Research

Measuring Facial Appearance in the Era of GLP-1 Receptor Agonist Use With the FACE-Q Aesthetics Item Library

Anne F Klassen, DPhil; Lucas Gallo, MD, PhD; Steven Dayan, MD; Lotte Poulsen, MD, PhD; Manraj Kaur, PhD; Andrea L Pusic, MD, MSc; Stefan J Cano, PhD; and Charlene Rae, PhD

Aesthetic Surgery Journal
2026, Vol 00(0) 1–10
Accepted date: April 24, 2026; online publish-ahead-of-print May 15, 2026.

https://doi.org/10.1093/asj/sjag095
www.aestheticsurgeryjournal.com
OXFORD UNIVERSITY PRESS

Abstract

Background: A patient-reported outcome measure is needed to study facial appearance in the context of rapid weight loss following use of antiobesity medications. Such treatments can cause sunken cheeks, hollow eyes, saggy skin, and an aged appearance (eg, Ozempic face). The FACE-Q Aesthetics item library can be used to create short-form fit-for-purpose scales.

Objectives: The goal was to create a short-form FACE-Q scale and to examine its psychometric performance in the context of antiobesity medication use.

Methods: An international sample (Prolific Academic) was surveyed 3 times (initial, test–retest, 3-month follow-up). The sample included individuals who wanted to have, were having, or previously had a glucagon-like peptide-1 (GLP-1) or glucose-dependent insulinotropic polypeptide (GIP)/GLP-1 receptor agonist to lose weight. Participants completed the FACE-Q item library that measures satisfaction with facial appearance. Rasch Measurement Theory analysis was used to item-reduce and produce a short-form scale. Psychometric properties were examined, including hypothesis-based construct validity, responsiveness, and test–retest reliability. Distribution-based and anchor-based minimally important differences (MIDs) were computed.

Results: The sample included 632 individuals who wanted to have (30.9%), were having (50.6%), or previously had (18.5%) a GLP-1 or GIP/GLP-1 receptor agonist to lose weight. Data for a 15-item short-form scale fit the Rasch model ($\chi^2 = 130.1$, df = 135, $P = .60$). Reliability was > 0.83 for 3 reliability coefficients. Hypothesis-based construct validity was supported with 14/14 (100%) and 7/9 (77.8%) of initial and change score hypotheses confirmed, respectively. The MIDs for the new scale for improvements were 6 (distribution-based) and 10 (anchor-based) points.

Conclusions: This FACE-Q Aesthetics short-form scale demonstrated strong measurement properties in the context of GLP-1 and GIP/GLP-1 receptor agonist use for weight loss.

Level of Evidence: 4 (Diagnostic)

One billion people worldwide live with obesity.[1] Antiobesity medications (eg, semaglutide, tirzepatide, liraglutide) have been shown to lead to weight loss of 15% to 20% in clinical trials.[2] Demand for weight loss medication has soared; a recent RAND survey of 8783 adults showed that 11.8% of adults reported they had used a glucagon-like peptide-1 (GLP-1) receptor agonist in 2025, and 14%

Dr Klassen is a professor, and Dr Rae is a research associate, Department of Pediatrics, McMaster University, Hamilton, ON, Canada. Dr Gallo is a plastic surgery resident, Department of Surgery, McMaster University, Hamilton, ON, Canada. Dr Dayan is a plastic surgeon, Denovo Research, Chicago, IL, USA. Dr Poulsen is a plastic surgeon, Odense, Denmark. Dr Kaur is an investigator, and Dr Pusic is a plastic surgeon, Mass General Brigham, Boston, MA, USA. Dr Cano is chief scientific officer, Modus Outcomes (a Division of Thread), Cheltenham, UK.

Corresponding Author:
Anne F Klassen, DPhil, Department of Pediatrics, McMaster University, 1280 Main Street W, Hamilton, ON L8N 3Z5, Canada.
E-mail: aklass@mcmaster.ca; X: @anneklassen

of adults expressed interest in taking a GLP-1 receptor agonist for weight loss.[3]

Weight loss by taking medications, such as GLP-1 and glucose-dependent insulinotropic polypeptide (GIP)/GLP-1 receptor agonists, can be rapid and lead to discernable change in facial appearance. Change occurs as fat pads shrink and the skin fails to retract quickly enough, resulting in sunken cheeks, hollowed eyes, sagging skin, and a more aged appearance. The American Society of Plastic Surgeons (ASPS) has acknowledged that the increased uptake of antiobesity medication has driven greater demand for corrective minimally invasive and surgical procedures to address what has been termed in the media as "Ozempic face."[4]

To measure satisfaction with appearance following aesthetic treatments, patient-reported outcome measures (PROMs) are needed. In the context of weight loss and body contouring, a systematic review that applied COnsensus-based Standards for the selection of health Measurement INstruments (COSMIN) criteria to evaluate PROMs reported that the BODY-Q demonstrated the most robust measurement properties among 24 instruments.[5-7] Although BODY-Q measures appearance-related concerns, it does not measure facial appearance. Because antiobesity medication can change a person's facial appearance, a PROM is needed to measure satisfaction from the patient's perspective.

Our team previously developed a modular PROM called FACE-Q Aesthetics to evaluate outcomes for minimally invasive and surgical facial aesthetic treatments.[8,9] The most frequently used FACE-Q scale has 10 items that measure satisfaction with overall facial appearance.[10] Recently, to extend the measurement by this scale, 42 additional items were added, creating an item library that can be customized for fit-for-purpose short-form scales.[11] The objectives of the current study were to create a weight-loss-specific short-form scale from the item library, and to evaluate its psychometric properties in the context of taking a GLP-1 or GIP/GLP-1 receptor agonist for weight loss.

METHODS

Ethics

The authors obtained ethical study approval from the Hamilton Integrated Research Ethics Board in Hamilton, Canada (Project 18651). Participants were asked to read the study consent letter and provide electronic consent before completing the study surveys. The surveys were designed in REDCap, a secure platform for electronic data collection, and hosted by the Faculty of Health Sciences at McMaster University. Data were collected online, and the survey was anonymous. Participants were paid a prorated rate of 12 UK pounds per hour for survey completion.

Data Collection

Screening Survey

The sample was drawn from a larger study that examined different weight loss approaches, including metabolic bariatric surgery, GLP-1 and GIP/GLP-1 receptor agonist use, and lifestyle methods. The authors screened the Prolific platform (prolific.com, London, UK) in May and July 2025 to identify a sample of residents of Australia, Canada, Ireland, New Zealand, the United Kingdom, and the United States. To ensure high-quality data, on July 22, 2025, screened participants were asked to repeat the screen 12 days after the initial screen. Participants who provided the same responses for type of metabolic bariatric surgery and/or type, mode of administration, and treatment status (current vs past use) for GLP-1 and GIP/GLP-1 receptor agonists were invited in July 2025 to complete the initial survey. The survey remained open until August 9, 2025. For the current study, the subset of individuals from the larger sample who wanted to have, were having, or previously had a GLP-1 or GIP/GLP-1 receptor agonist to lose weight were included.

Data Collected

Prolific participants were asked to complete the FACE-Q item library and demographic and weight-related questions.[11] For construct validation, participants also answered the FACE-Q Aesthetics Age Visual Analogue scale (VAS), which asked how much older/younger (±15 years) they looked compared with their actual age; 2 SKIN-Q items that measured satisfaction (very dissatisfied, somewhat dissatisfied, somewhat satisfied, very satisfied) with elasticity and overall skin quality; self-reported Merz Assessment Scale photonumeric severity ratings (none, mild, moderate, severe, very severe) for saggy skin on the jawline, hollowness under the eyes, how sunken the upper and lower cheeks looked, and depth of nasolabial folds and marionette lines; and self-reported Allergan temple hollowing photonumeric ratings (convex, flat, minimal, moderate, severe).[12-15]

Test–retest (TRT) data were collected between 7 and 14 days after the initial survey. Participants were asked (yes/no) if there had been any change in how satisfied they were with the appearance of their face and if they had had any cosmetic treatments on their face or neck since completing the initial survey. Participants who reported change or a treatment were excluded from the TRT analysis. We aimed to recruit at least 100 participants for the TRT to obtain a "very good" rating according to COSMIN guidelines.[16]

At least 3 months after completion of the initial survey, participants were invited to complete a follow-up survey. The authors used the piping function in REDCap to comply with recommendations by Devji et al, who reported that individuals have difficulty recalling a previous health state after 4 weeks and should be reminded.[17] The authors used a 3-part anchor to assess change in satisfaction with facial appearance, including questions that asked about direction (less, same, more), magnitude (a little, somewhat, quite a bit, a lot), and importance (not at all, a little, somewhat, very, extremely). Follow-up data were collected between November 13 and December 7, 2025.

Analysis

Supplemental Table 1 (available online at https://doi.org/10.1093/asj/sjag095) shows the psychometric tests performed. Rasch Measurement Theory (RMT) analysis was employed to item-reduce the FACE-Q item library.[18-21] The authors utilized RUMM2030 software (RUMM Laboratory, 2010, Perth, Australia) with the unrestricted partial credit model for polytomous data. To determine if the FACE-Q short-form scale worked the same in the GLP-1 and GIP/GLP-1 receptor agonist sample compared with the original facial aesthetic item library field test sample, they computed differential item functioning (DIF).[11] DIF was also examined in the current sample for age (18-34, 35-44, 45-54, >55); gender (male, female); and BMI (normal, overweight, class 1, class 2, and class 3 obesity).

Following the RMT analysis, the raw scores for the FACE-Q short-form scale were converted into 0 (low) to 100 (high) scores using the Rasch logit values. The transformed scores were analyzed in IBM SPSS Statistics for Windows, version 30 (IBM Corp., Armonk, NY) for test–retest reliability and hypothesis-based construct validity. The authors tested 14 and 10 predetermined hypotheses for the

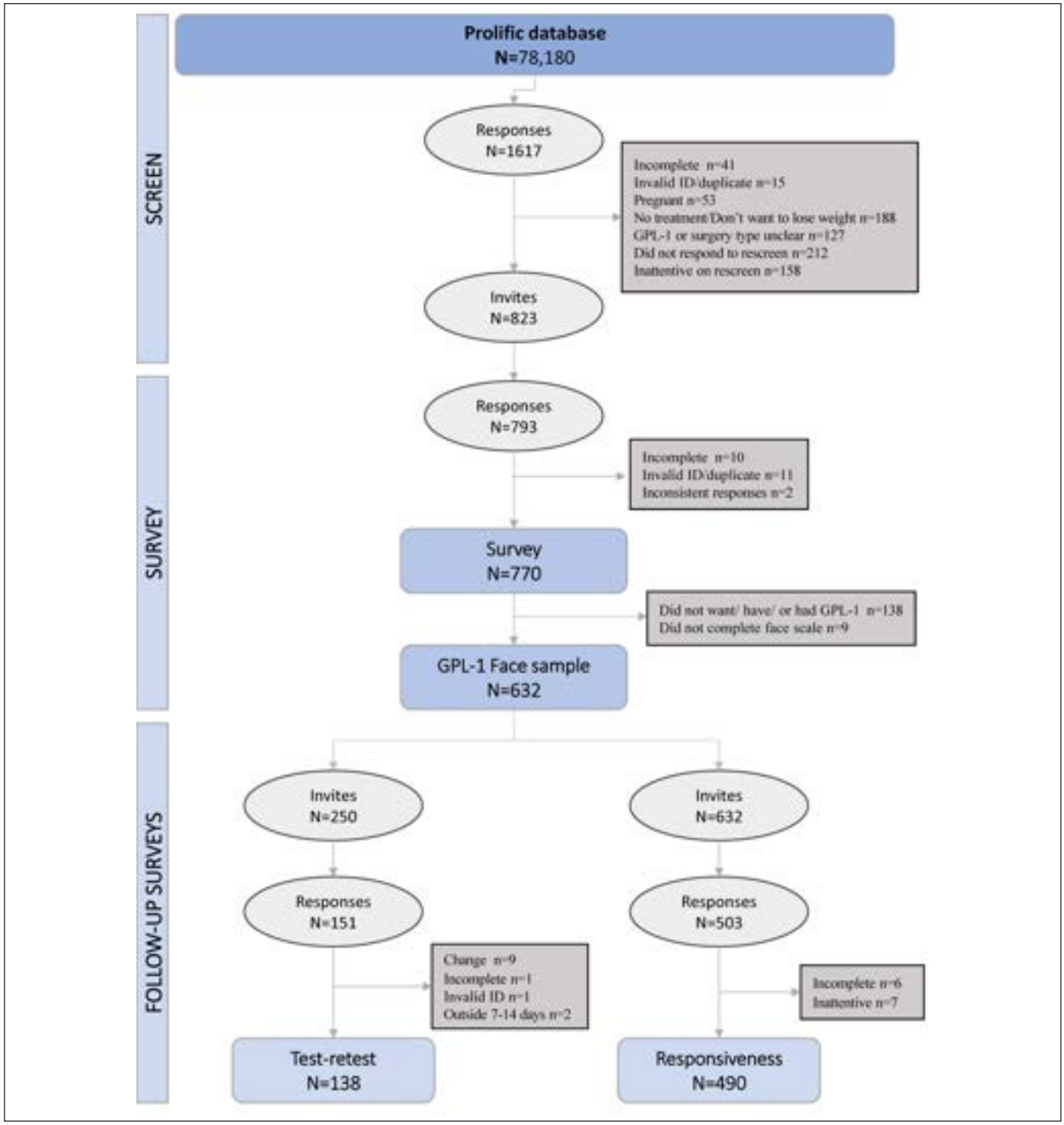


Figure 1. Recruitment for each component of the study.

initial and follow-up surveys, respectively. For the cross-sectional hypotheses, they calculated an effect size for group differences as eta squared (η^2) and interpreted as follows: < 0.01 negligible, $0.01 \leq \eta^2 < 0.06$ small, $0.06 \leq \eta^2 < 0.14$ medium, and ≥ 0.14 large.[21] These differences were hypothesized to be at least small in size ($\eta^2 \geq 0.01$). Likert scale directionality (eg, none, mild, moderate, severe) was hypothesized to relate to differences in scale scores. For change score hypotheses, the authors computed the mean difference and 95% CIs. For directionality, they expected the direction of the change to relate to the mean difference (ie, improvement = positive, worsening = negative, no change = 0). The goal for construct validity was to examine the evidence overall rather than focusing on the individual results for each hypothesis. Construct validity was considered sufficient if at least 75% of the hypotheses were confirmed.[22] For interpretation of scores, minimally important differences (MIDs) were computed. Distribution-based MIDs were calculated as half the effect size (ES)

Table 1. Participant Demographic and Weight-Related Characteristics

		Main survey		Follow-up	
		n = 632	%	*n* = 490	%
Country	USA	299	47.3	234	47.8
	Canada	64	10.1	49	10.0
	UK	235	37.2	185	37.8
	Australia	21	3.3	5	1.0
	Ireland	11	1.7	15	3.1
	New Zealand	2	0.3	2	0.4
Gender	Man	174	27.5	142	29.0
	Woman	447	70.7	342	69.8
	Nonbinary	9	1.4	4	0.8
	Other gender	2	0.3	2	0.4
Age (years)	18-34	170	26.9	124	25.3
	35-44	185	29.3	132	26.9
	45-54	159	25.2	127	25.9
	≥ 55	118	18.7	107	21.8
Racial group(s)	White	517	81.8	401	81.8
	Black	53	8.4	44	9.0
	South Asian	14	2.2	10	2.0
	Latin American	10	1.6	6	1.2
	Southeast Asian	6	0.9	5	1.0
	Middle Eastern	0	0.0	0	0.0
	East Asian	0	0.0	0	0.0
	Indigenous	3	0.5	2	0.4
	Mixed race	28	4.4	21	4.3
	Other	1	0.2	1	0.2
Education	Some high school	6	0.9	6	1.2
	Completed high school	67	10.6	50	10.2
	Some college or trade school or university	132	20.9	103	21.0
	Completed college/trade school/university	284	44.9	223	45.5
	Some master's or doctoral degree	33	5.2	27	5.5
	Completed master's or doctoral degree	110	17.4	81	16.5
Difficulty covering household expenses and bills in the past 3 months	Not at all difficult	179	28.3	144	29.4
	A little difficult	205	32.4	159	32.4
	Somewhat difficult	136	21.5	99	20.2
	Very difficult	67	10.6	53	10.8
	Extremely difficult	44	7.0	34	6.9

Table 1. Continued

		Main survey		Follow-up	
		n = 632	%	*n* = 490	%
	Prefer not to answer	1	0.2	1	0.2
Marital status	Married	306	48.4	237	48.4
	Never married	179	28.3	132	26.9
	Divorced	65	10.3	55	11.2
	Living common-law	47	7.4	37	7.6
	Separated	19	3.0	15	3.1
	Widowed	10	1.6	9	1.8
	Other	6	0.9	5	1.0
Fitzpatrick skin type	Always burn and never tan	75	11.9	62	12.7
	Usually burn and minimally tan	204	32.3	162	33.1
	Sometimes burn and uniformly tan	204	32.3	147	30.0
	Rarely burn and always tan	74	11.7	58	11.8
	Rarely burn and easily tan	51	8.1	41	8.4
	Never burn and never tan	24	3.8	20	4.1
BMI category	<24.9	65	10.3	3	0.6
	25-29.9	105	16.6	48	9.8
	30-34.9	124	19.6	74	15.1
	35-39.9	120	19.0	100	20.4
	>39.9	218	34.5	95	19.4
Currently trying to lose weight	No, but I want to	104	16.5	74	15.1
	No, don't want to	10	1.6	8	1.6
	Yes, trying	518	82.0	408	83.3
Weight change in last 12 months	Lost >100 pounds	12	1.9	8	1.6
	Lost 50-99 pounds	65	10.3	52	10.6
	Lost 30-49 pounds	115	18.2	90	18.4
	Lost <30 pounds	217	34.3	171	34.9
	No change	81	12.8	66	13.5
	Gained <30 pounds	109	17.2	78	15.9
	Gained 30-99 pounds	33	5.2	16	3.3
Location in weight loss journey	Not started	51	8.1	37	7.6
	Close to the start	259	41.0	206	42.0
	About halfway through	219	34.7	167	34.1
	Close to the end	95	15.0	74	15.1
	Finished	8	1.3	6	1.2

Table 1. Continued

		Main survey		Follow-up	
		n = 632	%	*n* = 490	%
Weight loss treatments to date	None but want treatment	135	21.4	109	22.2
	GLP-1	257	40.7	198	40.4
	GLP-1 + weight loss program	132	20.9	106	21.6
	GLP-1 + weight loss program + surgery	36	5.7	23	4.7
	GLP-1 + surgery	12	1.9	11	2.2
	Weight loss program	28	4.4	23	4.7
	Weight loss program + surgery	26	4.1	15	3.1
	Surgery	6	0.9	5	1.0

BMI, body mass index; GLP-1, glucagon-like peptide-1.

Table 2. GLP-1 and GIP/GLP-1 Receptor Agonists Taken by 437 Participants Who Reported Current or Past Use

Type	Main survey (*n* = 437)				Follow-up (*n* = 338)			
	Currently having		Had in past		Currently having		Had in past	
	n	%	*n*	%	*n*	%	*n*	%
Ozempic (Semaglutide)	87	19.9	96	22.0	68	20.1	73	21.6
Mounjaro (Tirzepatide)	138	31.6	45	10.3	110	32.5	35	10.4
Wegovy (Semaglutide)	42	9.6	57	13.0	34	10.1	38	11.2
Zepbound (Tirzepatide)	35	8.0	6	1.4	24	7.1	4	1.2
Liraglutide (Saxenda)	3	0.7	34	7.8	2	0.6	25	7.4
Trullicity (dulaglutide)	7	1.6	15	3.4	5	1.5	13	3.8
Victoza (liraglutide)	2	0.5	11	2.5	1	0.3	8	2.4
Setmelanotide (Imcivree)	2	0.5	1	0.2	1	0.3	0	0.0
Other	13	3.0	10	2.3	9	2.7	9	2.7

GIP, glucose-dependent insulinotropic polypeptide; GLP-1, glucagon-like peptide-1.

and one-half a standard deviation.[23] Anchor-based MIDs were calculated as the mean score change for those who reported a small but important difference.[24]

RESULTS

Figure 1 shows numbers of participants for the screen, main survey, TRT, and follow-up. The sample included more women (*n* = 447, 70.7%) than men (*n* = 174, 27.5%) and people who identified as another gender (*n* = 11, 1.7%). Table 1 shows sample characteristics. Most of the 632 participants came from the United States (47.3%) and identified as women (70.7%), White race (81.8%), married (48.4%), and college educated (67.5%). Participants ranged in age from 19 to 80 years (mean = 43; SD = 11.8). The sample BMI ranged from 14 to 90 (mean = 36.8; SD = 10.2). Most participants (87.1%) were dissatisfied with their current weight and most (82%) were trying to lose weight. Of the 632 individuals, 195 (30.9%) wanted to have, 320 (50.6%) were currently having, and 117 (18.5%) previously had a GLP-1 or GIP/GLP-1 receptor agonist to lose weight.

In terms of weight loss treatments, 69.1% of participants were having or had a GLP-1 or GIP/GLP-1 receptor agonist, 35.2% used a medically managed weight loss program, and 12.7% had metabolic bariatric surgery. A small proportion (5.7%) had utilized all 3 weight loss methods. Table 2 shows the type of GLP-1 and GIP/GLP-1 receptor agonists taken by the sample. Participants who perceived themselves as having "Ozempic face" (*n* = 143) reported it as being slight (*n* = 73, 51%), a little (*n* = 37, 25.9%), some (*n* = 21, 14.7%), quite a bit (*n* = 10, 7%), and a lot (*n* = 2, 1.4%).

Rasch Analysis

Supplemental Table 2 (available online at https://doi.org/10.1093/asj/sjag095) shows the item level fit statistics and DIF results. The item library was used to form a 15-item scale that included the following concepts that change with weight loss: natural, well-proportioned, healthy, age, smooth, youthful, full, lifted, hydrated, radiant, even-toned, attractive, rested, photos, and profile. Data from the sample fit the Rasch model (χ^2 = 130.1, df = 135, *P* = .60). Data from the sample fit the Rasch model (χ^2 = 130.1, df = 135, *P* = .60). All 15 items had ordered thresholds (Supplemental Figure 1), fit the Rasch model with nonsignificant *P* values after Bonferroni adjustment, and 12 of 15 had fit residuals within ±2.5. None of the items evidenced DIF by sample (field test vs GLP-1 and GIP/GLP-1 receptor agonist).[11] In the the present survey sample, 5 items evidenced DIF: 4 items for age, 3 items for gender, and 1 item for the BMI category. When the items with DIF were split on the characteristic, the Pearson correlations between the person locations from the original and split analyses indicated no impact of DIF on the scoring (correlations = 1.00).

Reliability was high with person separation index (PSI) values of 0.95, 0.94, and Cronbach alpha values of 0.96, 0.95, with and without extremes, respectively. Item fit residuals for 1 pair of items ("The age your face makes you look" and "How youthful your face looks") was 0.31, suggestive of some local dependency. However, a subtest showed that the impact of the residual correlations on the reliability coefficients was low (< 0.01). The scale was well-targeted to the sample, with 96% of participants scoring within the range of measurement provided. Floor (2.2%) and ceiling (1.7%) effects were low.

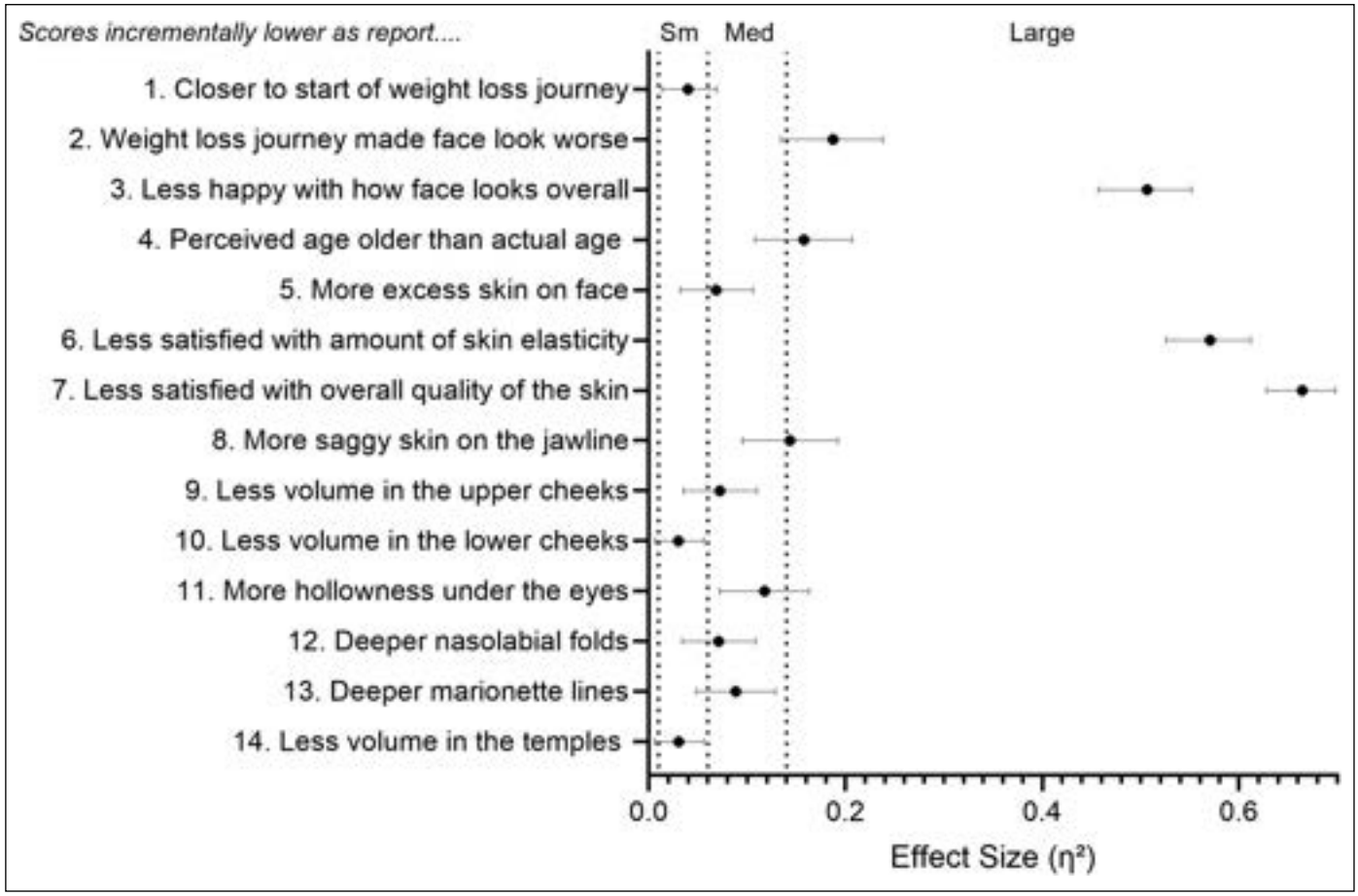


Figure 2. Effect size and 95% Confidence Intervals for cross-sectional hypotheses. Med, medium; Sm, small.

Supplemental Figure 2 shows the person-item threshold distribution; the top histogram shows that most participants (607/632) scored within the range of measurement for the scale (bottom histogram).

Test–Retest

Of 250 participants invited to complete the TRT, 151 responded. Of these, 13 were excluded for the following reasons: reported change in satisfaction with facial appearance or had a facial treatment ($n = 9$), completed the survey outside of 7 to 14 days ($n = 2$), did not finish the survey ($n = 1$), or gave an invalid Prolific ID ($n = 1$). Table 3 shows the TRT. The average ICC values were 0.91 ($n = 138$) and 0.95 ($n = 132$) with and without outliers, respectively.

Hypothesis-Based Construct Validity

For the initial survey, 14 hypotheses were tested and all 14 were accepted (100%). Figure 2 shows the effect size for each hypothesis with the 95% confidence intervals. The eta squared for all 14 hypotheses was > 0.01, meeting our hypotheses for magnitude and directionality. The 3 large effect sizes were for the overall items (face and skin) rather than items for specific facial areas. Detailed results are shown in Supplemental Table 3, available online at https://doi.org/10.1093/asj/sjag095. Significantly lower scores were associated with the following: being closer to the start of the weight loss journey; reporting that the weight loss journey made their face look worse; being less happy with how their face looked overall; looking older than one's chronological age; reporting more loose skin on the face; being less satisfied with skin elasticity and overall skin quality; higher photonumeric severity ratings for saggy skin on the jawline, sunken upper cheeks, eye hollowness, nasolabial folds, marionette lines, and temple hollowness.[13-15]

Table 1 shows the participant characteristics for the follow-up sample. Participants completed the follow-up survey on average 3.8 (SD = 0.2) months after the initial survey (range 3.2 to 4.5 months). Of the 632 participants invited, 503 responded. After excluding 13 participants who did not complete the survey ($n = 6$) or failed an attention check ($n = 7$), the follow-up sample included 490 participants. Of 10 change score hypotheses tested, the sample size was too small for 1 hypothesis. Of the remaining 9, 7 (77.8%) were accepted. Figure 3 shows the mean difference and 95% CIs; detailed results are shown in Supplemental Table 4, available online at https://doi.org/10.1093/asj/sjag095. In comparison to time 1, lower change scores were associated with the following: reporting more excess skin on the face; being less happy with how their face looked; looking older than their actual age; being less satisfied with skin elasticity and overall skin quality; being less satisfied with how the face looked (eg, youthful, attractive, smooth, rested); reporting more severity for saggy skin on the jawline and hollowness under the eyes.[13,14] Two hypotheses were not significant, including amount of excess skin on the face looking better, and perceived change (ie, more, the same, or less) in Ozempic face.

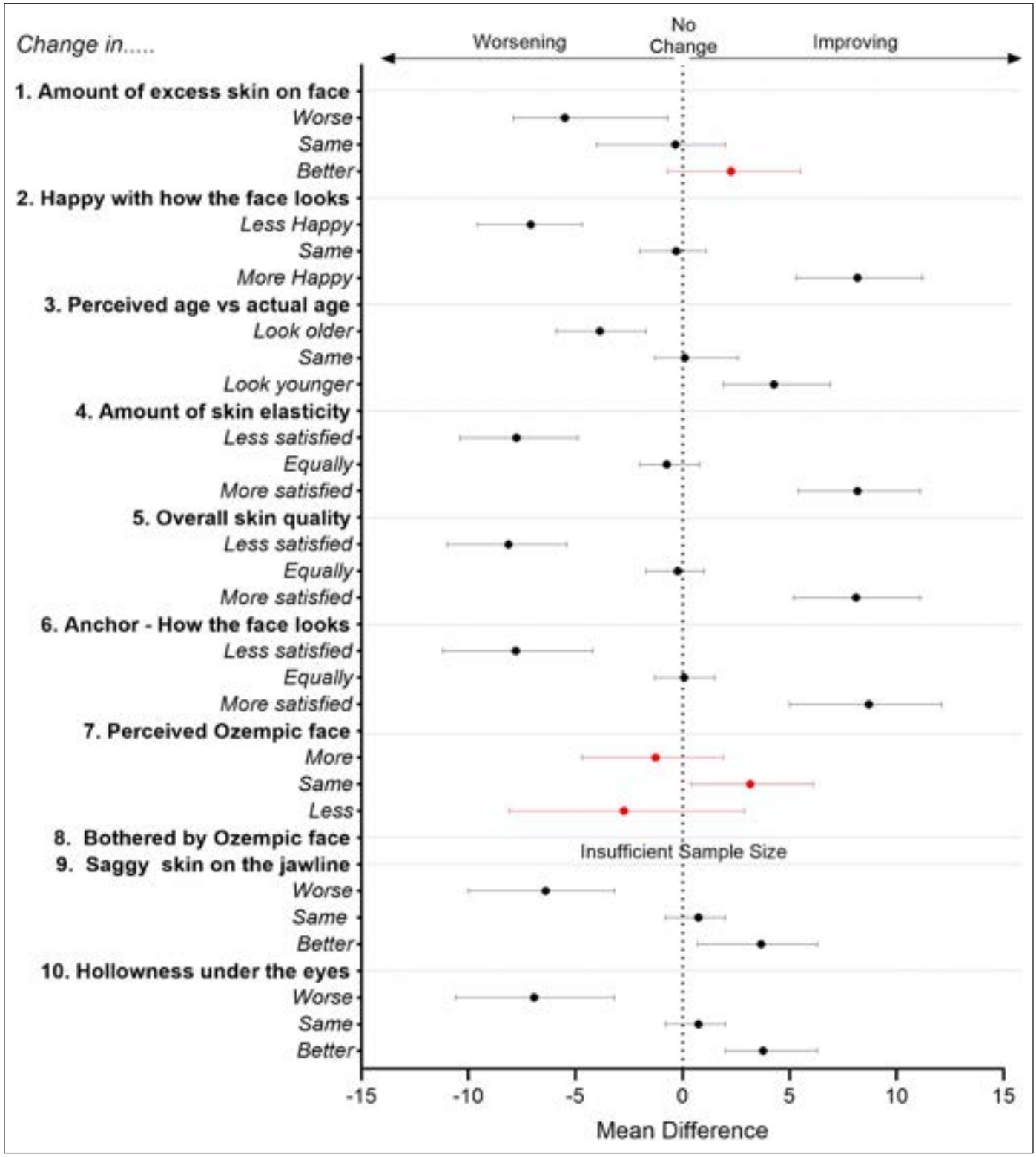


Figure 3. Mean difference and 95% Confidence Intervals for change score hypotheses. Red indicates nonsignificant findings.

Minimally Important Differences

Distribution-based MID values based on one-half SD were 6.9 points and 6.3 points for worse and improved satisfaction, respectively, and 7.2 based on all cases (Table 4). Anchor-based MID values were higher, with an 11-point change for those who reported a negative change in satisfaction that was small but important to them, and a 10-point change for those reporting a positive change in satisfaction that was small but important to them. The anchor-based MID on the full sample was 13.

Table 3 shows the smallest detectable change (SDC) results. The SDC at the group level was 1.4 and 1.0 with and without extremes respectively, which was smaller than the MID values.

Table 3. Test–Retest Results and Smallest Detectible change

Sample	*n*	Single				Average				Means for scales by time point						SDC		
		ICC	95% CI		Sig	ICC	95% CI		*P* value	T1 mean	T1 SD	T2 mean	T2 SD	Mean diff	Mean diff SD	SEM	SDC individual	SDC group
			LB	UB			LB	UB										
Face + outliers	138	0.84	0.78	0.88	<.001	0.91	0.88	0.94	<.001	55.2	19.5	53.1	19.1	2.0	10.9	5.8	16.0	1.4
Face – outliers	132	0.90	0.85	0.93	<.001	0.95	0.92	0.96	<.001	55.6	18.2	53.6	18.3	2.1	8.1	4.3	11.9	1.0

CI, confidence interval; diff, difference; ICC, intraclass correlation coefficient; LB, lower bound; SD, standard deviation; SDC, smallest detectible change; SEM, standard error of measurement; Sig, significance; T1, time 1; T2, time 2; UB, upper bound.

Table 4. Group Level Indicators of Small/Important Change as Well as Overall Cases

	Distribution-based MID											Anchor-based MID				
	Time point	Mean	SE	SD	*n*	SE diff for a person	ES (T2-T1)/ SD1	Mean change[a]	SD change	SRM	ES/ SRM	0.5 SD change	*n*	Mean	SD	LB	UB
Report worse[a]	T1	44.7	2.3	18.4	63	3.0	−0.4	−7.7	13.9	−0.6	0.8	6.9	29	−11	15.3	−16.8	−5.2
	T2	37.0	2.0	15.5	63												
Report better[a]	T1	57.7	2.6	18.7	51	3.7	0.5	8.5	12.6	0.7	0.7	6.3	14	10	13.0	2.3	17.3
	T2	66.2	2.6	18.5	51												
All cases	T1	52.7	0.9	19.1	490	1.2	0.001	−0.03	14.3	−0.002	−0.7	7.2	43	13	11.9	9.6	16.9
	T2	52.7	0.9	19.1	490												

[a]Based on participants who reported a small but important difference. Diff, difference; ES, effect size; LB, lower bound; MID, minimally important difference; SD, standard deviation; SE, standard error; SEM, standard error of measurement; SRM, standardized response mean; T1, time 1; T2, time 2; UB, upper bound.

DISCUSSION

The rapid adoption of weight loss medications, such as GLP-1 and GIP/GLP-1 receptor agonists, has introduced new challenges in facial aesthetic medicine. Accelerated weight loss can alter facial fat pads, leading to sunken cheeks, hollowed eyes, and an accentuated aged appearance, referred to as "Ozempic face".[4] These changes have driven increased demand for corrective surgical and nonsurgical facial aesthetic procedures. To evaluate patient-reported outcomes following these interventions, the authors developed and validated the psychometric properties of a novel FACE-Q Aesthetic short-form scale specifically for patients using GLP-1 and GIP/GLP-1 receptor agonists for weight loss.

This study provides evidence of strong psychometric performance for the 15-item fit-for-purpose weight loss scale created from the FACE-Q item library to measure outcomes that matter to patients.[11] Although this study was conducted to provide evidence of the FACE-Q scale in the context of antiobesity weight loss medication, a sizable number of participants had metabolic bariatric surgery (12.7%) and/or utilized a medically managed weight loss program (34.2%). Therefore, this new scale could be used to measure change in facial appearance for different weight loss treatments.

The new FACE-Q scale evidenced high reliability with 3 coefficients that exceeded COSMIN criteria.[22] The authors tested an extensive number of a priori hypotheses and found that 100% and 77.8% were supported for the cross-sectional and change hypotheses, respectively. Furthermore, in the RMT analysis, the data for the weight loss sample and all 15 items fit the Rasch model, providing further evidence of the scale's validity. Taken together, the findings provide strong evidence that this short-form scale reliably measures what it purports to measure.

This study provides researchers and clinicians with benchmarks to aid in the interpretation of the new FACE-Q scale scores. For example, the findings identify a distribution-based MID of 6 points and an anchor-based MID of 10 points, representing a small but important positive change in FACE-Q scores. As defined by Guyatt et al, the MID reflects "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".[25] These thresholds can be used to contextualize study findings and to inform sample size calculations for future observational and randomized studies in which the FACE-Q scale is included as an outcome measure.

Furthermore, this study estimates the SDC of the FACE-Q to be 12 for individual-level data and 1 for group-level data with outliers excluded. The SDC represents the minimum magnitude of change required between repeated measurements to ensure that an observed difference reflects a true change rather than measurement error.[16] In clinical practice, the individual-level SDC can be used to determine whether preintervention to postintervention changes in FACE-Q scores for a given patient exceed the threshold of measurement error. Conversely, the group-level SDC is appropriate for use in research, such as to compare outcomes between intervention and control groups.[26] The MID values were similar to, or slightly smaller than, the individual-level SDC, suggesting that individual-level change scores should be interpreted with caution. On the other hand, the MID exceeded the group-level SDC, indicating that the scale can detect meaningful change at the group level.

This study contains several key limitations. First, only English-speaking residents of Australia, Canada, Ireland, New Zealand, the United Kingdom, and the United States were included. Findings may not be transferable to patients in other countries. Second, participants self-selected to complete the study surveys through the online Prolific platform (prolific.com) and received monetary compensation for their involvement. Therefore there is a risk of volunteer/reporting bias impacting results. Third, as participant data were self-reported, responses to clinical and demographic questions could not be independently verified. Finally, the anchor-based MID estimates provided in this analysis are limited due to small samples of participants who reported a small but important difference in scale scores. Additional research with larger sample sizes is needed to verify these estimates.

CONCLUSIONS

As uptake of GLP-1 and GIP/GLP-1 receptor agonists increases demand for corrective procedures that address sunken cheeks, hollow eyes, saggy skin, and an aged appearance, this new FACE-Q fit-for-purpose short-form scale provides a means to measure change in satisfaction with facial appearance. This study showed that the FACE-Q scale demonstrated adequate reliability and validity for assessing patient-important facial aesthetic outcomes in the context of use of medication for weight loss. The authors reported both distribution- and anchor-based MID values, which facilitate interpretation of the scale's scores and can inform sample size calculations in future observational and randomized study designs. Further research is warranted to confirm the psychometric properties of the scale in additional patient populations and contexts of use. The FACE-Q is available for licensing for clinical and research use at https://qportfolio.org/face-q/aesthetics.

Supplemental Material

This article contains supplemental material located online at www.aestheticsurgeryjournal.com.

Disclosures

Drs Klassen, Cano, and Pusic are codevelopers of FACE-Q Aesthetics and receive a share of license revenues as royalties based on their institutions' inventor sharing policy. Stefan Cano is a principal, advisor (consultant) in health outcomes research for Modus Outcomes, a Division of Thread. Anne Klassen provides research consulting services to the pharmaceutical industry through EVENTUM Research. All other authors have nothing to disclose.

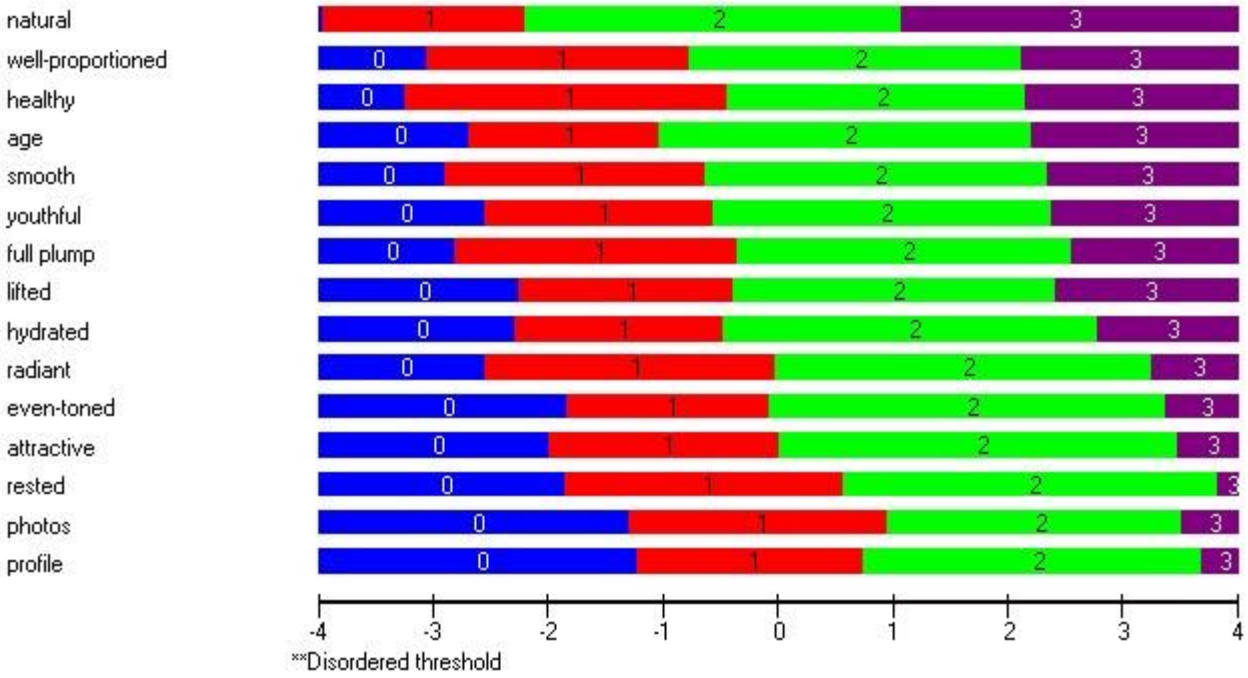
Funding

This study was funded by research funds provided to A Klassen from the Department of Pediatrics, McMaster University.

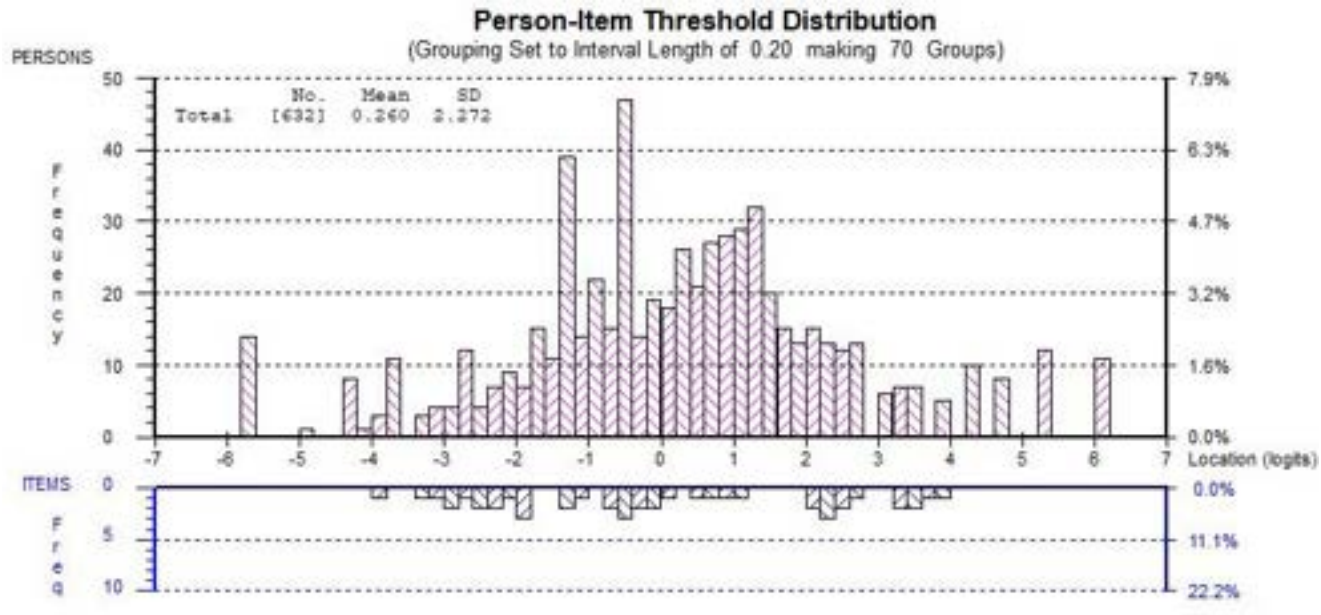
REFERENCES

1. NCD Risk Factor Collaboration (NCD-RisC). Worldwide trends in underweight and obesity from 1990 to 2022: a pooled analysis of 3663 population-representative studies with 222 million children, adolescents, and adults. *Lancet*. 2024;403:1027-1050. doi: 10.1016/S0140-6736(23)02750-2
2. Moiz A, Filion KB, Toutounchi H, et al. Efficacy and safety of glucagon-like peptide-1 receptor agonists for weight loss among adults without diabetes: a systematic review of randomized controlled trials. *Ann Intern Med*. 2025;178:199-217. doi: 10.7326/ANNALS-24-01590
3. Pollard M, Baird MD. *The RAND American Life Panel: Technical Description*. RAND Corporation; 2017. Accessed June 30, 2025. https://www.rand.org/pubs/research_reports/RR1651.html
4. American Society of Plastic Surgeons. *How plastic surgery can address "Ozempic face."* Accessed June 30, 2025. https://www.plasticsurgery.org/news/articles/how-plastic-surgery-can-address-ozempic-face
5. Klassen AF, Cano SJ, Alderman A, et al. The BODY-Q: a patient-reported outcome instrument for weight loss and body contouring treatments. *Plast Reconstr Surg Glob Open*. 2016;4:e679. doi: 10.1097/GOX.0000000000000665
6. de Vries CEE, Mou D, Poulsen L, et al. Development and validation of new BODY-Q scales measuring expectations, eating behavior, distress, symptoms, and work life in 4004 adults from 4 countries. *Obes Surg*. 2021;31:3637-3645. doi: 10.1007/s11695-021-05462-2
7. de Vries CEE, Kalff MC, Prinsen CAC, et al. Recommendations on the most suitable quality-of-life measurement instruments for bariatric and body contouring surgery: a systematic review. *Obes Rev*. 2018;19:1395-1411. doi: 10.1111/obr.12710
8. 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. doi: 10.1055/s-0030-1262313
9. 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;40:249-260. doi: 10.1016/j.cps.2012.12.001
10. Gallo L, Kim P, Churchill IF, et al. Measuring the impact of surgical and nonsurgical facial cosmetic interventions using FACE-Q Aesthetic Module scales: a systematic review and meta-analysis. *Plast Surg*. 2025;33:403-418. doi: 10.1177/22925503231225480
11. Klassen AF, Pusic AL, Kaur M, et al. Extending the range of measurement for minimally invasive treatments by adding new concepts to FACE-Q Aesthetics scales. *Plast Reconstr Surg Glob Open*. 2024;12:e5736. doi: 10.1097/GOX.0000000000005736
12. 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. doi: 10.1177/1090820X13510170
13. Klassen AF, Rae C, Gallo L, et al. Measuring satisfaction with minimally invasive aesthetic treatments with the SKIN-Q Treatment Outcome Scale. *Facial Plast Surg Aesthet Med*. 2024;26:247-255. doi: 10.1089/fpsam.2023.0204
14. Stella E, Di Petrillo A. Standard evaluation of the patient: the Merz scale. In: Goisis M, ed. *Injections in Aesthetic Medicine*. Springer; 2014:33-50.
15. Carruthers J, Jones D, Hardas B, et al. Development and validation of a photonumeric scale for evaluation of volume deficit of the temple. *Dermatol Surg*. 2016;42:S251-S258. doi: 10.1097/DSS.0000000000000856
16. Mokkink LB, Prinsen CAC, Patrick DL , et al. *COSMIN Study Design checklist for patient-reported outcome measurement instruments. Version July 2019*. COSMIN; 2019. Accessed 17 April 2026. https://www.cosmin.nl.
17. Devji T, Carrasco-Labra A, Qasim A, et al. Evaluating the credibility of anchor-based estimates of minimal important differences for patient-reported outcomes: instrument development and reliability study. *BMJ*. 2020;369:m1714. doi: 10.1136/bmj.m1714
18. Rasch G. *Probabilistic Models for Some Intelligence and Attainment Tests*. Danish Institute for Educational Research; 1960.
19. Hobart JC, Cano S, et al. Improving the evaluation of therapeutic interventions in multiple sclerosis: the role of new psychometric methods. *Health Technol Assess*. 2009;13:1-177. doi: 10.3310/hta13120
20. Tennant A, Conaghan PG. The Rasch measurement model in rheumatology: what is it and why use it? When should it be applied, and what should one look for in a Rasch paper? *Arthritis Rheum*. 2007;57:1358-1362. doi: 10.1002/art.23108
21. Cohen J. *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed. Lawrence Erlbaum Associates; 1988.
22. Mokkink LB, Elsman EBM, Terwee CB. *COSMIN manual for conducting systematic reviews of PROMs: Version 2.0*. COSMIN; 2025. Accessed 25 May 2026. https://www.cosmin.nl/wp-content/uploads/COSMIN-manual-V2_final.pdf
23. Norman GR, Sloan JA, Wyrwich KW. Interpretation of changes in health-related quality of life: the remarkable universality of half a standard deviation. *Med Care*. 2003;41:582-592. doi: 10.1097/01.MLR.0000062554.74615.4C
24. Devji T, Guyatt G, Carrasco-Labra A, et al. Mind the methods of determining minimal important differences: three critical issues to consider. *Evid Based Ment Health*. 2021;24:77-81. doi: 10.1136/ebmental-2020-300164
25. Guyatt GH, Osoba D, Wu AW, Wyrwich KW, Norman GR. Methods to explain the clinical significance of health status measures. *Mayo Clin Proc*. 2002;77:371-383. doi: 10.4065/77.4.371
26. 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:e0216065. doi: 10.1371/journal.pone.0216065

Supplemental Figure 1: Threshold maps for FACE-Q scale



Supplemental Figure 2: Person-item threshold distributions for FACE-Q scale



Supplemental Table 2: Item fit statistics and Differential Item Functioning (DIF) results

Items	Item fit							DIF			
	Location	SE	Fit Residual	DF	χ^2	DF	p-value	Sample	Age	Gender	BMI
Natural	-1.69	0.08	2.71	564	26.40	9	0.00	No	No	Yes	No
Well-proportioned	-0.57	0.07	1.65	564	8.40	9	0.49	No	Yes	No	No
Healthy	-0.51	0.07	-1.40	564	6.82	9	0.66	No	Yes	Yes	Yes
Age face	-0.50	0.07	-1.67	564	4.55	9	0.87	No	No	No	No
Smooth	-0.39	0.07	-1.94	564	6.77	9	0.66	No	No	No	No
Youthful	-0.25	0.07	-1.54	564	4.36	9	0.89	No	Yes	No	No
Full	-0.21	0.07	0.33	564	7.87	9	0.55	No	No	No	No
Lifted	-0.08	0.07	1.45	564	6.36	9	0.70	No	Yes	Yes	No
Hydrated	0.01	0.07	1.04	558	3.73	9	0.93	No	No	No	No
Radiant	0.23	0.07	-4.37	564	15.35	9	0.08	No	No	No	No
Even-toned	0.49	0.07	2.11	558	6.92	9	0.65	No	No	No	No
Attractive	0.50	0.07	-1.79	564	9.80	9	0.37	No	No	No	No
Rested	0.85	0.07	-2.81	562	8.06	9	0.53	No	No	No	No
Photos	1.06	0.07	0.86	562	8.11	9	0.52	No	No	No	No
Profile	1.07	0.07	0.11	562	6.62	9	0.68	No	No	No	No

Abbreviations: SE = Standard Error, DF = Degrees of Freedom, χ^2 = Chi-square; DIF = Differential Item Functioning

Supplemental Table 3: Detailed results for cross-sectional hypotheses

	N	Mean	SD	SE	95% CI		p-value/ES
					LB	UB	
1. Scores incrementally lower as report being closer to start of weight loss journey							
Start	310	49	20	1	46.6	51.0	<0.001 η^2=0.038
Middle	219	55	18	1	52.1	57.0	
End	103	58	17	2	55.1	61.7	
2. Scores incrementally lower as report weight loss journey made face look worse							
My face looks a lot worse	25	24	15	3	17.8	30.4	<0.001 η^2=0.190
My face looks a little worse	84	43	15	2	39.7	46.3	
My face looks the same	279	52	19	1	49.4	54.0	
My face looks a little better	203	57	15	1	55.0	59.2	
My face looks a lot better	41	70	17	3	64.2	75.1	
3. Scores incrementally lower as report being less happy with how face looks overall							
Extremely unhappy	47	23	18	3	17.2	28.0	<0.001 η^2=0.511
Very unhappy	72	39	17	2	35.2	43.2	
Somewhat unhappy	185	47	12	1	45.0	48.6	
Somewhat happy	214	57	12	1	55.1	58.5	
Very happy	95	71	13	1	68.0	73.2	
Extremely happy	19	88	10	2	83.0	92.7	
4. Scores lower as report perceived age older than actual age [13]							
Younger	296	58	17	1	56.5	60.4	<0.001 η^2=0.158
Same age	161	54	18	1	51.4	56.9	
Older	175	40	18	1	37.6	43.1	
5. Scores incrementally lower with more excess skin on face							
None	411	56	19	1	53.7	57.4	<0.001 η^2=0.068
A little	160	49	15	1	46.6	51.2	
A moderate amount	41	40	22	3	33.1	46.8	
A lot	20	39	28	6	26.0	52.4	
6. Scores incrementally lower as report less satisfied with amount of skin elasticity [14]							
Very dissatisfied	95	27	16	2	23.8	30.5	<0.001 η^2=0.574
Somewhat dissatisfied	170	46	11	1	44.4	47.9	
Somewhat satisfied	274	57	11	1	55.5	58.2	
Very satisfied	86	79	14	2	75.6	81.7	
7. Scores incrementally lower as report less satisfied with overall quality of skin [14]							
Very dissatisfied	82	24	15	2	20.7	27.4	<0.001 η^2=0.667
Somewhat dissatisfied	161	43	10	1	41.4	44.4	
Somewhat satisfied	303	58	10	1	56.6	58.9	
Very satisfied	79	81	13	1	78.0	83.9	

8. Scores incrementally lower as report more saggy skin on the jawline [15]							
No sagging	210	59	20	1	55.8	61.2	<0.001 η^2=0.144
Mild sagging	241	54	16	1	52.3	56.4	
Moderate sagging	142	45	17	1	42.7	48.3	
Severe/very severe sagging	37	30	19	3	24.0	37.0	
9. Scores incrementally lower as report less volume in upper cheeks [15]							
Full upper cheeks	348	55	19	1	53.3	57.4	<0.001 η^2=0.072
Mildly sunken upper cheeks	184	52	17	1	50.0	54.9	
Moderately sunken upper cheeks	74	44	17	2	40.0	48.0	
Severe/very severe sunken upper cheeks	24	34	24	5	23.8	43.9	
10. Scores incrementally lower as report less volume in lower cheeks [15]							
Full lower cheeks	340	54	19	1	51.4	55.6	<0.001 η^2=0.029
Mildly sunken lower cheeks	191	53	17	1	51.0	55.9	
Moderately sunken lower cheeks	74	49	19	2	44.6	53.2	
Severe/very sunken lower cheeks	25	38	24	5	28.1	48.1	
11. Scores incrementally lower as report more hollowness under the eyes [15]							
No hollowness	192	60	20	1	57.6	63.2	<0.001 η^2=0.118
Mild hollowness	269	52	17	1	50.3	54.4	
Moderate hollowness	133	44	16	1	41.4	46.8	
Severe/Very severe hollowness	36	40	23	4	32.2	47.6	
12. Scores incrementally lower as report deeper nasolabial folds [15]							
No folds	166	58	20	2	55.1	61.3	<0.001 η^2=0.070
Mild folds	240	54	18	1	51.9	56.4	
Moderate folds	160	48	17	1	44.9	50.3	
Severe/very severe folds	63	42	20	3	37.5	47.5	
13. Scores incrementally lower as report deeper marionette lines [15]							
No lines	222	58	20	1	55.3	60.6	<0.001 η^2=0.088
Mild lines	220	53	16	1	50.9	55.3	
Moderate lines	120	48	19	2	44.4	51.3	
Severe/very severe lines	67	40	18	2	35.3	43.9	
14. Scores incrementally lower as report less volume in the temples [16]							
Convex	142	56	19	2	52.8	59.1	<0.001 η^2=0.029
Flat	221	54	18	1	51.6	56.4	
Minimal	198	51	18	1	48.0	53.1	
Moderate/severe	66	45	24	3	38.9	50.9	

Eta squared (η^2) interpreted as <0.01 negligible, 0.01 ≤η^2<0.06 small, 0.06 ≤η^2<0.14 Medium, ≥0.14 Large;
SD – standard deviation, SE – standard error, CI – confidence interval, LB – lower bound, UB -upper bound, ES -Effect size

Supplemental Table 1: Psychometric tests performed

Test	Description
Response threshold order	To determine if the response categories were properly ordered [19-20]. We examined thresholds between response options.
Item fit	To determine if the data fit the Rasch model, we examined 3 item fit statistics: Chi-square after Bonferroni adjustment, fit residuals and item characteristic curves [19-20].
Local dependency	To determine if item residuals were closely related to each other, we will examine residual correlations between pairs of items [19-20]. For any greater than 0.30 a subtest was performed to determine their impact on reliability.
Targeting	To determine how well the scale measures the experience of the sample, we inspected person-item threshold plots to identify clustering of items and gaps on the scale [19]. We also computed the proportion of participants to score on the scale's range of measurement and floor and ceiling effects.
Differential item functioning (DIF)	We examined DIF to determine whether items responded differently by subgroups within the sample. We examined DIF by sample (original aesthetic vs. current weight loss), age-group, gender and BMI category. For each analysis, we selected random samples to ensure subgroups are equivalent in size and repeated the analysis 3 times to determine if the result was stable. When DIF was identified, we split the data based on the characteristic, and person locations from the original and split analyses were correlated to inspect the impact of DIF on scale scoring [19-20].
Reliability	We computed person separation index [19-20] and Cronbach alpha values with and without extremes included. For test-retest reliability, intraclass correlation coefficients (ICC) with a two-way random effects model were computed after excluding participants who reported change for the FACE-Q scale on retest. To determine the amount of measurement error in the score between the test and retest, we computed the standard error of measurement. Coefficients $\geq$0.70 were used to indicate acceptable reliability [22]. The smallest detectable change (SDC) was computed at the individual [$1.96*\sqrt{(2)}*SEM$] and group [$SDC_{ind}/\sqrt{n}$] level. The Standard Error of the mean (SEM) for SDC calculations was determined as follows: (SD_{T1}) + (SD_{T2})/2)*$\sqrt{(1-ICC)}$ [22].
Construct validity	To determine how well the scale accurately measures satisfaction with facial appearance. We tested 14 predetermined hypothesis-based construct validity tests for the initial surveys using group differences. ANOVA was used to test for differences, with effect sizes calculated as Eta-squared (η^2). Categories with small sample sizes were merged, including the photo-numeric Merz Assessment Scales' severe/very severe categories [14], and the Allergan Temple Hollowing scale moderate/severe categories [15].
Responsiveness	To determine if the scale can measure change. We tested 10 predetermined hypothesis-based construct validity tests of the change score using ANOVA. The Mean Differences and 95% confidence limits were calculated.

Interpretation	To interpret the change scores, we computed minimally important differences using anchor and distribution-based methods. For change scores, Likert questions and the Merz Assessment scales scores [14] were categorized into 3 groups to show the direction of the change (e.g., less, same, more). Change scores for the 2 SKIN-Q items [13], which have 4 response options, could range from -3 to +3.

Supplemental Table 4: Detailed results for change score hypotheses

	N	Mean Diff	SD	SE	95% CI		p-value
					LB	UB	
1. Amount of excess skin on face compared with time 1							
Better	77	2	17	2	-2.4	5.5	0.012
Same	331	1	13	1	-0.7	2.0	
Worse	82	-4	16	2	-7.9	-0.7	
2. Happy with how face looks compared with time 1							
Less happy	100	-7	12	1	-9.6	-4.7	<0.001
Same	289	0	13	1	-2.0	1.1	
Happier	101	8	15	1	5.3	11.2	
3. Perceived age versus actual age compared with time 1 [13]							
Younger	136	4	15	1	1.9	6.9	<0.001
Same	165	1	13	1	-1.3	2.6	
Older	189	-4	15	1	-5.9	-1.7	
4. Amount of skin elasticity compared with time 1 [14]							
Less satisfied	107	-8	14	1	-10.4	-4.9	<0.001
Same	262	-1	11	1	-2.0	0.8	
More satisfied	116	8	15	1	5.4	11.1	
5. Overall skin quality compared with time 1 [14]							
Less satisfied	110	-8	15	1	-11.0	-5.4	<0.001
Same	251	0	11	1	-1.7	1.0	
More satisfied	124	8	15	1	5.2	10.5	
6. Anchor – how the face looks compared with time 1							
Worse	63	-8	14	2	-11.2	-4.2	<0.001
Same	376	0	14	1	-1.3	1.5	
Better	51	9	13	2	5.0	12.1	
7. Perceived Ozempic face in those who lost weight compared with time 1							
More Ozempic face	42	-1	11	2	-4.7	1.9	0.057
Same	104	3	15	1	0.4	6.1	
Less Ozempic face	23	-3	13	3	-8.1	2.9	
8. Bothered by Ozempic face in those who lost weight compared with time 1							
More bothered	12						IS
Same	20						
Less bothered	5						
9. Saggy skin on the jawline compared with time 1 [15]							
Worse	69	-6	17	2	-10.0	-2.0	<0.001
Same	374	1	13	1	-0.8	1.9	
Better	45	4	13	2	0.7	8.3	
10. Hollowness under the eyes compared with time 1 [15]							
Worse	64	-7	15	2	-10.6	-3.2	<0.001
Same	361	1	14	1	-0.8	2.0	
Better	64	3	14	2	-0.9	6.3	