Establishing Content Validity of the CLEFT-Q: A New Patient-reported Outcome Instrument for Cleft Lip/Palate

Elena Tsangaris, MSc*
Karen W.Y. Wong Riff, MD, MSc, FRCS†
Tim Goodacre, MBBS, FRCS‡
Christopher R. Forrest, MD, MSC, FRCS†
Marieke Dreise, BA§
Jonathan Sykes, MD¶
Tristan de Chalain, MB, ChB||
Karen Harman, MD**
Aisling O’Mahony, DDS††
Andrea L. Pusic, MD, MPH, FACS‡‡
Lehana Thabane, PhD*
Achilleas Thoma, MD, FRCSC, FACS§§
Anne F. Klassen, DPhil¶¶

Background: The CLEFT-Q is a new patient-reported outcome instrument designed to measure outcomes that matter to patients. The aim of this qualitative study was to establish content validity of the CLEFT-Q in patients who differ by age and culture.

Methods: Patients aged between 6 and 29 years were recruited from plastic surgery clinics in Canada, India, Ireland, the Philippines, the Netherlands and the United States. Healthcare providers and other experts participated in a focus group or provided individual feedback. Input was sought on all aspects of the CLEFT-Q (item wording, instructions, and response options), and to identify missing content. Patient interviews and expert feedback took place between September 2013 and September 2014.

Results: Sixty-nine patients and 44 experts participated. The first draft of the CLEFT-Q consisted of 163 items measuring 12 constructs. The first round of feedback identified 92 items that required revision. In total, 3 rounds of interviews, and the involvement of an artist to create pictures for 17 items, were needed to establish content validity. At the conclusion of cognitive interviews, the CLEFT-Q consisted of 13 scales (total 171 items) that measure appearance, health-related quality of life, and facial function. The mean Flesch-Kincaid readability statistic for items was 1.4 (0 to 5.2).

Conclusion: Cognitive interviews and expert review allowed us to identify items that required re-wording, re-conceptualizing, or to be removed, as well as any missing items. This process was useful for refining the CLEFT-Q scales for further testing. (Plast Reconstr Surg Glob Open 2017;5:e1305; doi: 10.1097/GOX.0000000000001305; Published online 25 April 2017.)

Copyright © 2017 The Authors. Published by Wolters Kluwer Health, Inc. on behalf of The American Society of Plastic Surgeons. This is an open-access article distributed under the terms of the Creative Commons Attribution-Non Commercial-No Derivatives License 4.0 (CCBY-NC-ND), where it is permissible to download and share the work provided it is properly cited. The work cannot be changed in any way or used commercially without permission from the journal.

DOI: 10.1097/GOX.0000000000001305

Disclosure: 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.

Supplemental digital content is available for this article. Clickable URL citations appear in the text.
Outcomes of CL/P treatment have typically been appraised objectively using observer-reported or clinician-reported assessments. However, because the goal of CL/P treatment is to improve a patient’s physical health and quality of life, these outcomes would be difficult to measure through the use of observer- or clinician-reported outcome assessments alone. Patient-reported outcome (PRO) instruments are a unique tool that can be used to accurately evaluate a patient’s perspective of their own health outcomes. Understanding the patients’ perspective using carefully designed PRO instruments could provide evidence-based information to inform clinical practice and future research.

A systematic review by Eckstein et al. identified 44 PRO instruments that had been used to measure quality of life or satisfaction in patients with CL/P. Although 5 questionnaires were validated in a CL/P population, none were developed with a CL/P focus. The authors of this review call for the development of a new valid and reliable cleft-specific PRO instrument that could be used in patient care and clinical research to evaluate the impact of surgery and treatment on patients’ quality of life.

Best practice for the development of a PRO instrument is iterative and involves item generation, item reduction, and psychometric evaluation. Phase 1 should involve the development of a conceptual framework, and the generation of items via a literature review, qualitative interviews with patients, and expert feedback. To finalize phase 1, cognitive interviews with patients are necessary to ensure that the items forming scales are relevant to patients and are appropriately understood, to minimize error that may result from item misinterpretation during data collection. Expert feedback in conjunction with cognitive interviews may also help to provide information about the clinical relevance of specific items.

The CLEFT-Q is a specific and unique PRO instrument developed to measure outcomes that matter to children and young adults with CL/P. The aim of the present study was to establish content validity of the CLEFT-Q. Content validity is a measurement property that appraises whether items in a scale are comprehensive and adequately reflect the perspective of the population of interest. For the present study, we used cognitive interviews with patients and expert feedback to determine the content validity of the CLEFT-Q for patients who vary by age and culture.

METHODS

Research Ethics Board Approval

The study was approved by the Research Ethics Board at participating hospitals and was conducted in accordance with policies for ethical conduct in research involving humans. All participants or their legal guardians provided written informed consent or assent for participation according to each center’s policy.

The CLEFT-Q

The CLEFT-Q was developed in accordance with internationally recommended guidelines for the development of a new PRO instrument. Findings from a systematic review were used to develop a preliminary conceptual framework. Concepts within the framework were developed into questions that formed an interview guide used in a series of qualitative interviews. A total of 136 interviews were conducted with 138 patients (including 2 sets of twins who were interviewed together) with CL/P from 6 countries including Canada, England, India, Kenya, the Philippines, and the United States. Interviews were recorded, translated/transcribed into English, coded, and analyzed, resulting in the refinement of the CLEFT-Q conceptual framework, which included 5 domains and 12 minor themes as follows: appearance (of the face, nose, teeth, lips, jaws, and cleft lip scar), HRQOL (psychological, social, school, and speech-related distress), and facial function (speech and eating/drinking). From the coded data, we also created a comprehensive item pool. For each minor theme, the item pool was used to develop a set of items that together mapped out a scale on a continuum, from more to less of the construct. Whenever possible, items were created using positive or neutral wording that could be understood by patients as young as 6 years old. For each scale, 4 labeled response options were chosen to align with published guidelines. Each CLEFT-Q scale was designed to be independently functioning. This approach aimed to reduce patient burden, as only the scales relevant to the research or clinical purpose need to be completed.

The first draft of the CLEFT-Q was translated into Assamese (India), Tagalog (the Philippines), and Dutch (the Netherlands) to facilitate cognitive interviews in multiple countries. Translations were conducted in accordance with the International Society for Pharmacoeconomics and Outcomes Research guidelines for the translation and cultural adaptation of PRO instruments.

Study Design, Data Collection, and Analysis

Table 1 illustrates the cognitive interviewing approach used, which was adapted from Willis. Cognitive interviewing involved a series of one-on-one semi-structured interviews with patients, using a cognitive interview guide (see PDF, Supplemental Digital Content 1, which displays the cognitive interview guide, http://links.lww.com/PRSOG/A427). The objectives for cognitive interviewing were to determine challenges with participant comprehension of the item wording, instructions, and response options, using the think aloud approach and to identify items that were thought to be duplicate, missing, irrelevant, or insensitive. Verbal probing was used concurrently, whereby the interviewer asked specific questions about content that was found to be problematic in preceding interviews. This combined approach made it possible to identify problems with item interpretation and response selection. For items expressed as ambiguous or difficult to understand, upon being informed of the item’s meaning, participants were encouraged to suggest revisions to item wording. Similarly, experts provided feedback on the relevance and comprehensibility of the CLEFT-Q items, instructions, and response options, and were encouraged to suggest missing content.
Cognitive interviews and expert feedback were conducted in rounds. Figure 1 highlights the sequence of cognitive interviews and expert feedback by country and round. Round 1 interviews involved patients from McMaster Children’s Hospital (Hamilton) and the Hospital for Sick Children (Toronto) in Canada; round 2 interviews involved patients from Canada, UC Davis Health System (Sacramento, Calif.) in the United States, and the Operation Restore Hope New Zealand medical mission trip to the Philippines; and round 3 interviews involved patients from St. James Hospital in Dublin, Ireland, Operation Smile Comprehensive Cleft Care Center in Guwahati, India, and the University of Groningen in the Netherlands. Expert feedback was interspersed throughout each round and varied by country and expertise.

To obtain expert feedback, in round 1, focus groups were held in Canada and the United States (facilitated by E.T. and K.W.R., respectively). Each focus group involved health-care providers, whose clinical focus was CL/P, meeting together to review and discuss the CLEFT-Q content. During each round, the CLEFT-Q was distributed to a range of experts who provided written feedback on the instrument. In the third round, the CLEFT-Q was circulated for feedback to 16 members of the National Cleft Psychologists Special Interest Group in England. This group discussed its content at a meeting and provided written feedback as a group.

Patient interview and expert feedback data were entered into a Microsoft Excel (2016) worksheet for analysis. A reparative approach to data analysis was employed, which involved examining findings and revising items
concurrently. Changes were made to the CLEFT-Q after each round, where the evidence collected was brought to the research team to assess the compiled results and make revisions before moving on to the next round of item testing. Willis recommended conducting 2–4 sets of interviews with 5–15 people in each. Data were analyzed using the “text summary” approach, which involved summarizing notes from cognitive interviews and expert feedback, to identify consistent themes.

In each round of revisions, considerations were necessary to ensure that the CLEFT-Q items had the lowest possible Flesch–Kincaid (F–K) readability level. The F–K readability score indicates the grade-reading level of an item. Readability cutoffs were determined in accordance with the reading comprehension literature.

### Sampling

Eligible participants were patients aged 6 years and older with CL/P, who could read and understand any of the target languages. Age eligibility was based on research reporting that children as young as 5 years of age are able to self-report on age-appropriate questionnaires, and children as young as 8 years of age are able to self-report on wellbeing, psychosocial health, and health-promoting behaviors.

In Canada, Ireland, India, the Netherlands, and the United States, a member of the health-care team approached patients in clinic or by telephone and invited them to consider participation in the study. Interviews were set up either by a health care or research team member and were conducted in the hospital, at the patients’ home, or over the phone, depending on the site or patient preference. In the Philippines, interviews took place during a week-long Operation Restore Hope surgical mission trip to Batangas, Philippines. The intake nurse informed patients of the study, and patients willing to participate were interviewed before or after their surgery.

Experts in the field of CL/P were recruited through our team’s professional network. Feedback was obtained face-to-face during focus groups or through individual feedback by e-mail.

### RESULTS

Translation and data collection for the CLEFT-Q cognitive interviews and expert feedback took place between September 2013 and September 2014. Sixty-nine patients participated in our cognitive interviews. Mean age of participants at the time of recruitment was 13.2 years (range, 6–26 years old), with 58% of the sample aged 13 years or younger. More females (52%) participated, and most patients had CL/P (77%; Table 2). Forty-four experts provided individual feedback, of which 13 participated in 1 of 2 focus groups. Most experts were psychologists (48%) and were from England (41%; Table 3).

During the cognitive interviews and expert feedback, items that were identified to be hard to understand, or considered irrelevant, were either revised to improve comprehensibility or dropped. Items that were dropped were mainly because the item (1) represented a difficult concept in which the meaning could not be clarified, (2) was not considered to be important by patients, or (3) was considered clinically irrelevant by experts.

We conducted 3 rounds of revisions involving patients and experts. At the start of round 1, the CLEFT-Q consisted of 163 items in 12 scales. Feedback from 17 patients and 15 experts (Fig. 1) led to 25 items that remained the same, 92 revised, 46 deleted, and 48 added. Most of the revisions in this round were needed to ensure that item wording worked with the new response options (n = 58 items), as feedback led to us change the response options for all 7 appearance scales, ie, from “Strongly Disagree, Disagree, Agree, Strongly Agree” to “Like a Lot, Like a Little, Dislike a Little, Dislike a Lot.” This change was required because patients and experts expressed some difficulty applying the response options to the items. Thirty-four items within the HRQOL and facial function scales also required revision. By completion of round 1, the CLEFT-Q consisted of 165 items.

A total of 25 patients and 5 experts provided feedback in round 2 (Fig. 1). In this round, the vast majority of items (total 140, 85%) remained the same, 23 items were revised, 2 were dropped, and 6 were added. Items were revised mainly to clarify their meaning or to add an example.

In round 3, 169 items were tested in 27 patients, with 24 experts providing feedback (Fig. 1). At this stage, very minor changes were required. A total of 137 items remained the same, 28 were revised, 4 were dropped, and 6 were added. Items that patients had some difficulty with at this stage tended to be specific appearance items (total 17, 61%). To ensure that the appearance items were as easy as possible to understand, we had an artist create 17 images that illustrated specific parts of the face [eg, how the tip of your nose looks (the very end of your nose?)], or how specific parts of the face look during movement [eg, how much you can move your lips (like to whistle or kiss?)]. We also included a picture of nos-

### Table 2. Demographic Characteristics of the Patient Who Participated in the Cognitive Interviews

<table>
<thead>
<tr>
<th>Demographic Variables</th>
<th>No. Patients, Count (%), N = 69</th>
</tr>
</thead>
<tbody>
<tr>
<td>Country</td>
<td></td>
</tr>
<tr>
<td>Canada</td>
<td>20 (29.0)</td>
</tr>
<tr>
<td>India</td>
<td>4 (5.8)</td>
</tr>
<tr>
<td>Ireland</td>
<td>5 (7.5)</td>
</tr>
<tr>
<td>The Philippines</td>
<td>13 (18.8)</td>
</tr>
<tr>
<td>The Netherlands</td>
<td>18 (26.1)</td>
</tr>
<tr>
<td>The United States</td>
<td>9 (13.0)</td>
</tr>
<tr>
<td>Sex</td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>33 (47.8)</td>
</tr>
<tr>
<td>Female</td>
<td>36 (52.2)</td>
</tr>
<tr>
<td>Age (y)</td>
<td></td>
</tr>
<tr>
<td>6–9</td>
<td>16 (23.2)</td>
</tr>
<tr>
<td>10–13</td>
<td>24 (34.8)</td>
</tr>
<tr>
<td>14–17</td>
<td>13 (18.8)</td>
</tr>
<tr>
<td>18–21</td>
<td>11 (16.0)</td>
</tr>
<tr>
<td>22–29</td>
<td>5 (7.2)</td>
</tr>
<tr>
<td>Type of cleft</td>
<td></td>
</tr>
<tr>
<td>Cleft lip only</td>
<td>9 (13.0)</td>
</tr>
<tr>
<td>Cleft palate only</td>
<td>6 (8.7)</td>
</tr>
<tr>
<td>Cleft lip and palate</td>
<td>53 (76.8)</td>
</tr>
<tr>
<td>Cleft lip and alveolus</td>
<td>1 (1.5)</td>
</tr>
</tbody>
</table>
trils, jaws, and a cleft lip scar to ensure that participants knew exactly which part of the face the scales referred to. At the completion of the cognitive interviews and expert review, the CLEFT-Q comprised 171 items within 13 scales that reflected the original conceptual framework (Table 4).

**Instructions, Response Options, and Readability**

Although minor changes were needed to finalize the instructions for the CLEFT-Q scales, the 3 rounds provided time to explore different sets of response options for each scale. At the end of the process, the 7 scales that measure appearance ask respondents to answer each item thinking of how their face (or specific area of their face) looks now. Respondents are then asked to answer for each item “how much do you like …” using the 4 response options listed above. All other CLEFT-Q scales include a series of statements, with instructions asking respondents to answer each item in relation to the past week, and in terms of frequency: “Never,” “Sometimes,” “Often,” and “Always.”

Mean F-K readability for the 171 items was 1.4 (range, 0–5.2) for scales (Table 4). The readability scores were below the fifth-grade reading level, with the exception of 2 items in the psychological scale: “I feel okay about myself” and “I feel confident,” which scored slightly higher (F-K readability score = 5.2).

Table 4 provides an example of how a CLEFT-Q scale (Cleft Lip Scar) was modified after each round. This scale differed from all other appearance scales, as the original items had either negative content or used negative phrasing. Experts who participated in round 1 suggested that we reconceptualize the scale to create a version with items that used positive or neutral content and phrasing. From the original set of 10 items, 8 items were dropped, 2 were revised and retained, and 9 new items were included from the original item pool. Further changes made in relation to feedback from subsequent rounds involved minor wording changes to improve clarity.

**DISCUSSION**

Cognitive interviews with patients and expert feedback were used to establish content validity of the CLEFT-Q for patients who varied by age and culture. This psychometric property was achieved by obtaining feedback from a large international sample of patients, who helped us refine instructions and a set of items for each scale, and to choose appropriate response options. Content validity was also established through feedback from an international sample of experts in CL/P who provided insights about the suitability and perceived difficulty of items. Input received in round 1 led to substantial revisions of the first draft of the CLEFT-Q. Once we made these revisions, subsequent rounds largely involved the “fine tuning” of items.
Table 5. Example Modifications to the Cleft Lip Scar Scale after Each Round of Cognitive Interviews and Expert Feedback

<table>
<thead>
<tr>
<th>Round 1 Items</th>
<th>Decision</th>
<th>Round 2 Items</th>
<th>Decision</th>
<th>Round 3 Items</th>
<th>Decision</th>
<th>Final Items for Pilot-testing</th>
</tr>
</thead>
<tbody>
<tr>
<td>My cleft scar looks big.</td>
<td>Revise</td>
<td>... the size of your cleft scar?</td>
<td>Retain</td>
<td>... the size of your cleft scar?</td>
<td>Revise</td>
<td>... the size of your cleft lip scar?</td>
</tr>
<tr>
<td>My cleft scar looks odd when I smile.</td>
<td>Revise</td>
<td>... how your cleft scar looks when you smile?</td>
<td>Retain</td>
<td>... how your cleft scar looks when you smile?</td>
<td>Revise</td>
<td>... how your cleft lip scar looks when you smile?</td>
</tr>
<tr>
<td>I do not like how my cleft scar looks.</td>
<td>Drop</td>
<td>Add ... the width of your cleft lip scar?</td>
<td>Drop</td>
<td>Add ... the width of your cleft lip scar?</td>
<td>Drop</td>
<td>Add ... the width of your cleft lip scar?</td>
</tr>
<tr>
<td>My cleft scar is easy for people to see.</td>
<td>Drop</td>
<td>Add ... the shape of your cleft scar?</td>
<td>Retain</td>
<td>... the shape of your cleft scar?</td>
<td>Revise</td>
<td>... the shape of your cleft lip scar?</td>
</tr>
<tr>
<td>My cleft scar looks uneven.</td>
<td>Drop</td>
<td>Add ... the color of your cleft scar?</td>
<td>Retain</td>
<td>... the color of your cleft scar?</td>
<td>Revise</td>
<td>... the color of your cleft lip scar?</td>
</tr>
<tr>
<td>My cleft scar looks abnormal.</td>
<td>Drop</td>
<td>Add ... how your cleft scar looks when you laugh?</td>
<td>Retain</td>
<td>... how your cleft scar looks when you laugh?</td>
<td>Revise</td>
<td>... how your cleft lip scar looks when you laugh?</td>
</tr>
<tr>
<td>My cleft scar looks tight.</td>
<td>Drop</td>
<td>Add ... how your cleft scar looks in photos?</td>
<td>Retain</td>
<td>... how your cleft scar looks in photos?</td>
<td>Revise</td>
<td>... how your cleft lip scar looks in photos?</td>
</tr>
<tr>
<td>My cleft scar looks ugly.</td>
<td>Drop</td>
<td>Add ... how your cleft scar looks in the mirror?</td>
<td>Retain</td>
<td>... how your cleft scar looks in the mirror?</td>
<td>Revise</td>
<td>... how your cleft lip scar looks in the mirror?</td>
</tr>
<tr>
<td>My cleft scar looks weird.</td>
<td>Drop</td>
<td>Add ... how much your cleft scar has faded?</td>
<td>Retain</td>
<td>... how much your cleft scar has faded?</td>
<td>Revise</td>
<td>... how much your cleft lip scar has faded over time?</td>
</tr>
<tr>
<td>My cleft scar looks wide.</td>
<td>Drop</td>
<td>Add ... how much the color of your cleft lip scar matches your skin color?</td>
<td>Add ... how much the color of your cleft lip scar matches your skin color?</td>
<td>Add ... how much the color of your cleft lip scar matches your skin color?</td>
<td>Add ... how much the color of your cleft lip scar matches your skin color?</td>
<td>Add ... how much the color of your cleft lip scar matches your skin color?</td>
</tr>
</tbody>
</table>
to improve comprehension as much as possible. After round 3, the CLEFT-Q was determined to be ready for pilot testing, in which the results were used to further refine the CLEFT-Q items and response options.

Assessment of outcomes in CL/P care has primarily relied on objective evaluations by health-care providers, with very little patient input. Engaging patients in the assessment of treatment outcomes may provide an important perspective in research that measures the impact of CL/P care globally. In 2000 and 2001, the World Health Organization’s international consensus meetings on craniofacial anomalies called for outcome measures that capture issues that “... matter to ordinary people rather than sophisticated surrogate measures that may have little relevance in everyday life.” More recently, Mossey et al. restated the World Health Organization recommendation in a seminal paper, where they stressed the need for standardized PRO instruments. Specifically, a need was expressed for “… psychological and QOL measures and economic outcomes.”

Neither our team’s literature reviews nor the updated review by Eckstein et al. identified a CL/P-specific PRO instrument. Five PRO instruments, including the Youth Quality of Life–Facial Differences, Pediatric Voice-Related Quality of Life, Cleft Audit Protocol for Speech–Augmented, Child Oral Health Impact Profile, and Child Oral Health Quality of Life were validated in a population of patients with CL/P. These instruments capture a range of issues that are important to patients with CL/P, including oral health (Child Oral Health Quality of Life), speech-related issues (Pediatric Voice-Related Quality of Life), and craniofacial specific quality of life concerns (Youth Quality of Life–Facial Differences). However, an important limitation is that 4 of these PRO instruments did not include qualitative patient input in their initial development (exception Youth Quality of Life–Facial Differences). The exclusion of qualitative patient input in the initial development of the various instruments may explain the absence of important cleft-specific concepts. To our knowledge, the CLEFT-Q is the first self-report CL/P-specific PRO instrument developed according to published guidelines for PRO instrument development. The CLEFT-Q covers cleft-specific issues from the perspective of patients who varied by age and culture.

Most studies developing a new PRO instrument do not typically use as large a sample as we did, either for the initial development (n = 138) or cognitive interview (n = 69) stages. Our team decided it was crucial to ensure that the content of the CLEFT-Q and the final wording of each item resonated with young children and young adults with different types of clefts and from multiple countries, including low- and middle-income countries. Therefore, an important strength of our study is the inclusion of a large, heterogeneous, international sample of patients who took part in the cognitive interviews to refine the scales. Additionally, consistent cognitive interview methodology procedures were used throughout the process. This consistency was maintained by having 1 experienced qualitative researcher conduct all English interviews and oversee 22 interviews conducted in non-English languages.

Although we specifically set the inclusion criteria for age to be at the lower range of what is possible for self-report (6 years old), a limitation of our study is that only 2 children were 6 or 7 years of age. Although both patients were able to read, and accurately interpret and respond to items in the CLEFT-Q scales, further research is required to determine if the CLEFT-Q (or specific scales) can be completed by children under the age of 8 years. In addition, the use of interviewer notes rather than tape recordings may have resulted in incomplete data because the interviewer may not have been as comprehensive in note taking. Therefore, this approach to data collection may have resulted in key comments from participants being missed. However, given the number of participants in our study, and the numerous countries involved, we anticipate that this potential problem is unlikely to be clinically important.

CONCLUSIONS

No changes were required to the CLEFT-Q conceptual framework. Cognitive interviews and expert review allowed us to identify items that required rewording, re-conceptualizing, or to be removed, as well as any missing items. This process was useful for refining the CLEFT-Q scales. The CLEFT-Q has now been field tested in an international study that involved 30 hospitals in 12 countries. In addition, a subset of the CLEFT-Q scales has been included in the International Consortium for Health Outcomes Measurement standard set.

ACKNOWLEDGMENTS

We are grateful for the operating grant we received from the Canadian Institutes for Health Research. In addition, Anne F. Klassen held a CIHR Mid-Career Award in Women’s Health during the conduct of this research. Andrea L. Pusic received support through the National Institutes of Health/National Cancer Institute Cancer Center Support Grant P30 CA008748.

REFERENCES


Anne F. Klassen, DPhil
Department of Pediatrics
McMaster University
Room HSC 3N27
1280 Main Street West
Hamilton, Ontario L8S 4K1, Canada
E-mail: aklass@mcmaster.ca


