of life, health status, health-related quality of life, and functional status are often used interchangeably. Although there is a lack of conceptual clarity regarding these terms,[53] there is broad agreement on the core minimum set of health concepts that should be measured.[54] These concepts include physical health, mental health, social functioning, role functioning, and general health perceptions
Item	An individual question, statement, or task that is evaluated by the patient to address a particular concept
PRO instrument	A questionnaire used in a clinical or research setting in which responses are collected directly from patients. These questionnaires quantify aspects of health-related quality of life and/or significant outcome variables (eg, patient satisfaction, symptoms) from the patient's perspective.[17] PRO instruments provide a means of quantifying the way patients perceive their health and the impact treatments have on their quality of life
Reliability	An important property of a PRO instrument because it is essential to establish that any changes observed in patient groups are attributable to the intervention or disease and not to problems in the measure. Test-retest reliability may be evaluated by having individuals complete a questionnaire on more than 1 occasion over a time period when no changes in outcome are expected to have occurred. Commonly reported reliability statistics include the Cronbach alpha[39] and intraclass correlation coefficients[16]
Responsiveness	The ability of an instrument to accurately detect change. Responsiveness is an important psychometric property when evaluating change as the result of a health care intervention or when following patients over time. Responsiveness is usually examined by comparing preintervention and postintervention scores using standardized change indicators, such as effect size statistics[41]
Scale	The system of numbers or verbal anchors by which a value or score is derived. Examples include visual analog scales, Likert scales, and rating scales
Scientific soundness	Refers to the demonstration of reliable, valid, and responsive measurement of the outcome of interest
Score	A number derived from a patient's response to items in a questionnaire. A score is computed based on a prespecified, validated scoring algorithm and is subsequently used in statistical analyses of clinical study results. Scores can be computed for individual items, domains, or concepts, or as a summary of items, domains, or concepts
Validity	The ability of an instrument to measure what is intended to be measured. Establishment of validity may be considered an ongoing process. A PRO instrument is examined from various angles, including an assessment of the development process, consideration of known group differences, evaluation of internal consistency, and evaluation of both convergent and discriminant validity relative to other existing related measures

Adapted from Food and Drug Administration. Patient reported outcome measures: use in medical product development to support labeling claims. 2009;11:31–3. Available at: www.fda.gov/cber/gdlns/prolbl.pdf; and Cano S, Klassen A, Pusic A. The science behind quality-of-life measurement: a primer for plastic surgeons. Plast Reconstr Surg 2009;123:99–102e; with permission.

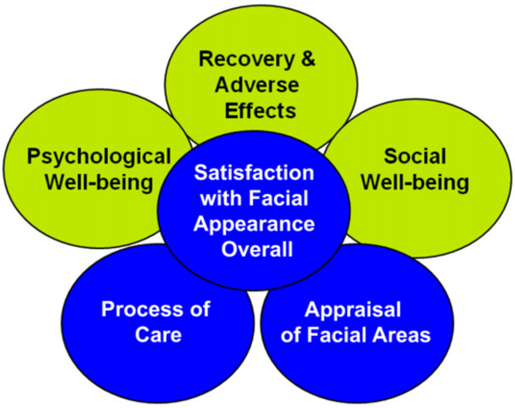
Recovery & Adverse Effects
Psychological Well-being
Social Well-being
Satisfaction with Facial Appearance Overall
Process of Care
Appraisal of Facial Areas

Fig. 1. FACE-Q conceptual framework.

Phase 2: Quantitative Phase

Data were collected and analyzed to identify the items representing the best indicators for each scale based on their performance against a standardized set of psychometric criteria. Data came from 2 separate studies, and were compiled for the purpose of psychometric analyses. Results presented in this article relate only to the Satisfaction with Facial Appearance scale. This scale was developed for use in research and clinical practice to compare outcomes across procedure types and/or to measure change before and after any facial aesthetic procedure. Future publications will present psychometric findings for the other FACE-Q scales.

Study 1

Data were collected from patients of 10 plastic surgeons and 2 dermatologists representing 10 different practices in the United States (New York, Washington, St Louis, Dallas, and Atlanta) and Canada (Vancouver) between June 2010 and June 2012. Eligible participants included those who could read English; were 18 years of age or older; and were planning to undergo, or had already undergone, any surgical or nonsurgical facial aesthetic procedure.

Given the large number of FACE-Q scales that were developed in the initial phases of research, we grouped scales into booklets based on common surgical and nonsurgical procedures and distributed these to the participating practices. All booklets included the Satisfaction with Facial Appearance scale. Instructions for this scale asked patients to answer a series of items based on "how you look right now" and to complete each item with their "entire face in mind." The 4 response options were as follows: very dissatisfied, somewhat dissatisfied, somewhat satisfied, and very satisfied. Patient responses to items in each scale are converted to a summary score which ranges from 0 to 100. A higher score indicates higher satisfaction or better quality of life.

Patients from 6 surgical practices were recruited at the time of their appointment and asked to complete a questionnaire booklet in the waiting room before their appointment. Patients from 4 practices were invited to participate in a postal survey. To ensure a high response rate, a personalized letter from the relevant health care provider

Table 2
FACE-Q scales

Appearance appraisal scales	Facial appearance overall[a,*] Skin Lines overall Forehead lines Forehead and eyebrows Lines between eyebrows Eyes (overall, double eyelid, upper and lower eyelids) Crow's feet Eyelashes Cheekbones Cheeks Ears Nasal bridge Nose Nasolabial folds Lips Lip lines Marionette lines Chin Lower face/jawline Under Chin Neck
Quality of life scales*	Psychological wellbeing Social well-being Age appraisal Expectations and motivations Psychological distress Recovery early life impact*
Adverse effect checklists for treatment	Recovery early symptoms Skin Forehead, scalp and eyebrows Eyes Nose Lower face and neck Lips Ears
Process of care scales*	Decision Doctor Information Office staff Office appearance

[a] see Table 4 for scale's content.
* Relevant scales for all patients.

was included with the appropriate FACE-Q booklet and up to 3 mailed reminders were sent as necessary.[20,21] All patients invited into the study were given a gift card ($5) to thank them for their participation.

Study 2

A medical device company was provided with the Satisfaction with Facial Appearance scale alongside other FACE-Q scales relevant to measuring the concerns of patients having facelifts for a clinical trial involving 100 patients from France, Germany, the United Kingdom, and Israel. Patients completed FACE-Q scales before and after surgery. MAPI (MArchés et Prospectives Internationaux [International Prospects and Markets in English]) Research Trust[22] provided translations and linguistic validation of the FACE-Q scales. This process ensured that the concepts measured by the FACE-Q scales are equivalent across languages (ie, English, German, French, and Hebrew) and easily understood by the people in the target country. In brief, MAPI uses a process based on translation principles as detailed by the European Regulatory Issues and Quality of Life Assessment (ERIQA) group[23] and the International Society of Pharmacoeconomics and Outcomes Research[24,25] and recommended by the US Food and Drug Administration.[11]

Rasch measurement theory and analysis We analyzed the Satisfaction with Facial Appearance scale data using Rasch measurement theory methods.[26,27] These methods are increasingly used in health outcome research.[28] Unlike traditional methods, Rasch analysis indicates the extent to which rigorous measurement is achieved by examining the difference (or fit) between the observed scores (patients' responses to items) and the expected values predicted from the data by a single mathematical model called the Rasch model. The criteria for measurement in Rasch analysis are evaluated interactively using the Rasch model.[29,30] Thus, a range of evidence is used to evaluate each questionnaire item in a scale. This evidence is then used to make a judgment about the overall quality of the scale.

Rasch analyses were performed on the Satisfaction with Facial Appearance scale using RUMM2030 software.[31] Results were interpreted using published criteria wherever possible as follows:

Item fit validity The items of the Satisfaction with Facial Appearance scale must work together (fit) as a conformable set both clinically and statistically. When items do not work together (misfit) in this way, it is inappropriate to sum item responses to reach a total score, and the validity of a scale is questioned. Three main indicators were examined to assess item fit[27,29]:

1. Log residuals (item-person interaction)
2. Chi-square values (item-trait interaction)
3. Item characteristic curves

There are no absolute criteria for interpreting fit statistics. It is more meaningful to interpret them together and in the context of their clinical usefulness as an item set. However, as a guide, fit residual should be between -2.5 and $+2.5$ with associated nonsignificant chi-square values (significance interpreted after Bonferroni adjustment).

Each of the items of the Satisfaction with Facial Appearance scale has multiple response categories (ie, very dissatisfied, somewhat dissatisfied, somewhat satisfied and very satisfied), which reflect an ordered continuum. Although this ordering may seem clinically sensible at the item level, it must also work together when the items are combined to form a set. Item fit validity analysis tests this statistically and graphically by threshold locations and plots. As such, the threshold values between adjacent pairs of response options for each item are expected to be ordered by magnitude (less to more). Thresholds are visible in graphical plots, in which the highest areas of the probability distributions of each response category should not be below adjacent category plots. When response options work as expected, important evidence for the validity of the scale is obtained.[32]

Targeting Scale-to-sample targeting concerns the match between the range of satisfaction with facial appearance measured by the Satisfaction with Facial Appearance items and the range of satisfaction with facial appearance as reported by a sample of patients. Targeting can be observed by examining the spread of person and item locations (ie, define the relative distributions of transformed total scores against the locations of the individual items across the continuum of satisfaction with facial appearance) in these two relative distributions. Targeting analysis informs about how suitable the sample is for evaluating the Satisfaction with Facial Appearance scale and how suitable the scale is for measuring the sample. Better targeting equates to a better ability to interpret the psychometric data with confidence.[27,33]

Reliability Person measurements (estimates) are examined with the Person Separation Index (PSI), a reliability statistic that is comparable with the Cronbach alpha.[34] The PSI quantifies the error associated with the measurements of people in a sample. Higher PSI values indicate better reliability (>0.70 indicates adequate reliability[33]).

Stability Scale performance (specifically item performance) should be stable across clinically important scenarios in which systematic differences between subgroups that may lead to bias in responding to items are not expected. Stability analysis enables an explicit test of scale performance in the form of an examination of differential item functioning (DIF). We examined DIF for gender, age, and ethnicity. As a guide, statistically significant chi-square values indicate potential DIF and therefore problems in scale performance (significance interpreted after Bonferroni adjustment).[35]

Traditional psychometric methods analysis Traditional psychometric methods primarily use correlation or descriptive analyses to evaluate scaling assumptions (legitimacy of summing items) and scale reliability and validity, which are described in detail elsewhere.[33] We examined data quality (percent missing data for each item), scaling assumptions (similarity of item means and variances; magnitude and similarity of corrected item-total correlations[36–38]), scale-to-sample targeting (score means; standard deviation [SD]; floor and ceiling effects), and internal consistency reliability (Cronbach alpha,[39] homogeneity coefficients[40]).

Responsiveness analysis The responsiveness of the Satisfaction with Facial Appearance scale to detect clinical change was examined in the largest subgroup in our sample (patients having facelifts) at the group level by comparing pretreatment and posttreatment Rasch transformed scores using paired *t*-tests and calculating the following 2 standard indicators: effect size (ES) calculations (Kazis ES[41]); and standardized response mean (SRM).[42] Larger ESs/SRMs indicate greater responsiveness, and it is standard practice to interpret the magnitude of the change using Cohen arbitrary criteria (0.20, small; 0.50, moderate; and 0.80, large). Preliminary minimal importance difference (MID) values were generated as follows: (1) calculating half standard deviation of the pretreatment mean score, and (2) extrapolation of a change score based on a 0.5 ES.

The responsiveness of the Satisfaction with Facial Appearance scale was also compared at the individual person level. This change score was achieved by computing, for each person, the significance of their own change in measurement (sig change).[43] First, we computed a change score for each person (before surgery to after surgery). Second, we computed the standard error associated with each person's change score (ie, the square root of the sum of the squared standard error values before and after surgery). Third, we computed the significance of the change for each person by dividing their change score by the standard error of the difference (SE_{diff}; ie, how large was their change in standard error units). Fourth, we categorized the significance of each person's change score into 1 of 5 groups according to the size and direction of the change score. We then counted the numbers of people achieving each level of significance of change. The formulae are as follows:

$$\text{Sig change} = \frac{\text{Postsurgery transformed score} - \text{Presurgery transformed score}}{\text{SE}_{\text{diff}}}$$

$$\text{where } \text{SE}_{\text{diff}} \text{ for a person} = \sqrt{(\text{SE presurgery transformed score})^2 + (\text{SE postsurgery transformed score})^2}$$

Significance of change values obtained from this formula was categorized into the following 5 groups:

Significant improvement = Sig change $\geq +1.96$
Nonsignificant improvement = 0 < Sig Change $\leq +1.95$
No change = Sig change = 0
Nonsignificant worsening = $-1.95 \leq$ Sig change < 0
Significant worsening = Sig change ≤ -1.96

RESULTS

Phase 1: Qualitative Phase

As described earlier and in our previous publication,[10] the qualitative work resulted in the development of a conceptual framework (see **Fig. 1**) and a series of independent scales that capture the important concerns described by facial aesthetics patients (see **Table 2**). The Satisfaction with Facial Appearance scale was specifically developed to be relevant to all aesthetic facial patients regardless of the number or type of procedures undergone. This scale is composed of 10 items that ask about satisfaction using descriptors (eg, symmetry, balance, proportion) as well as scenarios (eg, in photographs, under bright lights). The item set is easy to understand and complete with a Flesch Kincaid grade level of 0.8, and all items lower than grade 6 (range 0–5.2).

Phase 2: Quantitative Phase

A total of 360 patients were invited to participate through face-to-face recruitment, and 332 responded. A further 283 patients were sent the FACE-Q in the mail, and 167 responded. The overall response rate was 78%. Participants ranged in age from 18 to 85 years; 64 were men and 409 were women (**Table 3**). Participants completed from 1 to 3 copies of the FACE-Q at various time points.

Rasch analyses

Overall, the results of Rasch analysis supported the Satisfaction with Facial Appearance scale as a reliable and valid measure of satisfaction with facial appearance. All 10 items had ordered thresholds, which supports the appropriateness of the number and type of response options we created (**Fig. 2**). Just 1 of the 10 items had fit residuals marginally outside of the −2.5 to +2.5 range (Q4). However, no item had a significant chi-square value (**Table 4**). Distributions of item thresholds and person estimates were well matched, taking into consideration some gaps at the extremes of the continuum (lowest/highest satisfaction) (**Fig. 3**). The PSI was 0.92, showing good reliability. Analysis of the data set showed no statistical DIF by gender, age, or ethnicity (**Table 5**).

Traditional psychometric analysis

The results of traditional analysis also supported the Satisfaction with Facial Appearance scale as a reliable and valid measure (**Table 6**). The criteria were satisfied for all psychometric properties evaluated. Data quality was high (missing data range ≤2%, scale scores were computable for 98% of respondents) and scaling assumptions were satisfied (similar mean item scores, corrected item-total correlations range = 0.75–0.82). Scale-to-sample targeting was good (scale scores spanned the scale range and were not notably skewed; the scale midpoint, and ceiling effects were negligible), and internal consistency reliability was high (Cronbach alpha = 0.95; mean item-item correlation = 0.65 [0.52–0.89]).

Table 3
Patient characteristics from field tests

	Study 1	Study 2
N	399	100
Age (y)		
Mean (SD)	48.9 (14.8)	54.3 (7.8)
Range	18–85	37–77
Gender		
Female (%)	323 (85.7)	86 (89.6)
Male (%)	54 (14.3)	10 (10.4)
Ethnicity		
White non-Hispanic (%)	269 (73.5)	100 (100)
Asian (%)	21 (5.8)	—
South Asian (%)	18 (4.9)	—
Native American (%)	17 (4.6)	—
White Hispanic (%)	13 (3.6)	—
Black non-Hispanic (%)	10 (2.7)	—
Other (%)	18 (4.9)	—
Country		
United States (%)	197 (49.5)	—
Canada (%)	201 (50.5)	—
France (%)	—	15 (15)
Germany (%)	—	50 (50)
Israel (%)	—	20 (20)
United Kingdom (%)	—	15 (15)
Timing of Booklet		
Before only (%)	61(15.3)	12 (12)
After only (%)	294 (73.6)	—
Before and 1 after (%)	26 (6.5)	88 (88)
Before and 2 after (%)	9 (2.3)	—
2 after only	9 (2.3)	—
Booklet Type		
Fillers (%)	57 (14.2)	—
Botulinum toxin (%)	75 (18.8)	—
Skin resurfacing (%)	17 (4.3)	—
Lip injections (%)	11 (2.8)	—
Facelift (%)	97 (24.3)	100 (100)
Blepharoplasty (%)	65 (16.3)	—
Rhinoplasty (%)	45 (11.3)	—
Chin surgery (%)	22 (5.5)	—
Brow lift (%)	4 (1.0)	—
Cheeks (%)	6 (1.5)	—

Responsiveness

Group-level findings Ninety-seven patients completed presurgery and 6-month postsurgery versions of the scale. The responsiveness data generated by the analysis of interval measurements showed that the scale quantified significant change

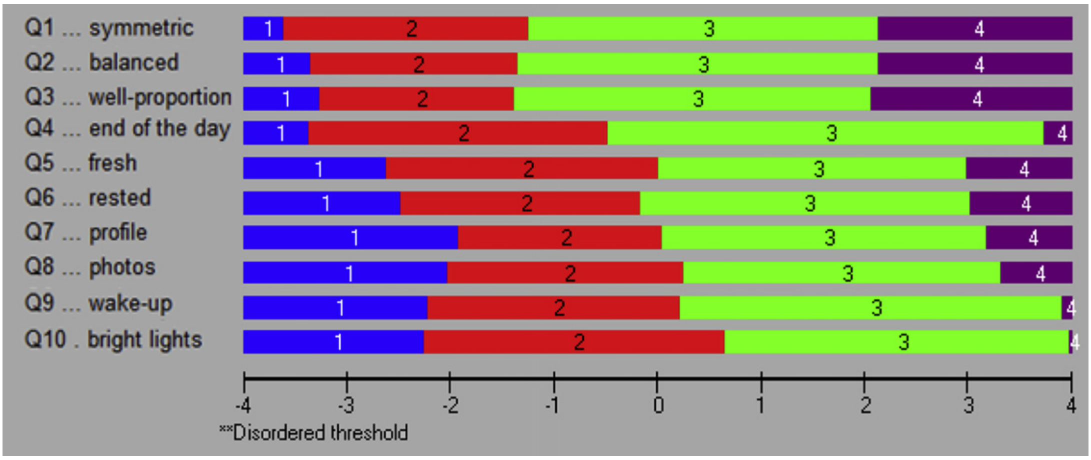

Fig. 2. Ordering of item response thresholds (location order). Threshold map for all items in the FACE-Q Satisfaction with Facial Appearance scale. The x-axis symbolizes the construct (satisfaction with facial appearance), with satisfaction increasing to the right. The y-axis shows the items' response categories: 1, very unsatisfied (*blue block*); 2, somewhat dissatisfied (*red block*); 3, somewhat satisfied (*green block*); 4, very satisfied (*purple block*).

at the group level. Patients' satisfaction with their facial appearance on a 0–100 scale was significantly higher following facelift-related treatment than it was before treatment (mean, SD = 45, 16 vs 56, 21, respectively, *P*<.0001). These statistically significant change scores were associated with moderate effect sizes (ES = 0.68, SRM = 0.50). In addition, preliminary MID analyses suggested an 8-point difference in total scores. This difference was exceeded in our analysis (mean change, SD = 11, 22).

Individual-level findings 94 out of 97 patients who had facelifts reported significant improvement in satisfaction with facial appearance, with the remaining 3 patients reporting nonsignificant improvement. This finding supports the scale's ability to measure important change following treatment.

DISCUSSION

Satisfaction with appearance and improved quality of life are arguably the most important outcomes for patients undergoing facial aesthetic procedures.[4,44] Despite this, research in facial aesthetics has been hindered by a lack of reliable and valid condition-specific PRO instruments. The FACE-Q is developed to address this void. In this study, the FACE-Q Satisfaction with Facial Appearance scale is a short, easy to complete, reliable, valid and responsive measurement tool.

Table 4
Statistical indicators of fit (fit residual; chi-square)

	Items	Item Location	SE	Fit Residual	Chi-Square	*P*
Q1	…symmetric	−0.91	0.08	0.26	3.53	.474
Q2	…balanced	−0.86	0.08	−2.21	7.15	.128
Q3	…well proportioned	−0.85	0.08	−0.79	9.44	.051
Q4	…end of day[a]	−0.03	0.08	−2.59	10.56	.032
Q5	…fresh	0.13	0.08	−1.66	4.77	.311
Q6	…rested	0.13	0.08	−1.64	8.03	.090
Q7	…profile	0.44	0.07	0.93	3.83	.430
Q8	…photos	0.51	0.07	1.75	3.80	.433
Q9	…wake-up	0.64	0.08	0.59	0.88	.927
Q10	…bright lights	0.80	0.08	−2.01	7.58	.108

Items are in serial order.
Abbreviation: SE, standard error.
[a] Fit residual outside + 2.5 criteria.

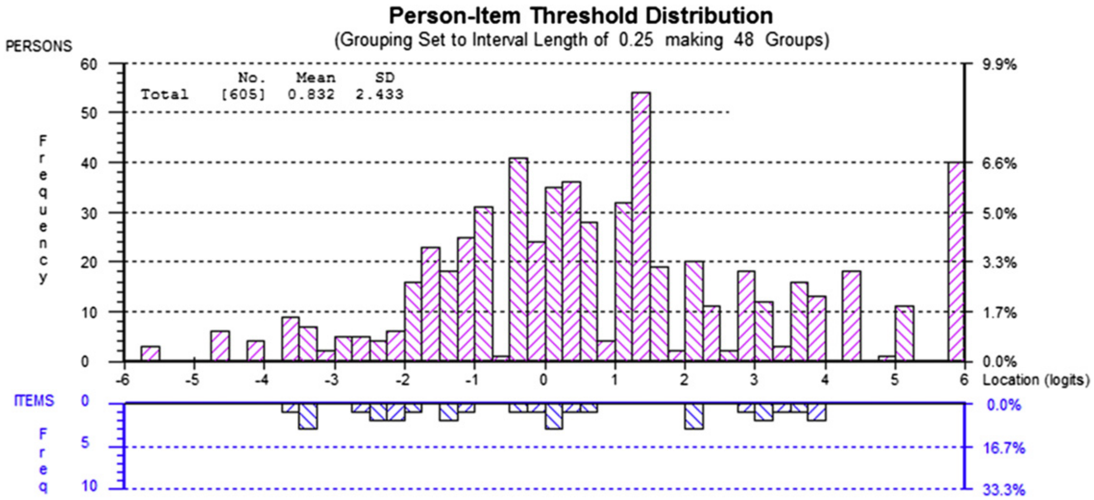

Fig. 3. Targeting of scale to sample (person-item threshold locations spread). The x-axis symbolizes the construct (satisfaction with facial appearance), with satisfaction increasing to the right. The y-axis shows the frequency of person measure locations (*top histogram*) and item locations (*bottom histogram*).

Our study provides the first empirical support for the use of this scale to measure satisfaction in patients undergoing any type of surgical or nonsurgical facial aesthetic procedure.

There are several strengths to this research. The FACE-Q Satisfaction with Facial Appearance scale was developed from qualitative research that involved in-depth interviews with a varied sample of patients as well as extensive expert input.[10] Careful qualitative work was instrumental to establishing a strong conceptual framework and a valid set of scales with items capable of measuring the unique concerns of patients having facial aesthetic surgery. Thus, unlike generic PRO instruments that have been used in plastic surgery studies in the past, our Satisfaction with Facial Appearance scale is well calibrated to measure preprocedure to postprocedure change. As a further strength, psychometric evaluation of the Satisfaction with Facial Appearance scale involved a large heterogeneous patient sample and our DIF results indicate that the scale performed the same in subgroups of patients that varied by age, gender, and ethnicity.

Accurate and reliable quantification of patient-centered outcomes is critical to ongoing practice improvement and technical advancement in facial plastic surgery. Such quantification requires the use of high-quality PRO instruments that accurately measure patients' subjective perception of

Table 5
DIF: gender, age, and ethnicity

		Gender			Age			Ethnicity		
		MS	F	*P*	MS	F	*P*	MS	F	*P*
Q1	…symmetric	2.12	2.25	.135	1.31	1.42	.197	1.29	1.41	.210
Q2	…balanced	0.00	0.00	.994	2.25	2.99	.005	0.71	0.95	.462
Q3	…well proportioned	0.94	1.04	.309	1.93	2.14	.038	0.83	0.95	.458
Q4	…end of day	0.05	0.07	.792	1.14	1.53	.155	0.53	0.73	.629
Q5	…fresh	0.78	0.95	.330	0.93	1.12	.348	1.14	1.41	.208
Q6	…rested	0.13	0.15	.696	1.36	1.64	.123	0.40	0.47	.827
Q7	…profile	0.61	0.63	.428	0.55	0.56	.786	1.20	1.20	.305
Q8	…photographs	1.23	1.19	.276	2.18	2.16	.037	1.01	0.98	.442
Q9	…wake-up	0.03	0.03	.859	0.71	0.74	.640	0.29	0.30	.938
Q10	…bright lights	0.56	0.68	.410	2.24	2.82	.007	1.22	1.50	.175

Abbreviations: F, F-statistic; MS, mean square.

Table 6
Data quality, scaling assumptions, and targeting

		Data Quality	Scaling Assumptions					Targeting	
		Item Missing Data (%)	Possible Range (Midpoint)	Score Range	Mean Score	SD	CITC	Floor/ Ceiling Effects (%)[a]	Skewness
Q1	…symmetric	1	1–4	1–4	3.07	0.80	0.75	4/32	−0.58
Q2	…balanced	1	1–4	1–4	3.08	0.82	0.78	5/33	−0.66
Q3	…well proportioned	1	1–4	1–4	3.09	0.82	0.77	5/33	−0.70
Q4	…end of day	1	1–4	1–4	2.82	0.82	0.81	6/20	−0.32
Q5	…fresh	1	1–4	1–4	2.78	0.91	0.82	9/25	−0.23
Q6	…rested	1	1–4	1–4	2.79	0.92	0.81	9/25	−0.29
Q7	…profile	2	1–4	1–4	2.71	0.94	0.78	12/22	−0.24
Q8	…photographs	2	1–4	1–4	2.67	0.93	0.75	12/21	−0.17
Q9	…wake-up	1	1–4	1–4	2.65	0.89	0.77	11/18	−0.17
Q10	…bright lights	2	1–4	1–4	2.59	0.90	0.81	11/17	−0.04
Total		2	10–40	10–40	28.3	7.3	—	1/8	−0.18

Abbreviation: CITC, corrected item-total correlation.
[a] Calculated as the percentage of people scoring either floor or ceiling.

outcomes and provide clinically meaningful interval-level data. Interval-level means that the scores derived from the FACE-Q Satisfaction with Facial Appearance scale have defined units and that the distance between each unit is the same.[45] Such interval-level measurement is analogous to measurements used in clinical practice, such as temperature in Celsius or millimeters on a ruler used in the operating room.[46]

The ability to move beyond raw scores to linearized measures is one of the benefits of Rasch measurement methods. Previous PRO instruments developed using traditional psychometric methods provide ordinal rather than interval-level measurement and, as such, have inherently limited clinical meaning. As an example, the Derriford scale is an older-generation PRO instrument developed to measure quality of life among patients having aesthetic surgery and developed using traditional methods providing only ordinal-level measurement. When a patient moves from a score of 100 to 120 on the Derriford scale following surgery, improvement has occurred; however, when another patient moves from 120 to 140, it cannot be assumed that both patients experienced the same magnitude of improvement from a 20-point change in score.[47] The FACE-Q is an example of a new generation of PRO instruments that can overcome the limitations of older-generation measures and provides clinical meaningful outcomes data. The advantages of Rasch measurement theory in PRO instrument development include the ability to compare directly patients' total scores and the item locations on the same metric; the improved potential to diagnose item-level psychometric problems; and the ability to move to a more accurate picture of individual person measurements derived from PRO instruments.[48]

Our current study has 3 main limitations. First, in both our development and psychometric evaluation of the FACE-Q, most patients were female. Although this mirrors the patient population seen in clinical practice (and DIF analysis indicated that item performance was stable across genders), future research to explore the psychometric properties of the scale when used with male patients is warranted. Second, our sample included patients who may have had multiple procedures (both surgical and nonsurgical). Although this reflects real-world practice, and hence increases the validity of our findings, the impact of specific procedures in this particular study cannot confidently be delineated. In future clinical trials with stringent inclusion criteria, the impact of specific procedures on patient satisfaction and quality of life may be examined. Third, there is a potential for selection bias in our research. Although our response rate was high among patients who received the questionnaire while in clinic, it was lower when administered by mail and this may contribute to bias. In addition, practices where clinicians volunteer to recruit patients for PRO research may be different from practices that do not volunteer, and we did not have control over which patients the office staff recruited into the study.

The FACE-Q Satisfaction with Facial Appearance scale is an example of a new-generation condition-specific PRO instrument capable of providing highly reliable, valid, and responsive patient assessments. The scale has strong psychometric properties and the potential to provide clinically meaningful scores. By providing scientifically sound and clinical interpretable outcomes data, this scale (and others in the FACE-Q PRO instrument suite) will be able to inform technical advancement and ongoing practice improvement in future studies and in individual clinical care.

ACKNOWLEDGMENTS

The authors acknowledge and thank the following clinicians for their invaluable assistance with the recruitment of patients and countless hours spent as expert reviewers: Vancouver, BC, Canada: Drs Nick Carr, Francis Jang, Nancy VanLaeken, Alistair Carruthers, Jean Carruthers, Richard Warren. Washington DC: Dr Stephen Baker; Dallas, TX: Drs Jeffery Kenkel, Rod Rohrich; Atlanta GA: Dr Foad Nahai; St Louis, MO: Dr Leroy Young; New York, NY: Drs David Hidalgo, David Rosenberg, Philip Miller, Alexes Hazen, and Haideh Hirmand.

REFERENCES

1. American Society for Aesthetic Plastic Surgery. 2012. Available at: http://www.plasticsurgery.org/News-and-Resources/2011-Statistics-.html. Accessed Sept 21, 2012.
2. Pusic A, Lemaine V, Klassen A, et al. Patient-reported outcome measures in plastic surgery: use and interpretation in evidence-based medicine. Plast Reconstr Surg 2011;127:6.
3. Fitzpatrick R, Jenkinson C, Klassen A, et al. Methods of assessing health-related quality of life and outcome for plastic surgery. Br J Plast Surg 1999;52:251–5.
4. Cano S, Browne J, Lamping D. Patient-based measures of outcome in plastic surgery: current approaches and future directions. Br J Plast Surg 2004;57:1–11.
5. Cano S, Hobart J. The problem with health measurement. Patient Prefer Adherence 2011;5:279–90.
6. Kosowski T, McCarthy C, Reavey P, et al. A systematic review of patient-reported outcome measures after facial cosmetic surgery and/or nonsurgical facial rejuvenation. Plast Reconstr Surg 2009;123:1819–27.
7. Cano S, Klassen A, Scott A, et al. The BREAST-Q ©: further validation in independent clinical samples. Plast Reconstr Surg 2012;129:293–302.
8. Klassen A, Pusic A, Scott A, et al. Satisfaction and quality of life in women who undergo breast surgery: a qualitative study. BMC Womens Health 2009;9:11–8.
9. Pusic A, Klassen A, Scott A, et al. Development of a new patient-reported outcome measure for breast surgery: the BREAST-Q. Plast Reconstr Surg 2009;124:345–53.
10. Klassen A, Cano S, Scott A, et al. Measuring patient-reported outcomes in facial aesthetic patients: development of the FACE-Q. Facial Plast Surg 2010;26:303–9.
11. Food and Drug Administration. Patient reported outcome measures: use in medical product development to support labeling claims. 2009. Available at: www.fda.gov/cber/gdlns/prolbl.pdf. Accessed Sept 21, 2012.
12. Cano S, Hobart J. Watch out, watch out, the FDA are about. Dev Med Child Neurol 2008;50:108–9.
13. Mokkink L, Terwee C, Patrick D, et al. The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: an international Delphi study. Qual Life Res 2010;19:539–49.
14. Scientific Advisory Committee of the Medical Outcomes Trust. Assessing health status and quality of life instruments: attributes and review criteria. Qual Life Res 2002;11:193–205.
15. Lasch K, Marquis P, Vigneuz M, et al. PRO development: rigorous qualitative research as crucial foundation. Qual Life Res 2010;19:9.
16. Hays R, Anderson R, Revicki D. Psychometric considerations in evaluating health-related quality of life measures. Qual Life Res 1993;2:441–9.
17. Cano S, Klassen A, Pusic A. The science behind quality-of-life measurement: a primer for plastic surgeons. Plast Reconstr Surg 2009;123:98e–106e.
18. Qualitative Solutions Research International: NVivo 8. Australia: QSR International; 2008.
19. Flesch R. A new readability yardstick. J Appl Psychol 1948;32:12.
20. Dillman D. Mail and telephone surveys: the total design method. New York: Wiley; 1978.
21. Dillman D. Mail and internet surveys: the tailored design method. 2nd edition. Toronto: Wiley; 2000.
22. MAPI Research Trust: France. 2004-2012. Available at: http://www.mapitrust.org/. Accessed Sept 21, 2012.
23. Chassany O, Sagnier P, Marquis P, et al. Patient-reported outcomes: the example of health-related quality of life — A European guidance document for the improved integration of health-related quality of

life assessment in the drug regulatory process. DIA J 2002;36:209–38.

24. Wild D, Grove A, Martin M, et al. Principles of good practice for the translation and cultural adaptation process for patient-reported outcomes (PRO) measures: report of the ISPOR Task Force for Translation and Cultural Adaptation. Value Health 2005;8: 94–104.
25. Wild D, Eremenco S, Mear I, et al. Multinational trials-recommendations on the translations required, approaches to using the same language in different countries, and the approaches to support pooling the data: the ISPOR Patient-Reported Outcomes Translation and Linguistic Validation Good Research Practices Task Force report. Value Health 2009; 12(4):430–40.
26. Andrich D. Controversy and the Rasch model: a characteristic of incompatible paradigms? Med Care 2004;42:17–116.
27. Wright B, Masters G. Rating scale analysis: Rasch measurement. Chicago: MESA; 1982.
28. Massof R. Understanding Rasch and item response theory models: applications to the estimation and validation of interval latent trait measures from responses to rating scale questionnaires. Ophthalmic Epidemiol 2011;18:19.
29. Andrich D. Rasch models for measurement. Beverley Hills (CA): Sage Publications; 1988.
30. Rasch G. Probabilistic models for some intelligence and attainment tests. Copenhagen (Denmark): Danish Institute for Education Research; 1960.
31. Andrich D, Sheridan B. RUMM 2030. Perth (Australia): RUMM Laboratory; 1997–2011.
32. Andrich D. Rating scales and Rasch measurement. Expert Rev Pharmacoecon Outcomes Res 2011; 11:14.
33. Hobart J, Cano S. Improving the evaluation of therapeutic intervention in MS: the role of new psychometric methods. Health Technol Assess 2009;13:1–200.
34. Andrich D. An index of person separation in latent trait theory, the traditional KR20 index and the Guttman scale response pattern. Educ Psychol Res 1982;9:9.
35. Hagquist C, Andrich D. Is the Sense of Coherence instrument applicable on adolescents? A latent trait analysis using Rasch modelling. Pers Indiv Differ 2004;36:13.
36. McHorney C, Haley S, Ware JJ. Evaluation of the MOS SF-36 Physical Functioning Scale (PF-10): II. Comparison of relative precision using Likert and Rasch scoring methods. J Clin Epidemiol 1997;50: 451–61.
37. Likert R. A technique for the measurement of attitudes. Arch Psychol 1932;140:50.
38. Ware J, Harris W, Gandek B, et al. MAP-R for Windows: multi-trait/multi-item analysis program—revised user's guide. Boston: Health Assessment Laboratory; 1997.
39. Cronbach L. Coefficient alpha and the internal structure of tests. Psychometrika 1951;16:297–334.
40. Eisen M, Ware JJ, Donald C, et al. Measuring components of children's health status. Med Care 1979;17:19.
41. Kazis L, Anderson J, Meenan R. Effect sizes for interpreting changes in health status. Med Care 1989;27:178–89.
42. Liang M, Fossel A, Larson M. Comparisons of five health status instruments for orthopedic evaluation. Med Care 1990;28:10.
43. Hobart J, Cano S, Thompson A. Effect sizes can be misleading: is it time to change the way we measure change? J Neurol Neurosurg Psychiatry 2010;81:4.
44. Ching S, Thoma A, McCabe R, et al. Measuring outcomes in aesthetic surgery: a comprehensive review of the literature. Plast Reconstr Surg 2003; 111:11.
45. Wright B, Linacre J. Observations are always ordinal: measurements, however, must be interval. Arch Phys Med Rehabil 1989;70:857–60.
46. Bond T, Fox C. Applying the Rasch model. Fundamental measurement in the human sciences. 2nd edition. Mahwah, New Jersey: Lawrence Erlbaum Associates; 2007.
47. Harris D, Carr A. The Derriford Appearance Scale (DAS59): a new psychometric scale for the evaluation of patients with disfigurements and aesthetic problems of appearance. Br J Plast Surg 2001;54: 216–22.
48. Wright B. Solving measurement problems with the Rasch model. J Educ Meas 1977;14:97–116.
49. Branski R, Cukier-Blaj S, Pusic A, et al. Measuring quality of life in dysphonic patients: a systematic review of content development in patient-reported outcomes measures. J Voice 2010;24:193–8.
50. Klassen A, Stotland M, Skarsgard E, et al. Clinical research in pediatric plastic surgery and systematic review of quality-of-life questionnaires. Clin Plast Surg 2008;35:251–67.
51. Pusic A, Chen CM, Cano S, et al. Measuring quality of life in cosmetic and reconstructive breast surgery: a systematic review of patient-reported outcomes instruments. Plast Reconstr Surg 2007;120:823–37 [discussion: 838–9].
52. Garratt A, Schmidt L, Mackintosh A, et al. Quality of life measurement: bibliographic study of patient assessed health outcome measures. BMJ 2002;324: 1417.
53. Hunt S. The problem of quality of life. Qual Life Res 1997;6:205–12.
54. Lamping D. Methods for measuring outcomes to evaluate interventions to improve health-related quality of life in HIV. Psychol Health 1994;9:31–9.