Journal of Plastic, Reconstructive & Aesthetic Surgery 95 (2024) 231–238

Establishing test-retest reliability and the smallest detectable change of FACE-Q Aesthetic Module scales

Lucas Gallo [a,*], Charlene Rae [b], Patrick J. Kim [a], Sophocles H. Voineskos [c], Achilles Thoma [d,e], Andrea L. Pusic [f], Anne F. Klassen [b], Stefan J. Cano [g]

[a] *Division of Plastic Surgery, McMaster University, Hamilton, ON, Canada*
[b] *Department of Pediatrics, McMaster University, Hamilton, ON, Canada*
[c] *Division of Plastic and Reconstructive Surgery, University of Toronto, Toronto, ON, Canada*
[d] *Department of Health Research Methods, Evidence & Impact, McMaster University, Hamilton, ON, Canada*
[e] *Department of Surgery, Division of Plastic Surgery, McMaster University, Hamilton, ON, Canada*
[f] *Division of Plastic Surgery, Brigham and Women's Hospital, Boston, MA, USA*
[g] *Modus Outcomes, Statfold, United Kingdom*

Received 2 April 2024; Accepted 2 June 2024

KEYWORDS
FACE-Q;
Quality of life;
Patient-reported outcome measures;
Cosmetic;
Aesthetic;
Reliability

Summary *Background:* The test-retest (TRT) reliability of FACE-Q Aesthetic scales is yet to be assessed. The aim of this study was to establish the TRT reliability of 17 FACE-Q Aesthetic scales and determine the smallest detectable change (SDC) that can be identified using these scales.
Methods: Data were collected from an online international sample platform (Prolific). Participants ≥20 years old, who had been to a dermatologist or plastic surgeon for a facial aesthetic treatment within the past 12 months were asked to provide demographic and clinical information and complete an online REDcap survey consisting of 17 FACE-Q Aesthetic scales. Participants were asked if they would be willing to complete the survey again in 7 days. Only the participants who reported no important change in the scale construct and completed the retest within 14 days were included.
Results: A total of 342 unique participants completed the TRT survey. The mean age of the sample was 36.6 (± 11.5) years, and 82.4% were female. With outlier data removed, all FACE-Q scales demonstrated an intraclass correlation coefficient > 0.70 indicating "good" TRT reliability. The standard error of measurement for the included scales ranged from 3.37 to 11.87, corresponding to a range of SDC_{group} from 0.95 to 3.23 and SDC_{ind} from 9.34 to 32.91.

* Correspondence to: Division of Plastic Surgery, Department of Surgery, McMaster University, 3N27, 1280 Main Street W, Hamilton, ON L8N 3Z5, Canada.
E-mail address: lucas.gallo@medportal.ca (L. Gallo).

https://doi.org/10.1016/j.bjps.2024.06.002

Conclusion: All included FACE-Q scales demonstrated sufficient TRT reliability and stability overall after the outlier data were removed. Moreover, the authors calculated the values for the SDC for these scales.

The FACE-Q Aesthetic Module is a validated, patient-reported outcome measure (PROM) that evaluates the outcomes that are important to patients receiving facial aesthetic interventions.[1,2] FACE-Q Aesthetics was designed to include 34 independently functioning scales and 6 checklists that measure facial appearance (i.e., satisfaction or how bothered an individual is by their appearance), health-related quality of life (HRQL), and the adverse effects of treatment.[3,4] More recently, FACE-Q Aesthetics was expanded to include a new module to measure the concept of natural from the patient's perspective and 2 item libraries that provide an innovative means to measure satisfaction with face and psychological wellbeing.[5,6] As FACE-Q Aesthetics is not intervention specific, and it can be used to measure and compare outcomes following a variety of surgical and minimally invasive facial aesthetic procedures.

Since the first publication of the FACE-Q Aesthetic Module in 2010, a series of publications have established the psychometric validity and internal consistency of these scales through Rasch measurement theory (RMT) and classical test theory (CTT) analyses.[1-3,7-12] However, establishing the validity and reliability of PROMs is an ongoing and iterative process. Notably, researchers are yet to establish the test-retest (TRT) reliability of key FACE-Q Aesthetic scales—an important criterion for evaluating the quality of PROMs as per COnsensus-based Standard for the selection of health Measurement INstruments (COSMIN) guidelines.[13]

Specifically, the TRT reliability of a scale is estimated by administering the same test to the same group of respondents at different times, when no change in the construct being measured is expected.[2] The correlation between the two scores, indicates the stability of the instrument. Using these data for each scale, one can further establish the standard error of measurement (SEM; the standard error in an observed score that obscures the true score) as well as the smallest detectable change (SDC; the smallest measurement change, that can be interpreted as a real difference).[14]

The aim of this study was to establish the TRT reliability of 17 FACE-Q Aesthetic scales and provide commentary on two different methods for assessing TRT reliability. Additionally, this study aimed to estimate the SDC that can be identified by these scales in an online, international community-based sample.

Methods

The study was coordinated at the McMaster University (Canada). Ethics board approval (#13603) was obtained from the Hamilton Integrated Ethics Board (Canada).

Participant sample and recruitment

An online screening survey was conducted in December 2022, using the crowd-sourcing platform Prolific (www.prolific.com).[15] Following a REDCap (Vanderbilt University, Nashville, Tennessee) pilot study where 144 individuals were invited to complete a survey consisting of FACE-Q Aesthetic scales, a sample of 1895 Prolific participants were then invited to participate. At that time, residents of Canada and the USA fluent in English in the Prolific sample totaled 121,170. The participants were paid the equivalent of 10.80 GBP per hour.

A flow diagram (Figure 1) illustrates the methods used for participant invitation and selection. For the original REDcap survey, individuals were asked to provide clinical and demographic information as well as complete 17 FACE-Q Aesthetic scales. These 17 scales were selected from the 34 available scales within the module owing to their relevance with minimally invasive facial aesthetic interventions. Participants were not required to complete all 17 FACE-Q Aesthetic scales. Specifically, scale administration was based on applicable branching logic (e.g., participants who endorsed crow's feet lines were asked through branching logic to complete the Lines: Crow's Feet FACE-Q Aesthetic scale).

At the end of the original survey, participants were asked (yes/no) if they would be willing to complete a second survey again in 7 days for a TRT study (Figure 1). Participants were only asked to complete relevant FACE-Q scales for the TRT if they were completed as part of the original survey.

As per the COSMIN guidelines, we aimed to invite approximately 100 participants per FACE-Q Aesthetic scale for the TRT analysis.[13] Participants were included in the TRT survey if they: 1) were ≥20 years old; 2) lived in Canada or the United States; and 3) had been to a dermatology or a plastic surgery clinic in the past 12 months to receive 1 of 14 facial aesthetic treatments (Supplementary Appendix 1). For the TRT analysis, only the participants who reported "no important change" in the scale construct being measured were included. Participants who reported an "important change" or completed the survey after 14 days (i.e., from the time of the initial survey) were excluded from this analysis.

FACE-Q Aesthetic scales

The FACE-Q Aesthetic scales (Table 1) convert a raw ordinal score into a continuous outcome that ranges from 0 (worst) to 100 (best), where *higher* scores correspond to *improved* appearance or HRQL outcomes. The description of individual scales can be found in the FACE-Q Aesthetics module user's guide (https://qportfolio.org/face-q/aesthetics/).[4]

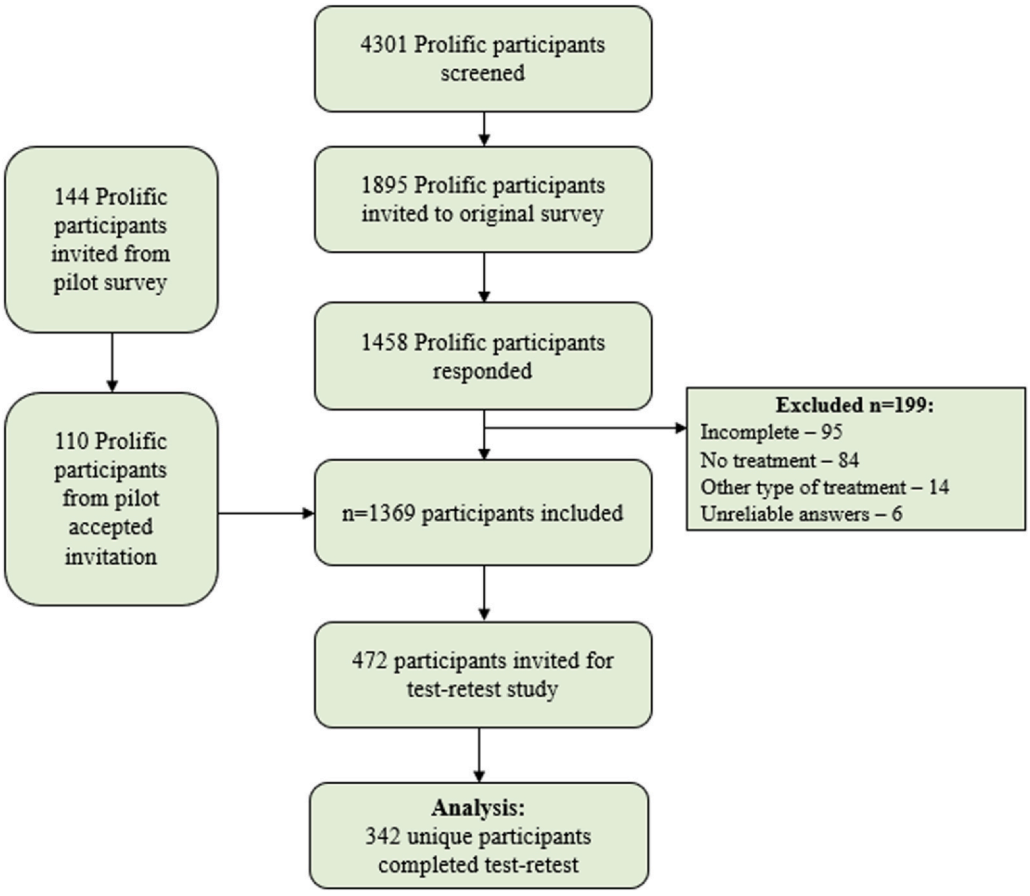

Figure 1 Participant sample and recruitment flow diagram.

Data analysis

To assess TRT reliability, the two-way mixed-effect model evaluating absolute agreement was used to calculate intraclass correlation coefficient (ICC) with 95% confidence intervals. ICC was considered acceptable if it was ≥0.70 as per the International Society of Quality of Life Research (ISOQOL) and COSMIN criteria.[3,16] Stability of the measure was further assessed using the Pearson's correlation coefficient to evaluate the correlation between initial and retested FACE-Q scale scores and were interpreted according to the thresholds outlined by Hinkle et al.[17] Extreme outliers were identified using boxplots and an apriori planned sensitivity analyses was performed with and without outliers for the TRT analysis.

To determine the SEM (i.e., standard error in an observed score that obscures the true score) for individual FACE-Q scale scores, the authors used the formula SEM = $(SD_{T1}) + (SD_{T2})/2)^*\sqrt{(1-ICC)}$, where standard deviations (SD) for test (T1) and retest (T2) were used.[14,18] After calculating the SEM, the SDC (i.e., the smallest measurement change in the scale that can be interpreted as a real difference and not due to measurement error) was determined at the individual [$SDC_{ind} = 1.96^*\sqrt{(2)}^*SEM$] and group [$SDC_{group} = SDC_{ind}/\sqrt{n}$, where "n" corresponds to the size of the sample] levels.[14,18]

Statistical significance was considered at $p < 0.05$. All analyses were performed using SPSS ® version 26.0 (IBM Corporation, Armonk NY, USA for Windows ®). Missing data were handled using an available case analysis.

Results

Among the original survey sample of n = 1369 individuals (Figure 1), a total of 472 participants were invited to complete the online REDcap TRT survey, with 342 unique participants providing data. The number of participants per FACE-Q Aesthetic scale ranged from 91-119 (Table 1). The mean age of the sample was 36.6 (± 11.5) years with 82.4% female, 15.9% male, and 1.7% gender diverse/other individuals. Participants completed a range of 1 to 11 unique FACE-Q Aesthetic scales for this analysis. Additional participant demographic characteristics are presented in Table 2.

Reliability

With outliers included, all FACE-Q Aesthetic scales with the exception of the Skin Appearance scale (ICC = 0.68), demonstrated an ICC > 0.70 (0.72-0.90) indicating "good" TRT reliability. When outlier data were excluded through the sensitivity analysis, all FACE-Q scale ICC values were ≥0.72 (0.72-0.98). The associated 95% ICC confidence intervals for each scale are provided in Table 1.

Table 1 FACE-Q Aesthetic Module scales TRT reliability and SDC analysis.

FACE-Q scales	Outliers (+/−)	N	T1 mean score	SD	T2 mean score	SD	Mean score difference	SD	ICC	ICC 95% CI (lower)	ICC 95% CI (upper)	Pearson's *r*	SEM	SDC (individual)*	SDC (group)
Aging appraisal	−	96	75.54	25.91	76.18	25.76	0.64	6.68	0.98	0.97	0.99	0.97	3.37	9.34	0.95
	+	112	73.52	28.43	75.41	25.78	1.89	17.85	0.88	0.82	0.92	0.79	9.43	26.13	2.47
Cheeks	−	100	70.30	22.13	73.05	22.45	2.75	14.88	0.87	0.81	0.91	0.78	7.97	22.10	2.21
	+	107	69.55	23.83	73.02	22.37	3.47	20.03	0.77	0.66	0.84	0.63	11.20	31.04	3.00
Lines: Crow's feet	−	105	65.97	25.56	73.20	24.30	7.23	16.30	0.86	0.76	0.92	0.79	9.29	25.76	2.51
	+	111	66.50	25.50	71.77	25.22	5.26	20.60	0.79	0.70	0.86	0.67	11.62	32.21	3.06
Lines: between eyebrow	−	106	67.30	24.70	76.10	22.56	8.80	16.82	0.82	0.66	0.90	0.75	9.91	27.47	2.67
	+	107	67.26	24.58	75.39	23.63	8.13	18.12	0.81	0.67	0.88	0.72	10.51	29.12	2.82
Early life impact	−	81	91.09	11.88	92.46	11.10	1.37	6.24	0.92	0.87	0.95	0.86	3.29	9.12	1.01
	+	99	88.82	13.88	90.64	14.41	1.82	11.72	0.79	0.69	0.86	0.66	6.47	17.93	1.80
Face overall	−	97	60.82	21.09	64.36	21.76	3.54	13.17	0.89	0.83	0.93	0.81	7.11	19.70	2.00
	+	107	60.49	22.36	65.91	22.50	5.42	19.55	0.75	0.63	0.83	0.62	11.12	30.83	2.98
Lines: forehead	−	109	68.06	24.18	76.98	23.40	8.93	19.75	0.76	0.60	0.85	0.66	11.63	32.24	3.09
	+	112	67.85	24.16	76.13	24.61	8.29	22.10	0.72	0.56	0.81	0.59	12.97	35.95	3.40
Lower face and jawline	−	104	61.24	28.56	70.55	27.10	9.31	20.50	0.82	0.68	0.89	0.73	11.87	32.91	3.23
	+	112	60.54	29.72	70.29	26.90	9.74	25.62	0.72	0.56	0.82	0.59	15.03	41.67	3.94
Lines: overall	−	101	64.99	23.63	70.74	23.49	5.75	12.95	0.91	0.83	0.94	0.85	7.26	20.13	2.00
	+	113	64.06	23.85	69.27	25.39	5.20	19.53	0.80	0.71	0.87	0.69	10.90	30.21	2.84
Lines: lips	−	86	70.13	24.74	78.92	24.03	8.79	17.97	0.81	0.65	0.89	0.73	10.52	29.15	3.14
	+	91	71.31	24.75	77.70	24.11	6.40	20.12	0.78	0.66	0.86	0.66	11.41	31.62	3.31
Lips	−	98	75.49	22.50	78.29	21.69	2.80	10.63	0.94	0.90	0.96	0.89	5.63	15.62	1.58
	+	113	74.36	23.04	78.74	21.38	4.38	17.41	0.81	0.72	0.87	0.70	9.65	26.76	2.52
Lines: Marionette	−	95	62.58	27.09	72.24	25.60	9.66	17.14	0.85	0.68	0.92	0.79	10.17	28.19	2.89
	+	101	63.18	27.75	71.54	26.43	8.37	23.79	0.74	0.60	0.83	0.62	13.79	38.21	3.80
Lines: nasolabial folds	−	98	67.69	25.83	75.52	23.37	7.83	16.47	0.85	0.73	0.91	0.78	9.50	26.32	2.66
	+	108	67.77	26.91	73.52	23.76	5.75	20.55	0.79	0.69	0.86	0.68	11.53	31.95	3.07
Skin	−	116	57.38	18.91	70.20	19.24	12.82	15.25	0.72	0.22	0.87	0.68	10.18	28.22	2.62
	+	119	57.22	19.56	69.74	20.18	12.52	17.53	0.68	0.29	0.83	0.61	11.31	31.35	2.87
Psychological	−	102	67.61	24.52	71.80	24.29	4.20	14.97	0.89	0.83	0.93	0.81	8.09	22.44	2.22
	+	109	67.10	25.56	72.12	23.89	5.02	18.91	0.82	0.73	0.88	0.71	10.49	29.07	2.78
Social	−	97	56.71	22.24	55.86	21.55	−0.86	9.55	0.95	0.93	0.97	0.91	4.90	13.57	1.38
	+	106	57.33	23.17	56.79	22.20	−0.54	15.93	0.86	0.80	0.91	0.75	8.49	23.53	2.29
Outcome	−	113	69.12	19.28	70.29	19.21	1.18	10.77	0.92	0.88	0.94	0.84	5.61	15.55	1.46
	+	115	69.54	19.39	70.10	19.10	0.56	11.69	0.90	0.85	0.93	0.82	6.12	16.95	1.58

N, sample size; T1, first scale administration; T2, second scale administration; SD, standard deviation; ICC, intraclass correlation coefficients; 95% CI, 95% confidence interval; SEM, standard error of measurement; SDC, smallest detectable change; +, outlier data included; -, outlier data excluded.
*The SDC_ind is applied at the patient level and should be interpreted with caution (see discussion).

Table 2 Participant sample demographics.

Participant demographics	N = 472	%
Age, years		
20-30 years	174	36.9%
31-40 years	154	32.6%
> 41 years	144	30.5%
Gender		
Female	389	82.4%
Male	75	15.9%
Gender Diverse/Other	8	1.7%
BMI, kg/m^2		
< 18.5	17	3.6%
18.5-24.9	233	49.4%
25.0-29.9	112	23.7%
≥30.0	97	20.6%
Missing	13	2.8%
Country		
USA	399	84.5%
Canada	72	15.3%
Missing	1	0.2%
Ethnicity		
Caucasian	315	66.7%
African American	32	6.8%
Latin American	24	5.1%
East Asian	29	6.1%
Southeast Asian	8	1.7%
South Asian	12	2.5%
Middle Eastern	4	0.8%
Mixed	44	9.3%
Other	4	0.8%
Marital status		
Single/Never married	201	42.6%
Separated/Divorced	37	7.8%
Married/Common-law	227	48.1%
Other	7	1.5%
Level of education		
Completed some/all of high school	26	5.5%
Completed some/all of college or university	304	64.4%
Completed some/all of masters or doctoral degree	142	30.1%
Financial difficulty		
Not at all difficult	186	39.4%
A little difficult	146	30.9%
Somewhat difficult	97	20.6%
Very difficult	23	4.9%
Extremely difficult	18	3.8%
Prefer not to answer	2	0.4%

For the Pearson's correlation coefficient, 65% (n = 11/17) of FACE-Q scales had *r* between 0.50-0.69 indicating "moderate" correlation and 35% (n = 6/17) demonstrated an *r* between 0.70-0.89 indicating "high" correlation. When outliers were removed, 76% (n = 13/17) of FACE-Q Aesthetic scales demonstrated an *r* between 0.70-0.89 indicating "high" correlation. Two scales had r > 0.90 indicating "very high" correlation and 2 scales had *r* between 0.50-0.69 indicating "moderate" correlation (Table 1).

Measurement error

The SEM for the included scales demonstrated a range from 6.12-15.03, corresponding to a range of SDC_{ind} from 16.95 to 41.67 and SDC_{group} from 1.58 to 3.94. With outlier data removed, the SEM for the included scales demonstrated a range from 3.37-11.87, corresponding to a range of SDC_{ind} from 9.34 to 32.91 and SDC_{group} from 0.95 to 3.23 (Table 1).

Discussion

The FACE-Q Aesthetic scales are rigorously designed and validated PROMs which measure outcomes that matter to patients who undergo surgical or minimally invasive facial cosmetic procedures.[1,2] Since its initial development, these scales have been used to measure primary and secondary outcomes within clinical studies. Notably, in April 2022, the US FDA qualified the most used FACE-Q Aesthetics scales as medical device development tools (MDDTs).[19,20] In the present study, the authors established a TRT reliability of the 11 MDDT qualified scales as well as 6 additional scales selected for their relevance to minimally invasive facial aesthetic interventions. When the outlier data were removed, the analysis confirmed that these scales are stable overall, exceeding the COSMIN guidelines.[13] This finding was determined by administering the same FACE-Q Aesthetic scales to the same group of respondents at different times (i.e., between 7 to 14 days following initial completion), when no change in the outcome/construct being measured was expected. These results are consistent with TRT reliability studies in related FACE-Q Aesthetic instruments including new natural scales and item libraries.[5,6]

Although clinicians frequently recognize the significance of validity when critically appraising PROMs, the importance of reliability is often less understood. Notably, validity refers to a PROM's ability to measure what it intends to measure, while reliability refers to the amount of random and systemic error that exists when attempting to measure an outcome. These properties can be equated to the "accuracy" and the "precision" of an instrument, respectively.[21] For example, a PROM that is valid but not reliable may accurately report the construct being evaluated, but these scores may vary considerably depending on the occasion and context when the PROM is administered. Therefore, an unreliable scale would have limited utility in the clinical/research setting. Given this, reliability is considered to be a necessary, but insufficient component of validity.

Reliability can be further broken down into internal consistency reliability and TRT reliability. TRT reliability measures the ability of a scale to produce consistent results across multiple time points.[22] A measure with sufficient TRT reliability suggests that any change observed within the PROM score is attributable to the construct being measured rather than the error in the measurement tool. Notably, the literature presents conflicting guidance on which measure (s) should be used to assess TRT reliability; thus, the authors appraised 2 commonly used measures—the ICC and Pearson's correlation coefficient. Although Pearson's correlation

Table 3 Summary of stability and measurement error statistics.

Stability statistics	Description	Interpretation	Findings	Limitations
Peason's correlation coefficient, *r*	Measure of linear correlation between 2 variables. For test-retest reliability, correlation is evaluated between scores at time 1 and time 2.	$-1 \leq r \leq 1$ Linear relationship is stronger as *r* approaches 1 (or inverse linear relationship as *r* approaches −1). No widely accepted cutoff values for test-retest reliability. Interpretation as per Hinkle et al.[17] used in this article: *±0.9-1.0: very high correlation* *±0.7-0.9: high correlation* *±0.5-0.7: moderate correlation* *±0.3-0.5: low correlation* *±0.0-0.3: negligible correlation*	All FACE-Q Aesthetics module scales had moderate to very high correlation depending on whether outliers were removed.	Does not account for agreement. Values can be perfectly linearly correlated but have little to no agreement. Conversely, values can have good agreement without linear correlation.[23]
ICC	Provides information about whether the responses in 2 groups are similar (agreement). Multiple models exist. This study uses a two-way mixed-effects model evaluating absolute agreement. The rationale for the selection of ICC models and the formula used are reported elsewhere.[23]	≥0.70 considered acceptable by COSMIN and ISOQOL for test-retest reliability.[3,16]	All FACE-Q Aesthetics module scales but the Skin Appearance scale had good test-retest reliability without outliers removed. All FACE-Q scales had good test-retest reliability when outliers were removed.	Values are dependent on the variance of the assessed population. If there is high variance in disease states between time 1 and 2, agreement will be poor regardless of the measure's actual stability.[25]

Measurement Error	Description	Interpretation	Findings	Limitations
SEM	Standard error in an observed score that obscures the true score. Formula: $(SDT1) + (SDT2)/2)^*\sqrt{(1-ICC)}$	Smaller values indicate more precise measures.	SEM was 6.12-14.35 and 3.25-11.03 with and without outliers, respectively.	The SEM assumes that individual scores at the center of the scale have the same level of precision as those at the floor and ceiling of the scale (i.e., where scores are the least precise). The SEM is also dependent on the score distribution (i.e., standard deviation, SD) of the individual sample. Thus, if the SD of the sample is large, the SEM may be large - limiting the precision of the measurement.[26]
SDC_{ind}	Minimum change in score needed to be confident that differences between an individual's scores are due to meaningful change, not systematic errors of the measurement tool. Formula: $1.96^*\sqrt{(2)}^*SEM$	Smaller values indicate greater ability of the measure to detect meaningful change.	SDC_{ind} was 16.95-39.78 and 9.34-30.57 with and without outliers, respectively.	
SDC_{group}	Minimum change in score needed to be confident that differences between groups' scores are due to meaningful change and not systematic errors of the measurement tool. Formula: $SDC_{ind}/\sqrt{n}$	Smaller values indicate greater ability of the measure to detect meaningful change.	SDC_{group} was 1.58-3.76 and 0.95-3.00 with and without outliers, respectively.	

ICC, intraclass correlation coefficient; ISOQOL, International Society of Quality of Life Research; COSMIN, COnsensus-based Standards for the selection of health Measurement INstruments; SEM, standard error of measurement; SDC_{ind}, smallest detectable change for individual subjects; SDC_{group}, smallest detectable change for groups.

coefficient is sometimes used as a measure of TRT reliability, this test is limited in that it measures the relationship between 2 variables without accounting for their level of agreement.[23] Specifically, the test and retest scores may have a perfectly linear relationship, but their actual agreement may be poor. Therefore, this method of evaluating TRT is not ideal.

More appropriately, the most commonly used measure of TRT reliability is the ICC, where ICC >0.70 is widely considered to be the threshold for acceptable TRT reliability.[3,16] ICC overcomes the shortcomings of the Pearson's correlation coefficient as it is designed to detect the extent of agreement between raters (or in this case, the test and retest).[24] However, ICC is dependent on the variance of the assessed population. Therefore, if there is high variance in disease states between test and retest times, agreement will be poor regardless of the measure's actual stability.[25] Practically, while the Pearson's correlation coefficient tends to produce similar results to the ICC, these concepts should be considered when interpreting the reliability of a PROM. See Table 3 for a summary of these stability statistics.

Furthermore, this is the first study to estimate the SDC for the FACE-Q Aesthetic scales. The SDC provides a value for the minimum change that should be observed in repeated measures of the FACE-Q Aesthetic scales to be confident that the change score is not a product of measurement error.[14,18] In this study, the SDC was calculated for individual subjects/patients (SDC_{ind}) as well as for mean scores of groups (SDC_{group}). In practice, the SDC_{ind} may be applied in the clinical setting at the individual patient level to determine whether the pre- and post-intervention change in FACE-Q Aesthetic scale scores is beyond what is expected by measurement error. Alternatively, SDC_{group} may more aptly be applied to the research setting, such as when comparing a new treatment or intervention group against a control sample.[14,18]

The SDC can also be interpreted in the context of the minimal important difference (MID). As defined by Guyatt et al. in 2002, MID is "the smallest difference in score in the domain of interest that patients perceive as important, either beneficial or harmful, and that would lead the clinician to consider a change in the patient's management."[26] Thus, a PROM should have an SDC that is smaller than the MID; this ensures that the PROM is sensitive enough to detect meaningful differences in the domain of interest.[26]

Notably, the error around an individual's PROM score (i.e., SEM) is determined to be a constant value regardless of the person's location on the scale's continuum. Thus, the calculation of the SEM requires several key assumptions. Specifically, it assumes that those individuals at the center of the scale (i.e., patients with the most precise scores) have the same level of precision as those at the floor and ceiling of the scale (i.e., patients with the least precise scores).[27] Additionally, as the computation of the SEM involves calculation of the standard deviation of the sample (i.e., the value dependent on the score distribution of the specific participant sample), the SEM may be large thus limiting the precision of the measurement. Given these limitations, the SEM and SDC are typically only recommended for use in the assessment of group level data (SDC_{group}) rather than for individual patient decision making (SDC_{ind})—as both values are presented in this study for reference.[27] See Table 3 for a summary of measurement error statistics.[28] Owing to the relatively small sample size obtained in this study and associated large SD in FACE-Q scale scores as well as 95% ICC confidence intervals, further research with even larger patient samples may be needed to verify these findings and narrow precision estimates for use in a clinical setting with individual patients. Ideally, ICC values >0.90 are required for applications in individual patients, which was obtained in $N=1$ scale when the outlier data were included and in $N=6$ scales without outliers.[28]

Limitations

This study has several limitations. First, this study included only English-speaking participants from the USA and Canada, and therefore these findings may not be directly applicable to other populations. Second, participants self-selected to participate in this study through an online platform (www.prolific.com) and received monetary compensation for their involvement.[7] Therefore, there may be an element of reporting/volunteer bias due to monetary compensation, which incentivizes participants to complete multiple studies and may impact these results (i.e., leading to outlier data due to speed of study completion). Finally, as the clinical and demographic data provided by participants were self-reported, data could not be independently verified by the study authors.

Conclusions

The FACE-Q Aesthetic Module is a validated PROM that evaluates outcomes that are important to patients receiving surgical and nonsurgical facial aesthetic interventions.[1-3] Reliability is an important component of the overall PROM quality. In this study, all included FACE-Q scales demonstrated sufficient TRT reliability when outlier data were removed and were determined to be stable measures overall. Moreover, the authors provided values for the SDC for these scales—the minimum change required to be confident that the observed change in individual (SDC_{ind}) and group (SDC_{group}) scores is not a product of measurement error. Going forward, future research should be conducted to examine the reliability of the FACE-Q Aesthetics in different populations and contexts. Additionally, there is a need to identify an anchor-based MID for key FACE-Q Aesthetic scale scores.

Ethical approval

The study was coordinated at McMaster University (Canada). Ethics board approval (#13603) was obtained from the Hamilton Integrated Ethics Board (Canada).

Funding

None.

Conflict of interest

Drs Cano, Pusic, and Klassen are co-developers of FACE-Q Aesthetics and receive a share of any license revenue from Memorial Sloan Kettering Cancer Center's inventor sharing policy. All remaining authors have no conflicts of interest to disclose.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.bjps.2024.06.002.

References

1. 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**(4):303–9. https://doi.org/10.1055/s-0030-1262313.
2. Pusic AL, Klassen AF, Scott AM, Cano SJ. Development and psychometric evaluation of the FACE-Q satisfaction with appearance scale: a new patient-reported outcome instrument for facial aesthetics patients. *Clin Plast Surg* 2013;**40**(2): 249–60. https://doi.org/10.1016/j.cps.2012.12.001.
3. Klassen AF, Cano SJ, Schwitzer JA, Scott AM, Pusic AL. FACE-Q scales for health-related quality of life, early life impact, satisfaction with outcomes, and decision to have treatment: development and validation. *Plast Reconstr Surg* 2015;**135**(2): 375–86. https://doi.org/10.1097/PRS.0000000000000895.
4. FACE-Q® AESTHETICS. A user's guide for researchers and clinicians. qportfolio.org. January 2023. ⟨https://qportfolio.org/wp-content/uploads/2023/01/FACE-Q-AESTHETICS-USERS-GUIDE.pdf⟩ [Accessed October 16, 2023].
5. Klassen AF, Cano S, Mansouri J, et al. "I want it to look natural": development and validation of the FACE-Q Aesthetics Natural Module. *Aesthet Surg J* 2024;**44**(7):733–43. https://doi.org/10.1093/asj/sjad374.
6. Klassen AF, Pusic AL, Kaur M, et al. The SKIN-Q: an innovative patient-reported outcome measure for evaluating minimally invasive skin treatments for the face and body. *Facial Plast Surg Aesthet Med* 2024;**26**(3):247–55. https://doi.org/10.1089/fpsam.2023.0204.
7. 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**(8):1099–109. https://doi.org/10.1177/1090820×13510170.
8. 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**(1):21–30. https://doi.org/10.1097/01.prs.0000436814.11462.94.
9. Klassen AF, Cano SJ, East CA, et al. Development and psychometric evaluation of the face-Q scales for patients undergoing rhinoplasty. *JAMA Facial Plast Surg* 2016;**18**(1):27–35. https://doi.org/10.1001/jamafacial.2015.1445.
10. Klassen AF, Cano SJ, Schwitzer JA, et al. Development and psychometric validation of the face-Q skin, lips, and facial rhytids appearance scales and adverse effects checklists for cosmetic procedures. *JAMA Dermatol* 2016;**152**(4):443–51. https://doi.org/10.1001/jamadermatol.2016.0018.
11. Klassen AF, Cano SJ, Grotting JC, et al. Face-Q eye module for measuring patient-reported outcomes following cosmetic eye treatments. *JAMA Facial Plast Surg* 2017;**19**(1):7–14. https://doi.org/10.1001/jamafacial.2016.1018.
12. Klassen AF, Cano SJ, Alderman A, et al. Self-report scales to measure expectations and appearance-related psychosocial distress in patients seeking cosmetic treatments. *Aesthet Surg J* 2016;**36**(9):1068–78. https://doi.org/10.1093/asj/sjw078.
13. Mokkink L, Prinsen C, Patrick D, et al. COSMIN Study Design checklist for patient-reported outcome measurement instruments. ⟨https://www.cosmin.nl/⟩. July 2019. ⟨https://www.cosmin.nl/wp-content/uploads/COSMIN-study-designing-checklist_final.pdf⟩ [Accessed October 16, 2023].
14. King MT. A point of minimal important difference (MID): a critique of terminology and methods. *Expert Rev Pharmacoecon Outcomes Res* 2011;**11**(2):171–84. https://doi.org/10.1586/erp.11.9.
15. Prolific: Run research with confidence based on the highest-quality data. Prolific·Quickly find research participants you can trust. ⟨https://www.prolific.com/⟩ [Accessed October 16, 2023].
16. Reeve BB, Wyrwich KW, Wu AW, et al. ISOQOL recommends minimum standards for patient-reported outcome measures used in patient-centered outcomes and comparative effectiveness research. *Qual Life Res* 2013;**22**:1889–905.
17. Hinkle DE, Wiersma W, Jurs SG. *Applied statistics for the behavioral sciences*. Boston: Houghton Mifflin; 2003.
18. Geerinck A, Alekna V, Beaudart C, et al. Standard error of measurement and smallest detectable change of the sarcopenia quality of Life (sarqol) questionnaire: an analysis of subjects from 9 validation studies. *PLoS One* 2019;**14**(4) :e0216065. https://doi.org/10.1371/journal.pone.0216065.
19. Gallo L, Kim P, Yuan M, et al. Best practices for face-Q aesthetics research: a systematic review of study methodology. *Aesthet Surg J* 2023;**43**(9):NP674–86. https://doi.org/10.1093/asj/sjad141.
20. MDDT summary of evidence and basis of qualification decision for FACE-Q Aesthetics. ⟨www.fda.gov⟩. ⟨https://www.fda.gov/media/157956/download⟩ [Accessed October 16, 2023].
21. Thoma A, Cornacchi SD, Lovrics PJ, Goldsmith CH. *Patient-important outcome measures in surgical care*. Evidence-Based surgery users' guide to the surgical literature: How to assess an article on health-related quality of life. Springer International Publishing; 2008. p. 71–83.
22. Voineskos SH, Nelson JA, Klassen AF, Pusic AL. Measuring patient-reported outcomes: key metrics in reconstructive surgery. *Annu Rev Med* 2018;**69**:467–79.
23. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J Chiropr Med* 2016;**15**(2):155–63.
24. Ranganathan P, Pramesh CS, Aggarwal R. Common pitfalls in statistical analysis: measures of agreement. *Perspect Clin Res* 2017;**8**(4):187–91.
25. Costa-Santos C, Bernardes J, Ayres-de-Campos D, Costa A, Costa C. The limits of agreement and the intraclass correlation coefficient may be inconsistent in the interpretation of agreement. *J Clin Epidemiol* 2011;**64**(3):264–9.
26. Guyatt GH, Osoba D, Wu AW, Wyrwich KW, Norman GR, Clinical Significance Consensus Meeting Group. Methods to explain the clinical significance of health status measures. *Mayo Clin Proc* 2002;**77**:371–83.
27. Hobart J, Cano S. Improving the evaluation of therapeutic interventions in multiple sclerosis: the role of new psychometric methods. *Health Technol Assess* 2009;**13**(12):1–8. https://doi.org/10.3310/hta13120.
28. de Vet HCW, Terwee CB, Mokkink LB, Knol DL. *Reliability*. Measurement in medicine a practical guide. Cambridge University Press; 2018. p. 142–5.

SUPPLEMENTARY APPENDIX 1: Screening questions used in Prolific

AESTHETICS TREATMENTS - FACE	
In the PAST 12 MONTHS, have you been to a DERMATOLOGY or a PLASTIC SURGERY CLINIC to have a FACIAL AESTHETIC treatment?	0, No 1, Yes
In the PAST 12 MONTHS, have you had any of these FACIAL AESTHETIC Treatments: Choose all that apply.	0, NONE 1, BOTULINUM TOXIN A - ie, Botox, Dysport, Xeomin or Jeuveau, Xeomin 2, FILLER - eg, Restylane, Juvederm, Radiesse, Sculptra 3, FAT REDUCTION - eg, Kybella to treat a double chin 4, SKIN BOOSTER (eg, Prophilo) (*asked in field-test screen) 5, PLATELET RICH PLASMA (PRP) injections 6, SKIN TIGHTENING with ultrasound - eg, Ultherapy 7, SKIN TIGHTENING with Radio-frequency - eg, Thermage, Morpheus8, Exilis, Profound RF 8, CHEMICAL PEEL 9, MICRODERMABRASION 10, LASER - eg, CO2, Vbeam, Fraxel 11, INTENSE PULSED LIGHT Light (IPL) - eg, Lynton Lumina IPL 12, MICRONEEDLING 13, HYDRAFACIAL 14, THREADLIFT 15, Other

You said you had BOTOX injected. What was the MAIN REASON for having BOTOX?	1, Cosmetic reasons - to look better, younger, refreshed 2, Medical reasons - to treat migraines, to stop grinding teeth 3, Other reason 88, None of the above
You said you a SOFT TISSUE FILLER injected. The last time you had filler, where was the filler injected? Choose all that apply	1, Cheeks - to add volume and restore fullness 2, Lips - to plump or to smooth out lip lines 3, Other