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

ELSEVIER

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

FACE-Q Craniofacial Module: Part 1 validation of CLEFT-Q scales for use in children and young adults with facial conditions

Anne F Klassen[a,*], Charlene Rae[a], Karen WY Wong Riff[b], Neil Bulstrode[c], Rafael Denadai[d], Jesse Goldstein[e], Marinka LF Hol[f], Dylan J Murray[g], Shirley Bracken[g], Douglas J Courtemanche[h], Justine O'Hara[c], Daniel Butler[c], Ali Tassi[i], Claudia C Malic[j], Ingrid M Ganske[k], Yun S Phua[l], Damian D Marucci[m], David Johnson[n], Marc C Swan[o], Eleonore E Breuning[p], Tim EE Goodacre[o], Andrea L Pusic[q], Stefan Cano[r]

[a] *Department of Pediatrics, McMaster University, Hamilton, ON, Canada*
[b] *Department of Surgery, Hospital for Sick Children, Toronto, Ontario, Canada*
[c] *Division of Plastic and Reconstructive Surgery, Great Ormond Street Hospital, London, UK*
[d] *Institute of Plastic and Craniofacial Surgery, SOBRAPAR Hospital, Campinas, Sao Paulo , Brazil*
[e] *Department of Plastic Surgery, Children's Hospital of Pittsburgh, Pittsburgh, PA, USA*
[f] *Department of Otolaryngocoloy and Head and Neck Surgery, University Medical Center Utrecht, Prinses Maxima Center for Childhood oncology, Utrecht, The Netherlands*
[g] *National Paediatric Craniofacial Centre, Children's Health Ireland at Temple Street, Dublin, Ireland*
[h] *Division of Plastic Surgery, BC Children's Hospital, Vancouver, BC, Canada*
[i] *Division of Graduate Orthodontics, Schulich School of Medicine and Dentistry, Western University, London, ON, Canada*
[j] *Department of Surgery, Children's Hospital of Eastern Ontario (CHEO), Ottawa, ON, Canada*

Meetings: N/A

* Corresponding author: Department of Pediatrics, McMaster University, Room HSC 3N27, 1280 Main Street West, Hamilton, Ontario, L8S 4L8, Canada.

E-mail addresses: aklass@mcmaster.ca (A.F. Klassen), crae@mcmaster.ca (C. Rae), karenw.wong@sickkids.ca (K.W. Wong Riff), Neil.Bulstrode@gosh.nhs.uk (N. Bulstrode), denadai.rafael@hotmail.com (R. Denadai), jesse.goldstein@chp.edu (J. Goldstein), M.l.f.hol-12@umcutrecht.nl (M.L. Hol), dylanmurray@plasticsurgeon.ie (D.J. Murray), research@craniofacial.ie (S. Bracken), douglas.courtemanche@ubc.ca (D.J. Courtemanche), drohara@drjustineohara.com.au (J. O'Hara), dan.butler@doctors.org.uk (D. Butler), ali.tassi@schulich.uwo.ca (A. Tassi), CMalic@cheo.on.ca (C.C. Malic), Ingrid.ganske@childrens.harvard.edu (I.M. Ganske), yun.phua@me.com (Y.S. Phua), dmarucci@hotmail.com (D.D. Marucci), david.johnson@ouh.nhs.uk (D. Johnson), marc.swan@nds.ox.ac.uk (M.C. Swan), Elly.Breuning@alderhey.nhs.uk (E.E. Breuning), tim.goodacre@queens.ox.ac.uk (T.E. Goodacre), apusic@bwh.harvard.edu (A.L. Pusic), stefan.cano@modusoutcomes.com (S. Cano).

https://doi.org/10.1016/j.bjps.2021.05.040

[k] Department of Plastic and Oral Surgery, Boston Children's Hospital, Boston, MA, USA
[l] Department of Plastic and Reconstructive Surgery, Queensland Children's Hospital, Brisbane, Australia
[m] The Children's Hospital at Westmead Clinical School, Faculty of Medicine and Health, The University of Sydney, Westmead, NSW, Australia
[n] Oxford Craniofacial Unit, Oxford University Hospitals NHS Foundation Trust, Oxford, United Kingdom
[o] Spires Cleft Centre, Oxford University Hospitals NHS Foundation Trust, Oxford, United Kingdom
[p] Department of Plastic Surgery, Alder Hey Children's Hospital, Liverpool, United Kingdom
[q] Division of Plastic and Reconstructive Surgery, Brigham and Women's Hospital, Boston, MA, USA
[r] Modus Outcomes, Letchworth Garden City, United Kingdom

Received 29 July 2020; accepted 27 May 2021
Available online xxx

KEYWORDS
FACE-Q;
Craniofacial;
Cleft lip and/or palate;
Patient- reported outcome measure;
PROM;
Psychometrics;
CLEFT-Q;
Facial conditions;
Noncleft facial conditions

Summary *Background:* The CLEFT-Q includes 12 independently functioning scales that measure appearance (face, nose, nostrils, teeth, lips, jaws), health-related quality of life (psychological, social, school, speech distress), and speech function, and an eating/drinking checklist. Previous qualitative research revealed that the CLEFT-Q has content validity in noncleft craniofacial conditions. This study aimed to examine the psychometric performance of the CLEFT-Q in an international sample of patients with a broad range of facial conditions.
Methods: Data were collected between October 2016 and December 2019 from 2132 patients aged 8 to 29 years with noncleft facial conditions. Rasch measurement theory (RMT) analysis was used to examine Differential Item Function (DIF) by comparing the original CLEFT-Q sample and the new FACE-Q craniofacial sample. Reliability and validity of the scales in a combined cleft and craniofacial sample (n=4743) were examined.
Results: DIF was found for 23 CLEFT-Q items when the datasets for the two samples were compared. When items with DIF were split by sample, correlations between the original and split person locations showed that DIF had negligible impact on scale scoring (correlations $\geq$0.995). In the combined sample, RMT analysis led to the retention of original content for ten CLEFT-Q scales, modification of the Teeth scale, and the addition of an Eating/Drinking scale. Data obtained fit with the Rasch model for 11 scales (exception School, p=0.04). Person Separation Index and Cronbach alpha values met the criteria.
Conclusion: The scales described in this study can be used to measure outcomes in children and young adults with cleft and noncleft craniofacial conditions.

Introduction

Patient-reported outcome measures (PROMs) are carefully designed questionnaires that measure outcomes important to patients from their perspective.[1-3] Facial appearance and facial function are concepts that are largely absent from most PROMs used in children and young adults with conditions associated with facial differences.[4-5] To address this gap, our team created a PROM for cleft lip and/or palate,[6] the most common craniofacial condition.[7] Treatments for this condition include procedures to improve appearance, speech, dentition, and hearing. The CLEFT-Q was designed to measure outcomes for these treatments from the patient perspective.[6,8]

The CLEFT-Q measures the concepts given in Table 1. We hypothesized that most questionnaire items in the CLEFT-Q would be applicable to other patients with facial differences considering their shared lived experience. Therefore, instead of developing a new PROM to measure concepts already covered by the CLEFT-Q, we aimed to determine which items covered concerns important to patients with noncleft conditions.

Full details of the methods used to develop the CLEFT-Q are published elsewhere.[8] Briefly, interviews with 138 patients from 6 countries were used to elicit content for scales.[9] The CLEFT-Q was then refined with input from 69 patients and feedback from 44 experts,[10] translated into 5 languages,[11-12] and field-tested in a sample of 2434 patients from 12 countries.[6,8] Psychometric analysis involved Rasch Measurement Theory (RMT) analysis, a modern psychometric approach that uses a set of tests and criteria to identify anomalies with items and within scales.[13-14] For example, in RMT analysis, it is possible to examine Differential Item Functioning (DIF) to determine if any items are biased in favor of any subgroups in a sample. DIF analysis for the CLEFT-Q showed that scales worked the same for participants who differed in terms of age group, gender, and language. These findings supported the use of a common

Table 1 CLEFT-Q© scales, including the number of items, appropriate ages and cleft type, recall period for completing

Name of scale	Items	Age	Cleft type	Recall period
Face	9	8 to 29	All diagnoses	now
Nose	12	8 to 29	All diagnoses	now
Nostrils	6	8 to 29	All diagnoses	now
Teeth	8	8 to 29	All diagnoses	now
Jaws	7	12 to 29	All diagnoses	now
Lips	9	8 to 29	All diagnoses	now
Cleft lip scar	7	8 to 29	CLP, CLA, CL	now
Psychological	10	8 to 29	All diagnoses	past week
School	10	8 to 18	All diagnoses	past week
Social	10	8 to 29	All diagnoses	past week
Speech Distress	10	8 to 29	CLP, CP, CLA	past week
Speech Function	12	8 to 29	CLP, CP, CLA	past week
Eating and Drinking	9	8 to 29	All diagnoses	past week

CLP = cleft lip and/or palate; CLA = cleft lip and alveolus; CL = cleft lip only; CP = cleft palate only; All diagnoses = CL, CLP, CLA, CP

scoring algorithm for each scale that could facilitate international benchmarking.[6]

Building on the CLEFT-Q study, we presented the scales to 84 patients aged 8 to 29 years with a broad range of facial conditions.[15] These interviews provided evidence that all, but one CLEFT-Q scale (i.e., Cleft Lip Scar scale), measured issues that mattered to patients with noncleft facial conditions.[15]

The aims of this study (Part 1), which focuses on validating the CLEFT-Q scales for use in conditions associated with a facial difference, were two-fold: (1) to conduct DIF testing to determine if the *original* CLEFT-Q scales function the same in patients with cleft and noncleft facial conditions; and (2) to examine the psychometric properties of the CLEFT-Q scales in a combined sample of cleft and noncleft facial conditions. In part 2, we describe the psychometric findings for the *new* FACE-Q Craniofacial Module scales not covered by the CLEFT-Q.[16]

Methods

This study was approved by the coordinating site ethics board (Hamilton Integrated Research Ethics Board) and by the ethics board at each participating site. Written and informed assent and/or consent was obtained from participants and guardians.

Data Collection

The analysis included data from three studies as follows:

1. **Craniofacial**

The FACE-Q Craniofacial Module phase 2 field-test study collected data from patients (n=2036) aged 8 to 29 years with a visible and/or functional facial difference. Recruitment took place at 31 sites in 10 countries between October 2016 and December 2019. Patients' consent for the study was obtained by a member of the healthcare team or research assistant during a clinic visit. Data were collected using electronic (tablets) or paper-and-pencil (booklets). A clinical form was completed by the site recruiter for each participant. The form asked about facial areas (e.g., jaw, lips, nose) and functions (e.g., eat/drink, speak) that corresponded to each scale. For each area and problem, a severity rating (none, minor, major) was given. The form also asked for the child's age, gender and diagnoses. Answers were used with branching logic to ensure that participants completed only relevant scales. In addition, through social media, members of Microtia UK, the US Moebius Syndrome Foundation, Bell's Palsy and Facial Paralysis Foundation, and Facial Palsy UK were sent study recruitment materials inviting patients to complete the REDCap® survey online. Table 2 shows the number of items from the CLEFT-Q and new FACE-Q items that were tested.

2. **Cancer**

Data from an international (French, Dutch, UK, USA) follow-up study of patients aged 8 to 29 years treated for head and neck tumors – when they were aged 0 to 18 years – were also included in the FACE-Q field-test sample. Questionnaire booklets were used for data collection in outpatient clinics. Participants were invited to complete the following scales/checklists: Face, Nose, Lips, Teeth, Jaws, Eating/Drinking, Speech Function, Speech Distress, Psychological, Social, and School.

Cleft lip and/or palate

To determine if CLEFT-Q scales function the same in patients with cleft and noncleft facial conditions (aim 1), we compared FACE-Q and CLEFT-Q[6] field-test samples. These analyses were used to determine if the content of the CLEFT-Q scales worked the same (i.e., no DIF) in patients with noncleft conditions. The CLEFT-Q field-test study included 2434 participants aged 8 to 29 years with cleft lip, cleft palate, cleft lip and palate, or cleft lip and alveolus.[6]

To determine the psychometric performance of the CLEFT-Q scales in the combined cleft and noncleft sample (aim 2), data from an additional CLEFT-Q phase 3 study[17] were included. This study measured changes 6 months after rhinoplasty, orthognathic surgery, lip revision, and alve-

Table 2 Number of items from the field-test version of the CLEFT-Q and new items tested in the FACE-Q field-test study

Domain	Scale	Source and number of items tested and retained				
		CLEFT-Q distribution version	CLEFT-Q discarded items retested	FACE-Q new items added	Items tested	Items retained
Appearance	Face	9	1	0	10	9
	Jaws	7	1	5	13	7
	Lips	9	0	2	11	9
	Nose	12	1	0	13	12
	Nostrils	6	0	0	6	6
	Teeth	8	2	9	19	12
Health related quality of life	Psychological	10	0	0	10	10
	School	10	0	0	10	10
	Social	10	1	0	11	10
	Speech Distress	10	0	0	10	10
Function	Speech	12	1	1	14	12
	Eating/Drinking	7*	0	6	13	9
Total		110	7	23	140	116

*Eating/Drinking in CLEFT-Q version has 2 additional items that are cleft-specific and not included

olar bone graft. Patients (n=177) were aged 8 to 29 years. Recruitment took place at seven cleft centers in Canada, the USA, and the UK, between January 2018 and October 2019. CLEFT-Q scales measuring outcomes most relevant to each procedure were completed using electronic (tablets) or paper-and-pencil (booklets).

Statistical Analysis

All data were entered into a secure REDCap database[18-19] hosted at McMaster University (Canada). SPSS Version 26 (IBM Corporation, Armonk NY, USA for Windows®/Apple Mac®) and RUMM2030 (RUMM version 2030, RUMM Laboratory Pty Ltd., Duncraig, Western Australia (RUMM Laboratory Pty Ltd, 1998-14) were used for data analysis. The following RMT analyses were performed:

DIF: CLEFT-Q versus FACE-Q field-test samples

To determine statistically if the content of CLEFT-Q scales worked the same for the FACE-Q Craniofacial field-test sample, we computed DIF to determine if the items of a scale were invariant (stable) across the trait that the scale was designed to measure in different subgroups.[20-21] DIF exists if one group shows a systematic difference in their responses to an item across the range of a scale, which is indicated by a significant main effect. For this analysis, the CLEFT-Q field-test sample was compared with the FACE-Q Craniofacial sample. For each scale, we identified the smaller subgroup and used the feature provided in RUMM2030 to select a random sample of the same size from the other subgroup(s). DIF analysis was repeated thrice to determine if the results based on random samples were stable. Any item that evidenced DIF (i.e., significant Chi-square p-values after Bonferroni adjustment) from the unadjusted analyses, were recalibrated as a separate item for each subgroup. For scales completed by a large sample of patients, we conducted DIF analyses with and without adjusting the overall sample to 500.[22] To determine the impact of DIF on scoring, the original and re-calibrated person locations were correlated (Spearman correlations).

Psychometric findings: pooled sample

A common set of scales for use across all craniofacial conditions would be advantageous. Therefore, a pragmatic decision was taken such that new items needed to perform well statistically and add important clinical value, for them to be retained. The RMT analysis was performed on the full sample of 4743 participants.

In the RMT analysis, the following set of statistical and graphical tests were performed to examine whether the observed data fit the Rasch model providing valid and reliable measurement:[13-14,22]

Item fit: We examined the item fit to determine if the items of each scale worked together clinically and statistically. The sample size was amended to 500 for tests of fit statistics.[22] Item fit was examined by inspecting item response options (ordering of the item thresholds),[23] and graphical (item characteristic curves) and statistical ((log residuals (item-person interaction) and Chi-square values (item-trait interaction)) indicators of fit. Ideal fit residuals fall between -2.5 and +2.5, with Chi-square values non-significant after Bonferroni adjustment.[14]

Targeting: This analysis determined whether items were spread over a range that matched the range of the construct reported by the sample. Targeting was examined graphically (person-item threshold distribution) and statistically (proportion of the sample to score outside the range of each scale's measurement). Ideally, scales should include items that provide information for all levels of the concept as experienced by the sample.[22]

DIF: We also conducted DIF on the combined sample for age (8-10, 11-13, 14-17, 18-29 years), gender (male versus female), and language (English versus other). As described above, DIF analysis was repeated thrice on a random sample to determine if results based on subgroups the same

size were stable. The analysis was conducted with and without adjusting the sample to 500. Items with significant Chi-square p-values after Bonferroni adjustments were split on the sample characteristics, and the new and original person locations were correlated (Spearman correlations) to determine the impact of DIF on scoring.[14]

Reliability: Scale reliability was examined in terms of Person Separation Index (PSI) and Cronbach alpha.[24] A reliability coefficient greater than or equal to 0.70 was considered sufficient.[25] When responses to an item are influenced by responses to other items, scale reliability can be artificially inflated. We examined residual correlations between items over 0.20 and performed a subtest to measure the impact of residual correlations (i.e., dependency) on the PSI value.[23]

Results

Table 3 shows characteristics for the 4743 participants and Table 4 shows the number of assessments for each scale provided by the sample.

DIF: CLEFT-Q versus FACE-Q field-test samples

Table 5 summarizes the results of the DIF analysis for the 103 items that formed the 11 CLEFT-Q scales we tested. A total of 23 items across nine scales (exceptions Jaws and Teeth) evidenced DIF in the unadjusted analysis, and nine items in four scales evidenced DIF in the adjusted analysis. Pearson correlations between the original and adjusted person locations for items that evidenced DIF were 0.995 or higher. These findings provided evidence that the participants in the FACE-Q field-test responded the same way to items from the CLEFT-Q as did participants from the original cleft field-test sample. Considering the lack of DIF between the two study samples, the RMT analysis proceeded with the full combined sample.

Psychometric findings: pooled sample

Table 6 reveals the RMT summary results based on the full sample of 4743 participants. The item-level fit statistics and DIF results are given in **Appendix 1**. The RMT analysis led to the original CLEFT-Q content for 10 of the 11 scales being retained verbatim. For the Teeth scale, three CLEFT-Q items were dropped owing to poor item fit, and seven new items were retained forming a 12-item scale. All 107 items for these 11 scales had ordered thresholds and nonsignificant Chi-square p-values after Bonferroni correction. Item fit residuals were outside $\pm$ 2.5 for 52 items. Most participants scored within the range of measurement for each scale (range 74% to 91%).

A new 9-item Eating/Drinking scale was formed from seven CLEFT-Q items and two new items. The RMT analysis for this scale excluded the cleft samples since we previously presented that the CLEFT-Q version functioned like a problem checklist instead of a scale.[6] We rescored the scale by collapsing across the two middle response options since two items had disordered thresholds. The rescored data were used in the RMT analysis. All item fit residuals for this scale were within $\pm$ 2.5 and had nonsignificant Chi-square p-values after Bonferroni correction.

Figure 1 demonstrates an example of targeting for the Teeth scale. The figure shows the distribution of person measurement (top histograms) and item locations (bottom histograms) by how participants answered the question: "How much do you like how your teeth look overall?" Most participants (88.8%) scored inside the range for which this scale provided measurement. Those participants who scored outside the range of the scale (right side) tended to report that they 'Very Much' liked how their teeth look overall. Finally, the bell curve shows that the best point of measurement on this scale is located in the center. Figure 2 shows a second example of targeting for the Eating/Drinking scale. Most participants (74.2%) scored inside the range for which the scale provided measurement. Participants who scored outside the range tended to be those who 'Never' experienced an eating/drinking problem.

Appendix 1 gives the DIF results for gender, age, and language. Across the 12 scales, significant DIF was evident for 37 items in seven scales for age group, 21 items in seven scales for gender, and 31 items in nine scales by language. When the sample was adjusted to 500, DIF was only evident for four items in three scales by age group, and two items in two scales by language. Correlations between the original person locations and the locations after the items were split by DIF were $\geq$0.995.

The observed data fit the expectations of the Rasch model for 11 scales (refer Table 6). The p-value for the School scale was marginally significant (p=0.04). PSI values with and without extremes were $\geq$0.77 and $\geq$0.80 for the 12 scales, respectively. Cronbach alpha values with and without extremes were $\geq$0.90 and $\geq$0.86, respectively. Item residual correlations were greater than 0.20 for seven pairs of items in five scales. Subtests performed to examine the impact of the residual correlations on scale reliability exhibited a drop in the PSI values with or without extremes of 0.01 (Face, Social, Speech Distress), 0.02 (Speech Function), and 0.05 (Eating/Drinking).

Discussion

The treatment of craniofacial conditions aims to improve appearance, function, and health-related quality of life. Rather than creating a new PROM for noncleft facial conditions, we hypothesized that the shared experience of living with a facial difference meant that CLEFT-Q content would cover concerns relevant to patients with other facial conditions. Our qualitative efforts provided support for this hypothesis.[15] The findings here provide quantitative support that the CLEFT-Q scales can be used to measure outcomes in noncleft facial differences. Specifically, we discovered that contents from 11 CLEFT-Q scales were invariant (stable) across both field-test samples. This finding was necessary before combining participants for the RMT analysis. In the combined sample, the content of each scale, including the two that differ from CLEFT-Q, met RMT specifications demonstrating good "fit" of the data to the Rasch model.

The findings in our study fit well with the literature on psychological adjustment in patients with craniofacial

Table 3 Characteristics (Number, %) for the 4743 participants

	FACE-Q Phase 2		Paediatric Cancer		CLEFT-Q Phase 2		CLEFT-Q Phase 3		Total	
	N	%	N	%	N	%	N	%	N	%
Total	2036	42.9	96	2.0	2434	51.3	177	3.7	4743	100
Country										
Australia	38	1.9	-	-	25	1.0	-	-	63	1.3
Brazil	178	8.7	-	-	-	-	-	-	178	3.8
Canada	857	42.1	-	-	624	25.6	69	39.0	1550	32.7
Chile	7	0.3	-	-	89	3.7	-	-	96	2.0
China	360	17.7	-	-	-	-	-	-	360	7.6
Colombia	-	-	-	-	210	8.6	-	-	210	4.4
France	-	-	21	21.9	-	-	-	-	21	0.4
India	-	-	-	-	232	9.5	-	-	232	4.9
Ireland	113	5.6	-	-	100	4.1	-	-	213	4.5
Netherlands	-	-	40	41.7	206	8.5	-	-	246	5.2
Spain	28	1.4	-	-	93	3.8	-	-	121	2.6
Sweden	-	-	-	-	100	4.1	-	-	100	2.1
Turkey	-	-	-	-	54	2.2	-	-	54	1.1
United Kingdom	318	15.6	29	30.2	339	13.9	70	39.5	756	15.9
United States	134	6.6	6	6.3	362	14.9	38	21.5	540	11.4
Other	3	0.1	-	-	-	-	-	-	3	0.1
Language										
Chinese	360	17.7	-	-	-	-	-	-	360	7.6
Dutch	-	-	40	41.7	206	8.5	-	-	246	5.2
English	1470	72.2	35	36.5	1450	59.6	177	100.0	3132	66.0
French	-	-	21	21.9	-	-	-	-	21	0.4
Hindi	-	-	-	-	232	9.5	-	-	232	4.9
Portuguese	178	8.7	-	-	-	-	-	-	178	3.7
Spanish	28	1.4	-	-	392	16.1	-	-	420	8.9
Swedish	-	-	-	-	100	4.1	-	-	100	2.1
Turkish	-	-	-	-	54	2.2	-	-	54	1.1
Age in years										
8-10	482	23.7	15	15.6	656	27.0	63	35.6	1216	25.6
11-13	557	27.4	20	20.8	553	22.7	18	10.2	1148	24.2
14-17	553	27.2	25	26.0	678	27.9	37	20.9	1293	27.3
18-29	444	21.8	36	37.5	546	22.4	59	33.3	1085	22.9
Missing	-	-	-	-	1	0	-	-	1	0
Gender										
Male	1010	49.6	53	55.2	1351	55.5	107	60.5	2521	53.2
Female	1022	50.2	43	44.8	1081	44.4	70	39.5	2216	46.7
Other	3	0.1	-	-	-	-	-	-	3	0.1
Missing	1	0	-	-	2	0.1	-	-	3	0.1
Main Condition*										
BIRTHMARK										
Congenital melanocytic naevus	39	1.9	-	-	-	-	-	-	39	0.8
Haemangioma	66	3.2	-	-	-	-	-	-	66	1.4
Sebaceous naevus	17	0.8	-	-	-	-	-	-	17	0.4
Vascular malformation	76	3.7	-	-	-	-	-	-	76	1.6
Birthmark other	4	0.2	-	-	-	-	-	-	4	0.1
CLEFT										
Cleft lip	-	-	-	-	263	10.8	9	5.1	272	5.7
Cleft palate	-	-	-	-	568	23.3	2	1.1	570	12.0
Cleft lip and palate	-	-	-	-	1399	57.5	140	79.1	1539	32.4
Cleft lip and alveolus	-	-	-	-	204	8.4	24	13.6	228	4.8
Cleft - type not specified	-	-	-	-	-	-	2	1.1	2	0.0

(*continued on next page*)

Table 3 (*continued*)

	FACE-Q Phase 2		Paediatric Cancer		CLEFT-Q Phase 2		CLEFT-Q Phase 3		Total	
	N	%	N	%	N	%	N	%	N	%
EAR CONDITION										
Microtia	549	27.0	-	-	-	-	-	-	549	11.6
Prominent ears	146	7.2	-	-	-	-	-	-	146	3.1
Ear other	34	1.7	-	-	-	-	-	-	34	0.7
SKELETAL										
Acquired Skeletal	55	2.7	-	-	-	-	-	-	55	1.2
Craniofacial microsomia	79	3.9	-	-	-	-	-	-	79	1.7
Craniofrontonasal condition	27	1.3	-	-	-	-	-	-	27	0.6
Craniosynostosis non-syndromic	168	8.3	-	-	-	-	-	-	168	3.5
Craniosynostosis syndromic	105	5.2	-	-	-	-	-	-	105	2.2
Fibrous dysplasia	30	1.5	-	-	-	-	-	-	30	0.6
Mandibular condition	39	1.9	-	-	-	-	-	-	39	0.8
Multiple bony anomalies	19	0.9	-	-	-	-	-	-	19	0.4
Post-traumatic bony defect	42	2.1	-	-	-	-	-	-	42	0.8
Other Congenital Skeletal	21	1.0	-	-	-	-	-	-	21	0.4
SOFT TISSUE										
Acquired soft tissue	30	1.5	-	-	-	-	-	-	30	0.6
Congenital soft tissue	14	0.7	-	-	-	-	-	-	14	0.3
Neurofibromatosis type 1	31	1.5	-	-	-	-	-	-	31	0.7
Parry-Romberg Syndrome	44	2.2	-	-	-	-	-	-	44	0.9
Soft tissue other	15	0.7	-	-	-	-	-	-	15	0.3
TRAUMA										
Bite	10	0.5	-	-	-	-	-	-	10	0.2
Burn	20	1.0	-	-	-	-	-	-	20	0.4
Fracture	71	3.5	-	-	-	-	-	-	71	1.5
Laceration	13	0.6	-	-	-	-	-	-	13	0.3
Trauma other	23	1.1	-	-	-	-	-	-	23	0.5
OTHER CONDITION										
Cancer	12	0.6	96	100	-	-	-	-	108	2.3
Facial paralysis	62	3.0			-	-	-	-	62	1.3
Orthodontic	153	7.5	-	-	-	-	-	-	153	3.2
Other syndrome	22	1.1	-	-	-	-	-	-	22	0.5

*Condition listed represents the main diagnosis, classifications may have varied by site. 9.6% of participants had multiple conditions.

anomalies.[26-28] First, Stock and Feragen recommended that the same outcome measures be used in cleft and noncleft craniofacial conditions because of the shared experience of patients.[26-27] Our findings support this recommendation as CLEFT-Q scales were revealed to be psychometrically valid in patients with cleft and noncleft craniofacial conditions. Using the same outcome tools for both populations would allow researchers to combine and compare data and findings where appropriate. Second, the use of separately functioning scales targeting specific constructs rather than a single overall total score was highlighted as important in a literature review of psychological adjustment.[28] The authors suggested that looking at each domain separately would make it easier to identify strengths and difficulties experienced by patients.[28]

Many patients with craniofacial conditions require orthodontic treatment. A recent systematic review identified the use of 18 PROMs in the orthodontic population of which three were commonly used.[29] The three PROMS, however, lack content that asks about the appearance of the teeth (an important outcome of orthodontic care). The content of the FACE-Q Teeth scale includes five CLEFT-Q items plus seven new items. By using the RMT approach, it is possible to directly compare the original CLEFT-

Table 4 Number of assessments for each scale by study sample and total

	FACE-Q Phase 2	Paediatric Cancer	CLEFT-Q Phase 2	CLEFT-Q Phase 3
Face	1356	96	2402	305
Jaws	406	71	1476	46
Lips	282	93	2213	229
Nose	232	94	2297	304
Nostrils	216	5	2278	300
Teeth	432	93	2311	186
Psychological	2103	94	2254	297
Social	2099	93	2242	294
School	1606	57	1662	183
Speech Distress	162	0	999	0
Speech Function	163	0	984	0
Eating/drinking	391	0	0	0

*Relevant scales may not have been completed by all participants in the field-test because they may have chosen to not complete all scales; some participants contributed more than one assessment.

Table 5 Items that evidenced DIF for each random sample (unadjusted and adjusted) and Spearman correlations between person locations before and after adjusting for DIF

Scale	Number in each sample	Items	DIF		Correlations for each random sample		
			Unadjusted	Adjusted	1	2	3
Face	1452	Look best	1,3	-	0.999	0.999	0.999
		Go out	1	-			
		Laugh	1,2,3	-			
		Match	2	-			
Lips	375	Smile	1,2,3	1,2,3	0.996	0.996	0.997
		Photos	1	-			
		Laugh	2	-			
Nose	326	Middle	3	-	-	-	0.999
Nostrils	221	Size	3	3	-	-	0.995
Psychological	2197	Feel okay	1,2,3	-	0.997	0.997	0.997
		Like self	1,2,3	-			
		Proud of self	2,3	-			
School	1663	Happy	1	-	0.996	-	-
Social	2192	Treat same	1,2,3	-	0.998	0.998	0.998
		Feel same	1,2,3	-			
		People listen	2	-			
Speech Distress	162	Frustrated	NA	3	0.998	0.998	0.997
		Embarrassed	NA	1			
		Repeat	NA	2			
		Not understood	NA	1,2			
Speech Function	163	Family	NA	1,2,3	0.998	0.998	0.998
		Friends	NA	1,2,3			
		Some words	NA	2			

Q version with the new 12-item version through a technique called equating.[22,30] The clinical benefit is that the shorter and longer versions can be used with different patient populations and can provide directly comparable scores.

Like orthodontics, PROMs for childhood survivors of head and neck cancer lack content about appearance and facial function.[31] Our team previously conducted cognitive interviews with head and neck cancer patients and discovered that the FACE-Q Craniofacial Module has content validity for this patient population.[32] Considering that it can be hard to accrue a large sample to validate a PROM in rare conditions, the inclusion of a cohort of cancer patients in the field test represents the strength of our study. The FACE-Q can be used to measure outcomes in this cancer population.

One of the strengths of our study is the large, international heterogeneous sample. As with the original CLEFT-

Table 6 Rasch Measurement Theory scale level statistics

Scale	Full sample	Sample in RMT analysis	% scored on scale	Chi-square	DF	p-value	PSI +ext	PSI-ext	Cronbach alpha +ext	Cronbach alpha -ext
Face	4159	3777	90.8	48.9	72	0.98	0.87	0.87	0.92	0.89
Nose	2928	2559	87.4	52.7	108	0.99	0.92	0.92	0.96	0.94
Nostrils	2799	2192	78.3	23.6	54	0.99	0.90	0.87	0.95	0.90
Lips	2817	2316	82.2	35.3	81	0.99	0.91	0.90	0.96	0.93
Jaws	1999	1480	74.0	24.5	56	0.99	0.91	0.89	0.96	0.92
Teeth	3022	2684	88.8	54.1	96	0.99	0.86	0.85	0.95	0.93
Psychological	4748	3811	80.3	64.4	90	0.98	0.86	0.87	0.94	0.91
School	3508	2742	78.2	115.0	90	0.04	0.78	0.81	0.91	0.87
Social	4728	3865	81.7	50.9	80	1.00	0.80	0.82	0.90	0.87
Speech Distress	1161	1041	89.7	89.8	90	0.49	0.85	0.82	0.90	0.86
Speech Function	1147	1047	91.3	103.5	84	0.07	0.88	0.86	0.92	0.89
Eating/Drinking	391	290	74.2	20.1	27	0.83	0.77	0.80	0.91	0.86

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

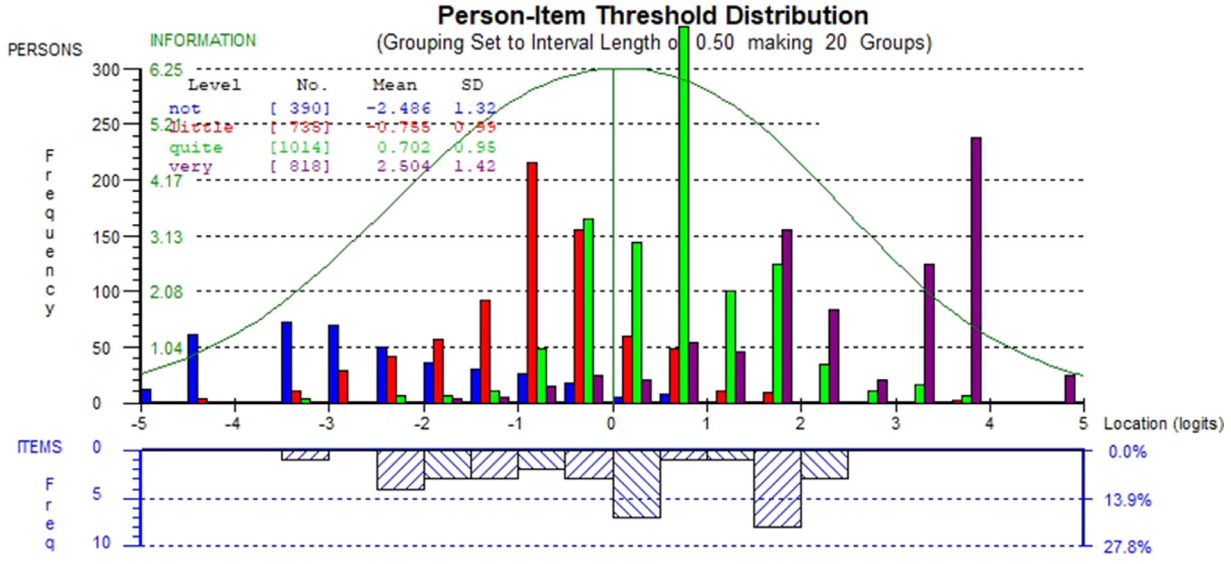

Fig. 1 Person-item threshold distribution for Teeth scale by how much participants liked how their teeth look overall.

Q findings,[6] we did not find evidence of bias by age, gender, or language. These findings support the international use of a common scoring algorithm for each scale. A difference with the CLEFT-Q findings[6] is that the Eating/Drinking scale worked psychometrically for noncleft conditions. The reliability was lower than other scales, but still acceptable by COSMIN standards.[25]The present study recruited patients with a broad range of noncleft facial conditions. Elsewhere we report that the FACE-Q Craniofacial Module can be used with children and adults of any age with facial nerve paralysis. Since the scales were designed for patients aged 8 to 29 years, the first step required showing that the scales covered concepts important to older adults.[33] This step was followed by a field-test and psychometric analyses in a combined sample of 235 patients aged 8 to 81 years.[34] Further research could examine the psychometric properties of the FACE-Q Craniofacial Module in similar studies of patients of different ages with specific craniofacial conditions.

This study has several limitations. We relied upon sites recruiting patients from busy clinic settings. No information was collected about eligible patients who were missed, nor about patients invited to take part who declined participation. COSMIN criteria of psychometric properties of PROMs unexamined in our study include test-retest reliability, responsiveness, and correlation with other instruments.[25] While many of the comparisons in the study were between cleft and noncleft conditions, it is important to note that 249 participants had both cleft and craniofacial conditions. Most of these participants (n=222) were in the CLEFT-Q field-test sample.

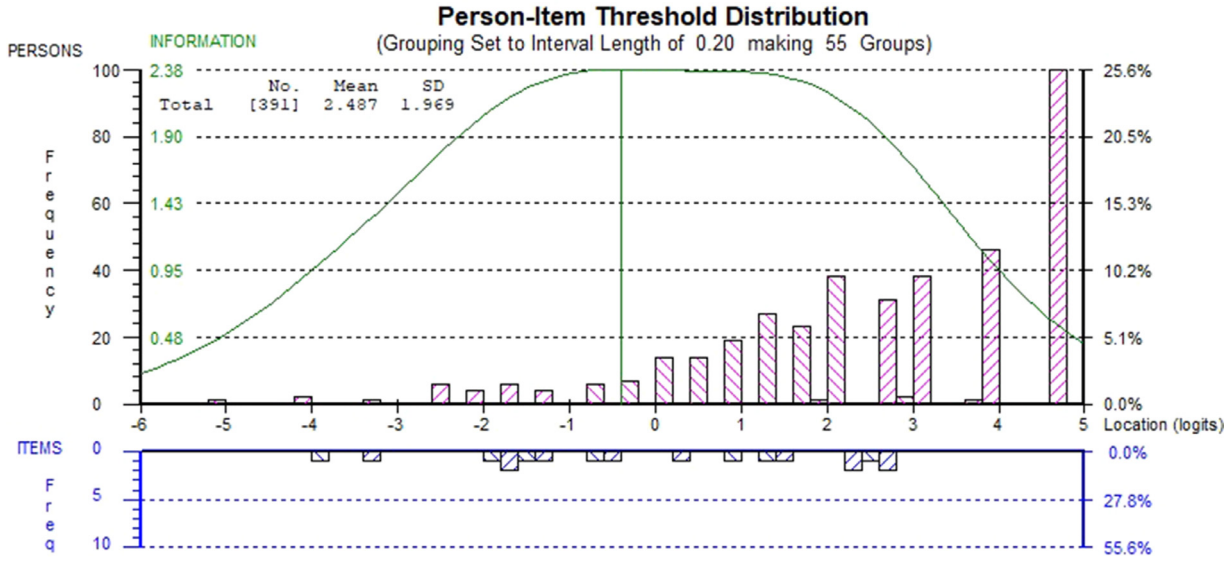

Fig. 2 Person-item threshold distribution for Eating/Drinking scale

Conclusion

Efforts to streamline data collection in cleft and craniofacial conditions continue,[35,36] particularly in the context of patient-centered care. To measure change and improve the care provided to patients with facial differences, carefully designed PROMs are needed. The scales described in this study provide clinicians and researchers with rigorously developed scales that can be used to measure outcomes in patients aged 8 to 29 years with any condition associated with a facial difference. These scales are made available free of charge to nonprofit users on signing a licensing agreement (qportfolio.org).

Conflict of Interest Statement

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

Funding Statement

The research described in this study was supported by grant # FRN 148779 from the Canadian Institute of Health Research (CIHR). 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 CIHR. 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.2021.05.040.

References

1. US Food and Drug Administration. Guidance for Industry Patient-Reported Outcome Measures: Use in Medical Product Development to Support Labeling Claims. Available at: https://www.fda.gov/downloads/drugs/guidances/ucm193282.pdf. Accessed February 8, 2021.
2. Black N. Patient reported outcome measures could help transform healthcare. *BMJ* 2013:346.
3. Meadows KA. Patient-reported outcome measures: an overview. *Br J Community Nurs* 2011;**16**(3):146–51.
4. 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.
5. 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.
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. Tanaka SA, Mahabir RC, Jupiter DC, et al. Updating the epidemiology of cleft lip with or without cleft palate. *Plast Reconstr Surg* 2012;**129** 511e–518e.
8. 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.
9. Wong Riff KWY, Tsangaris E, Goodacre TEE, et al. 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.
10. Tsangaris E, Wong Riff KWY, Goodacre T, et al. Establishing con-

tent validity of the CLEFT-Q: a new patient-reported outcome instrument for cleft lip/palate. *Plast Reconstr Surg Glob Open* 2017;**5**(04):e1305.
11. Tsangaris E, Riff KWY, Vargas F, et al. Translation and cultural adaptation of the CLEFT-Q for use in Colombia, Chile, and Spain. *Health Qual Life Outcomes* 2017;**15**(1):228.
12. Tsangaris E, Riff KW, Dreise M, et al. Translation and cultural adaptation of the CLEFT-Q into Arabic, Dutch, Hindi, Swedish, and Turkish. *Europ J Plas Surg* 2018;**41**(6):643–52.
13. Rasch G. Probabilistic models for some intelligence and attainment tests. *Vol. 1 of studies in mathematical psychology*. Copenhagen: Danmarks Paedagogiske Institut; 1960.
14. Andrich D. Rasch models for measurement. *Sage university papers series quantitative applications in the social sciences*, 07-068. Thousand Oaks (CA): Sage; 1988.
15. 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**(5):499–508.
16. Klassen AF, Rae C, Riff KW, et al. FACE-Q Craniofacial Module: Part 2 Psychometric properties of newly developed scales for children and young adults with facial conditions. *J Plast Reconstr Aesthet Surg* 2021. doi:10.1016/j.bjps.2021.03.009.
17. Miroshnychenko A, Rae C, Wong Riff K, et al. Psychometric validation of the CLEFT-Q patient reported outcome measure: a prospective study to examine cross-sectional construct validity *Proceedings of the American Cleft Palate – Craniofacial Association Virtual Annual Meeting*; 2021. Apr 29 - May 1. Abstract nr 2243.
18. 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.
19. Harris PA, Taylor R, Minor BL. The REDCap consortium: Building an international community of software partners. *J Biomed Inform* 2019;**95**:103208. doi:10.1016/j.jbi.2019.103208.
20. Andrich D, Sheridan B, Luo G. Interpreting RUMM2030. Perth: RUMM Laboratory Pty Ltd; 200
21. RUMM Laboratory. Welcome to RUMM. Current version RUMM 2030. http://www.rummlab.com.au/ [Accessibility verified July 20, 2020]
22. Hobart J, Cano S. Improving the evaluation of therapeutic intervention in MS: the role of new psychometric methods. *Health Technol Assess* 2009;**13**(iii, ix–x):1–177.
23. Wright BD, Masters G. *Rating scale analysis*. Chicago: MESA Press; 1982.
24. Cronbach LJ. Coefficient alpha and the internal structure of tests. *Psychometrika* 1951;**16**:297–334.
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.
26. Stock NM, Feragen KB. Assessing Psychological Adjustment to Congenital Craniofacial Anomalies: An Illustration of Methodological Challenges. *Cleft Palate Craniofac J* 2019;**56**(1):64–73. doi:10.1177/1055665618769661.
27. 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.
28. Feragen KB, Stock NM. Psychological adjustment to craniofacial conditions (excluding oral clefts): a review of the literature. *Psychol Health* 2017;**32**(3):253–88.
29. Zaror C, Pardo Y, Espinoza-Espinoza G, et al. Assessing oral health-related quality of life in children and adolescents: a systematic review and standardized comparison of available instruments. *Clin Oral Invest* 2018;**22**:1–15.
30. Smith SC, Hendriks AAJ, Cano SJ, Black N. Proxy reporting of health-related quality of life for people with dementia: a psychometric solution. *Health Qual Life Outcomes* 2020;**18**:148.
31. Anthony SJ, Selkirk E, Sung L, et al. Considering quality of life for children with cancer: a systematic review of patient-reported outcome measures and the development of a conceptual model. *Qual Life Res* 2014;**23**(3):771–89.
32. Wang Y, Rae C, Wong Riff K, et al. Validation of a PROM for use in Pediatric Head and Neck Cancer: The FACE-Q Craniofacial Module. Qual Life Res 2021; (Abstract #3070).
33. Allori AC, Kelley T, Meara JG, et al. A standard set of outcome measures for the comprehensive appraisal of cleft care. *Cleft Palate Craniofac J* 2017;**54**(5):540–54.
34. International Consortium for Health Outcomes Measurement (ICHOM). Standard sets: Start measuring outcomes that matter most to patients. https://www.ichom.org/standard-sets/. [Accessibility verified July 20, 2020]
35. Norris J, Longmire NM, Kilcoyne S, et al. Exploring patient experience of facial nerve palsy to inform the development of a patient-reported outcome measure: a qualitative study. *Plast Reconstr Surg Glob Open* 2019;**7**(1):e2072.
36. Klassen AF, Rae C, Gallo L, et al. Psychometric Validation of FACE-Q Craniofacial Module scales for Adults and Children with Facial Nerve Paralysis. *Facial Plast Surg Aesthet* 2021. doi:10.1089/fpsam.2020.0575.