Measuring Outcomes That Matter to Face-Lift Patients: Development and Validation of FACE-Q Appearance Appraisal Scales and Adverse Effects Checklist for the Lower Face and Neck

Anne F. Klassen, D.Phil.
Stefan J. Cano, Ph.D.
Amie M. Scott, M.P.H.
Andrea L. Pusic, M.D., M.H.S.

Hamilton, Ontario, Canada; Plymouth, United Kingdom; and New York, N.Y.

Background: The FACE-Q is a new patient-reported outcome instrument to evaluate a range of outcomes for patients undergoing any type of facial cosmetic operation, minimally invasive cosmetic procedure, or facial injectable. This article describes the development and validation of FACE-Q scales relevant to face-lift patients.
Methods: The FACE-Q was developed by following international guidelines for patient-reported outcome instrument development. For outcomes following a face lift, the authors developed five appearance appraisal scales (i.e., Satisfaction with Cheeks, Satisfaction with Lower Face and Jawline, Appraisal of Nasolabial Folds, Appraisal of Area Under the Chin, and Appraisal of the Neck) and an adverse effects checklist. A field test of these scales was performed in a sample of 225 face-lift patients, and were evaluated using both modern and traditional psychometric methods.
Results: The five FACE-Q appearance appraisal scales were found to be clinically meaningful, reliable, valid, and responsive to clinical change. These findings were supported by Rasch measurement theory analysis (e.g., overall chi-square values of $p \geq 0.18$; Person Separation Index ≥ 0.88). Responsiveness analyses showed that patient scores for facial appearance improved significantly after treatment ($p < 0.001$); changes in scores were associated with moderate effect sizes (range effect size, 0.40 to 0.79; range standardized response mean, 0.37 to 0.69). Traditional psychometric statistics provided further support (e.g., Cronbach's alpha values ≥0.94)
Conclusions: The FACE-Q appearance appraisal scales are scientifically sound and clinically meaningful and can be used with the adverse effects checklist to measure patient-reported outcomes following a face lift. (*Plast. Reconstr. Surg.* 133: 21, 2014.)
CLINICAL QUESTION/LEVEL OF EVIDENCE: Diagnostic, III.

Increased societal acceptance of cosmetic surgery has resulted in an increased number of patients seeking facial rejuvenation. A face-lift is one of most popular procedures used to combat the appearance of aging, with over 119,026 face-lifts performed in the United States in 2011, 5 percent more than in 2010.[1] Face lifts are the fifth

From McMaster University; Plymouth University Peninsula Schools of Medicine and Dentistry; and Memorial Sloan-Kettering Cancer Center.
Received for publication April 19, 2013; accepted July 30, 2013.

DOI: 10.1097/01.prs.0000436814.11462.94

Disclosure: *The FACE-Q is owned by Memorial Sloan-Kettering Cancer Center. Stefan J. Cano, Ph.D., Anne F. Klassen, D.Phil., and Andrea L. Pusic, M.D., M.H.S., are co-developers of the FACE-Q and, as such, receive a share of any license revenues based on Memorial Sloan-Kettering Cancer Center's inventor sharing policy.*

This work was supported by
THE PLASTIC SURGERY FOUNDATION

most common surgical cosmetic procedure in the United States.[1] Satisfaction with facial appearance is undoubtedly the most important outcome to face-lift patients, but there exists limited research evaluating the patient perspective.[2,3]

Measurement of the patient's view of his or her facial appearance has been hampered by a lack of clinically meaningful and scientifically valid questionnaires. A systematic review published by our team found only one patient-reported outcome instrument developed to evaluate appearance in face-lift patients.[4] The Facelift Outcomes Evaluation scale, published over a decade ago, is a six-item scale that measures appearance, functional outcome, and social acceptance.[5,6] Given its limited content and lack of published information regarding its development and scientific properties, it is not a surprise that this patient-reported outcome instrument has not been broadly adopted.

To address the lack of available patient-reported outcome instruments for patients undergoing any type of facial cosmetic operation, minimally invasive cosmetic procedure, or facial injectable, our research team developed the FACE-Q. This is a patient-reported outcome instrument made up of independently functioning scales and checklists measuring concepts important to facial aesthetic patients, including quality of life, appearance, and process of care (Table 1). Each scale provides a standalone score ranging from 0 to 100, with higher scores indicating a better outcome. There is also an adverse effects checklist that includes questions about postsurgical symptoms for different facial areas. Depending on the surgical or nonsurgical procedure, only those FACE-Q scales and/or checklists relevant to a particular patient or procedure(s) need be completed. The aim of this article is to describe the development and psychometric evaluation of five appearance scales and the adverse effects checklist for use with face-lift patients.

Table 1. FACE-Q Scales

Conceptual Framework	Scales and Checklists
Appearance appraisal scales	Facial appearance overall
	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 and jawline*
	Under chin*
	Neck*
Quality-of-life scales	Psychological well-being
	Social well-being
	Aging appraisal
	Expectations and motivations
	Psychological distress
	Recovery early life impact
Adverse effect checklists	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

*FACE-Q scales examined in this study.

METHODS

Development of FACE-Q Scales and Checklists

Ethics review board approval was obtained before the study was started. FACE-Q scales and checklists were developed by our team following an approach that adheres to internationally recommended guidelines for patient-reported outcome instrument development.[7,8]

Phase 1: Qualitative Research Methods

In the first phase, we developed a conceptual framework to account for outcomes of importance to facial aesthetic patients that is composed of the following four major domains: appearance, quality of life, process of care, and adverse effects. These domains were identified using a mixed methods approach that included a systematic review, qualitative interviews, and expert input, and is reported in detail elsewhere.[9]

Phase 2: Quantitative Research Methods

Data were collected as part of two separate studies, and compiled for the purpose of analyses. The following FACE-Q appearance scales and checklist were evaluated in this article:

1. Satisfaction with Cheeks (i.e., sides of the face below the cheekbones): measures satisfaction using items that ask, for example, about symmetry, contour, and fullness.

2. Satisfaction with Lower Face and Jawline: measures satisfaction with items that ask, for example, about how sculpted and how prominent the jawline appears.
3. Appraisal of Nasolabial Folds (i.e., the deep lines that run downward from the sides of the nose): asks how bothered a patient is with his or her nasolabial folds with items such as how deep or noticeable the folds are, and how the folds appear during certain facial expressions.
4. Appraisal of Area Under the Chin: asks how bothered a patient is with this area of the face with items that ask, for example, about loose skin and fat, fullness, and contour.
5. Appraisal of the Neck: asks how bothered someone is with his or her neck with items that ask, for example, about hanging skin, wrinkles, and having to cover up the neck.
6. Adverse effects checklist: asks how bothered a patient is with a range of postsurgical symptoms.

Flesch-Kincaid scores for 29 of 30 items in the five appearance scales were lower than grade 6 (range, 0 to 6.7). For the adverse effects checklist, 12 of 15 items were lower than grade 6 (range, 0 to 10.2).[10]

Study 1: Data Collection

Ten plastic surgery and dermatology practices in the United States and Canada recruited patients between June of 2010 and June of 2012. Eligible participants were 18 years of age or older who had undergone or were waiting to undergo any surgical or nonsurgical facial aesthetic procedure. For the purposes of this article, we used the data provided by the subsample of patients in the FACE-Q field test who had undergone or were waiting to undergo a face lift. Patients from six practices were recruited in person, and patients from four practices were recruited through a postal survey, with up to three mailed reminders as necessary.

Study 2: Data Collection

A medical device company used the FACE-Q scales for an international clinical trial. The Mapi Research Trust[11] provided translations and linguistic validation of the FACE-Q scales. Participants completed FACE-Q scales before and after surgery.

Statistical Analyses

For the five FACE-Q appearance-related scales, decisions about item inclusion/exclusion were based on their performance against a standardized set of psychometric criteria. The adverse effects checklist was not analyzed in this way, as it is a descriptive tool (i.e., each item is an individual clinically important issue, and a total score is not computed).

Rasch Measurement Theory

We analyzed the FACE-Q scale data using Rasch measurement theory methods[12,13] in RUMM2030 software.[14] Rasch measurement theory analysis examines the differences between observed and predicted item responses to determine the extent to which the data for a set of items accord with ("fit") a mathematical model. When data fit the Rasch model, the measurement theory (i.e., that a scale measures a specific construct) is supported by the data. Rasch measurement theory analysis examines the difference (or fit) between the observed scores (patients' responses to items) and the expected values predicted by the Rasch model, which is evaluated interactively using a range of statistical and graphic tests to examine each item in a scale.[15,16] This combined evidence is used to make a judgment about the overall quality of the scale. Results for our scales were interpreted with reference to published criteria wherever possible as follows:

Thresholds for item response options: The use of response categories scored with successive integer scores implies a continuum (e.g., increasing satisfaction with facial appearance). We tested this assumption by examining the ordering of thresholds (or points of crossover between adjacent response categories).

Item fit statistics: The items of a scale must work together (fit) as a conformable set both clinically and statistically. When items do not work together (misfit), it would be inappropriate to sum the individual item responses to reach a total score. We examined the following three indicators of fit: log residuals (item-person interaction), chi-square values (item-trait interaction), and item characteristic curves. Fit statistics are usually interpreted together in the context of their clinical usefulness as an item set, but as a guide, fit residual should fall between −2.5 and +2.5, and chi-square values should be nonsignificant after Bonferroni adjustment.

Item locations: The items of a scale define a continuum, and inspecting where items are located on the continuum shows how well the items map out a construct. Items should be spread evenly over a reasonable range.

Person Separation Index: This reliability statistic is comparable to Cronbach's alpha[17] and quantifies the error associated with the measurements of people in a sample. Higher values indicate greater reliability.

Responsiveness analysis: The ability to detect clinical change was examined at the group level by comparing pretreatment and posttreatment Rasch transformed scores using paired *t* tests and calculating two standard indicators of change as follows: effect size calculations (Kazis' effect size)[18] and standardized response mean.[19] The magnitude of the change can be interpreted using Cohen's arbitrary criteria (small, 0.20; moderate, 0.50; and large, 0.80). Preliminary minimal importance difference values were generated as follows: (1) calculating ½ SD of the pretreatment mean score and; (2) extrapolation of a change score based on a 0.5 effect size.

Responsiveness at the person level for each scale was computed by determining the significance of the change in their individual measurement.[20] First, we computed a change score for each person (before surgery to after surgery) and the standard error for the change score. Then, we computed the significance of the change for each person by dividing his or her change score by the standard error of the difference. Finally, we categorized the significance of each person's change score into one of five groups and counted the numbers of people achieving each level of significance of change. The five groups were as follows: significant improvement (change ≥ 1.96), nonsignificant improvement ($0 <$ change ≤ 1.95), no change (change = 0), nonsignificant worsening ($-1.95 \leq$ change < 0), and significant worsening (change, < -1.96).

Traditional Test Theory Analysis

Traditional psychometric methods are described more fully elsewhere.[21] For each FACE-Q scale, we examined the following: data quality (percentage missing data for each item), scaling assumptions (similarity of item means and variances; magnitude and similarity of corrected item-total correlations),[22–24] scale-to-sample targeting (score means, standard deviation, floor and ceiling effects), and internal consistency reliability (Cronbach's alpha[17] and homogeneity coefficients).[25]

Aspects of validity were assessed in two ways. First, we computed intercorrelations between FACE-Q scales to examine the extent to which subscales measured separate but related constructs.[26] We predicted that these intercorrelations would range between $r = 0.30$ and $r = 0.70$, as the scales of the FACE-Q purport to measure distinct but related clinical variables.[27] Second, we examined the ability of the FACE-Q to detect differences between predefined subgroups. Specifically, all patients completed the FACE-Q Patient-Perceived Age Visual Analogue Scale, which asks them to indicate how many years younger or older they think they look compared with their actual age. The scale anchors for the Patient-Perceived Age Visual Analogue Scale are –15 years to +15 years. We categorized patient responses into the following five groups and compared the mean score for each FACE-Q scale using analysis of variance: (1) looks more than 5 years older than actual age, (2) looks between 1 and 5 years older than actual age, (3) looks actual age, (4) looks 1 to 5 years younger than actual age; and (5) looks more than 5 years younger than actual age. We hypothesized that FACE-Q Appearance Appraisal Scale scores would be incrementally higher in the younger subgroups compared with the older subgroups.

RESULTS

Phase 1: Qualitative Research Results

Through the qualitative phase of our study, we developed and refined the final set of FACE-Q scales and checklists shown in Table 1. Each of

Table 2. Patient Characteristics

	Study 1 (%)	Study 2 (%)
No.	125	100
Age, yr		
Mean ± SD	61.4 ± 6.1	54.3 ± 7.8
Range	36–75	37–77
Sex		
Female	117 (95.1)	88 (88)
Male	6 (4.9)	12 (12)
Ethnicity		
Caucasian	110 (91.7)	100 (100)
Other	10 (8.3)	—
Country		
United States	87 (69.6)	—
Canada	38 (30.4)	—
France	—	15 (15)
Germany	—	50 (50)
Israel	—	20 (20)
United Kingdom	—	15 (15)
Timing of booklet		
Before surgery only	6 (4.8)	1 (1)
After surgery only	117 (93.6)	1 (1)
Before and one after surgery	1 (0.8)	74 (74)
Before and two after surgery	—	24 (24)
Two after surgery only	1 (0.8)	—

Table 3. Overall Fit to the Rasch Model and Person Separation Index for Each Scale

Scale	χ^2	Degrees of Freedom	*p*	Person Separation Index
Cheeks	21.0	16	0.18	0.88
Lower face and jawline	18.7	20	0.56	0.88
Nasolabial folds	24.3	20	0.23	0.88
Under chin	19.1	20	0.51	0.89
Neck	33.8	40	0.74	0.90

the five appearance appraisal scales for face-lift patients has four response options. Instructions ask the participants to complete the items for each scale based on how they *look right now*, and to indicate how much in the past week they have been either "bothered by" or "satisfied with" the particular facial area.

Phase 2: Quantitative Research Results

Data Collection

In study 1, 360 patients were recruited face-to-face, 332 of whom responded (response rate, 92 percent); and 283 patients were recruited by mail, of whom 167 responded (response rate, 59 percent). The overall response rate was 78 percent. Table 2 lists characteristics of the 225 face-lift patients included in the following analyses.

Rasch Measurement Theory

Table 3 shows the summary fit statistics to the Rasch model (i.e., how closely the observed data match those expected by the model). A nonsignificant chi-square value supported the fit to the Rasch model for the five scales. Targeting was good, with minimal floor/ceiling effects, and all items in each of the five scales displayed ordered thresholds, indicating that respondents were able to distinguish between the four response options (data available on request). Table 4 shows the individual item fit statistics. The findings provide further support for each of the five scales as reliable and valid measures of their respective constructs. From the five scales, only one item had a fit residual marginally outside the recommended criteria of −2.5 to +2.5. This item was retained given that all other fit statistics were satisfied. The Person Separation Index values (Table 3) for the five scales were greater than or equal to 0.88, indicating good reliability.

Table 4. Rasch Measurement Theory Statistical Indicators of Fit

Scale	Items*	Item Location	SE	Fit Residual	χ^2	*p*
Cheeks	Symmetric	−0.50	0.12	0.71	7.36	0.12
	Smooth	−0.49	0.13	−0.50	2.62	0.46
	Attractive†	−0.19	0.13	−2.70	5.28	0.15
	Contour	0.13	0.12	−0.53	2.22	0.53
	Fullness	1.06	0.12	−1.54	3.56	0.31
Nasolabial folds	Deep	−0.39	0.11	−1.24	4.70	0.32
	Relaxed	−0.15	0.11	−1.70	3.70	0.45
	Old	−0.12	0.11	−0.46	2.39	0.67
	Smile	−0.11	0.11	0.27	1.36	0.85
	Compared	0.78	0.11	1.14	12.18	0.02
Lower face and jawline	Prominent	−0.55	0.12	0.04	7.21	0.13
	Sculpted	−0.15	0.11	−1.98	3.55	0.47
	Profile	0.06	0.11	−1.42	2.38	0.67
	Nice	0.15	0.12	0.35	1.03	0.91
	Smooth	0.50	0.11	−0.61	4.27	0.37
Under chin	Profile	−0.58	0.12	−0.72	4.98	0.29
	Loose	−0.43	0.13	1.02	3.32	0.51
	Sagging	−0.27	0.13	−1.99	2.16	0.71
	Contour	0.33	0.12	−1.17	3.98	0.41
	Fullness	0.96	0.13	1.47	4.70	0.32
Neck	Sagging	−0.69	0.11	−0.19	4.031	0.40
	Old	−0.51	0.11	−2.37	8.515	0.07
	Wrinkled	−0.30	0.11	0.82	0.833	0.93
	Profile	−0.28	0.11	−0.10	2.381	0.67
	Grimace	−0.10	0.11	1.74	2.104	0.72
	Hanging	0.14	0.11	−1.87	3.56	0.47
	Collars	0.15	0.11	0.38	1.699	0.79
	Lines	0.27	0.11	1.95	6.247	0.18
	Compared	0.58	0.12	−1.04	2.377	0.67
	Cover up	0.75	0.12	−0.94	2.054	0.73

*Items are in serial order for each scale.
†Indicates items with fit residual ±2.5.

Table 5. FACE-Q Scales (Linearized Data) before and 6 Months after Surgery

	Cheeks	Lower Face and Jawline	Nasolabial Folds	Under Chin	Neck
FACE-Q scale scores (linearized measurements; 0–100)					
No.	98	97	98	96	90
Before surgery					
Mean	45	44	38	48	55
SD	20	24	21	28	24
½ SD	10	12	11	14	
6 mo after surgery					
Mean	60	60	53	60	65
SD	26	26	25	30	25
½ SD	13	13	13	15	
Before to 6 mo after surgery					
Mean	16	16	16	11	11
SD	26	25	30	30	27
Indicators of group-level responsiveness (before to 6 mo after surgery)					
t test					
t	5.92	6.25	5.10	3.65	3.75
p	<0.001	<0.001	<0.001	<0.001	<0.001
RE	95%	100%	82%	58%	60%
1-way ANOVA					
F	46.2	20.5	45.6	16.2	18.1
p	<0.001	<0.001	<0.001	<0.001	<0.001
RP	100%	21%	99%	35%	39%
Effect size					
Kazis	0.79	0.67	0.74	0.40	0.43
SRM	0.60	0.64	0.69	0.37	0.39
Indicators of individual person-level responsiveness (before to 6 mo after surgery), no. (%)					
Significance of change					
Significant improvement	35 (36)	37 (38)	40 (41)	31 (32)	34 (37)
Nonsignificant improvement	30 (31)	26 (27)	20 (21)	25 (25)	21 (23)
No change	12 (12)	15 (16)	14 (14)	18 (19)	7 (8)
Nonsignificant worsening	15 (15)	11 (11)	18 (18)	10 (10)	20 (23)
Significant worsening	6 (6)	8 (8)	6 (6)	13 (14)	8 (9)

RP, relative precision; RE, relative efficiency; ANOVA, analysis of variance; SRM, standardized response mean.

Responsiveness

Table 5 shows that patient satisfaction/appraisal with aspects of their facial appearance improved significantly after treatment. The associated statistically significant change scores were associated with moderate effect sizes. In addition, preliminary minimal importance difference analyses suggested a 10- to 14-point difference in total scores. This difference was exceeded in our analysis (range mean change ± SD, 11 ± 27 to 16 ± 30). For individual-level results, depending on the scale, between 32 and 41 percent of patients who had face-lift procedures reported significant improvement in satisfaction with facial appearance.

Traditional Test Theory Analysis

All scales exceeded criteria for acceptability, reliability, and validity (Table 6). Specifically, Cronbach's alpha coefficients (≥0.94) and intraclass correlation coefficients (≥0.74) supported scale reliability. Scale validity was supported by the high Cronbach's alpha coefficients and interscale correlations that ranged between $r = 0.30$ and $r = 0.71$, showing that each scale measures a distinct but clinically related variable (Table 7). Our examinations of clinical known group validity (Table 8) revealed that our hypotheses relating to the patterns and significance of scores across subgroups were supported (i.e., FACE-Q scores were higher in participants who indicated they appeared younger than their actual age). Overall, our findings indicated that the items in each scale constituted a statistically conformable group, and that these scores were reliable and valid measures. Finally, Table 9 shows the frequency table for the adverse effects checklist.

DISCUSSION

The FACE-Q was developed using rigorous qualitative research that involved in-depth interviews with a varied sample of patients, extensive expert input, and modern psychometric methods to identify the best indicators of outcome for each scale. Our overriding goal was to address the lack of available patient-reported outcome tools for patients who undergo facial cosmetic surgery, minimally invasive cosmetic procedures, and/or facial injectables. We chose to develop scales and

Table 6. Traditional Psychometric Methods Including Data Quality, Scaling Assumptions, Targeting, and Reliability

	Data Quality	Scaling Assumptions					Targeting		Reliability		
	Item Missing Data (%)	Possible Range (midpoint)	Actual Score Range	Mean Score	SD	CITC	Floor/ Ceiling Effects (%)	Skewness	Cronbach's Alpha	Mean IIC	Range IIC
Cheeks											
Symmetric	1	1–4	1–4	3.08	0.96	0.83	9/41	−0.80	—	—	—
Smooth	1	1–4	1–4	3.06	0.94	0.85	8/39	−0.72	—	—	—
Attractive	1	1–4	1–4	2.98	0.93	0.91	7/35	−0.49	—	—	—
Contour	1	1–4	1–4	2.95	0.98	0.87	10/35	−0.56	—	—	—
Fullness	1	1–4	1–4	2.78	1.03	0.88	15/30	−0.38	0.95	0.80	0.74–0.85
Nasolabial folds											
Deep	1	1–4	1–4	2.66	0.97	0.85	12/24	−0.07	—	—	—
Relaxed	1	1–4	1–4	2.73	0.99	0.86	12/27	−0.15	—	—	—
Old	1	1–4	1–4	2.75	1.00	0.84	13/27	−0.28	—	—	—
Smile	1	1–4	1–4	2.72	1.00	0.83	12/28	−0.14	—	—	—
Compared	1	1–4	1–4	2.98	1.00	0.75	8/41	−0.43	0.94	0.74	0.66–0.84
Lower face and jawline											
Prominent	1	1–4	1–4	2.86	0.91	0.84	9/26	−0.47	—	—	—
Sculpted	1	1–4	1–4	2.76	0.99	0.88	12/27	−0.27	—	—	—
Profile	1	1–4	1–4	2.71	0.99	0.88	13/25	−0.23	—	—	—
Nice	1	1–4	1–4	2.77	0.98	0.84	12/25	−0.07	—	—	—
Smooth	1	1–4	1–4	2.60	1.01	0.86	15/24	−0.02	0.95	0.79	0.70–0.86
Under chin											
Profile	1	1–4	1–4	2.84	1.08	0.90	15/36	−0.41	—	—	—
Loose	1	1–4	1–4	2.87	1.03	0.97	13/35	−0.42	—	—	—
Sagging	1	1–4	1–4	2.89	1.05	0.92	13/37	−0.47	—	—	—
Contour	1	1–4	1–4	3.02	1.07	0.90	12/46	−0.65	—	—	—
Fullness	1	1–4	1–4	3.12	1.01	0.86	9/48	−0.79	0.96	0.83	0.78–0.89
Neck											
Sagging	1	1–4	1–4	3.02	1.02	0.88	10/42	−0.62	—	—	—
Old	1	1–4	1–4	3.06	1.00	0.88	10/44	−0.72	—	—	—
Wrinkled	1	1–4	1–4	3.10	0.96	0.85	8/43	−0.75	—	—	—
Profile	1	1–4	1–4	3.11	0.98	0.85	8/46	−0.76	—	—	—
Grimace	1	1–4	1–4	3.16	0.96	0.80	7/48	−0.82	—	—	—
Hanging	1	1–4	1–4	3.23	0.96	0.89	7/53	−0.97	—	—	—
Collars	1	1–4	1–4	3.22	0.96	0.85	7/53	−0.92	—	—	—
Lines	1	1–4	1–4	3.25	0.93	0.82	6/53	−1.01	—	—	—
Compared	1	1–4	1–4	3.29	0.92	0.84	5/56	−0.96	—	—	—
Cover up	1	1–4	1–4	3.42	0.93	0.83	6/66	−1.40	0.97	0.74	0.62–0.87

CITC, Corrected Item-Total Correlation; ICC, intraclass correlation coefficient;

checklists for anatomical areas of the face rather than instruments to evaluate outcomes particular to surgical or nonsurgical procedure as others have done.[4]

For face-lift patients, the FACE-Q appearance appraisal scales were found to be clinically meaningful, valid, reliable, and responsive to change 6 months after treatment. In addition, the adverse effects checklist was useful for identifying the proportion of patients experiencing postsurgical symptoms. We suggest that this set of FACE-Q scales represents a promising new set of tools that can be used with face-lift patients in both research and clinical practice.

In addition to the scales presented in this article, researchers and clinicians measuring

Table 7. Convergent and Discriminant Construct Validity of the FACE-Q Scales

	Cheeks	Lower Face and Jawline	Nasolabial Folds	Under Chin	Neck
Cheeks	*				
Lower face and jawline	0.64†	*			
Nasolabial folds	0.38†	0.36†	*		
Under chin	0.40†	0.61†	0.32†	*	
Neck	0.48†	0.55†	0.35†	0.71†	*

*Correlation=1.0.
†Correlations consistent with predictions.

Table 8. Mean Scores (SD) for FACE-Q Appearance Scales by Patient-Perceived Age Visual Analogue Scale Categorized into Five Age Groups

	How many years younger or older do you think you look compared with your actual age?					
Scale	More than 5 years older	1 to 5 years older	I look my age	1 to 5 years younger	More than 5 years younger	*p*
Cheeks	40 (23)	48 (26)	59 (26)	82 (23)	90 (15)	<0.001
Lower face and jawline	40 (31)	47 (24)	50 (27)	75 (25)	82 (23)	<0.001
Nasolabial folds	38 (21)	51 (27)	56 (28)	67 (26)	72 (27)	<0.001
Under chin	50 (35)	58 (31)	59 (29)	78 (27)	84 (24)	<0.001
Neck	52 (26)	61 (25)	67 (26)	84 (17)	87 (17)	<0.001

Table 9. FACE-Q Lower Face and Neck Adverse Effects Checklist: Proportion of Patients Reporting a Problem at 6-Month Follow-Up

Items	Not at All (%)	A Little (%)	A Lot (%)
Parts of face feeling numb	49 (58)	26 (31)	9 (11)
Tightness	57 (68)	21 (25)	6 (7)
Not looking smooth	61 (74)	10 (12)	11 (13)
Parts of face feeling sensitive	62 (74)	19 (23)	3 (4)
Tingling	63 (75)	14 (17)	7 (8)
How scars feel	65 (77)	12 (14)	17 (8)
Discomfort	66 (79)	15 (18)	3 (4)
Itching	67 (80)	13 (16)	4 (5)
How scars look	69 (82)	9 (11)	6 (7)
Pulling	70 (84)	11 (13)	2 (2)
Swelling	72 (86)	12 (14)	0 (0)
Parts of face feeling hard	74 (88)	7 (8)	3 (4)
Difficulty with facial expressions	74 (88)	2 (2)	8 (10)
Bruising	81 (96)	2 (2)	1 (1)
Difficulty with facial movements	81 (96)	2 (2)	1 (1)

outcomes in face-lift patients might also want to include the FACE-Q 10-item core scale, which measures overall satisfaction with facial appearance. This scale can be used to compare outcomes across any procedure type and/or to measure change before and after any facial aesthetic procedure.[28] Our team also developed a seven-item aging appraisal scale, which provides an assessment of a patient's perception of his or her appearance in the context of facial aging.[29]

Our current study has some limitations. First, it is rare to find a face-lift patient who has not previously had other facial aesthetic treatments before undergoing a face lift. In fact, in study 1, only 5.6 percent of our sample had not had any facial aesthetic procedure before their face lift. This finding reflects the nature of facial aesthetic patients and the challenge that exists in measuring the benefit of any particular facial aesthetic treatment. Second, our sample was composed of more women than men. Future research could investigate the use of FACE-Q scales with male patients. Third, although our response rate for face-to-face recruitment was high, our response rate to the mailed survey was lower than we would have liked. Fourth, it is possible there could have been some bias introduced at the individual clinic level by office staff who recruited patients for us. Finally, we acknowledge that the sample size in several countries was small. We therefore recommend further research be carried out to add to the evidence base for the use of the scales and the generalizability of their measurement properties.

CONCLUSIONS

Based on the development process and these preliminary validation data, we argue that our scales and checklist are tools that can used to advance knowledge about the outcomes that matter the most to face-lift patients. Our scales are short and easy to complete and have high face validity, making them the type of tools that can easily be incorporated into routine clinical practice. Previous research has shown that integration of patient-reported outcomes into clinical practice improves patient-clinician communication and can enhance patient care and outcomes.[30–32]

In addition to their use in clinical practice, we envision FACE-Q scales as important new metrics that can be used to define the outcomes of facial aesthetics with broad application in clinical research. For example, incorporation of FACE-Q scales into clinical trials could help to guide future

surgical innovation and advance comparative effective research in facial aesthetic treatments. Given an ever-growing range of interventions and products in facial aesthetic surgery, the incorporation of patient-reported outcome instruments into clinical research is absolutely essential if we are to understand the profound impact that cosmetic treatments have on the appearance and quality of life of patients. For researchers who plan to use the FACE-Q in future studies, it is important to note that our scales are designed to function independently from each other. This means that researchers can choose to administer only those scales that are most appropriate for their research hypothesis or patient population. For example, in a face-lift study, the scales described in this article might be used, whereas in a blepharoplasty study, only scales related to the eye might be selected. This approach minimizes response burden and improves targeting. We would also stress that study design and timing of FACE-Q administration is entirely at the discretion of individual research teams. As an example, an investigator may elect to use the FACE-Q before and after treatment in a randomized clinical trial, whereas another might select a cross-sectional cohort study design.

Andrea L. Pusic, M.D., M.H.S.
Plastic and Reconstructive Surgery
Memorial Sloan-Kettering Cancer Center,
Room MRI-1007
1275 York Avenue
New York, N.Y. 10065
pusica@mskcc.org

ACKNOWLEDGMENTS

This study was funded by grants from the Plastic Surgery Foundation. 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, British Columbia, Canada: Drs. Nick Carr, Francis Jang, Nancy VanLaeken, Alistair Carruthers, Jean Carruthers, and Richard Warren; Washington, D.C.: Dr. Stephen Baker; Dallas, Texas: Drs. Jeffery Kenkel and Rod Rohrich; Atlanta, Georgia: Dr. Foad Nahai; St. Louis, Missouri: Dr. Leroy Young; New York, New York: Drs. David Hidalgo, David Rosenberg, Philip Miller, Alexes Hazen, and Haideh Hirmand. The authors thank Jonathan Switzer for his assistance with final manuscript preparation.

REFERENCES

1. American Society of Plastic Surgeons. Available at: http://www.plasticsurgery.org/News-and-Resources/2011-Statistics-.html. Accessed November 29, 2013.
2. Honigman RJ, Phillips KA, Castle DJ. A review of psychosocial outcomes for patients seeking cosmetic surgery. *Plast Reconstr Surg.* 2004;113:1229–1237.
3. Shridharani SM, Magarakis M, Manson PN, Rodriguez ED. Psychology of plastic and reconstructive surgery: A systematic clinical review. *Plast Reconstr Surg.* 2010;126:2243–2251.
4. Kosowski TR, McCarthy C, Reavey PL, et al. A systematic review of patient-reported outcome measures after facial cosmetic surgery and/or nonsurgical facial rejuvenation. *Plast Reconstr Surg.* 2009;123:1819–1827.
5. Alsarraf R, Larrabee WF Jr, Anderson S, Murakami CS, Johnson CM Jr. Measuring cosmetic facial plastic surgery outcomes: A pilot study. *Arch Facial Plast Surg.* 2001;3:198–201.
6. Alsarraf R. Outcomes research in facial plastic surgery: A review and new directions. *Aesthetic Plast Surg.* 2000;24:192–197.
7. U.S. Food and Drug Administration. Guidance for Industry. Patient-Reported Outcome Measures: Use in Medical Product Development to Support Labeling Claims. Available at: http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM193282.pdf. Accessed November 29, 2013.
8. 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.
9. Klassen AF, Cano SJ, Scott A, Snell L, Pusic AL. Measuring patient-reported outcomes in facial aesthetic patients: Development of the FACE-Q. *Facial Plast Surg.* 2010;26:303–309.
10. Flesch R. A new readability yardstick. *J Appl Psychol.* 1948;32:221–233.
11. MAPI Research Trust (Web site). 2004–2012. Available at: http://www.mapitrust.org/. Accessed November 29, 2013.
12. Andrich D. Controversy and the Rasch model: A characteristic of incompatible paradigms? *Med Care* 2004;42(Suppl):1–16.
13. Wright BD, Masters G. Rasch measurement. In: *Rating Scale Analysis.* Chicago: Mesa; 1982.
14. Andrich D, Sheridan B, Luo G. *RUMM 2030.* 1997–2013. Perth, Australia: RUMM Laboratory.
15. Andrich D. *Rasch Models for Measurement.* Newbury Park, Calif: Sage; 1988.
16. Rasch G. *Probabilistic Models for Some Intelligence and Attainment Tests.* Copenhagen: Danish Institute for Education Research; 1960.
17. Cronbach LJ. Coefficient alpha and the internal structure of tests. *Psychometrika* 1951;16:297–334.
18. Kazis LE, Anderson JJ, Meenan RF. Effect sizes for interpreting changes in health status. *Med Care* 1989;27(Suppl):S178–S189.
19. Liang MH, Fossel AH, Larson MG. Comparisons of five health status instruments for orthopedic evaluation. *Med Care* 1990;28:632–642.
20. Hobart JC, Cano SJ, Thompson AJ. Effect sizes can be misleading: Is it time to change the way we measure change? *J Neurol Neurosurg Psychiatry* 2010;81:1044–1048.
21. Hobart JC, Cano SJ. Improving the evaluation of therapeutic intervention in MS: The role of new psychometric methods. *Health Technol Assess.* 2009;13:1–200.
22. McHorney CA, Haley SM, Ware JE Jr. 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;50:451–461.
23. Likert R. A technique for the measurement of attitudes. *Arch Psychol.* 1932;140:1–55.
24. Ware JE, Harris WJ, Gandek B, et al. *MAP-R for Windows: Multi-trait/Multi-Item Analysis Program—Revised User's Guide.* Boston: Health Assessment Laboratory; 1997.

25. Eisen M, Ware JE Jr, Donald CA, Brook RH. Measuring components of children's health status. *Med Care* 1979;17:902–921.
26. McHorney CA, Ware JE Jr, Lu JF, Sherbourne CD. The MOS 36-Item Short-Form Health Survey (SF-36): III. Tests of data quality, scaling assumptions and reliability across diverse patient groups. *Med Care* 1994;32:40–66.
27. Bohrnstedt GW. Measurement. In: Rossi PH, Wright JD, Anderson AB, eds. *Handbook of Survey Research*. New York: Academic Press; 1983:69–121.
28. Pusic AL, Klassen AK, Scott A, Cano SJ. Development and psychometric evaluation of the FACE-Q satisfaction with appearance scale: A new patient-reported outcome instrument for facial aesthetic patients. *Clin Plast Surg*. 2013;40:249–260.
29. Panchapakesan V, Klassen AF, Cano SJ, et al. Development and psychometric evaluation of the FACE-Q Aging Appraisal Scale and Patient-Perceived Age Visual Analog Scale. *Aesthet Surg J*. 2013;33:1099-1109.
30. Marshall S, Haywood K, Fitzpatrick R. Impact of patient-reported outcome measures on routine practice: A structured review. *J Eval Clin Pract*. 2006;12:559–568.
31. Valderas JM, Kotzeva A, Espallargues M, et al. The impact of measuring patient-reported outcomes in clinical practice: A systematic review of the literature. *Qual Life Res*. 2008;17:179–193.
32. Greenhalgh J, Meadows K. The effectiveness of the use of patient-based measures of health in routine practice in improving the process and outcomes of patient care: A literature review. *J Eval Clin Pract*. 1999;5:401–416.