FACE-Q Scales for Health-Related Quality of Life, Early Life Impact, Satisfaction with Outcomes, and Decision to Have Treatment: Development and Validation

Anne F. Klassen, D.Phil.
Stefan J. Cano, Ph.D.
Jonathan A. Schwitzer, B.A.
Amie M. Scott, B.Sc., M.P.H.
Andrea L. Pusic, M.D.

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

Background: An ever-growing range of facial cosmetic products and treatments are available, but little clinical research is being performed to determine treatment outcomes from the patient's perspective. The FACE-Q is a patient-reported outcome instrument composed of more than 40 independently functioning scales and checklists. The aim of this article is to describe the development and psychometric evaluation of five new FACE-Q scales.
Methods: FACE-Q scales were developed according to international guidelines for patient-reported outcome instrument development. The following FACE-Q scales and a single symptom checklist (Recovery Early Symptoms) were evaluated in this study: Psychological Wellbeing, Social Function, Satisfaction with Decision to Have Treatment, Satisfaction with Outcome of Treatment, and Early Life Impact of Treatment. Modern and traditional psychometric methods were used to examine reliability, validity, and responsiveness.
Results: The sample included 702 participants from three studies. The FACE-Q scales were found to be reliable, valid, and responsive to clinical change. These findings were supported by Rasch measurement theory (e.g., overall chi-square values, $p \geq 0.06$; Person Separation Index ≥0.81), traditional psychometric (e.g., Cronbach alpha values ≥0.90) and responsiveness (i.e., significant improvement following face lift and lip treatment) analysis.
Conclusions: The FACE-Q measures concepts and symptoms important to facial aesthetic patients. The five scales and single symptom checklist described here can be used to measure what patients think about cosmetic treatments in a scientifically sound manner. As the cosmetics industry continues to expand, the patient perspective of treatment outcomes should be measured and reported. (*Plast. Reconstr. Surg.* 135: 375, 2015.)
CLINICAL QUESTION/LEVEL OF EVIDENCE: Diagnostic, III.

As worldwide acceptance and range of accessible cosmetic treatments continue to expand, a better understanding of the outcomes of such treatments from the perspective of the patient is needed. Measuring what patients think about cosmetic treatments in a scientifically sound manner can be accomplished by asking them to complete patient-reported outcome instruments before and after treatment.[1] Patient-reported outcome instruments

From McMaster University; Plymouth University Peninsula Schools of Medicine and Dentistry; and Memorial Sloan Kettering Cancer Center.
Received for publication April 9, 2014; accepted June 6, 2014.
Preliminary findings for the Psychological Wellbeing and Social Function scales were presented orally at the 29th Annual Scientific Meeting of the British Association of Aesthetic Plastic Surgeons, in London, United Kingdom, September 27, 2013.

DOI: 10.1097/PRS.0000000000000895

Disclosure: *The FACE-Q is owned by Memorial Sloan Kettering Cancer Center. Drs. Cano, Klassen, and Pusic 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. The other authors have no financial interests to disclose.*

This work was supported by THE PLASTIC SURGERY FOUNDATION.

are questionnaires that measure concepts such as symptoms, satisfaction with appearance, and health-related quality of life.[2] Such measures were initially designed for use in academic and industry research but are increasingly being used for other purposes, including clinical care, patient/consumer education, benchmarking, and quality improvement.[3] For example, in the United Kingdom, the BREAST-Q, which measures satisfaction and health-related quality of life, was included in a voluntary nationwide audit of over 8000 mastectomy and breast reconstruction patients treated between 2008 and 2009.[4]

A literature review commissioned by the United Kingdom Department of Health and published in 2013[5] identified that there are nine patient-reported outcome instruments for cosmetic treatments developed with patient input that have demonstrated adequate psychometric properties.[6–16] Of these nine, the review authors stated that only three meet current recommendations for the development and validation of patient-reported outcome instruments (i.e., BREAST-Q,[8,9] FACE-Q,[13] and Skindex[16]). The review concluded that research dedicated to the evaluation of patient-reported outcome instruments in cosmetic surgery is urgently required.

The FACE-Q encompasses a set of more than 40 independently functioning scales and checklists measuring concepts and symptoms important to facial aesthetic patients in relation to different facial areas (Table 1). FACE-Q scales can be administered to any facial aesthetic patient (surgical and nonsurgical) to measure what patients think about their appearance, quality of life, adverse effects, and process of care. Each scale provides a standalone score from 0 to 100, with higher scores indicating a better outcome. 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. Given the large number of FACE-Q scales addressing various facial anatomical areas and aspects of outcome, results for the scales are being published as a series of articles, each of which describes clinically relevant scale grouping (e.g., FACE-Q scales for face-lift patients)[17] and aspect of outcomes (satisfaction, health-related quality of life, adverse effect).

The aim of the present article is to describe the development and psychometric evaluation of FACE-Q scales that evaluate quality-of-life outcomes and decisional satisfaction. These include five scales (Psychological Wellbeing, Social Function, Satisfaction with Decision to Have Treatment, Satisfaction with Outcome of Treatment, and Early Life Impact of Treatment) and a single symptom checklist (Recovery Early Symptoms).

Table 1. FACE-Q Scales

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 function*
Aging appraisal
Expectations and motivations
Psychological distress
Early life impact*
Adverse effect checklists
Recovery early symptoms*
Skin
Forehead, scalp, and eyebrows
Eyes
Nose
Lower face and neck
Lips
Ears
Patient experience scales
Outcome*
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 starting the study. The FACE-Q was developed following internationally recommended guidelines for patient-reported outcome instruments.[2,18–21] The methods are described in detail elsewhere[13,17,22,23] and described briefly below.

Phase 1: Qualitative Research Methods

In the first phase, we developed a conceptual framework using a mixed methods approach that involved a literature review,[24] expert input, and interviews with 50 patients who had undergone one or more surgical and/or nonsurgical facial aesthetic procedures. The framework includes four domains (appearance, quality of life, adverse effects, and process of care). The qualitative data were used to develop a comprehensive set of items, which in turn were used to form the set of FACE-Q scales. The following five scales and symptom checklist are described in this article:

1. Psychological Wellbeing. This scale measures psychological well-being in terms of a series of positively worded statements (e.g., "I feel happy"). Instructions ask that respondents answer with their facial appearance in mind and to think about the past week. Four response options are provided (i.e., definitely agree, somewhat agree, somewhat disagree, and definitely disagree).
2. Social Function: This scale has a series of positively worded statements (e.g., "I am comfortable meeting new people") that measure social confidence. Instructions ask that respondents answer with their facial appearance in mind and to think about the past week. Four response options are provided (i.e., definitely agree, somewhat agree, somewhat disagree, and definitely disagree).
3. Satisfaction with Decision to Have Treatment: This scale has a series of positively worded items that ask respondents to indicate how they feel about the outcome of their most recent facial aesthetic procedure (e.g., "It was worth the time and effort"). Four response options are provided (i.e., definitely agree, somewhat agree, somewhat disagree, and definitely disagree).
4. Satisfaction with Outcome of Treatment: This scale has a series of positively worded items that ask respondents to indicate how they feel about the result of their most recent facial aesthetic procedure (e.g., "The result was just as I expected"). Four response options are provided (i.e., definitely agree, somewhat agree, somewhat disagree, and definitely disagree).
5. Early Life Impact: This scale is composed of statements that ask about recovery from a facial aesthetic procedure. Respondents are asked to indicate how often in the past 2 days each statement applies to them (e.g., avoided certain facial movements). Four response options are provided (i.e., not at all, a little of the time, some of the time, and most of the time).
6. Recovery Early Symptoms: This checklist asks about physical symptoms following a facial aesthetic procedure (e.g., discomfort, feeling tired). Instructions ask respondents to indicate how much in the past 2 days they have been bothered by each symptom. Four response options are provided (i.e., extremely, moderately, a little, and not at all).

Phase 2: Quantitative Research Methods

Study 1: Data Collection

Plastic surgeons and dermatologists in the United States and Canada recruited patients between June of 2010 and November of 2013. Eligible participants were 18 years of age or older who had undergone or were waiting to undergo any surgical or nonsurgical facial aesthetic procedure. Patients were recruited in person and through a mail-out survey. The field-test data were used to develop and validate the Psychological Wellbeing, Social Function, Satisfaction with Decision to Have Treatment, and Satisfaction with Outcome scales.

Studies 2 and 3: Data Collection

Two companies used the Early Life Impact scale and the Recovery Early Symptom checklist in international clinical trials. Study 2 was a face-lift study where the scale and checklist were administered on days 2, 3 through 5, 7, and 30 after treatment. Study 3 was a minimally invasive lip treatment study where the scale and checklist were administered on days 1 and 14. From both studies, data collected in the first week after treatment (up to day 7) were used to develop and validate the Early Life Impact scale.

Study 2 also administered the Psychological Wellbeing and Social Function scales at baseline and days 14, 30, and 90, and the Satisfaction with Decision to Have Treatment and the Satisfaction with Outcome of Treatment scales on days 14, 30, and 90. Data from study 2 were included in the analysis to assess construct validity and responsiveness.

Data Analyses

The Recovery Early Symptom checklist data were examined descriptively in terms of the proportion of participants that answered to each

response option on each posttreatment assessment day. Rasch measurement theory,[25,26] conducted within RUMM2030 software (RUMM Laboratory, Perth, Australia),[27] was used to identify a set of items that represented the best indicators of outcome for each of the five scales. The analysis uses a number of statistical and graphic tests to examine each item in a scale.[28–30] The evidence is considered together to make decisions about the overall quality of a scale. When data fit the Rasch model, estimates from the model are considered appropriate because the measurement theory is supported by the data. The following tests and criteria were used by our team:

1. Thresholds for item response options: The use of response categories scored with successive integer scores implies a continuum (e.g., increasing psychological well-being). This assumption was tested by examining the ordering of thresholds, which are points of crossover between adjacent response categories (e.g., between somewhat disagree and somewhat agree).
2. Item fit statistics: The items that compose a scale must concurrently satisfy statistical and clinical meaning. We examined three indicators of fit as follows: (1) log residuals (item-person interaction); (2) chi-square values (item-trait interaction); and (3) item characteristic curves. These fit statistics should be interpreted together in the context of their clinical usefulness. Criteria for fit residuals should fall between −2.5 and +2.5, and chi-square values should be nonsignificant after Bonferroni adjustment.
3. Item locations: The items of a scale are meant to define a continuum. Inspecting where items are located on the continuum in relation to each other indicates how well the items map out a construct; items should be evenly spread over a reasonable range.
4. Person Separation Index: This reliability statistic is comparable to the Cronbach alpha[31] and is used to quantify error associated with the measurements of people in a sample. Higher values indicate greater reliability.

Traditional psychometric tests[28] included the following: data quality (percentage missing data for each item), scaling assumptions (similarity of item means and variances, and magnitude and similarity of corrected item-total correlations),[32–34] scale-to-sample targeting (score means, standard deviation, and floor and ceiling effects), and internal consistency reliability (Cronbach alpha[31] and homogeneity coefficients).[35] The following aspects of construct validity were assessed to examine scale performance:

1. We correlated scores for the five scales to determine the extent to which they measured separate but related constructs.[36] We hypothesized that these intercorrelations would range between $r = 0.30$ and $r = 0.70$, as these scales were developed to measure distinct but related clinical variables.[37]
2. We correlated scores for five scales with patient characteristics [i.e., age, sex, and ethnicity (where available)] to determine the extent to which a scale may be susceptible to bias by these sociodemographic variables. We predicted that these correlations would be low (<0.30).[38,39]
3. We computed a new variable to represent the total number of symptoms (zero to 17) on the Recovery Early Symptom checklist and correlated this with the Early Life Impact of Treatment scale score. We hypothesized that more symptoms would correlate with lower Early Life Impact of Treatment scores (i.e., worse health-related quality of life during recovery).

We examined responsiveness by comparing mean Rasch-transformed scores using analysis of variance and *t*-test statistics, and computing the effect sizes as described by Kazis et al.[40] and standardized response means.[41] The magnitude of the change for the latter two indicators of change can be interpreted using Cohen's arbitrary criteria (i.e., small, 0.20; moderate, 0.50; and large, 0.80). A minimally importance difference was calculated as a half standard deviation of the pretreatment mean score and by extrapolation of a change score based on a 0.5 effect size.[42]

RESULTS

In study 1, 228 participants were recruited face-to-face, of whom 220 responded (response rate, 96.5 percent); 199 participants were recruited by mail, and 103 responded (response rate, 51.7 percent). The overall response rate was 75.6 percent. Table 2 lists characteristics of the 702 participants.

In the Rasch measurement theory analysis, all items for four scales had ordered thresholds. The exception was the Early Life Impact of Treatment scale, where thresholds for seven of the 12 items were disordered. Specifically, participants

Table 2. Patient Characteristics

	Study 1 (%)	Study 2 (%)	Study 3 (%)
No. of participants	323	100	279
Age, yr			
Mean ± SD	45.6 ± 16.8	54.3 ± 7.8	47.9 ± 12.2
Range	18–89	37–77	18–76
Sex			
Female	259 (80.2)	88 (88)	274 (98.2)
Male	51 (15.8)	12 (12)	5 (1.8)
Missing	13 (4.0)	—	—
Country			
United States	308 (95.4)		—
Canada	15 (4.6)		—
France	—	15 (15)	98 (35.1)
Germany	—	50 (50)	—
Israel	—	20 (20)	—
United Kingdom	—	15 (15)	181 (64.9)
Procedure			
Fillers and/or botulinum toxin	17(5.3)	—	—
Skin resurfacing	33 (10.2)	—	—
Lip injections	8 (2.5)	—	279 (100)
Face lift	96 (29.7)	100 (100)	—
Rhinoplasty	87 (26.9)	—	—
Chin and/or jaw surgery	36 (11.2)	—	—
Cheeks	9 (2.8)	—	—
Blepharoplasty	12 (3.7)	—	—
Multiple surgical procedures	25 (7.7)	—	—
No. of participants and assessments by scale/checklist			
Psychological	286/305	—	278/1075
Social	288/307	—	278/1076
Decision	181/189	—	271/797
Outcome	180/199	—	271/798
Early Life Impact of Treatment	—	94/365	278/549
Recovery Early Symptoms	—	94/365	274/544

were not able to distinguish between the two middle response options (i.e., a little of the time and some of the time). To simplify the scoring and ensure that all items had ordered thresholds, we rescored all 12 items into three response options by collapsing into one category the two middle response options. Subsequent Rasch measurement theory analyses for this scale used the rescored data.

Table 3 lists the summary fit statistics. A nonsignificant chi-square value was achieved for all five scales. This finding supported the data fit to the Rasch model for each scale. The Person Separation Index values for the five scales ranged from 0.81 to 0.90, indicating good reliability.

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
Psychological	108.96	88	0.06	0.89
Social	66.06	64	0.41	0.90
Decision	52.99	48	0.29	0.81
Outcome	62.6	50	0.11	0.85
Early Life Impact	131.21	108	0.06	0.81

Table 4 lists the individual item-fit statistics. Fit residuals for 37 of the 42 items were within the recommended criteria of −2.5 to +2.5. The five items that were found to fall marginally outside of the criteria range were retained, given that the other fit statistics were satisfied. The chi-square *p* values for all 42 items were nonsignificant, indicating good item fit.

Items within each of the five scales in Table 4 reflect the serial order of each scale. For example, in the Satisfaction with Outcome of Treatment scale, the easiest item for participants to strongly agree with was "I am pleased with the result," which had an item location of −1.15, and the hardest item for participants to strongly agree with was "The result is miraculous," which had an item location of 1.87.

Tables 5 and 6 list the traditional psychometric results. The scale level Flesch-Kincaid grade ranged from 0 to 6. All scales exceeded criteria for acceptability, reliability, and validity. Cronbach alpha coefficients (≥0.90) and interitem correlations (≥0.43) supported scale reliability.

Scale validity was supported by the high Cronbach alpha coefficients and interscale correlations. The interscale correlations (Table 7) for

Table 4. Rasch Measurement Theory Statistical Indicators of Fit

Scale and Items*	Item Location	SE	Fit Residual	χ^2	p
Psychological					
Like	–1.43	0.16	–1.86	16.30	0.06
Positive	–0.83	0.15	–2.45	12.30	0.14
Okay	–0.46	0.15	1.56	23.08	0.01
Happy	–0.22	0.14	0.26	4.53	0.87
Comfortable	–0.05	0.14	–1.95	4.09	0.91
Accepting	–0.01	0.14	0.45	3.60	0.89
Good†	0.04	0.14	–3.99	13.70	0.13
Confident	0.39	0.14	–0.87	6.48	0.69
Attractive	1.07	0.13	1.20	10.65	0.30
Great	1.50	0.13	–0.61	14.24	0.11
Social					
First impressions	–1.02	0.15	1.64	11.77	0.16
Confident new person	–0.95	0.15	–2.04	10.60	0.23
Comfortable meeting people	–0.92	0.15	–0.83	9.01	0.34
Make friends easily	–0.16	0.14	0.87	9.06	0.34
Group situations	0.38	0.14	–0.14	4.55	0.80
New social situations	0.46	0.14	–1.32	5.37	0.72
Relaxed around strangers	0.67	0.14	–1.81	12.62	0.13
Room full of people	1.54	0.14	–0.78	3.08	0.93
Decision					
Worth time/effort	–1.85	0.20	–1.78	13.73	0.09
Money well spent	–0.48	0.15	1.55	4.12	0.85
Just what wanted†	–0.05	0.15	–2.95	12.82	0.12
Just what needed	0.48	0.15	–1.19	7.07	0.53
Look how I want to	0.59	0.15	1.51	3.40	0.91
Life better†	1.31	0.14	2.78	11.86	0.16
Outcome					
Pleased result	–1.15	0.15	–0.71	15.31	0.08
Result great†	–0.89	0.14	–3.02	10.65	0.22
Result expected†	–0.14	0.14	3.34	20.06	0.01
Look good mirror	0.10	0.14	0.16	2.72	0.95
Result fantastic	0.21	0.14	–1.35	6.38	0.70
Result miraculous	1.87	0.14	1.32	7.48	0.49
Early Life Impact					
Regret	–1.64	0.11	–0.08	3.26	0.95
Anxious	–0.84	0.09	0.16	3.36	0.95
Sleeping	–0.79	0.09	1.18	12.75	0.17
Worthwhile	–0.75	0.09	1.42	12.36	0.19
Tired	–0.46	0.09	0.95	13.12	0.16
Head movements	0.27	0.08	–1.05	10.00	0.35
Usual activities	0.51	0.07	–2.45	19.89	0.02
Facial movements	0.54	0.08	0.27	6.03	0.74
Drinking	0.58	0.08	0.67	11.35	0.25
Eating	0.63	0.08	–0.80	7.98	0.54
Social situations	0.79	0.07	–2.22	18.85	0.03
Intimacy	1.14	0.07	–0.39	12.26	0.20

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

Table 5. Summary Statistics for Flesch-Kincaid Reading Grade and Traditional Test Theory Results

Scale	Flesh-Kincaid Reading Grade	Cronbach Alpha	IIC	
			Mean	Range
Psychological	2.0	0.96	0.73	0.58–0.87
Social	3.9	0.96	0.76	0.69–0.82
Decision	0	0.90	0.62	0.43–0.78
Outcome	1.7	0.91	0.64	0.48–0.79
Early Life Impact	6.0	0.90	0.43	0.26–0.77

IIC, interitem correlations.

four scales (exception, Early Life Impact of Treatment) were in the hypothesized range; these scales measured a distinct but clinically related variable. The Early Life Impact of Treatment correlated less than 0.30 with the other scales, indicating that the construct measured by this scale was distinct. The correlations between the scales and participant characteristics were low, as predicted, suggesting that little bias was present.

The correlation between total number of symptoms experienced following treatment and the Early Life Impact of Treatment scores in the first week following treatment were strong, as predicted (lip filler participants, $r = -0.66$, $p < 0.01$; face-lift participants, $r = -0.46$, $p < 0.01$).

Table 8 shows the proportion of participants to report being bothered by a symptom on the

Table 6. Traditional Test Theory Results Including Data Quality, Scaling Assumptions, Targeting, and Reliability

		Data Quality	Scaling Assumptions					Targeting	
Scale and Items	FK Grade Level	Missing Data (%)	Possible Range	Actual Range	Mean Score	SD	CITC	Floor/Ceiling Effects	Skewness
Psychological									
Like	0	2.3	1–4	1–4	3.55	0.64	0.83	1/62	−1.44
Positive	6.4	0	1–4	1–4	3.49	0.68	0.87	2/58	−1.31
Okay	4.4	1.0	1–4	1–4	3.52	0.69	0.74	2/61	−1.49
Happy	0	1.0	1–4	1–4	3.47	0.72	0.82	3/58	−1.41
Comfortable	4.4	0.3	1–4	1–4	3.40	0.75	0.86	2/54	−1.15
Accepting	2.4	0.3	1–4	1–4	3.36	0.75	0.82	2/51	−0.94
Good	0.5	1.0	1–4	1–4	3.39	0.76	0.90	3/53	−1.15
Confident	0.7	0.7	1–4	1–4	3.32	0.78	0.85	3/48	−1.01
Attractive	0.7	0.7	1–4	1–4	3.20	0.84	0.82	4/43	−0.83
Great	2.4	0.7	1–4	1–4	3.12	0.87	0.85	5/39	−0.69
Social									
First impressions	0.6	0.3	1–4	1–4	3.44	0.73	0.81	2/57	−1.22
New person	3.6	0.3	1–4	1–4	3.48	0.79	0.86	2/64	−1.36
Meeting people	5.6	0	1–4	1–4	3.45	0.79	0.86	2/62	−1.27
Make friends easily	1.2	0.3	1–4	1–4	3.37	0.80	0.83	3/54	−1.14
Group situations	7.7	0.3	1–4	1–4	3.32	0.85	0.85	5/52	−1.18
New social	6.0	0.3	1–4	1–4	3.25	0.85	0.88	4/48	−0.90
Strangers	2.6	0.7	1–4	1–4	3.22	0.86	0.89	4/46	−0.81
Room full of people	4.6	1.0	1–4	1–4	3.09	0.90	0.86	7/39	−0.75
Decision									
Worth time/effort	0	1.6	1–4	1–4	3.74	0.58	0.73	2/79	−2.60
Money well spent	0	0.5	1–4	1–4	3.53	0.77	0.70	3/67	−1.67
Just what wanted	0	0	1–4	1–4	3.34	0.85	0.86	4/55	−1.14
Just what needed	0.6	2.1	1–4	1–4	3.21	0.87	0.81	5/45	−0.93
Look how I want	0	1.6	1–4	1–4	3.17	0.86	0.73	5/42	−0.81
Life better	0	1.6	1–4	1–4	2.97	0.97	0.65	9/36	−0.56
Outcome									
Pleased result	0	1.6	1–4	1–4	3.50	0.79	0.72	4/64	−1.61
Result great	0	1.1	1–4	1–4	3.35	0.87	0.86	4/57	−1.16
Result expected	3.7	0.5	1–4	1–4	3.12	0.91	0.66	5/43	−0.64
Look good mirror	1.8	2.7	1–4	1–4	3.01	0.90	0.78	6/35	−0.49
Result fantastic	2.8	2.1	1–4	1–4	3.02	0.99	0.82	9/41	−0.62
Result miraculous	7.6	3.2	1–4	1–4	2.53	1.03	0.73	20/20	−0.09
Early Life Impact									
Regret	8.1	1.1	1–4	1–4	1.36	0.74	0.50	2/78	1.99
Anxious	2.4	1.1	1–4	1–4	1.50	0.85	0.56	5/68	1.67
Sleeping	0.7	0.2	1–4	1–4	1.50	0.88	0.53	6/70	1.66
Worthwhile	6.2	1.1	1–4	1–4	1.57	0.88	0.53	5/65	1.38
Tired	0	1.1	1–4	1–4	1.65	0.94	0.55	7/61	1.24
Head movements	6.0	0.4	1–4	1–4	1.95	1.16	0.67	17/53	0.73
Usual activities	7.6	0.5	1–4	1–4	2.05	1.24	0.71	22/52	0.62
Facial movements	9.6	1.3	1–4	1–4	2.08	1.18	0.68	19/47	0.55
Drinking	6.6	0.2	1–4	1–4	2.11	1.07	0.65	15/37	0.50
Eating	6.6	0	1–4	1–4	2.15	1.11	0.70	17/38	0.42
Social situations	9.5	0.2	1–4	1–4	2.18	1.26	0.71	26/46	0.44
Intimacy	10.7	0.2	1–4	1–4	2.37	1.29	0.67	32/40	0.19

FK, Flesch-Kincaid; CITC, corrective item-total correlation.

Recovery Early Symptom checklist. Common early (day 2) symptoms reported by face-lift participants included discomfort (96 percent), feeling your face is tight (95 percent), tenderness (91 percent), and numbness (loss of feeling) (90 percent). For lip filler participants, common early (day 1) symptoms included swelling (92 percent), tenderness (85 percent), discomfort (73 percent), and feeling sore (71 percent).

Table 9 shows the responsiveness results. For the Early Life Impact of Treatment scale, for both face-lift and lip participants, a large improvement in scores was reported by face-lift participants between their first (day 3–5) and final (day 30) assessments (effect size, 1.7; standardized response mean, 1.5), and for lip treatment participants between day 1 and day 14 (effect size, 1.0; standardized response mean, 1.0). For lip treatment participants (first/last assessment = baseline/90 days) showed that there were moderate to large changes in the Psychological Wellbeing (effect size, 0.8; standardized response mean, 0.7) scale and moderate changes in the Social Function (effect size, 0.5; standardized response mean,

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

	Psychological		Social		Decision		Outcome		Early Life Impact	
	r	*p*	*r*	*p*	*r*	*p*	*r*	*p*	*r*	*p*
Psychological										
Social	0.67	<0.01								
Decision	0.55	<0.01	0.42	<0.01						
Outcome	0.51	<0.01	0.31	<0.01	0.66	<0.01				
Early Life Impact	0.24	<0.01	0.21	<0.01	0.29	<0.01	0.24	<0.01		
Ethnicity	0.05	0.37	0.05	0.44	0.13	0.09	0.13	0.09		
Age	0.14	0.02	0.18	0.01	0.22	<0.01	0.12	0.10	0.13	<0.01
Sex	0.05	0.44	0.06	0.31	−0.02	0.78	−0.14	0.06	−0.01	0.82

Table 8. Proportion of Participants to Report a Problem for Each Symptom on the Early Symptom Checklist

	Lips		Face Lift			
Symptoms and Responses	**Day 1 (*n* = 278)**	**Day 14 (*n* = 272)**	**Day 2 (*n* = 94)**	**Days 3–5 (*n* = 93)**	**Day 7 (*n* = 88)**	**Day 30 (*n* = 92)**
Swelling						
Not at all	5	78	15	22	33	61
A little	48	13	39	38	42	24
Moderate	26	7	26	26	17	12
Extreme	18	1	20	13	6	2
Missing	2	1	0	2	2	1
Tenderness						
Not at all	13	79	7	11	10	27
A little	54	16	30	37	52	52
Moderate	23	4	39	37	28	19
Extreme	8	0.4	22	15	7	2
Missing	2	1	1	1	2	0
Discomfort						
Not at all	26	84	4	7	13	36
A little	46	12	36	53	58	46
Moderate	18	3	44	25	24	13
Extreme	9	0	16	13	2	3
Missing	1	1	0	3	3	2
Feel bruised						
Not at all	30	86	12	20	31	57
A little	46	10	39	41	43	33
Moderate	17	3	28	25	21	10
Extreme	6	0.7	19	10	2	0
Missing	1	1	2	4	3	1
Feel sore						
Not at all	28	85	46	56	66	77
A little	43	11	18	20	22	17
Moderate	21	2	25	17	10	5
Extreme	7	0.7	10	2	1	0
Missing	2	2	2	4	1	0
Feeling your face is tight						
Not at all	51	92	4	5	7	34
A little	32	4	20	31	43	42
Moderate	10	2	44	38	32	21
Extreme	5	0	31	24	16	3
Missing	1	2	1	1	2	0
Pain						
Not at all	55	92	19	32	39	65
A little	28	4	32	36	38	28
Moderate	10	2	33	25	18	5
Extreme	5	0.4	16	7	2	0
Missing	2	1	0	1	3	1
Numbness (loss of feeling)						
Not at all	62	89	9	12	21	32
A little	27	9	26	39	35	46
Moderate	5	0.4	39	29	30	21
Extreme	5	0	25	18	8	2
Missing	1	1	2	2	6	0

(*Continued*)

Table 8. (*Continued*)

Symptoms and Responses	Lips		Face Lift			
	Day 1 (*n* = 278)	Day 14 (*n* = 272)	Day 2 (*n* = 94)	Days 3–5 (*n* = 93)	Day 7 (*n* = 88)	Day 30 (*n* = 92)
Stinging						
Not at all	71	92	38	45	56	62
A little	19	5	21	27	24	29
Moderate	6	2	31	20	17	7
Extreme	3	0	10	7	2	2
Missing	1	1	0	1	1	0
Tingling						
Not at all	65	89	68	72	77	84
A little	25	7	21	20	16	14
Moderate	6	2	9	7	5	2
Extreme	3	0.7	2	0	0	0
Missing	1	1	0	1	2	0
Throbbing						
Not at all	69	93	61	60	64	73
A little	20	5	23	26	26	20
Moderate	7	1	16	12	9	7
Extreme	3	0.4	0	2	0	1
Missing	1	1	0	1	1	0
Burning						
Not at all	80	95	34	39	38	41
A little	12	3	28	30	40	38
Moderate	4	1	25	23	16	15
Extreme	3	0	14	8	6	5
Missing	1	1	0	1	1	0
Feeling tired						
Not at all	76	91	71	62	65	69
A little	13	6	26	29	25	29
Moderate	6	1	3	7	9	2
Extreme	4	1	0	1	0	0
Missing	1	1	0	1	1	0
Itching						
Not at all	79	92	66	73	72	78
A little	14	5	22	22	23	16
Moderate	4	2	11	2	5	5
Extreme	2	0.4	1	2	0	0
Missing	1	1	0	1	1	0
Feeling light-headed						
Not at all	91	97	40	51	60	76
A little	4	2	33	29	24	17
Moderate	2	0	16	15	9	5
Extreme	1	0.4	11	4	2	1
Missing	1	1	0	1	5	0
Headache						
Not at all	84	94	75	83	83	94
A little	10	3	15	10	10	5
Moderate	3	1	6	5	6	1
Extreme	1	0.7	3	1	0	0
Missing	2	1	1	1	1	0
Feeling feverish						
Not at all	93	96	80	67	66	73
A little	3	2	15	20	23	24
Moderate	2	1	4	10	9	2
Extreme	1	0.4	1	1	1	1
Missing	1	1	0	2	1	0

0.5) scale. The low but positive effect size/standardized response means across the assessments for the Satisfaction with Outcome of Treatment and Satisfaction with Decision to Have Treatment scales indicate that these are relatively stable over time. Comparing first/last assessment change scores across all scales against the most conservative (largest) minimum important differences in each instance revealed that these were surpassed in all instances except for Satisfaction with Outcome of Treatment and Satisfaction with Decision to Have Treatment scales. Early Life Impact of Treatment and Psychological Wellbeing scales showed the largest differences.

Table 9. FACE-Q Scale Mean Scores, Standard Deviation, and *p* Value for Difference between Groups, Incremental Effect Size Statistics, and Minimum Important Differences for Each Assessment Time Point*

Scale and Assessment Days	No.	Mean	SD	*p*	Effect Size		MID		
					ES	SRM	0.5 SD	0.5 ES	0.5 SRM
Early Life Impact									
2	94	53.6	15.2		—	—	7.6	—	—
4	91	59.8	18.6		0.4	0.5	9.3	7.6	6.2
7	86	65.2	17.9		0.3	0.4	9.0	9.0	5.2
30	92	80.1	16.2	<0.01	0.8	0.9	8.1	9.0	7.9
Early Life Impact									
1	278	73.6	18.5		—	—	9.2	7.6	8.7
14	271	92.6	12.7	<0.01	1.0	1.0	6.3	8.9	10.4
Social Function									
0	278	58.8	18.7		—	—	9.3	—	—
14	270	66.8	22.5		0.4	0.4	11.3	9.4	10.1
30	264	70.0	19.8		0.2	0.2	9.9	11.3	8.3
90	264	68.1	19.2	<0.01	–0.1	–0.1	9.6	10.0	7.3
Psychological Wellbeing									
0	278	53.6	16.3		—	—	8.1	—	—
14	270	65.6	21.3		0.8	0.6	10.7	8.2	9.9
30	263	69.2	19.7		0.2	0.2	9.9	10.7	9.1
90	264	66.4	19.2	<0.01	–0.2	–0.2	9.6	9.9	8.6
Satisfaction with Outcome of Treatment									
14	271	62.6	21.9		—	—	11.0	—	—
30	264	62.5	21.1		–0.0	–0.0	10.6	11.1	8.7
90	263	56.6	22.9	<0.01	–0.3	–0.3	11.5	10.6	9.0
Satisfaction with Decision to Have Treatment									
14	271	73.9	20.7		—	—	10.4	—	—
30	265	73.1	21.9		–0.0	–0.0	10.9	10.3	9.2
90	261	70.3	23.1	0.13	–0.1	–0.2	11.6	10.8	8.9

MID, minimum important difference; ES, effect size; SRM, standardized response mean.
*The ES and SRM were computed to reflect the incremental change between assessment days. Positive ES and SRMs reflect an improvement in scale scores, whereas negative ES and SRMs represent a worsening in scale scores.

DISCUSSION

The FACE-Q is being developed and validated using a rigorous mixed methods approach to provide the research and clinical community with a comprehensive outcomes measurement system. Our goal in designing the FACE-Q was to address the lack of available patient-reported outcome tools applicable to facial aesthetics patients. As with previously published FACE-Q scales,[22,23] the scales examined in the present article were found to be clinically meaningful, valid, and reliable. In addition, the Recovery Early Symptom checklist identified a range of postoperative symptoms following face lifts and lip treatment.

The scales were also responsive to clinical change following treatment. As would be expected, face-lift participants reported more early life impact compared with the lip treatment sample, although both groups showed clinically important change over the course of recovery. The trial data provide evidence that lip treatment has an effect on quality of life, with large (Psychological Wellbeing) and moderate (Social Function) effect sizes noted. Given the nature of minimally invasive lip treatments, which disappear gradually over time, Psychological Wellbeing, Social Function, and Satisfaction with Outcome of Treatment scores declined by day 90, as expected. Interestingly, participants' scores for Satisfaction with Decision to Have Treatment did not decline.

Although there is an ever-growing range of facial cosmetic products and treatments available, very little clinical research is being performed in either academia or industry to understand outcomes from the patient's perspective. This oversight is despite the fact that such information is exactly what potential consumers want and need to know (i.e., what other patients who have undergone facial aesthetics treatment experience). Furthermore, clinicians need such information to educate their patients about the benefits they may or may not derive from any particular treatment. As the cosmetics industry continues to grow and diversify, the application and use of patient-reported outcome instruments in research needs to keep pace. The patient perspective in outcomes research is increasingly being taken seriously.[3] It is, after all, a crucial indicator of the success of aesthetic treatments and therefore should be routinely measured and reported.

It is important to keep in mind limitations in our study design, which we have described previously and will review here in brief. First, plastic surgery and dermatology office staff invited an array of patients to complete the FACE-Q scales. An advantage to this recruitment strategy is that we are able to ensure that the final version of each scale is well-targeted to the general population of facial aesthetic patients rather than a subgroup of patients. The disadvantage is that our data set is not able to answer in a definitive way how patient outcomes relate to specific patient or clinical characteristics. Second, our sample is composed of more women than men and reflects the fact that facial aesthetics is still overwhelmingly pursued by women. Third, our response rate for the mailed survey portion of our study was lower than we would have liked but represents only a small proportion of the sample. Fourth, we acknowledge that there could have been some bias introduced at the individual clinic level by office staff who recruited their patients for us. For example, it is possible that staff may have overlooked "unsatisfied" clients in their recruitment. Future studies that recruit a consecutive sample of patients from a random sample of clinics would enable us to determine how well our scales' content is targeted to capturing the full range of experiences of cosmetic patients.

CONCLUSIONS

This new set of FACE-Q instruments can be smoothly incorporated into research and/or routine clinical practice to better understand the recovery process and quality-of-life impact of facial aesthetics procedures and what patients think about their decision to undergo surgery. As plastic surgeons seek to continually improve aesthetic outcomes alongside the patient experience, we envision that future studies implementing FACE-Q scales will contribute to a greater appreciation of comparative effectiveness between various new and existing techniques.

Andrea L. Pusic, M.D., M.H.S.
Plastic and Reconstructive Surgery
Memorial Sloan Kettering Cancer Center
1275 York Avenue, Room MRI-1007
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/or countless hours spent as expert reviewers: Drs. Nick Carr, Francis Jang, Nancy VanLaeken, Alistair Carruthers, Jean Carruthers, and Richard Warren (Vancouver, British Columbia, Canada); Dr. Stephen Baker (Washington, D.C.); Drs. Jeffery Kenkel and Rod Rohrich (Dallas, Texas); Dr. Foad Nahai (Atlanta Ga.); Dr. Leroy Young (St. Louis, Mo.); and Drs. Anne Chapas, David Hidalgo, David Rosenberg, and Philip Miller (New York, N.Y.).

REFERENCES

1. Cano SJ, Klassen A, Pusic AL. The science behind quality-of-life measurement: A primer for plastic surgeons. *Plast Reconstr Surg.* 2009;123:98e–106e.
2. International Society For Pharmacoeconomics and Outcomes Research. Guidance for industry: Patient-reported outcome measures: Use in medical product development to support labeling claims. Available at: http://www.ispor.org/workpaper/FDA%20PRO%20Guidance.pdf. Accessed March 10, 2014.
3. Black N. Patient reported outcome measures could help transform healthcare. *BMJ* 2013;346:f167.
4. Jeevan R, Cromwell D, Browne J, et al. The NHS Information Centre: Fourth annual national mastectomy and breast reconstruction audit 2011. Available at: http://www.hscic.gov.uk/catalogue/PUB02731/clin-audi-supp-prog-mast-brea-reco-2011-rep1.pdf. Accessed April 1, 2014.
5. Morley D, Jenkinson C, Fitzprick R. A structured review of patient-reported outcome measures used in cosmetic surgical procedures: Report to Department of Health, 2013. Available at: http://phi.uhce.ox.ac.uk/pdf/Cosmetic%20Surgery%20PROMs%20Review2013.pdf. Accessed April 1, 2014.
6. Sigurdson L, Kirkland SA, Mykhalovskiy E. Validation of a questionnaire for measuring morbidity in breast hypertrophy. *Plast Reconstr Surg.* 2007;120:1108–1114.
7. Kerrigan CL, Collins ED, Striplin D, et al. The health burden of breast hypertrophy. *Plast Reconstr Surg.* 2001;108:1591–1599.
8. Pusic AL, Klassen AF, Scott AM, Klok JA, Cordeiro PG, Cano SJ. Development of a new patient-reported outcome measure for breast surgery: The BREAST-Q. *Plast Reconstr Surg.* 2009;124:345–353.
9. Cano SJ, Klassen AF, Scott AM, Cordeiro PG, Pusic AL. The BREAST-Q: Further validation in independent clinical samples. *Plast Reconstr Surg.* 2012;129:293–302.
10. Harris DL, Carr AT. 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;54:216–222.
11. Carr T, Moss T, Harris D. The DAS24: A short form of the Derriford Appearance Scale DAS59 to measure individual responses to living with problems of appearance. *Br J Health Psychol.* 2005;10:285–298.
12. Cox SE, Finn JC, Stetler L, Mackowiak J, Kowalski JW. Development of the Facial Lines Treatment Satisfaction Questionnaire and initial results for botulinum toxin type A-treated patients. *Dermatol Surg.* 2003;29:444–449; discussion 449.
13. Pusic A, Klassen AF, Scott AM, Cano SJ. Development and psychometric evaluation of the FACE-Q Satisfaction with Appearance Scale: A new PRO instrument for facial aesthetics patients. *Clin Plast Surg.* 2013;40:249–260.

14. Brown BC, McKenna SP, Solomon M, Wilburn J, McGrouther DA, Bayat A. The patient-reported impact of scars measure: Development and validation. *Plast Reconstr Surg.* 2010;125:1439–1449.
15. Durani P, McGrouther DA, Ferguson MW. The Patient Scar Assessment Questionnaire: A reliable and valid patient-reported outcomes measure for linear scars. *Plast Reconstr Surg.* 2009;123:1481–1489.
16. Chren MM, Lasek RJ, Quinn LM, Mostow EN, Zyzanski SJ. Skindex, a quality of life measure for patients with skin disease: Reliability, validity, and responsiveness. *J Invest Dermatol.* 1996;107:707–713.
17. Klassen AF, Cano SJ, Scott AM, Pusic AL. 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. *Plast Reconstr Surg.* 2014;133:21–30.
18. 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.
19. Lasch KE, Marquis P, Vigneux M, et al. PRO development: Rigorous qualitative research as the crucial foundation. *Qual Life Res.* 2010;19:1087–1096.
20. Patrick DL, Burke LB, Gwaltney CJ, et al. Content validity: Establishing and reporting the evidence in newly developed patient-reported outcomes (PRO) instruments for medical product evaluation. ISPOR PRO good research practices task force report: Part 1—Eliciting concepts for a new PRO instrument. *Value Health* 2011;14:967–977.
21. Patrick DL, Burke LB, Gwaltney CJ, et al. Content validity: Establishing and reporting the evidence in newly developed patient-reported outcomes (PRO) instruments for medical product evaluation. ISPOR PRO Good Research Practices Task Force report: Part 2—Assessing respondent understanding. *Value Health* 2011;14:978–988.
22. Panchapakesan V, Klassen AF, Cano SJ, Scott AM, Pusic 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.
23. 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.
24. 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.
25. Andrich D. Controversy and the Rasch model: A characteristic of incompatible paradigms? *Med Care* 2004;42 (Suppl):I7–I16.
26. Wright B, Masters G. *Rating Scale Analysis.* Chicago: Mesa; 1982.
27. Andrich D, Sheridan B. *RUMM2030.* Perth, Australia: RUMM Laboratory; 1997–2011.
28. 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.
29. Andrich D. *Rasch Models for Measurement.* Newbury Park, Calif: Sage; 1988.
30. Rasch G. *Probabilistic Models for Some Intelligence and Attainment Tests.* Copenhagen: Danish Institute for Education Research; 1960.
31. Cronbach LJ. Coefficient alpha and the internal structure of tests. *Psychometrika* 1951;16:297–334.
32. 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.
33. Likert R. A technique for the measurement of attitudes. *Arch Psychol.* 1932;140:1–55.
34. 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.
35. Eisen M, Ware JE Jr, Donald CA, Brook RH. Measuring components of children's health status. *Med Care* 1979;17: 902–921.
36. 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.
37. Bohrnstedt GW. Measurement. In: Rossi PH, Wright JD, Anderson AB, eds. *Handbook of Survey Research.* New York: Academic Press; 1983:69–121.
38. Lamping DL, Schroter S, Kurz X, Kahn SR, Abenhaim L. Evaluation of outcomes in chronic venous disorders of the leg: Development of a scientifically rigorous, patient-reported measure of symptoms and quality of life. *J Vasc Surg.* 2003;37:410–419.
39. Lamping DL, Schroter S, Marquis P, Marrel A, Duprat-Lomon I, Sagnier PP. The community-acquired pneumonia symptom questionnaire: A new, patient-based outcome measure to evaluate symptoms in patients with community-acquired pneumonia. *Chest* 2002;122:920–929.
40. Kazis LE, Anderson JJ, Meenan RF. Effect sizes for interpreting changes in health status. *Med Care* 1989;27 (Suppl):S178–S189.
41. Liang MH, Fossel AH, Larson MG. Comparisons of five health status instruments for orthopedic evaluation. *Med Care* 1990;28:632–642.
42. Revicki DA, Cella D, Hays RD, Sloan JA, Lenderking WR, Aaronson NK. Responsiveness and minimal important differences for patient reported outcomes. *Health Qual Life Outcomes* 2006;4:70.