The face is the primary means by which individuals present themselves, interact with others, and communicate. As people age, significant body changes begin to occur, and concerns about appearance focus increasingly on the face, sometimes resulting in negative psychosocial sequelae. Anxiety about aging and concerns regarding body image are significant predictors of social motivations to pursue cosmetic surgery. In 2011, almost 7 million nonsurgical facial cosmetic procedures were performed in the United States, and aesthetic facial surgery accounted for approximately 30% of the 1.5 million-plus plastic surgery procedures performed in the same year.

For all cosmetic procedures, successful outcomes are intrinsically linked to patients’ expectations, goals, desires, and, ultimately, patient satisfaction. With regard to aesthetic facial procedures, patient satisfaction with appearance, improved body image functioning, and overall quality of life (QOL) are perhaps the most important.
outcomes.\textsuperscript{11,12,14-16} The measurement of such constructs is best captured by patient-reported outcome (PRO) instruments, which are considered vital to clinical research and are used to support decision making at all levels of health care.\textsuperscript{17}

Despite the importance of such measurement tools, a systematic review performed by our team identified a lack of reliable and valid PRO instruments suitable for use among facial aesthetic patients.\textsuperscript{18} To address this deficit, our team set out to develop the FACE-Q, a PRO instrument that includes a range of separate scales to measure outcomes important to patients who undergo facial aesthetic procedures. The methodology to develop this instrument has been used previously to create other plastic surgery–specific PRO instruments.\textsuperscript{19-21}

In the present article, we describe the development and psychometric evaluation of the FACE-Q Aging Appraisal Scale, a 7-item scale designed to provide overall global assessment of a patient’s perception of his or her appearance in the context of facial aging. We also discuss the FACE-Q Patient-Perceived Age Visual Analog Scale (VAS), which consists of a single item that asks patients to report their perceived age in comparison to their actual age.

**METHODS**

Approval for the study was obtained from the local Institutional Review Board at each of 5 study locations. The content of the FACE-Q Aging Appraisal Scale was created as part of a larger study to develop a suite of scales that cover a range of concepts important to facial aesthetic patients.\textsuperscript{22} The scales were constructed with strict adherence to recommended guidelines for PRO instrument development.\textsuperscript{23-27} These guidelines describe the 3 phases required to develop a scientifically credible and clinically meaningful instrument.

In brief, the first phase involved formally defining a conceptual framework and generating a pool of items. The items were developed from literature review, qualitative patient interviews, and expert opinion. A series of scales was developed from the item pool, which then underwent pilot testing on the target sample to clarify ambiguities in item wording, confirm appropriateness, and determine acceptability and completion time. This phase has been described in detail in a separate publication.\textsuperscript{28} The second phase involved field testing the scales in a large sample of subjects. Questions (also referred to as items) were retained or rejected based on their performance against a standardized set of psychometric criteria. Phase 3 (not yet completed) will entail further psychometric evaluation of the item-reduced scales.\textsuperscript{29,30}

**Phase 1: Qualitative Phase**

Qualitative methodology was used to assess patient perception of facial aesthetic concerns. Between January 2008 and February 2009, 50 patients were recruited from among 7 offices of plastic surgeons and dermatologists in the United States (New York) and Canada (Vancouver). The participants were 20 to 79 years of age (mean, 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), necklift (n = 8), browlift (n = 4), and chin implant (n = 2).

Open-ended questions were used for the interviews, which were digitally recorded, transcribed verbatim, and coded within QSR International, Burlington, Massachusetts\textsuperscript{31} using a line-by-line coding approach. Concurrent data collection and analysis and the constant comparative method were utilized to refine codes and categories as they emerged. Data analyses led to the development of a framework of concepts important to facial aesthetic patients.

Scales with items covering the concepts in the framework were then developed. Codes were examined in relation to specific patient characteristics. Attaching key patient characteristics to each code provided the information needed to develop both core items and unique items. A set of scales was created through examination of the item lists generated from the coded material. This enabled identification of a set of items that together described a continuum for each major concept. Flesch-Kincaid grade-level scores were assessed for each item\textsuperscript{32} and adjusted as necessary to ensure that the lowest possible reading level was used. Scale instructions and appropriate response options were developed for each scale.

Expert appraisal was performed by presenting the scales to 26 experts (15 plastic surgeons, 4 dermatologists, 3 psychologists, and 4 office staff). Following this, 35 facial aesthetic patients participated in 1-on-1 cognitive debriefing interviews to clarify ambiguous wording and confirm the appropriateness, acceptability, and completion time of the preliminary scales. A set of independently functioning scales that measure the concepts forming the framework was the end result of this process.

**Phase 2: Quantitative Phase (Field Test)**

Items were identified that represented the best indicators for each scale based on their performance against a standardized set of psychometric criteria. Data from patients recruited through 2 separate studies were compiled for psychometric analyses. The results presented here pertain only to the FACE-Q Aging Appraisal Scale. This scale was developed for use in research and clinical practice to compare outcomes related to facial rejuvenation across procedure types and/or to measure changes after any facial aesthetic procedure. In general, the goal of facial aesthetic procedures is to diminish the signs and appearance of aging.

**Study 1**

Between June 2010 and June 2012, data were collected for patients from 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). Eligible patients were those who could read English, were at least 18 years of age, and were planning to undergo or had already undergone any surgical or nonsurgical facial aesthetic procedure.

Numerous FACE-Q scales were developed in the initial phases of research. These scales were grouped into booklets based on common surgical and nonsurgical procedures and were distributed to the participating practices. Primarily patients undergoing facelift or blepharoplasty completed the FACE-Q Aging Appraisal Scale. In the instructions for this scale, the patient was asked to respond to each item based on “how you feel about the age your face looks” and to indicate 1 of the 4 response options: definitely agree, somewhat agree, somewhat disagree, or definitely disagree. A higher score on the scale equated to the person feeling that he or she looked younger in appearance, whereas a lower score denoted thinking that he or she looked older. Therefore, higher scores on this scale represented a better outcome.

Patients from 6 surgical practices were recruited at the time of their initial consultation or follow-up appointment and asked to complete a questionnaire booklet in the waiting room before the appointment. Patients from the other 4 practices were invited to participate in a postal survey. To ensure a high response rate, a modified Dillman protocol was utilized, involving personalized letters, multiple reminders, and a $5 gift card to thank patients for their participation.

Study 2
Between March 2010 and July 2011, the FACE-Q Aging Appraisal Scale and the Patient-Perceived Age VAS were administered to 100 patients who were planning to undergo or had already undergone facelift surgery. The scale was administered alongside other FACE-Q scales relevant to measuring patient concerns. This study included patients from France, Germany, the United Kingdom, and Israel. Translations and linguistic validation of the FACE-Q scales were provided by Mapi Research Trust (Lyon, France). The Mapi process is based on translation principles as detailed by the European Regulatory Issues and Quality of Life Assessment (ERIQA) group and the International Society of Pharmacoeconomics and Outcomes Research and recommended by the US Food and Drug Administration. This process ensured that the concepts measured by the FACE-Q scales were equivalent for all languages (ie, English, German, French, and Hebrew) and easily understood by the people in each participating country.

Rasch Measurement Theory Methods
Rasch measurement theory (RMT) methods were used to evaluate the scaling properties and construct validity of the FACE-Q Aging Appraisal Scale. RMT analysis is the formal testing of an outcome scale against a mathematical measurement model that was developed by the Danish mathematician Georg Rasch. During RMT analysis, responses to a set of items from a scale (observed scores) can be tested against response patterns predicted by the model. This approach differs from traditional psychometric methods. The pattern expected by the Rasch model has been described as a probabilistic version of the Guttman scale, which itself is a deterministic pattern that follows a strict hierarchical ordering of items. The formal tenets of measurement (order, unidimensionality, and additivity) are operationalized, enabling interval-level data to be obtained from questionnaires. The Rasch model is derived from theory and therefore is independent of the sample. This separation is of significant value in the development of measurement instruments, because discrepancies between what the model predicts and what the sample demonstrates can be analyzed, understood, and improved to optimize the instrument.

The properties that make RMT methods different from traditional psychometric methods enable RMT to supplement standard assessments of construct validity and reliability by also evaluating the fundamental scaling properties of an instrument. Assessment of validity involves examination of thresholds for item response options, item fit statistics, item locations, and differential item functioning. Assessment of reliability involves examination of the Person Separation Index (PSI). The overall picture given by the multiple tests in the Rasch model provides evidence for a judgment to be made about the overall quality of the scale. (A detailed explanation and a more comprehensive review of RMT methods may be found elsewhere.) RMT analyses were performed on the FACE-Q Aging Appraisal Scale using RUMM2030 software (RUMM Laboratory, Duncraig, Western Australia).

Validity
Item Response Category Thresholds
Each item of the FACE-Q Aging Appraisal Scale has multiple response categories (ie, definitely disagree, somewhat disagree, somewhat agree, and definitely agree), which reflect an ordered continuum that increases for the construct of interest. A threshold is the location at which the probability of responding in adjacent pairs of response options is 50%. When the categories are working as intended, the thresholds are ordered. “Disordered” thresholds imply that the response categories for that item are not working as intended. Disordered thresholds occur when respondents have difficulty consistently discriminating between response options. When response options work as expected, evidence of the validity of the scale is obtained.

Item Fit Statistics
The items of the FACE-Q Aging Appraisal Scale must work together (fit) to define a single continuum both clinically and statistically. Item fit is assessed by statistical and graphic indicators. Statistical indicators include fit residuals (item-person interaction) and χ² values (item-trait
interaction), whereas graphic indicators are the item characteristic curves (ICC). We expect the fit residuals to fall between −2.5 and +2.5, with associated nonsignificant χ² values (significance interpreted after Bonferroni adjustment). Fit statistics are most meaningful when interpreted together and in the context of their clinical usefulness as an item set.

**Targeting**
The items of any given questionnaire should be well targeted to the patient population being studied. Scale-to-sample targeting pertains to the match between the range of the construct measured by the scale and the range of construct in the study sample. Targeting analysis informs us of the sample’s degree of suitability for evaluating the FACE-Q Aging Appraisal Scale and the suitability of the scale for measuring the sample. Better targeting translates to better information as well as the ability to interpret psychometric data with confidence.

**Stability**
Stability analysis enables an explicit test of scale performance in the form of examining differential item functioning (DIF). DIF refers to the degree to which item performance remains stable across subgroups. DIF occurs when different subgroups of a sample (eg, males and females) respond differently to a certain question when the level of the underlying trait (eg, appearing old) is the same. As a guide, statistically significant χ² values indicate potential DIF and therefore problems in scale performance (significance interpreted after Bonferroni adjustment).

When DIF occurs, attempts are made to remedy it if possible by calibrating an item for each group. If not possible, the item is removed from the scale.

**Reliability**

**Person Separation Index**
The Person Separation Index (PSI) is analogous to Cronbach’s α and is an indicator of the degree of reliability of the fit characteristics. A value of 0.70 for group use, as well as 0.80 for individual patient use, is considered the minimum accepted level of PSI for adequate reliability.

**Traditional Psychometric Methods**

Traditional psychometric methods primarily use correlation or descriptive analyses to evaluate scaling assumptions (legitimacy of summing items) and scale reliability and validity. These are described in detail elsewhere. The following were examined in our study: data quality (percentage of missing data for each item), scaling assumptions (similarity of item means and variances), magnitude and similarity of corrected item-total correlations, scale-to-sample targeting (score means, SD, and floor/ceiling effects), and internal consistency reliability (Cronbach’s α).

Additional known-groups validity testing was performed by examining correlations between scores on the FACE-Q Aging Appraisal Scale and the single-item Patient-Perceived Age VAS, in which patients reported their perceived age compared with their actual age. Patients were asked to circle the category that reflected the number of years older or younger they look, ranging from 15 years younger to 15 years older. In particular, we examined the ability of the FACE-Q Aging Appraisal Scale to detect clinical differences between predefined subgroups. We compared scale scores with a global item on the self-reported age appraisal, ranging from 6 to 10 years older to 6 to 10 years younger than current age. We hypothesized that FACE-Q Aging Appraisal Scale scores would be statistically significant incrementally in “younger” subgroups compared with “older” subgroups. We also compared FACE-Q Aging Appraisal Scale scores between patients who completed the questionnaire before surgery and those who completed it afterward. We hypothesized that those who completed it after surgery would have higher scores, indicating that they believed they appeared younger. For comparative statistics, raw FACE-Q Aging Appraisal Scale scores were first transformed through RMT methods to interval-level data on a scale of 0 to 100.

**RESULTS**

**Phase 1: Qualitative Phase**

As described above and in our previous publication, the qualitative work resulted in the development of a conceptual framework and series of independently functioning scales that measure the important concerns of facial aesthetic patients. The FACE-Q Aging Appraisal Scale consists of 7 items whereby patients are asked to indicate how they feel about the age of their face using various descriptors (eg, worried, reminded, recognize) and scenarios (eg, in photos, in the mirror). The item set is easy to understand and complete with a Flesh-Kincaid grade level of 1.7, and each item in the scale has a grade level score below 3.6.

**Phase 2: Field Test**

The overall response rate was 78% (92% for face-to-face interviews, 59% for postal questionnaires). A total of 288 patients completed the questionnaire (122 face-to-face interviews, 166 postal questionnaires). The age range of respondents was 28 to 89 years. There were 251 women and 30 men; sex data were missing for the other 7 patients (Table 1).

**RMT Analyses**

Overall, the results of RMT analysis supported the FACE-Q Aging Appraisal scale as a reliable and valid measure of how people feel about the age that their face looks.
Validity was supported by several findings. First, all 7 items had ordered thresholds from the outset, which supported the appropriateness of the number and type of response options we created (Figure 1). Second, distributions of item thresholds and person estimates were well matched (targeting), indicating suitability of the sample for evaluating the scale as well as suitability of the scale for measuring the sample (Figure 2). Third, fit to the Rasch model was good. Only 1 of the 7 items (Q3) had a fit residual of 2.69, which was just slightly above the suggested range of −2.5 to +2.5. No item had a significant $\chi^2$ value after Bonferroni adjustment (Table 2). Scale reliability was supported by high PSI (0.86), demonstrating excellent reliability for individual patient use. Data set analysis showed no statistical DIF after Bonferroni adjustment ($P < .002$) with respect to sex, age (28-44, 45-54, 55-64, or 65+ years), timing of questionnaire (presurgery or postsurgery; Table 3), type of surgery (facelift, blepharoplasty, or other), previous procedures, or geographic location (United States, Germany, UK, France, Israel, or Canada).

### Traditional Psychometric Methods Analysis

The results of traditional analysis also supported the FACE-Q Aging Appraisal Scale as a reliable and valid measure (Table 4). The criteria were satisfied for all psychometric properties evaluated. Data quality was high (missing data 1.4%; scale scores were computable for 98% of respondents), and scaling assumptions were satisfied (similar mean item scores, corrected item-total correlations range of 0.57-0.87). Scale-to-sample targeting was good (scale scores spanned the scale range and were not notably skewed; scale midpoint and ceiling effects were negligible), and internal consistency reliability was high (Cronbach’s $\alpha = 0.92$).

Our validity-focused examinations of clinical known-groups validity showed that our hypotheses relating to the patterns and significance of scores across the subgroups were supported (Table 5). Scale validity was further supported by correlations between scores on the FACE-Q Aging Appraisal Scale and the FACE-Q Patient-Perceived Age VAS, as well as between scores on the FACE-Q Aging Appraisal Scale and patient age in years (Pearson’s $r = 0.58$ and $-0.28$, respectively). Patients who completed the FACE-Q Aging Appraisal Scale before surgery scored lower than those who completed it after surgery, indicating an ability to discriminate between the 2 groups (FACE-Q Aging Appraisal score: 51 [n = 111] vs 78 [n = 173], respectively; $P < .01$ on a $t$ test of independent samples).

### DISCUSSION

Body image, which plays an important role in both self-esteem and QOL, can be positively affected through

### Table 1. Patient Characteristics From Field Test

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of patients</td>
<td>288</td>
</tr>
<tr>
<td>Age, y</td>
<td></td>
</tr>
<tr>
<td>Mean (SD)</td>
<td>57.9 (14.8)</td>
</tr>
<tr>
<td>Range</td>
<td>28-89</td>
</tr>
<tr>
<td>Sex</td>
<td></td>
</tr>
<tr>
<td>Women</td>
<td>251 (87.2)</td>
</tr>
<tr>
<td>Men</td>
<td>30 (10.4)</td>
</tr>
<tr>
<td>Missing data</td>
<td>7 (2.4)</td>
</tr>
<tr>
<td>Ethnicity</td>
<td></td>
</tr>
<tr>
<td>White non-Hispanic</td>
<td>249 (86.4)</td>
</tr>
<tr>
<td>Asian</td>
<td>6 (2.1)</td>
</tr>
<tr>
<td>South Asian</td>
<td>1 (0.3)</td>
</tr>
<tr>
<td>Native American</td>
<td>10 (3.5)</td>
</tr>
<tr>
<td>White Hispanic</td>
<td>4 (1.4)</td>
</tr>
<tr>
<td>Black non-Hispanic</td>
<td>5 (1.7)</td>
</tr>
<tr>
<td>Other</td>
<td>3 (1.0)</td>
</tr>
<tr>
<td>Geographic location</td>
<td></td>
</tr>
<tr>
<td>United States</td>
<td>104 (36.1)</td>
</tr>
<tr>
<td>Canada</td>
<td>89 (30.1)</td>
</tr>
<tr>
<td>France</td>
<td>15 (5.2)</td>
</tr>
<tr>
<td>Germany</td>
<td>45 (15.6)</td>
</tr>
<tr>
<td>Israel</td>
<td>20 (6.9)</td>
</tr>
<tr>
<td>United Kingdom</td>
<td>15 (5.2)</td>
</tr>
<tr>
<td>Timing of booklet</td>
<td></td>
</tr>
<tr>
<td>Preoperative</td>
<td>114 (39.6)</td>
</tr>
<tr>
<td>Postoperative</td>
<td>174 (60.4)</td>
</tr>
<tr>
<td>Type of surgery</td>
<td></td>
</tr>
<tr>
<td>Facelift</td>
<td>217 (75.3)</td>
</tr>
<tr>
<td>Blepharoplasty</td>
<td>67 (23.3)</td>
</tr>
<tr>
<td>Other</td>
<td>4 (1.4)</td>
</tr>
<tr>
<td>Previous surgery</td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>177 (61.5)</td>
</tr>
<tr>
<td>No</td>
<td>111 (38.5)</td>
</tr>
</tbody>
</table>

Values are presented as number (%) unless otherwise indicated.
Dissatisfaction with body image has been shown to be a strong motivator for the pursuit of aesthetic treatments. The overall goal of most facial aesthetic procedures is to enhance appearance and reduce the signs of aging, in turn improving body image, self-esteem, and QOL. Results of aesthetic procedures are intrinsically tied to assessment of outcomes by the patient. With facial aesthetic procedures, it is paramount to assess patient-reported changes with regard to the appearance of aging. This requires a condition-specific PRO instrument, guided by rigorous methodology. Tips and pearls for clinicians to use when assessing, selecting, and utilizing PRO instruments are provided in Table 6.

The FACE-Q Aging Appraisal Scale was developed from extensive qualitative research that resulted in a detailed conceptual framework, which formed the cornerstone that guided development of a scale capable of measuring the unique aging concerns of patients who undergo facial aesthetic procedures. The RMT analyses confirmed that this scale is scientifically sound and clinically meaningful in measuring the construct of aging appearance in these patients. Traditional psychometric analyses provided additional support for the robustness of this measure. These findings were based on a large heterogeneous patient sample, and further results showed that the performance of the scale was consistent regardless of patient age, sex, geographic location, procedure type, whether previous procedures had been performed, and whether the scale was completed before or after surgery. Raw ordinal scores can be converted to interval-level scores, which makes the scale suitable for individual patient evaluation.

![Figure 1](https://academic.oup.com/asj/article-figures/33/8/1099/224106)

**Figure 1.** Threshold map for all items in the FACE-Q Aging Appraisal scale. The x-axis symbolizes the construct (satisfaction with aging appearance), with satisfaction increasing from left to right. The y-axis and rectangular bars represent the response category for each item: blue (1) = definitely agree, red (2) = somewhat agree, green (3) = somewhat disagree, and purple (4) = definitely disagree.

![Figure 2](https://academic.oup.com/asj/article-figures/33/8/1099/224106)

**Figure 2.** Person-item threshold distribution. The x-axis symbolizes the construct (feeling about age appearance of face), with satisfaction increasing from left to right. The y-axis shows the frequency of person measure locations (upper histogram) and item locations (lower histogram). In the upper histogram, the group has been divided into pretreatment (blue) and posttreatment (red) samples. This difference (in logits) between pretreatment and posttreatment scores was statistically significant ($P < .0001$).
### Table 2. Statistical Indicators of Fit (Fit Residual, $\chi^2$)

<table>
<thead>
<tr>
<th>Item No.</th>
<th>Item</th>
<th>Item Location</th>
<th>SE</th>
<th>Fit Residual</th>
<th>$\chi^2$ Value</th>
<th>P Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q1</td>
<td>Recognize</td>
<td>−1.98</td>
<td>0.13</td>
<td>0.08</td>
<td>12.05</td>
<td>.017</td>
</tr>
<tr>
<td>Q2</td>
<td>Don’t look like self</td>
<td>−0.88</td>
<td>0.11</td>
<td>2.1</td>
<td>5.84</td>
<td>.211</td>
</tr>
<tr>
<td>Q3</td>
<td>Bothered*</td>
<td>0.24</td>
<td>0.10</td>
<td>−2.69*</td>
<td>13.53</td>
<td>.009</td>
</tr>
<tr>
<td>Q4</td>
<td>Want to look</td>
<td>0.49</td>
<td>0.10</td>
<td>2.37</td>
<td>3.21</td>
<td>.523</td>
</tr>
<tr>
<td>Q5</td>
<td>Starting to look</td>
<td>0.59</td>
<td>0.10</td>
<td>0.57</td>
<td>3.69</td>
<td>.450</td>
</tr>
<tr>
<td>Q6</td>
<td>Reminded</td>
<td>0.76</td>
<td>0.11</td>
<td>−0.91</td>
<td>5.23</td>
<td>.264</td>
</tr>
<tr>
<td>Q7</td>
<td>Photos</td>
<td>0.76</td>
<td>0.10</td>
<td>−0.14</td>
<td>5.43</td>
<td>.246</td>
</tr>
</tbody>
</table>

*Denotes items with fit residual ±2.5 or significant P value.

### Table 3. Uniform Differential Item Functioning: Sex, Age, and Timing of Surgery

<table>
<thead>
<tr>
<th>Item No.</th>
<th>Item</th>
<th>Sex</th>
<th>MS</th>
<th>F</th>
<th>P Value</th>
<th>Age</th>
<th>MS</th>
<th>F</th>
<th>P Value</th>
<th>Presurgery vs Postsurgery</th>
<th>MS</th>
<th>F</th>
<th>P Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q1</td>
<td>Recognize</td>
<td></td>
<td>0.90</td>
<td>1.10</td>
<td>.29</td>
<td>1.59</td>
<td>1.90</td>
<td>.13</td>
<td></td>
<td></td>
<td></td>
<td>0.08</td>
<td>0.10</td>
</tr>
<tr>
<td>Q2</td>
<td>Don’t look like self</td>
<td></td>
<td>0.02</td>
<td>0.02</td>
<td>.01</td>
<td>2.13</td>
<td>1.84</td>
<td>.14</td>
<td></td>
<td></td>
<td></td>
<td>1.90</td>
<td>1.59</td>
</tr>
<tr>
<td>Q3</td>
<td>Bothered</td>
<td></td>
<td>0.02</td>
<td>0.03</td>
<td>.87</td>
<td>0.75</td>
<td>1.36</td>
<td>.26</td>
<td></td>
<td></td>
<td></td>
<td>1.00</td>
<td>1.75</td>
</tr>
<tr>
<td>Q4</td>
<td>Want to look</td>
<td></td>
<td>0.06</td>
<td>0.05</td>
<td>.81</td>
<td>1.39</td>
<td>1.19</td>
<td>.31</td>
<td></td>
<td></td>
<td></td>
<td>7.60</td>
<td>6.70</td>
</tr>
<tr>
<td>Q5</td>
<td>Starting to look</td>
<td></td>
<td>6.24</td>
<td>7.32</td>
<td>.01</td>
<td>1.98</td>
<td>2.22</td>
<td>.09</td>
<td></td>
<td></td>
<td></td>
<td>1.74</td>
<td>1.93</td>
</tr>
<tr>
<td>Q6</td>
<td>Reminded</td>
<td></td>
<td>0.12</td>
<td>0.17</td>
<td>.69</td>
<td>0.06</td>
<td>0.08</td>
<td>.97</td>
<td></td>
<td></td>
<td></td>
<td>0.23</td>
<td>0.31</td>
</tr>
<tr>
<td>Q7</td>
<td>Photos</td>
<td></td>
<td>0.31</td>
<td>0.38</td>
<td>.54</td>
<td>3.68</td>
<td>4.69</td>
<td>.003</td>
<td></td>
<td></td>
<td></td>
<td>2.04</td>
<td>2.50</td>
</tr>
</tbody>
</table>

Abbreviations: F, item fit residual; MS, mean square.

### Table 4. Traditional Psychometric Analyses: Data Quality, Scaling Assumptions, and Targeting

<table>
<thead>
<tr>
<th>Item No.</th>
<th>Item</th>
<th>Data Quality: % of Missing Data</th>
<th>Possible Range (Midpoint)</th>
<th>Actual Score Range</th>
<th>Mean Score</th>
<th>SD</th>
<th>CITC</th>
<th>Floor/Ceiling Effects, %a</th>
<th>Skewness</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q1</td>
<td>Recognize</td>
<td>1.4</td>
<td>1-4</td>
<td>1-4</td>
<td>3.68</td>
<td>0.66</td>
<td>0.57</td>
<td>1/77</td>
<td>−2.1</td>
</tr>
<tr>
<td>Q2</td>
<td>Don’t look like self</td>
<td>1.4</td>
<td>1-4</td>
<td>1-4</td>
<td>3.33</td>
<td>0.93</td>
<td>0.69</td>
<td>5/60</td>
<td>−1.12</td>
</tr>
<tr>
<td>Q3</td>
<td>Bothered</td>
<td>1.4</td>
<td>1-4</td>
<td>1-4</td>
<td>2.90</td>
<td>1.06</td>
<td>0.87</td>
<td>11/41</td>
<td>−0.38</td>
</tr>
<tr>
<td>Q4</td>
<td>Want to look</td>
<td>1.4</td>
<td>1-4</td>
<td>1-4</td>
<td>2.84</td>
<td>1.14</td>
<td>0.75</td>
<td>16/42</td>
<td>−0.35</td>
</tr>
<tr>
<td>Q5</td>
<td>Starting to look</td>
<td>1.4</td>
<td>1-4</td>
<td>1-4</td>
<td>2.80</td>
<td>1.10</td>
<td>0.80</td>
<td>15/38</td>
<td>−0.26</td>
</tr>
<tr>
<td>Q6</td>
<td>Reminded</td>
<td>1.4</td>
<td>1-4</td>
<td>1-4</td>
<td>2.75</td>
<td>1.08</td>
<td>0.83</td>
<td>14/34</td>
<td>−0.18</td>
</tr>
<tr>
<td>Q7</td>
<td>Photos</td>
<td>1.4</td>
<td>1-4</td>
<td>1-4</td>
<td>2.76</td>
<td>1.11</td>
<td>0.82</td>
<td>16/36</td>
<td>−0.24</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>1.4</td>
<td>7-28</td>
<td>7-28</td>
<td>21.05</td>
<td>5.94</td>
<td>—</td>
<td>0.4/24</td>
<td>−0.41</td>
</tr>
</tbody>
</table>

Abbreviations: CITC, corrected item-total correlation; SD, standard deviation.

*aCalculated as the percentage of people scoring either floor or ceiling level.
Aging Appraisal Scale would be statistically significantly higher for patients who thought they look younger was supported. Thus, the VAS allowed us to quantify, in the study sample, the differences between patients’ perceptions of appearing “older than,” “exactly,” or “younger than” their age, which translated to a difference of approximately 10 points between “older” and “their age” and approximately 20 points between “their age” and “younger.” The score differences between the defined subgroups can be used to help build hypotheses for expected differences and change scores for future clinical research. This analysis provides an early view about the meaning of scale scores, which, with further evidence and support, can be used to start identifying clinically meaningful differences between (and within) patients. This analysis also aids in interpreting and contextualizing the relative impact of treatments in these patients.

This research further supports the use of RMT in developing new scales, which moves beyond the limitations of traditional psychometric analyses. First, traditional psychometric analyses rely directly on the items and samples used to estimate them. Second, the psychometric properties of the items and scales vary depending on the sample. As a result, reliability and validity estimates of a specific measure may differ for different patient groups. Third, confidence intervals for individual patient scores tend to be large; therefore, PRO instruments developed with such methods provide total scores that can be used only for group comparisons—not for individual comparisons. Finally, most PRO instruments used in medical research that were developed from traditional psychometric methods utilize an ordinal scale to rate the responses. Ordinal scales have limited ability to provide a true quantitative assessment that represents patient status along a continuum. The Rasch model is derived from theory, and therefore reliability and validity estimates computed using RMT are much less sample dependent than those derived from traditional methods. The transformation of ordinal scores into interval-level measurements, enabling estimates suitable for individual person measurements, is facilitated, which provides more clinically meaningful comparisons between patients and within patients over time (such as comparisons of different procedures or assessing an individual patient before and after a procedure).

A limitation of our study is that the sample included patients who may have had multiple procedures previously. This is reflective of a pragmatic research approach and hence increases the validity of our findings, but the impact of specific procedures cannot be ascertained from this study. Moreover, the study population was primarily female. Again, this is pragmatic because more than 90% of cosmetic surgery patients in the United States are women, and many studies show that women express greater interest in, and greater willingness to undergo, cosmetic surgery than do men. It is noteworthy that we found no major differences in DIF based on sex; however, future studies should include more male patients. Further research should include evaluation of scale performance among patients who undergo facial aesthetic procedures other than facelifts or blepharoplasties and more detailed analyses of responsiveness (eg, the ability to quantify

### Table 5. Clinical Validity: Analysis of Variance Results for Age Appraisal Testing of Known Groups

<table>
<thead>
<tr>
<th>Category</th>
<th>How many years younger or older do you think you look compared with your actual age?</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>6-10 y Older</td>
</tr>
<tr>
<td>FACE-6 Aging Appraisal Scale, mean score (SD)</td>
<td>41 (33)</td>
</tr>
</tbody>
</table>

### Table 6. Tips and Pearls for Assessing, Selecting, and Utilizing Patient-Reported Outcome (PRO) Instruments

- **PRO instruments can be either generic or condition specific.** The former is useful in comparing outcomes across different conditions, whereas the latter tends to be more clinically relevant (ie, covers the issues important to the patient population) and appropriate for assessing treatment outcomes. Choose a measure that suits your purpose.

- **PRO instruments should undergo rigorous development and demonstrate reliability, validity, and responsiveness.** Features of well-designed PRO instruments include:
  - Patient input through qualitative methods
  - A conceptual framework on which the PRO instrument is based
  - Psychometric testing to demonstrate reliability, validity, and responsiveness

- In general, a PRO instrument should be used in patients similar to those of the study group for which it was designed.

- For measuring treatment outcomes, a PRO instrument can be utilized before, during, and after treatment.

- Choose PRO instruments that facilitate the ability to measure linear, continuous data (interval data [eg, on a scale of 0-100]) as opposed to raw scores alone (ordinal data), because only the former can measure and quantify changes between treatments in a meaningful manner.

- Use PRO instruments as recommended by the developers, without modification of instructions, response options, or items; such alterations would necessitate repeating the psychometric testing and would violate the copyright provisions.

- PRO instruments should be practically relevant in terms of the number of questions, length of time to complete, and ease of scoring.
significant change). Our initial group-level pre- and post-treatment analyses demonstrated significant change, which was further supported by minimal ceiling effects in the pretreatment sample and clear differences in comparisons with the VAS scale. These findings provide useful first steps toward understanding the meaning of score differences, but further research is warranted to elaborate on clinical meaning and minimally important clinical differences.

**CONCLUSIONS**

The FACE-Q Aging Appraisal Scale is a psychometrically sound, condition-specific PRO instrument with excellent reliability and validity. It can provide clinically meaningful scores, enabling accurate assessment of clinical comparisons of age appearance in patients who undergo facial rejuvenation. The Patient-Perceived Age VAS can be used to complement the FACE-Q Aging Appraisal Scale and further improve clinical interpretability. As such, scientifically sound and clinically interpretable outcomes data can be obtained through these tools, in turn leading to practice improvement for individual physicians and to improvement in the quality of outcome assessments.

**Disclosures**

The FACE-Q is owned in part by Memorial Sloan-Kettering Cancer Center (MSKCC). Based on MSKCC inventor-sharing policy, Drs Pusic, Cano, and Klassen receive a portion of licensing revenues when the FACE-Q is used in industry-sponsored clinical trials. Dr Panchapakesan and Ms Scott have no conflicts of interest to disclose.

**Funding**

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This study was funded in part by a grant from the Plastic Surgery Foundation (PSF).

**REFERENCES**


44. RUMM 2030 [computer program]. Perth, Australia: RUMM Laboratory Pty Ltd; 2009.


55. Likert R. A technique for the measurement of attitudes. Arch Psychol. 1932;140:50.


60. Hobart JC, Cano SJ, Zajicek JP, Thompson AJ. Rating scales as outcome measures for clinical trials in
neurology: problems, solutions, and recommendations. 