Journal of Plastic, Reconstructive & Aesthetic Surgery xxx (xxxx) xxx

JPRAS
An International Journal of Surgical Reconstruction
www.JPRASurg.com

FACE-Q craniofacial module: Part 2 Psychometric properties of newly developed scales for children and young adults with facial conditions

Anne F Klassen[a,*], Charlene Rae[a], Wong Riff[b], Rafael Denadai[c], Dylan J Murray[d], Shirley Bracken[e], Douglas J Courtemanche[f], Neil Bulstrode[g], Justine O'Hara[g], Daniel Butler[g], Jesse Goldstein[h], Ali Tassi[i], Marinka LF Hol[j], David Johnson[k], Ingrid M. Ganske[l], Lars Kölby[m], Susana Benitez[n], Eleonore E Breuning[o], Claudia C. Malic[p], Gregory C. Allen[q], Andrea L Pusic[r], Stefan Cano[s]

[a] *Department of Pediatrics, McMaster University, Hamilton, ON, Canada*
[b] *Department of Surgery, Hospital for Sick Children, ON Canada*
[c] *Institute of Plastic and Craniofacial Surgery, SOBRAPAR Hospital, Campinas, Sao Paulo, Brazil*
[d] *National Paediatric Craniofacial Centre, Children's Health Ireland at Temple Street, Dublin, Ireland*
[e] *National Paediatric Craniofacial Centre, Children's Health Ireland at Temple Street, Dublin, Ireland*
[f] *Division of Plastic Surgery, BC Children's Hospital Vancouver, BC, Canada*
[g] *Department of Plastic and Reconstructive Surgery, Great Ormond Street Hospital, London, United Kingdom*
[h] *Department of Plastic Surgery, Children's Hospital of Pittsburgh, Pittsburgh, PA United States*
[i] *Division of Graduate Orthodontics, Schulich School of Medicine and Dentistry, Western University, London ON, Canada*
[j] *Department of Otolaryngology and Head and Neck Surgery, University Medical Center Utrecht, Utrecht, The Netherlands AND Princess Maxima Center for Childhood oncology, Utrecht, Netherlands*
[k] *Oxford Craniofacial Unit, Oxford University Hospitals NHS Foundation Trust, Oxford, United Kingdom*
[l] *Department of Plastic and Oral Surgery, Boston Children's Hospital, Boston, MA United States*

[*] Corresponding author.
E-mail addresses: aklass@mcmaster.ca (A.F. Klassen), crae@mcmaster.ca (C. Rae), karenw.wong@sickkids.ca (W. Riff), dylanmurray@plasticsurgeon.ie (D.J. Murray), research@craniofacial.ie (S. Bracken), douglas.courtemanche@ubc.ca (D.J. Courtemanche), Neil.Bulstrode@gosh.nhs.uk (N. Bulstrode), drohara@drjustineohara.com.au (J. O'Hara), dan.butler@doctors.org.uk (D. Butler), jesse.goldstein@chp.edu (J. Goldstein), ali.tassi@schulich.uwo.ca (A. Tassi), M.l.f.hol-12@umcutrecht.nl (M.L. Hol), david.johnson@ouh.nhs.uk (D. Johnson), Ingrid.ganske@childrens.harvard.edu (I.M. Ganske), lars.kolby@surgery.gu.se (L. Kölby), Elly.Breuning@alderhey.nhs.uk (E.E. Breuning), CMalic@cheo.on.ca (C.C. Malic), Gregory.allen@childrenscolorado.org (G.C. Allen), apusic@bwh.harvard.edu (A.L. Pusic), stefan.cano@modusoutcomes.com (S. Cano).

https://doi.org/10.1016/j.bjps.2021.03.009

Please cite this article as: A.F. Klassen, C. Rae, W. Riff et al., FACE-Q craniofacial module: Part 2 Psychometric properties of newly developed scales for children and young adults with facial conditions, Journal of Plastic, Reconstructive & Aesthetic Surgery, https://doi.org/10.1016/j.bjps.2021.03.009

[m] *University of Gothenburg, The Sahlgrenska Academy, Institute of Clinical Sciences, Department of Plastic Surgery, Sahlgrenska, University Hospital, Gothenburg, Sweden*
[n] *Department of Plastic Surgery, Clinica Las Condes, Santiago, Chile*
[o] *Department of Plastic Surgery, Alder Hey Children's Hospital, Liverpool, United Kingdom*
[p] *University of Ottawa, Department of Surgery, Children's Hospital of Eastern Ontario, Ottawa, ON, Canada*
[q] *Department of Otolaryngology – Head & Neck Surgery, University of Colorado School of Medicine, Aurora, CO, United States*
[r] *Division of Plastic and Reconstructive Surgery, Brigham and Women's Hospital, Boston, MA, United States*
[s] *Modus Outcomes, Letchworth Garden City, United Kingdom*

Received 29 July 2020; accepted 11 March 2021
Available online xxx

KEYWORDS
FACE-Q;
CLEFT-Q;
Birthmark;
Craniofacial;
Quality of Life;
Appearance;
Patient reported outcome measure;
PROM;
Psychometrics

Summary *Background:* The FACE-Q Craniofacial Module is a patient-reported outcome measure designed for patients aged 8 to 29 years with conditions associated with a facial difference. In part 1, we describe the psychometric findings for the *original* CLEFT-Q scales tested in patients with cleft and noncleft facial conditions. The aim of this study was to examine psychometric performance of *new* FACE-Q Craniofacial Module scales.
Methods: Data were collected between December 2016 and December 2019 from patients aged 8 to 29 years with conditions associated with a visible or functional facial difference. Rasch measurement theory (RMT) analysis was used to examine psychometric properties of each scale. Scores were transformed from 0 (worst) to 100 (best) for tests of construct validity.
Results: 1495 participants were recruited with a broad range of conditions (e.g., birthmarks, facial paralysis, craniosynostosis, craniofacial microsomia, etc.) RMT analysis resulted in the refinement of 7 appearance scales (Birthmark, Cheeks, Chin, Eyes, Forehead, Head Shape, Smile), two function scales (Breathing, Facial), and an Appearance Distress scale. Person separation index and Cronbach alpha values met criteria. Three checklists were also formed (Eye Function, and Eye and Face Adverse Effects). Significantly lower scores on eight of nine scales were reported by participants whose appearance or functional difference was rated as a major rather than minor or no difference. Higher appearance distress correlated with lower appearance scale scores.
Conclusion: The FACE-Q Craniofacial Module scales can be used to collect and compare patient reported outcomes data in children and young adults with a facial condition.

Introduction

Patient-reported outcome measures (PROM) used in research with children and young adults with conditions associated with a facial difference lack content validity in terms of appearance and facial function.[1-2] A new PROM for such patients is needed to inform clinical care and to include the patient perspective in research efforts. Our team previously created the CLEFT-Q to address the most common craniofacial anomaly.[3] The CLEFT-Q was developed and refined using qualitative methods[4-5] and field-tested internationally with 2434 patients from 12 countries.[6] The CLEFT-Q includes an Eating/Drinking checklist and 12 scales designed to measure appearance (of the face, nose, nostrils, teeth, lips, jaws and cleft lip scar), health-related quality of life (psychological, school and social function and speech distress), and speech function.

After developing CLEFT-Q, in order to address noncleft craniofacial conditions, we interviewed 84 patients aged 8 to 29 years with 28 different congenital and acquired conditions (e.g., microtia, facial paralysis, craniosynostosis, craniofacial microsomia and birthmarks).[7] This qualitative study provided the evidence to support the use of CLEFT-Q scales with patients with noncleft craniofacial conditions. The qualitative study also identified the need for additional scales to measure constructs not covered by the CLEFT-Q.

Table 1 FACE-Q Craniofacial Module for children and young adults.

Appearance		Function	Health-Related Quality of Life	Adverse Effects
Birthmark*	Head Shape*	Breathing*	Appearance Distress*	Ears+
Cheeks*	Jaws‡	Eating/ Drinking‡	Psychological‡	Eye*
Chin*	Lips‡	Eye*	Social‡	Face*
Ears+	Nose‡	Facial*	School‡	
Eyes*	Nostrils‡	Speech‡	Speech Distress	
Face‡	Teeth‡			
Forehead*	Smile*			

* FACE-Q scales described in this paper.
‡ Scales originally part of CLEFT-Q; + Scales part of EAR-Q.

Our team used the qualitative data to design new scales measuring additional aspects of appearance, facial function and health-related quality of life not captured in the CLEFT-Q. The full set of scales that form the FACE-Q Craniofacial Module are shown in Table 1.

The psychometric findings for the scales that form the FACE-Q are published in separate papers. Elsewhere, we describe 2 scales developed for patients with a variety of ear conditions (i.e., EAR-Q).[8] In this journal, we have also published Part 1 that describes the findings for the validation of a set of CLEFT-Q scales/checklist used in patients with non-cleft facial conditions.[9] The aim of this paper (Part 2) is to describe the reliability and validity findings for 13 new FACE-Q scales tested in patients with a broad range of craniofacial conditions.

Methods

We obtained ethics board approval for the study coordinating site (Hamilton Integrated Research Ethics Board) and from the ethics board at each participating site prior to starting the study. Written and informed assent and/or consent was obtained from the study participants and their guardians.

Data collection

The psychometric analysis included data from two studies as follows

FACE-Q field-test study

Data were collected from patients aged 8 to 29 years with a wide range of craniofacial conditions as part of the FACE-Q Craniofacial Module field-test study. Participants included anyone with a congenital or acquired visible facial and/or facial function difference. Participants who could not complete the scales independently were excluded. For the Birthmark scale, recruitment included patients aged 8 to 29 years with birthmarks anywhere on the face or body. Data for these participants without a facial difference were only used in the validation of the Birthmark scale.

Patients were recruited from hospital clinics and social media sites. In the hospital clinics, data collection took place face-to-face during clinic visits using electronic (tablets) or paper-and pencil (booklets) means depending on each site's preference. Data collection took place between December 2016 and December 2019 at 24 sites in nine countries. We also recruited through social media sites (i.e., Microtia UK, US Moebius Syndrome Foundation, Bell's Palsy and Facial Paralysis Foundation, and Facial Palsy UK). Members were sent study recruitment materials and invited to complete the survey online.

A clinical form was used by site recruiters. The form comprised of a matrix that listed the facial areas (e.g., jaw, lips, nose) and functional concerns (e.g., eating, speaking) related to each FACE-Q scale, by the severity (no, yes-minor, yes-major) of each appearance or functional concern. Additional questions asked child's age and diagnoses, and whether the child had facial surgery in the past six months. The form was used to ensure participants completed only relevant scales. For example, the Cheeks scale was completed by participants with a minor or major difference in cheek appearance, and/or patients with specific diagnoses (i.e., Craniofacial Microsomia, and Syndromic Craniosynostosis). All data were collected and managed using the secure REDCap® electronic data capture tools[10-11] hosted at McMaster University (Canada).

Pediatric head and neck cancer study

FACE-Q Craniofacial Module scales were included in an international follow-up study of pediatric head and neck cancer. Participants, now aged 8 to 29 years, were aged 0 to 18 years and treated with chemotherapy, and local therapy consisting of surgery and/or radiotherapy for a head and neck tumor. This study collected data with questionnaire booklets during outpatient clinics held in the Netherlands, France, the United Kingdom, and United States. Participants were invited to complete a range of FACE-Q scales. Data were entered into the REDCap® database hosted at McMaster University (Canada).

Statistical analysis

Data were analyzed using SPSS® version 26.0 (IBM Corporation, Armonk NY, USA for Windows®/Apple Mac®) and RUMM2030 software (RUMM version 2030, RUMM Laboratory Pty Ltd., Duncraig, Western Australia, 1998-14). To examine reliability and validity, Rasch Measurement Theory (RMT) analysis was performed.[12-13] Specifically, a set of statistical

Table 2 Characteristics (Number,%) for the 1495 participants.

	N	%
Country		
Australia	38	2.5
Brazil	178	11.9
Canada	828	55.4
Chile	7	0.5
France	6	0.4
Ireland	113	7.6
Sweden	13	0.9
United Kingdom	185	12.4
United States	126	8.4
Other	1	0.1
Language		
English	1290	86.3
French	6	0.4
Portuguese	178	11.9
Spanish	7	0.5
Swedish	13	0.9
Age in years		
8-10	335	22.4
11-13	355	23.7
14-17	429	28.7
18-29	376	25.2
Gender		
Male	655	43.8
Female	835	55.9
Other	4	0.3
Missing	1	0.1
Main Condition[*]	**N**	**%**
BIRTHMARK		
Congenital melanocytic naevus	44	2.9
Haemangioma	73	4.9
Sebaceous naevus	18	1.2
Vascular malformation	142	9.5
Birthmark other	4	0.3
EAR CONDITION		
Microtia	45	3.0
Prominent ears	37	2.5
Ear other	10	0.7
SKELETAL		
Acquired Skeletal	55	3.7
Craniofacial microsomia	78	5.2
Craniofrontonasal condition	27	1.8
Craniosynostosis non-syndromic	175	11.7
Craniosynostosis syndromic	111	7.4
Fibrous dysplasia	30	2
Mandibular condition	39	2.6
Multiple bony anomalies	19	1.3
Post-traumatic bony defect	42	2.8
Other congenital skeletal	21	1.4
SOFT TISSUE		
Acquired soft tissue	30	2
Congenital soft tissue	15	1
Neurofibromatosis type 1	31	2.1
Parry-Romberg Syndrome	44	2.9
Soft tissue other	15	1
TRAUMA		
Bite	10	0.7
Fracture	71	4.7
Laceration	12	0.8
Burn	20	1.3
Trauma other	24	1.6
OTHER		
Cancer	18	1.1
Facial paralysis	61	4.1
Other syndrome	21	1.4
Orthodontic	153	10.2

[*] Condition listed represents the main diagnosis, classifications may have varied by site. 14.7% of participants had multiple conditions.

and graphical tests were conducted to examine whether the observed data fit the Rasch model for each scale.[12-14] The following tests were performed:

Item fit: To determine if the items of each scale worked together clinically and statistically, item fit was examined. We examined item response options to determine if the item thresholds were properly ordered.[15] We also examined graphical (item characteristic curves) and statistical (log residuals (item-person interaction) and Chi-square values (item-trait interaction)) indicators of item fit. Ideal fit residuals fall between -2.5 and $+2.5$ with Chi-square values that are nonsignificant after Bonferroni adjustment.[13] For the Appearance Distress scale, due to the large sample size, we amended the analysis to 500 for tests of fit statistics.[14]

Targeting: Scales should be designed such that they have a set of items that provide information for all levels of the concept as experienced by the sample.[14] We examined the items in each scale to determine their spread and whether that matched the range of the construct reported by the sample. Scales were examined graphically (person-item threshold distribution) and statistically (proportion of the sample to score outside the range of each scale's measurement).

Differential Item Function (DIF): We examined DIF for age, gender, and language (English versus other). DIF was computed for any scale when there were 150 or more participants per subgroup (to allow for 50 participants in each of three class intervals). Based on sample size, we were able to examine gender, language and age for four subgroups (8-10, 11-13, 14-17, 18-29 years) for Appearance Distress. For the remaining scales, we were able to examine gender and age for two subgroups (8-12, 13-17 years). DIF analysis was repeated three times, each time selecting a random sample to ensure the subgroups were of equal size. Since the analysis for Appearance Distress included a large sample size, we computed DIF with and without adjusting the sample size to 500. Items with significant chi-square p-values after Bonferroni adjustments were split on the sample characteristic that evidenced DIF, and the new and original person locations were correlated (Spearman Correlation) to determine the impact of DIF on scoring.[13]

Table 3 Number (%) of participant to report each eye and facial problem.

	Very much		Quite a bit		A little bit		Not at all		Missing	
	n	%	n	%	n	%	n	%	n	%
EYE FUNCTION										
…eyelids close unexpectedly	4	2.3	2	1.1	15	8.6	153	87.4	1	0.6
…opening eyelids	9	5.1	5	2.9	19	10.9	141	80.6	1	0.6
…blinking eyes	13	7.4	10	5.7	14	8.0	136	77.1	2	1.1
…seeing properly	13	7.4	7	4.0	32	18.3	122	69.7	1	0.6
…closing eyelids	22	12.6	10	5.7	14	8.0	127	72.6	2	1.1
…eyelids closed when asleep	20	11.4	13	7.4	21	12.0	117	66.9	4	2.3
…one eye works better	46	26.3	12	6.9	30	17.1	86	49.1	1	0.6
EYE ADVERSE EFFECTS										
…eyelids twitch	1	0.6	6	3.4	40	22.9	127	72.6	1	0.6
…eyes are sore (hurt)	1	0.6	10	5.7	45	25.7	119	68.0	0	0
…eyes are itchy	3	1.7	6	3.4	47	26.9	119	68.0	0	0
…whites of eyes are red	5	2.9	10	5.7	32	18.3	127	72.6	1	0.6
…something in eye(s)	6	3.4	15	8.6	33	18.9	121	69.1	0	0
…eyes water too much	7	4.0	17	9.7	47	26.9	103	58.9	1	0.6
…eyes are dry	14	8.0	13	7.4	31	17.7	117	66.9	0	0
FACE ADVERSE EFFECTS										
…face is bruised	2	1.2	6	3.6	25	15.0	134	80.2	0	0
…face feels sore	2	1.2	11	6.6	33	19.8	121	72.5	0	0
…face feels tingly	3	1.8	10	6.0	23	13.8	130	77.8	1	0.6
…face feels sensitive	3	1.8	13	7.8	38	22.8	112	67.1	1	0.6
…face feels itchy	3	1.8	14	8.4	33	19.8	116	69.5	1	0.6
…face feels numb	8	4.8	7	4.2	24	14.4	127	76.0	1	0.6
…face is puffy or swollen	8	4.8	12	7.2	27	16.2	119	71.3	1	0.6
…face feels uncomfortable	9	5.4	11	6.6	34	20.4	112	67.1	1	0.6
…face feels tight	8	4.8	13	7.8	29	17.4	116	69.5	1	0.6
…face feels firm	7	4.2	19	11.4	31	18.6	108	64.7	2	1.2

Reliability: Scale reliability was examined by computing Person Separation Index (PSI) and Cronbach alpha.[16] Reliability coefficients greater than or equal to 0.70 were considered adequate.[17] To determine whether items were influenced by responses to other items in a scale (which can artificially inflate reliability), we identified residual correlations between items over 0.20 and performed a subtest to measure their impact on the PSI value.[15]

To examine construct validity, we transformed the Rasch logit scores into 0 (worse) to 100 (best) to test specific hypothesis. P-values less than 0.05 were considered significant. Normality was assessed using Kurtosis (absolute >2) and Skewness (absolute >2),[18] and non-parametric statistics were applied if distributions were non-normal. First, we hypothesised that FACE-Q scale scores would be lower in patients with a major versus a minor or no difference in appearance and function. Second, based on published findings that most CLEFT-Q scale scores were lower for older patients, and some scales scores were lower for female gender,[6] we hypothesised that FACE-Q scale scores would also be lower in both older patients and female patients, and further that lower scores on the Appearance Distress scale would moderately correlate with lower scores on the appearance scales. Finally, we hypothesised that scale scores would correlate more strongly within their domain (e.g., appearance) than with scales in other domains. Correlation coefficients were interpreted as follows: <0.3 negligible, 0.30 to 0.49 low, 0.50 to 0.69 moderate, 0.70 to 0.89 high, 0.9 to 1.00 very high.[19]

Results

Table 2 shows characteristics for the 1495 participants who provided a total of 1509 assessments. Participants with a range of facial conditions were recruited. Of the 271 participants with a birthmark, 60 had the birthmark on their body and no facial condition. These participants were only included in the RMT analysis for the Birthmark scale.

RMT analysis provided evidence of reliability and validity for 10 of the 13 scales tested in this study. The three scales that did not work psychometrically were Eye Function, Eye Adverse Effects and Face Adverse Effects. Each scale had one or more items with disordered thresholds. After we rescored each scale's items across their two middle response options, and deleted seven items deemed redundant, the item fit statistics in the three scales were acceptable, but scale reliability was low in terms of the PSI values. Table 3 shows the three sets of items used as problem checklist.

The number of items tested across the remaining ten scales was reduced by 32 to 85 items. Thresholds were dis-

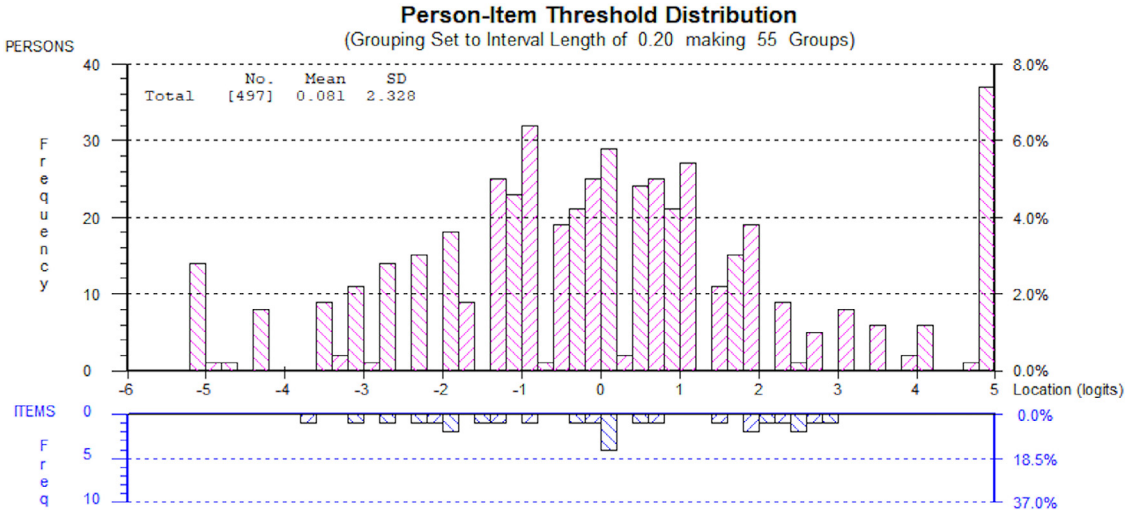

Figure 1 Person-item threshold distributions as examples of targeting for the Smile scale.

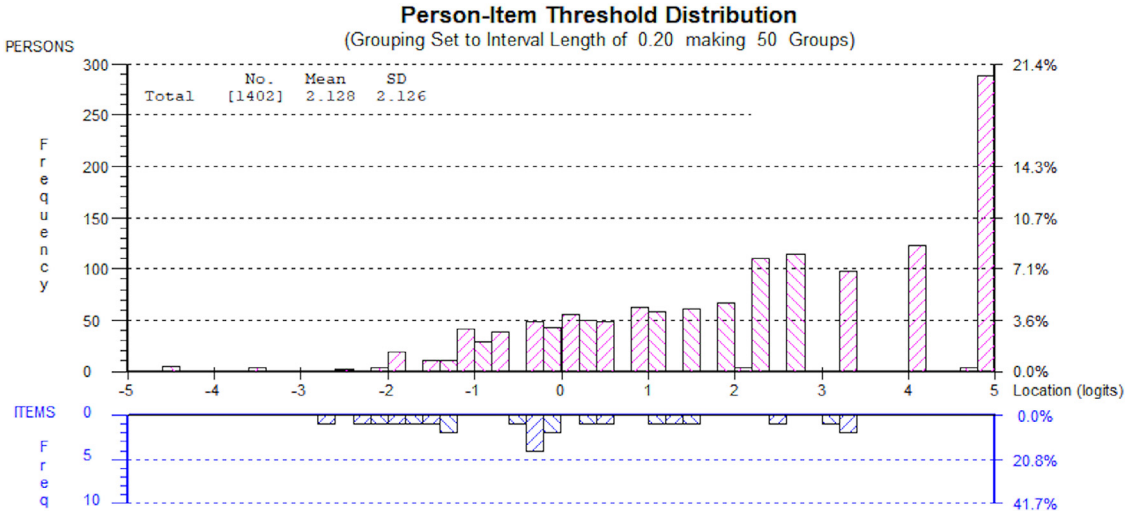

Figure 2 Person-item threshold distributions as examples of targeting for the Appearance Distress scale.

ordered for 9 of 10 items in the Facial Function scale. When items for this scale were rescored across the two middle options, all had ordered thresholds. The RMT analysis proceeded for this scale using the rescored data. All 85 items had nonsignificant Chi-square p values after Bonferroni adjustment (see **Appendix 1**). The item fit was within +2.5 for 74 items.

Figures 1-3 shows the distribution of person measurement and item location for an appearance (Smile), health-related quality of life (Appearance Distress) and function (Breathing) scale to illustrate targeting. The proportion of the sample to score within the range of each scale's measurement was 88.9% (Smile), 92.7% (Breathing) and 78.9% (Appearance Distress). Participants who scored outside the range (to the right in each figure) were participants with high scores on each scale indicating better outcomes.

Based on the sample sizes in subgroups, we were able to examine DIF for age, gender and language for the Appearance Distress scale, and by age for four appearance scales (Eyes, Forehead, Head Shape, Smile) and gender for five appearance scales (Cheeks, Eyes, Forehead, Head Shape, Smile). For the Appearance Distress scale, the unadjusted analysis DIF was evident for one item for gender (people stare), three items for age-group (self-conscious, people stare, unhappy) and four items by language (self-conscious, people stare, mirror, going out). In the adjusted analysis, there was evidence of DIF in one item by age (self-conscious). For the five appearance scales where DIF was examined, one item (match) in the Head Shape scale evidenced DIF by age. All items that evidenced DIF in the unadjusted analysis, had very negligible impact on the scoring when the items with DIF were split and person locations correlated (all $\geq$0.99).

Data from the sample fit the Rasch model with nonsignificant p-values for six scales (see Table 4). For the remaining four scales, the p-values showed slight misfit to the Rasch model. For the health-related quality of life and appearance scales, reliability was high with PSI values $\geq$0.83 with and without extremes, and Cronbach alpha values $\geq$0.87 with and without extremes. Reliability for the two func-

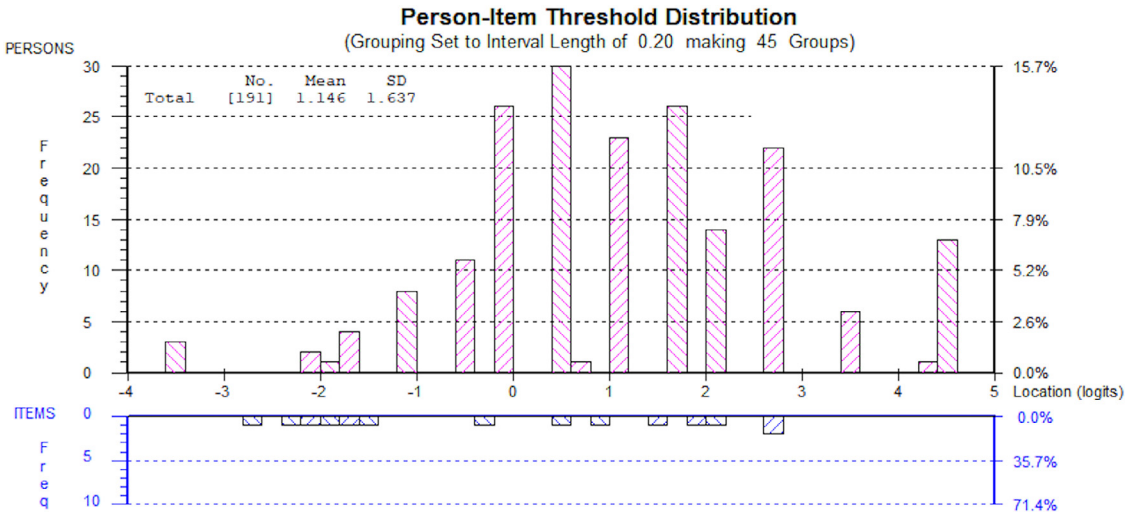

Figure 3 Person-item threshold distributions as examples of targeting for the Breathing scale.

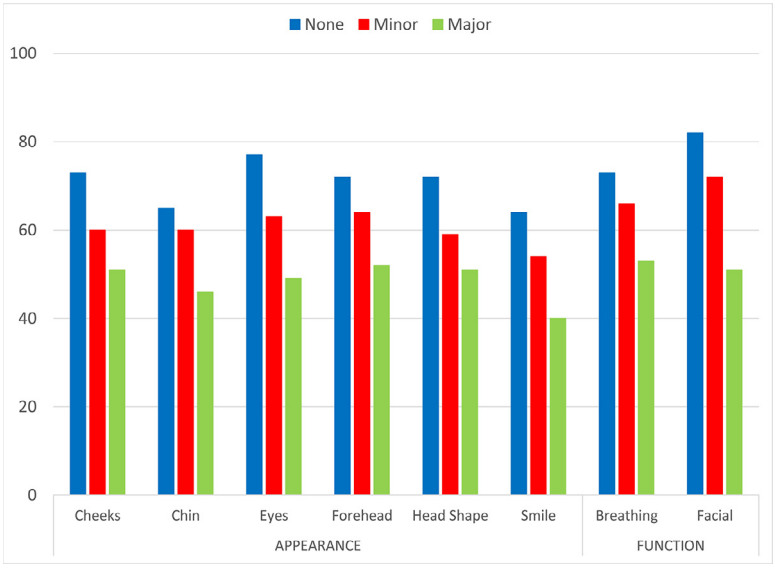

Figure 4 Mean score for each FACE-Q Craniofacial Module scale by severity of appearance or functional difference. Significant association between severity and scale score for 8 scales ($p \leq 0.001$);
Footnote- *Post hoc no significant differences none vs minor ($p \geq 0.184$) for Chin, Breathing and Facial Function; none vs major for Chin ($p = 0.085$) - sample size for chin difference 'none' category $n = 12$.

tion scales was lower, with PSI values ≥0.71 with and ≥0.69 without extremes, and Cronbach alpha values >0.80 with and >0.74 without extremes, respectively. Residuals in one or more item pairs in seven scales were correlated above 0.20. The impact of these correlations on the PSI values for five scales (Appearance Distress, Chin, Cheeks, Eyes, Smile) represented a drop in PSI value of ≥0.01. For the remaining two scales the drop in PSI was larger at 0.05 (Birthmark) and 0.09 (Facial Function).

Based on Skewness and Kurtosis values, all data were normally distributed and parametric statistics were applied. Figure 4 shows the mean score on each FACE-Q scale by the severity rating. The hypothesis that participants with a major difference in appearance and function would score lower on FACE-Q scales was supported. Differences between group means was significant for all scales ($p \leq 0.001$ on ANOVA).

Females reported lower scores on independent samples t-tests for the Appearance Distress (mean diff=4.3; SE 1.1; $p < 0.001$), Eyes (mean diff=6.2; SE=2.5 $p = 0.013$) and Smile (mean diff=4.3; SE=2.1; $p = 0.041$) scales. Differences between the three age groups (8-13 yrs;14-19 yrs; 20-29 yrs) were observed for all scales ($p < 0.001$ on ANOVA), except for the following scales: Birthmark ($p = 0.270$), Breathing ($p = 0.523$) and Facial Function ($p = 0.059$) (See Figure 5). Appearance Distress correlated moderately with the scores for appearances based on Pearson's correlation co-efficient (see Table 5). Finally, correlations between scales within domains were stronger, as hypothesised, than with other domains.

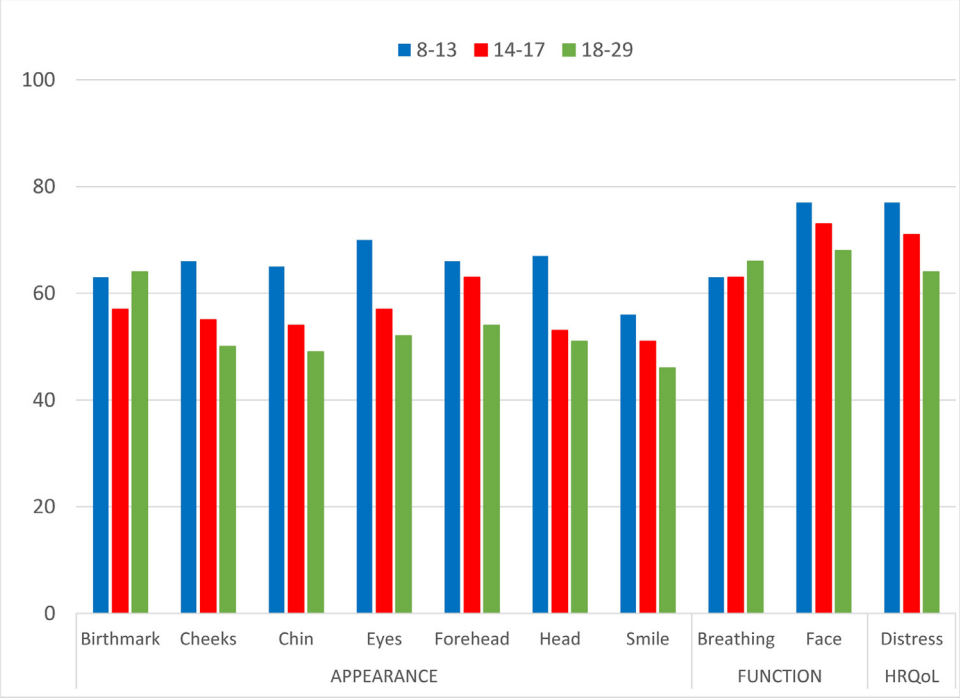

Figure 5 Mean score for each FACE-Q Craniofacial Module scale by age group. Significant association between age group and scale score for 7 scales ($p \leq 0.001$); No association between age group and the following scale scores: Birthmark ($p = 0.217$), Breathing ($p = 0.523$) and Facial Function ($p = 0.059$).
Footnote -Post hoc tests showed no significant difference between 8-13 and 14-17 ($p \geq 0.223$) for Forehead and Smile; and between 14-17 and 18-29 ($p \geq 0.089$) for Cheek, Chin, Eyes, Head Shape, and Smile.

Discussion

Surgical treatments for conditions associated with a facial difference are often complex and burdensome for patients. Outcome measures used to evaluate operations that aim to change how someone looks and/or their facial function should measure the patient perspective given the subjective nature of such outcomes. Our research here and elsewhere[7-9,20,21] shows that the FACE-Q Craniofacial Module for provides reliable and valid measurement of outcomes that matter to children and young adults with a broad range of facial conditions. The use of a modern psychometric approach (RMT analysis) made it possible to identify any problems within each scale. We dropped some items and rescored some response options after which the psychometric findings provided evidence of reliability and validity for ten scales. Each scale measured a clinical hierarchy for their concepts and worked as hypothesised, with lower scores associated with older age, female gender and having a major facial difference.

The Eye Function, Eye Adverse Effects, and Face Adverse Effects represented exceptions. While the Rasch approach aims to develop scales that measure unidimensional constructs via a set of items that map out a clinical hierarchy, contrary to our hypotheses, these three sets of items did not work together statistically. We reported a similar finding in the CLEFT-Q field-test, whereby Eating and Drinking did not function like a scale.[6] Although Eye Function, Eye Adverse Effects and Face Adverse Effects had acceptable Cronbach alpha values, we recommend their use as problem checklists since the overall findings do not support the summing of items to form scale scores. Even though the three checklists do not have a Rasch-based scoring algorithm, they can provide clinically important information, such as monitoring for post-operative complications.

Recent reviews have drawn attention to the challenge of assessing appearance and body image in patients with craniofacial conditions. Research has shown that patients generally having positive scores for satisfaction with appearance, and that dissatisfaction is generally associated only with the impacted facial area.[22,23] The FACE-Q Craniofacial Module addresses this issue by having feature specific appearance scales (e.g., eyes, nose lips). These specific scales can be used in conjunction with the Face scale to capture overall appearance as well as the facial features that are of most concern to the patient.

The uptake and use of PROMs are rapidly expanding around the world. PROMs provide a means to measure the burden of a condition and the impact of treatments provided to patients. Previously we reported findings about the impact of completing the CLEFT-Q from 2056 children and young adults. Specifically, the majority of participants reported that they liked completing the CLEFT-Q, most liked the questions about how they look (82%), and most felt the same or better about how they look after completing the CLEFT-Q (67%).[24] A small minority of participants reported that they felt worse about how they look after completion. These findings suggest that patients who complete the FACE-Q Craniofacial Module may have different experiences both positive and negative. Therefore, to

Table 4 Rasch Measurement Theory scale level statistics.

Scale	# items tested	# items retained	Full sample	Sample in RMT analysis	% scored on scale	Chi-square	DF	p-value	PSI +ext	PSI -ext	Cronbach alpha +ext	Cronbach alpha -ext
Appearance Distress	10	8	1402	1106	78.9	76.16	64	0.14	0.83	0.84	0.93	0.89
Birthmark	14	8	271	204	75.3	16.71	16	0.40	0.87	0.85	0.95	0.89
Chin	12	9	258	208	80.6	16.68	18	0.55	0.93	0.91	0.97	0.93
Cheeks	10	9	396	305	77.0	44.45	36	0.16	0.93	0.91	0.97	0.93
Eyes	14	9	468	351	75.0	53.78	36	0.03	0.89	0.89	0.96	0.91
Forehead	15	10	554	465	83.9	70.35	60	0.17	0.89	0.88	0.94	0.91
Head Shape	8	6	427	341	79.9	40.30	24	0.02	0.88	0.84	0.93	0.87
Smile	15	9	497	442	88.9	70.62	45	0.01	0.91	0.89	0.94	0.91
Breathing	7	7	191	177	92.7	17.90	14	0.21	0.74	0.69	0.80	0.74
Facial Function	12	10	132	109	82.6	38.36	20	0.01	0.71	0.72	0.89	0.85

DF - Degrees of freedom; PSI - Person Separation Index; ext - extremes.

minimize the negative impact of completing a PROM, it is important that researchers and cliniciansthoughtfully select which outcome tools to use. While the FACE-Q Craniofacial Module may appear long, no patient needs to complete all the scales. Healthcare professionals and researchers can pick-and-choose from the full set of independently functioning scales the subset best suited to address their specific questions or clinical need. To facilitate benchmarking, five of the FACE-Q scales are applicable to any patient with a facial condition, i.e., Face, Appearance Distress, Psychological, Social and School. The remaining scales are specific to facial area or specific facial functions and would be more useful in the evaluation of specific treatment outcomes.

Our study has several limitations. First, the sample accrued for the Facial Function scale was slightly less than 150. Rasch analysis uses Chi-square where a sample of 150 provides 50 participants in each of three class intervals for tests of item fit to the Rasch model. We did not collect information about the number of patients that the recruitment staff might have missed, nor about characteristics of patients who refused to participate. The severity ratings of major and minor difference in appearance and facial function were based on the judgement of the recruiter. The sample included a small number of participants with birthmarks who did not have a facial difference. However, data for these participants were excluded from the RMT analysis for any other scales to ensure that only patients with a facial difference were included. COSMIN criteria[25] for psychometric properties of PROMs includes tests that we did not perform in our study due to the length of the field-test questionnaire. These tests, which include test-retest reliability, responsiveness, and correlation with other PROMs, can be examined in future studies.

Conclusion

In order to improve care provided to patients with conditions associated with a facial difference, highly specific, carefully designed PROMs are needed. The FACE-Q Craniofacial Module provides healthcare professionals and researchers with a set of tools to measure the patient perspective of outcomes associated with craniofacial care for anyone aged 8 to 29 years.

Declaration of Competing Interest

Anne Klassen and Karen Wong are co-developers of the patient-reported outcome scales described in this publication and share in any license revenues as royalties based on their institutions' inventor sharing policy for their use in for-profit study. The other authors have no conflict of interest to declare in relation to this work.

Financial disclosure

The research described in this paper was supported by a grant from the Canadian Institute of Health Research (FRN

Table 5 Correlations between scales.

Scale	HRQOL	APPEARANCE							FUNCTION
	Distress	Birthmark	Cheeks	Chin	Eyes	Forehead	Head Shape	Smile	Breathing
Distress									
Birthmark	0.431**								
Cheeks	0.489**	0.228							
Chin	0.588**	0.392*	0.677**						
Eyes	0.556**	0.321	0.579**	0.535**					
Forehead	0.476**	0.442**	0.571**	0.588**	0.713**				
Head Shape	0.565**	0.519*	0.672**	0.591**	0.677**	0.734**			
Smile	0.500**	0.278	0.674**	0.503**	0.514**	0.491**	0.735**		
Breathing	0.370**	0.221	0.190*	0.413**	0.326**	0.315**	0.254**	0.218*	
Facial Function	0.379**	0.067	0.349**	0.305*	0.267*	0.330**	0.281	0.285**	0.382**

** Correlation is significant at the 0.01 level (2-tailed).;
* Correlation is significant at the 0.05 level (2-tailed). HRQOL - health-related quality of life.

148779). The authors have no financial interest to declare in relation to the content of this article. The Article Processing Charge was paid from the CIHR grant.

Acknowledgment

We are grateful for the operating grant we received from the Canadian Institutes for Health Research. We are also grateful to the many healthcare professionals and research staff in craniofacial sites around the world for their dedication and help with our research.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.bjps.2021.03.009.

References

1. Wickert NM, Riff KW, Mansour M, et al. Content validity of patient-reported outcome instruments used with pediatric patients with facial differences: a systematic review. *Cleft Palate Craniofac J* 2018;**55**(7):989–98.
2. Tapia VJ, Epstein S, Tolmach OS, Hassan AS, Chung NN, Gosman AA. Health-related quality-of-life instruments for pediatric patients with diverse facial deformities: a systematic literature review. *Plast Reconstr Surg* 2016;**138**(1):175–87.
3. Wong Riff KW, Tsangaris E, Goodacre T, et al. International multiphase mixed methods study protocol to develop a cross-cultural patient-reported outcome instrument for children and young adults with cleft lip and/or palate (CLEFT-Q). *BMJ Open* 2017;**7**(01):e015467.
4. Wong Riff KWY, Tsangaris E, Goodacre TEE. What matters to patients with cleft lip and/or palate: an international qualitative study informing the development of the CLEFT-Q. *Cleft Palate Craniofac J* 2018;**55**(3):442–50.
5. Tsangaris E, Wong Riff KWY, Goodacre T, et al. Establishing content validity of the CLEFT-Q: a new patient-reported outcome instrument for cleft lip/palate. *Plast Reconstr Surg Glob Open* 2017;**5**(04):e1305.
6. Klassen AF, Riff KWY, Longmire NM, et al. Psychometric findings and normative values for the CLEFT-Q based on 2434 children and young adult patients with cleft lip and/or palate from 12 countries. *CMAJ* 2018;**190**(15):E455–62.
7. Longmire NM, Wong Riff KWY, O'Hara JL, et al. Development of a new module of the FACE-Q for children and young adults with diverse conditions associated with visible and/or functional facial differences. *Facial Plast Surg* 2017;**33**:499–508.
8. Klassen A.F., Rae C., Bulstrode N.W., et al. An international study to develop the EAR-Q patient-reported outcome measure for children and young adults with ear conditions. *J Plast Reconstr Aesthet Surg*. 2021 Feb 5.
9. Klassen A.F., et al. FACE-Q craniofacial module: part 1 validation of CLEFT-Q scales for use in children and young adults with facial conditions. *J Plast Reconstr Aesthet Surg* [Submitted].
10. Harris PA, Taylor R, Thielke R, et al. Research electronic data capture (REDCap) — A metadata-driven methodology and workflow process for providing translational research informatics support. *J Biomed Inform* 2009;**42**:377–81.
11. Harris PA, Taylor R, Minor BL. The REDCap consortium: Building an international community of software partners. *J Biomed Inform* 2019;**95**:103208.
12. Rasch G. Probabilistic models for some intelligence and attainment tests. *Vol. 1 of Studies in Mathematical Psychology*. Copenhagen: Danmarks Paedagogiske Institut; 1960.
13. Andrich D. *Rasch Models for Measurement. Sage University Papers Series Quantitative Applications in the Social Sciences, Vol. 07-068*. Thousand Oaks (CA): Sage; 1988.
14. Hobart J, Cano S. Improving the evaluation of therapeutic intervention in MS: the role of new psychometric methods. *Health Technol Assess* 2009;**13**:1–177 iii, ix-x.
15. Wright BD MG. *Rating Scale Analysis*. MESA Press; 1982.
16. Cronbach LJ. Coefficient alpha and the internal structure of tests. *Psychometrika* 1951;**16**:297–334.
17. Nunnally JC. *Psychometric Theory*. 3rd Ed. New York, NY: McGraw-Hill; 1994.
18. Kim HY. Statistical notes for clinical researchers: assessing normal distribution (2) using skewness and kurtosis. *Restor Dent Endod* 2013;**38**(1):52–4.
19. Mukaka MM. A guide to appropriate use of correlation coefficient in medical research. *Malawi Med J* 2012;**24**(3):69–71.
20. Tassi A, Tan J, Piplani B, et al. Establishing content validity of an orthodontic subset of the FACE-Q Craniofacial Module in chil-

dren and young adults with malocclusion. *Orthod Craniofac Res* 2021 [EPub ahead of print].
21. Klassen AF, Rae C, Gallo L, et al. Psychometric Validation of the FACE-Q Craniofacial Module for Facial Nerve Paralysis. *Facial Plast Surg Aesthet Med* 2021.
22. Stock NM, Feragen KB. Psychological adjustment to cleft lip and/or palate: a narrative review of the literature. *Psychol Health* 2016;**31**(7):777–813.
23. Stock NM, Feragen KB. Comparing psychological adjustment across cleft and other craniofacial conditions: implications for outcome measurement and intervention. *Cleft Palate Craniofac J* 2019;**56**(6):766–72.
24. Klassen AF, Dalton L, Goodacre TEE, et al. Impact of completing a patient-reported outcome measure that asks about appearance: an international study to develop the CLEFT-Q. *Cleft Palate Craniofac J* 2020;**57**(7):840–8.
25. Prinsen CA, Mokkink LB, Bouter LM, et al. COSMIN guideline for systematic reviews of patient-reported outcome measures. *Qual Life Res* 2018;**27**(5):1147–57.