CLEFT-Q: CONSTRUCT VALIDITY AND RESPONSIVENESS

PSYCHOMETRIC VALIDATION OF THE CLEFT-Q PATIENT REPORTED OUTCOME MEASURE: A PROSPECTIVE STUDY TO EXAMINE CONSTRUCT VALIDITY AND RESPONSIVENESS FOLLOWING FOUR CLEFT-SPECIFIC OPERATIONS

By Anna Miroshnychenko, B.Sc.Hon.

A Thesis

Submitted to the School of Graduate Studies

in Partial Fulfillment of Requirements for

the Degree Master of Science

McMaster University MASTER OF SCIENCE (2020)

Hamilton, Ontario (Health Research Methodology)

TITLE: Psychometric Validation of the CLEFT-Q Patient Reported Outcome Measure: A Prospective Study to Examine Construct Validity and Responsiveness Following Four Cleft-Specific Operations

AUTHOR: Anna Miroshnychenko, H.B.Sc.

SUPERVISOR: Dr. Anne Klassen, Professor

NUMBER OF PAGES: x, 65

ABSTRACT

CHAPTER 1: Introduction: The most common craniofacial congenital anomaly is the cleft lip and/or palate (CLP). The CLEFT-Q is the first condition-specific comprehensive patient reported outcome instrument (PROM) for patients with CLP. Other measures used in assessment of patients with CLP are Child Oral Health Impact Profile (COHIP) and Cleft Hearing, Appearance and Speech Questionnaire (CHASQ). The development and validation of the CLEFT-Q have been completed in three phases. In phase I, 138 patients with CLP from six countries were interviewed, and data were used to form 13 scales measuring appearance, facial function and health-related quality of life (HR-QOL). In phase II, scales were field-tested internationally with 2434 patients to examine reliability and validity as well as develop a common scoring algorithm for international use. Phase III, the focus of this thesis, aimed to examine further construct validity and responsiveness of the CLEFT-Q scales.

CHAPTER 2: Methods: Patients were recruited at six cleft centres in Canada, USA and UK between January 2018 and October 2019. The sample included patients aged 8-29 seeking rhinoplasty, orthognathic, cleft lip scar revision and alveolar bone graft (ABG) operations. Before and six months after surgery, participants were asked to complete the CLEFT-Q scales relevant to their operation and two other PROMs frequently used in cleft research, i.e., COHIP and CHASQ. Cross-sectional construct validity was examined by testing prespecified hypotheses about expected relationships between CLEFT-Q, CHASQ and COHIP instruments. Internal responsiveness was examined using the distribution-based method. Data were analysed using paired sample t-tests and calculation of effect sizes (ESs) and minimally important differences (MIDs).

CHAPTER 3: Results: Examination of cross-sectional construct validity of the CLEFT-Q scales using the COHIP and CHASQ subscale resulted in 39/53 (74%) hypotheses having been supported by the results. The required sample size to examine responsiveness using the anchor-based approach was not reached. Assessment of internal responsiveness using the distribution-based approach demonstrated that the appearance scales were

highly responsive to change following cleft-specific surgeries, with statistically significant results and ESs ranging from 0.4 (small) to 1.8 (large). Change on the CLEFT-Q HR-QOL scales was not statistically significant. As predicted, the ESs on scales measuring facial aspects most affected by rhinoplasty and orthognathic surgeries were larger than the ESs on scales measuring facial aspects least affected by these surgeries. MIDs for each scale in each operation were determined.

CHAPTER 4: Discussion: Assessment of cross-sectional construct validity demonstrated that CLEFT-Q performs as it was intended when compared with other similar measures (i.e., CHASQ and COHIP). The CLEFT-Q appearance scales were responsive to change following rhinoplasty, orthognathic and cleft lip scar revision operations. As predicted, the CLEFT-Q appearance scales did not detect change following the ABG operation as this operation does not result in visible difference. As hypothesized, the CLEFT-Q HR-QOL scales were less responsive to change than appearance scales as HR-QOL is a more distal construct than appearance in relation to the cleft-related surgeries performed. The preliminary MIDs estimated by the distribution-based approach should be confirmed in studies with diverse CLP populations and larger sample sizes using the anchor-based approach. The findings of this phase III study build on the results of another CLEFT-Q validation study, which demonstrated the ability of the CLEFT-Q scales to detect differences between groups with varying surgical status, i.e., need surgery, have had surgery and never needed surgery.

CHAPTER 5: Conclusion: Cross-sectional construct validity of the CLEFT-Q scales was supported by most prespecified hypotheses. The CLEFT-Q scales were found to be responsive to change. MIDs were determined. The results of this phase III study should be confirmed in a larger and more culturally diverse patient population. Future studies to examine reproducibility and measurement error as well as external responsiveness of the CLEFT-Q scales may be beneficial to add to the psychometric evaluation process.

ACKNOWLEDGEMENTS

First and foremost, I would like to thank Dr. Anne Klassen for her outstanding mentorship and unconditional support throughout the last four years of me studying under her supervision. Dr. Klassen, you have always provided direction in the uncertain and challenging times in the matter of minutes. Your passion for and dedication to your work continues to inspire me tremendously and I feel beyond fortunate to have had the opportunity to learn from you. I hope to continue building my academic career under your mentorship. Furthermore, I am incredibly grateful to you for connecting me with many inspiring experts in the field, whom I now have the pleasure of working with on exciting projects. Thank you from the bottom of my heart for all your efforts and, importantly, for spoiling your students with delicious meals at our team meetings.

To my supervisory committee, Drs. Achilleas Thoma and Karen Harman, you have been invaluable sources of knowledge, guidance and support throughout my Master's degree. Dr. Thoma, thank you for taking me on as your student to complete my internship in health economics and plastic surgery. Thank you for nudging me to expand my methodological expertise by taking the health economics and health technology assessment courses. I am beyond grateful to have had the opportunity to produce publications and present at the Canadian Society of Plastic Surgeons under your supervision. Dr. Harman, your renowned expertise in cleft care have been integral to the development of the CLEFT-Q instrument. Thank you for sharing your knowledge, perspective and much needed guidance throughout this project. I feel very fortunate to have your support, Drs. Thoma and Harman.

To the Health Research Methodology (HRM) program, completing my Master's degree at the birthplace of evidence-based medicine among world-renowned experts has been a privilege. The Health Research Methods, Evidence and Impact (HEI) department is known for its impeccable ability to produce, synthesize and share the best available research evidence in the health and health-related fields. I am exceptionally thankful to the HRM program for giving me the opportunity to be a part of this work.

To Lorraine Carroll and Kristina Vukelic, thank you both for your tireless support and guidance with the logistical aspects of my Master's degree. Your dedication is tremendously appreciated and valued.

To Charlene Rae, the research coordinator within our team, thank you for your invaluable support and guidance with this project and others. Thank you for sharing your immense amount of statistical knowledge and for mentoring me through the complex statistical analyses that we have undertaken together.

To Dr. Karen Wong, who spearheaded the CLEFT-Q patient reported outcome measure development, thank you for creating such a valuable and robust instrument to improve the lives of many children and adults with craniofacial conditions. Thank you for providing your indispensable guidance and support throughout the third phase of the CLEFT-Q development.

To the wonderful research teams around the globe who have collaborated with us, thank you for your continued support. Drs. Goodacre, Swan, Goldstein, Slator and Shearer, thanks to you and your teams for being persistent with your efforts and diligent with your communication with us despite the distance. A special thank you to Drs. Christopher Forrest, Emily Ho and John Phillips and the rest of the Plastic and Reconstructive surgery team at the Hospital for Sick Children who have kindly supported me through the data collection process and provided the necessary tools to be as successful in reaching our recruitment goals as possible.

To my parents, I do not know where to begin thanking you for your dedication to raising my sister and I into the resilient, intelligent, capable and loving people that we are. I am forever indebted to you for your support throughout my many victories and challenges. I am forever grateful for your decision to leave your lives behind and start anew in Canada. I love you both beyond measure.

To my little sister who is growing up too fast, thank you for your dedicated support and advice. Although young, you are wise beyond your years and I am blessed to have you by my side. My love for you is boundless.

TABLE OF CONTENTS

LIST OF FIGURES

Figure Number and Title **Page**

LIST OF TABLES

Table Number and Title **Page**

LIST OF ABBREVIATIONS

Abbreviation	Definition
OFC	Orofacial cleft
CLP	Cleft lip and/or palate
ABG	Alveolar bone graft
PRO	Patient reported outcome
PROM	Patient reported outcome measure
YQOL-FD	Youth Quality of Life – Facial Differences Module
COHIP	Child Oral Health Impact Profile
QOL	Quality of life
OHR-QOL	Oral health-related quality of life
COHIP-SF-19	Child Oral Health Impact Profile–Short Form 19
CLEP	Cleft Lip Evaluation Profile
CHASQ	Cleft Hearing Appearance and Speech Questionnaire
SWA	Satisfaction with Appearance
ICHOM	International Consortium for Health Outcome Measurement
CAT	Computerised adaptive testing
CTT	Classical test theory
IRT	Item response theory
RMT	Rasch measurement theory
DIF	Differential item functioning
HR-QOL	Health-related quality of life
COSMIN	COnsensus-based Standards for the selection of health Measurement INstruments
MID	Minimally important difference
ES	Effect size
SD	Standard deviation
SRM	Standardized response mean

CHAPTER 1: INTRODUCTION

CLP Epidemiology

Orofacial clefts (OFCs) such as cleft lip, cleft palate and cleft lip and palate comprise the most common craniofacial congenital anomalies in live births [1]. The international incidence of OFC is approximately 10 to 12 children with OFC per 10,000 live births worldwide.2,3 The highest incidence rate is in Japan (19.05 per 10,000 live births) and the lowest is in South Africa (3.13 per 10,000 live births). In Canada, approximately 400 to 500 infants are born with OFC each year [2,3]. The prevalence of OFC in Canada ranges from 11 to 15.3 per 10,000 live births [2]. The distribution of cleft types for live births with OFC is approximately 17% for cleft, 41% for cleft palate and 42% for cleft lip and palate. The cleft lip and cleft lip and palate are more common in males, and cleft palate is more common in females [2]. Birth weight and gestational age is lower for newborns with orofacial clefting than in newborns with no cleft [2].

Complications Associated with CLP

There are several types of cleft lip (i.e., forme fruste unilateral cleft lip, incomplete unilateral cleft lip, complete unilateral cleft lip, incomplete bilateral cleft lip and complete bilateral cleft lip) and cleft palate (i.e., incomplete cleft palate, complete cleft palate and submucous cleft palate). These cleft lip and cleft palate types may result in a variety of issues related to feeding, speaking and hearing [4]. Majority of newborns with cleft lip and/or palate (CLP) have difficulty feeding due to failure to form sufficient negative intraoral pressure, which makes them unable to breast feed or bottle feed with normal bottles [5]. The feeding-related issue may lead to slow weight gain and malnourishment [5]. In turn, malnutrition may result in immunodeficiency and mortality caused by infectious disease in children less than five years old [5].

A prominent complication of CLP is abnormal speech production due to deviations in oronasal and orofacial structure, neuromotor patterns learned during early infancy, and disrupted psychosocial development [6]. Individuals with an unrepaired cleft palate cannot build up pressure in the mouth to make speech sounds that need the palate to close, known as “pressure consonants” [6]. The cleft-specific errors of speech sound

production are classified into two types: obligatory and compensatory [6]. The obligatory errors are mistakes in production due to interference of structural deformities such as a misaligned tooth, a residual cleft or an oronasal fistula [6]. Obligatory errors can be corrected primarily by fixing these structural abnormalities with surgery. Compensatory errors are errors in air flow direction learned by children during the developmental period [6]. These errors can be corrected generally through speech therapy [6].

Abnormal nasal resonance is another feature of CLP [6]. The resonance of speech is governed by the size and shape of the oral, nasal and pharyngeal cavities [6]. The abnormal nasal resonance in CLP could manifest as hypernasality or hyponasality. Hypernasality indicates an excessive nasal resonance that is perceived for vowels and oral consonants [6]. Hyponasality indicates a decreased nasal resonance for vowels and nasal consonants. Abnormal resonance can be caused by structural disturbances such as obstructions in the nasopharynx due to adenoid hypertrophy, swelling of the nasal passages secondary to allergic rhinitis or hypertrophic tonsils, large oronasal fistula and/or velopharyngeal dysfunction [6].

Another common complication of CLP is hearing loss. Hearing loss is the reduction of hearing in any degree that diminishes the comprehensibility of the spoken message for accurate interpretation or learning [7]. Any type of hearing loss can compromise language, learning, cognitive development and social inclusion [7]. The most frequent hearing-related disease in children with CLP is otitis media with effusion, which can be responsible for delayed acquisition of language as well as cognitive and psychosocial development [7].

Evidently, the multitude of complications associated with CLP affect the individual's appearance, facial function, psychological and social well-being, potentially for the entirety of their life.

Treatment of CLP

As a result of a multitude of issues that may arise in patients with CLP, the course of pediatric CLP treatment is extensive and complex, beginning at the time of diagnosis and continues to early adulthood. Presurgical options consist of lip taping, lip adhesion and palatal devices that assist in maximizing tissue positions prior to lip repair [8]. The first surgical procedure a child with CLP undergoes is lip repair, which takes place after two to three months of age [8,9]. The cleft palate repair is generally the second surgical procedure performed after first six to 12 months of age [8,9].

This thesis evaluates the outcome of four cleft-specific surgeries taking place between eight and 29 years of age including alveolar bone graft (ABG), orthognathic surgery, rhinoplasty and cleft lip scar revision. The ABG operation usually takes place between nine and 12 years of age mainly to provide a bone scaffold for cleft tooth eruption [8]. Furthermore, ABG maintains palatal width, completes the alveolar ridge, serves as a bone base to the nostril sill and ala, and effectively closes the oronasal fistulas [8]. The ABG operation is not meant to result in visible change to the facial appearance. Orthognathic surgery is performed between 12 and 21 years of age when the midface and mandibular growth have occurred, all teeth are in and the orthodontics have maximised tooth position for occlusal purposes [8]. The most common orthognathic operation is the Le Fort I osteotomy [8]. Severe cases of class III malocclusion, a gap greater than 7 to 10 mm, are treated with a combination of Le Fort I osteotomy and bilateral sagittal split osteotomy [8]. Clinical effectiveness following the orthognathic operation has been reported in numerous studies [10,11,12]. In the case of a cleft lip, certain features are characteristic of the nose, including (but not exclusive to) poor tip projection and definition, a widened nostril, alar malposition and flattening, uneven alar base, shortened columella, dislocated and flattened lower lateral cartilage and fibrofatty thickening of the tip-lobule complex [8]. These features are corrected with an open-technique rhinoplasty or septorhinoplasty. Clinical effectiveness following the rhinoplasty operation has been demonstrated in various studies [13,14,15,16]. As children with CLP enter social systems (i.e., school), cleft lip and nose deformities are sometimes revisited [8]. Revisions of the

lip and nose are considered to address widened scars, vermillion mismatch, shortened lip segments, flattened ala, flattened nasal tip, lip soft tissue paucity, whistle deformity as well as mucosal lip contour irregularities [8]. Several studies have been conducted to determine the clinical effectiveness of the cleft lip scar revision surgery [17,18]. Additional surgical procedures include speech correcting surgeries and various forms of orthodontic treatment that are performed at different ages.

Numerous techniques and their variations are used at different centers by different surgeons for each CLP-related procedure [19]. For instance, the most relevant and useful techniques to correct a cleft palate include von Langenbeck's bipedicle flap technique, Veu-Wardill-Kilner Pushback technique, Bardach's two-flap technique, Furlow Double opposing Z-Plasty, two-stage palatal repair, hole in one repair, raw area free palatoplasty, alveolar extension palatoplasty, primary pharyngeal flap, intravelar veloplasty, vomer flap and buccal myomucosal flap [19]. As a result of an abundance of existing surgical techniques, there is a substantial variation in the treatment protocols for management of patients with CLP between and within countries. This has been demonstrated by the Eurocleft study in the United Kingdom (UK) [20,21,22,23], the Americleft study in North America [24, 25, 26, 27, 28] and Scandicleft study in Denmark, Finland, Norway, Sweden and the UK [29,30,31,32,33,34,35,36,37,38].

Patient Reported Outcome Measures

Patient reported outcomes (PRO)s are reports that come directly from patients about how they function or feel in relation to a health condition and its therapy, without interpretation by a physician or anyone else [39]. Consequently, patient reported outcome measures (PROMs) are tools designed to collect PROs. There are two main types of PROMs: generic and condition-specific. Generic PROMs are designed for use in a broad range of medical conditions, thus allowing for comparisons across conditions or with population norms [40]. Condition-specific PROMs allow for assessment of concerns that are specific to a particular condition and its impact on outcome [41]. Often, a combination of both generic and condition-specific tools are used. A shift toward PROs has been

suggested for CLP care and development of condition-specific PROMs for CLP patients has been recommended [42].

A systematic review identified three generic PROMs that have been used to assess CLP such as PedsQL4.0, Child Health Questionnaire and KINDL-R [43]. The authors highlighted that these generic measures focus on mobility, energy and drive, and fail to measure outcomes specific to CLP such as appearance, facial mimics and function, and eating function. Several PROMs have been developed for children with craniofacial conditions including the Youth Quality of Life – Facial Differences Module (YQOL-FD) and the Child Oral Health Impact Profile (COHIP). The YQOL-FD is a multidimensional PROM for older children ages 11-18 that asks about issues specific to a range of craniofacial conditions [44].

The COHIP, used in this thesis, is composed of three domains (i.e., oral health, function and socio-emotional) and examines the impact of oral disease on quality of life (QOL) in children [45]. The COHIP was first published as a 34-item instrument that was validated with a diverse sample of school-aged (ages eight to 17) treatment-seeking children with varying oral conditions, health systems and ethnicities. The COHIP was created by following a multistage process that included psychometric testing, descriptive studies of patient populations, caregiver proxy and child comparisons, and construct validity testing using other PROM instruments [46,47,48,49]. The 34-item COHIP measures four domains (school environment, self-image, socio-emotional well-being and functional well-being) with five subscales (oral health, functional well-being, socio-emotional well-being, school environment and self-image). This oral health-related quality of life (OHR-QOL) instrument is applicable to children and adolescents ages eight to 17. Subsequently, a 19-item short-form version (COHIP-SF-19) was validated with a sample of children seeking pediatric, orthodontic and craniofacial treatment [50]. The COHIP-SF-19 contains three domains 1) oral health (5 items), 2) functional (4 items) and 3) socio-emotional (10 items). Oral health includes items about oral health symptoms (e.g., pain, spots on teeth). Functional well-being is comprised of items related to everyday activities (e.g., speaking, chewing). Socio-emotional well-being relates to peer-

interactions and mood states (e.g., been unhappy or sad, felt worried or anxious) (See Table 1 in Appendix). The response options are as follows: "never" = 0, "almost never" = 1, "sometimes" = 2, "fairly often" = 3 and "almost all the time" = 4. Reliability and validity testing demonstrated that the COHIP-SF 19 is a psychometrically sound instrument in a school-aged pediatric population [50]. Responsiveness of the COHIP-SF 19 has not yet been established. In this CLEFT-Q phase III study, the short-form version of COHIP (i.e., COHIP-SF 19) was used.

Several condition-specific tools for CLP have been developed, including the Cleft Lip Evaluation Profile (CLEP), Cleft Hearing Appearance and Speech Questionnaire (CHASQ), and the CLEFT-Q. The CLEP evaluates the cosmetic and functional outcome after cleft lip and nose operations [51]. The CHASQ, used in this thesis, measures patients' satisfaction with features of appearance, speech and hearing. The CHASQ is a modified version of the Satisfaction with Appearance (SWA) questionnaire [52]. The SWA questionnaire was developed by the Cleft Psychology Special Interest Group of the Craniofacial Society of Great Britain and Ireland for patients with facial disfigurement [53]. The CHASQ has two subscales. The first subscale (i.e., factor 1) includes nine items that ask about features typically affected by a cleft including "face", "whole appearance", "side view/profile", "how good-looking", "nose", "lips", "teeth", "speech" and "how noticeable" (See Table 1 in Appendix). The second subscale (i.e., factor 2) includes six items that ask about features not typically affected by a cleft such as "chin", "cheeks", "hair", "ears", "eyes" and "hearing". The second subscale was not included in this study. Each CHASQ item contains 10 (i.e., 1-10) response options ranging from "very happy" to "very unhappy". Items are summed to produce a total score for each subscale. While SWA questionnaire and CHASQ have been used to measure outcomes in several studies [54,55,56,57], evidence addressing their psychometric properties has not yet been published.

Use of Patient Reported Outcome Measures

The importance of PROMs is recognized worldwide for their versatile application. PROMs are valuable tools in health services research [58]. The most common use of PROMs is as primary and secondary end point measures in clinical trials to evaluate new drugs, procedures and technologies. Importantly, PROMs are proving to be an effective method for establishing gold standards and treatment protocols that are consistent within and between countries. A survey of 100,000 clinical trials reported use of PROMs in 27% of studies [59]. Additionally, PROMs are used in audits of programs of care. For instance, the BREAST-Q was used in the UK by the National Health System as the main outcome instrument in a large-scale prospective national audit of breast cancer surgery (mastectomy and reconstruction) [60]. Furthermore, PROMs may be used as prognostic tools, especially in cancer-related research, and are increasingly integrated into clinical care with electronic data collection and real-time generation of patient reports [61,62,63]. The use of PROMs in global benchmarking initiatives is increasing as well. The CLEFT-Q scales were included in the International Consortium for Health Outcome Measurement (ICHOM) standard set for CLP, craniofacial microsomia and pediatric facial palsy [64]. Hospitals around the world using the standard sets are encouraged to share collected data as part of a global benchmarking initiative. Innovative methods are being applied to administration of PROMs to reduce the burden associated with data collection. The computerised adaptive testing (CAT) version of the CLEFT-Q scales has been developed to ease their integration into practice and research initiatives through shortening of the number of questions by 61% (i.e., from 110 to a mean of 43.1 (range 34–60, SE < 0.55)), while maintaining 97% correlation between scores obtained with full-length scales and CAT [65]. As virtual care is becoming more prevalent, electronic administration of PROMs is essential for these tools to be effective in evaluating patients' health and quality of care.

Methods Used to Develop Patient Reported Outcome Measures

The process to develop a scientifically credible and clinically meaningful PROM is multi-phased, iterative and involves a mixed-method approach [66]. The first phase focuses on development of a conceptual model and generation of an item pool. The second phase pertains to testing of the PROM in a large sample of the target population. The items that most accurately predict the outcome of interest are selected according to their performance on a range of psychometric tests. Since validation of a PROM is an ongoing process, the third phase further tests the item-reduced instrument in the target population to examine the same or additional psychometric properties [66]. The main psychometric properties include validity, reliability and responsiveness.

Psychometric Theories Underly Patient Reported Outcome Measure Development

There are two main psychometric theories that underly development of PROMs [67]: classical test theory (CTT) and item response theory (IRT). The CTT approach operates on the assumption that the error score for each item in the instrument is uncorrelated with the true score (68,69). That is, the variation in error is equal for all values of the true score and the average error, summed over all items, is zero [68,69]. Under this assumption, reliability of the scale increases as the number of items and the correlation among these items increases [67]. There are several limitations associated with the CTT approach. An important limitation relates to sample dependency, where the item and scale statistics apply only to the specific group of participants who completed the test during the validation process. If the scale was administered to individuals with a different diagnosis, or if the scale was shortened, the psychometric properties of the scale would have to be re-established [70]. Another limitation with the CTT approach is the assumption of item equivalence [71]. The CTT approach assumes that each item contributes equally to the final score [71]. In other words, unless different weights are attached to each item, the total score is simply the sum of the scores of the individual items, regardless of how well each item correlates with the underlying construct. Item statistics and clinical judgement suggest that some items are more important in measuring

the attribute than others, however CTT does not account for this in the scale. According to the CTT approach, summing items to create a total score assumes that all items are measured on the same scale. However, this assumption is often untrue; items are ordinal rather than interval and therefore the distance between response options varies from one item to the next [72].

The CLEFT-Q was developed using the item response theory (IRT), more specifically, the Rasch measurement theory (RMT) approach. This approach overcomes limitations of CTT. The Rasch model is a particular type of the IRT model, i.e., the one parameter model [73]. In this approach, Rasch analysis is used to examine the difference between observed and predicted item responses to determine whether data collected from a sample fit the Rasch model. According to the RMT, data must fit the requirements of the Rasch model to provide meaningful measurement [73]. A scale developed with the RMT is analogous to a ruler with the items lined up in a clinical hierarchy from a low to high 'amount' of the construct. The mathematical model that underlies the Rasch model produces a scale with interval-level measurement properties. When data fit the Rasch model, the scale provides person estimates that are independent of the sampling distribution, therefore it can be used in different subsets of the target population [73]. In scale development, items that do not fit the Rasch model can be identified. For example, if differential item functioning (DIF) occurs, whereby one subset of a population answers items differently than another subset despite having the same amount of the trait, the items can be dropped or kept with adjustments made to the scoring. The RMT analysis makes it possible to identify the best subset of items to retain in a scale to maximize its psychometric properties [73].

Reliability

The reliability property diverges into three concepts of internal consistency, reproducibility and measurement error [74,75,76]. Internal consistency examines the extent to which all items in a scale measure the same concept [75]. When a scale measures a single concept, it is considered to be 'unidimensional'. Reproducibility is the ability of

a measure to provide reproducible results, which can be assessed through intra-rater reliability, inter-rater reliability and test-retest reliability [75]. Intra-rater reliability examines how similar the scale scores provided by an observer are on two or more different occasions [75]. Inter-rater reliability determines the level of agreement between two or more observers providing ratings on a scale [75]. Test-retest reliability assesses whether the participant will score similarly on two different occasions [75]. Measurement error indicates how precise the measurement of each of the three reliability tests is [75].

Internal consistency of the CLEFT-Q scales was examined in the second phase of development [77]. Reproducibility and measurement error of the CLEFT-Q scales have not yet been addressed.

Validity

Validity is the ability of an instrument to measure what it intends to measure [78]. Validity can be classified into three main concepts: content validity, criterion validity and construct validity [78]. Content validity of the CLEFT-Q scales has been established during the initial qualitative phase [66,79]. This attribute measures the extent to which the content of a measurement tool adequately represents the concepts of interest for a patient population [80,81,82]. Content validity is determined by examining if the content of an instrument is comprehensive, comprehensible and relevant [82]. Collecting input from patients throughout the developmental process ensures that the content of the scale is comprehensive and valid. An aspect of content validity is face validity, which determines if the scale items appear on the surface to be measuring what they are intended to measure [78].

Criterion validity refers to the correlation of a scale with another measure of the trait or disorder being studied, a ‘gold standard’, which has been used and accepted in the field. Criterion validity is divided into concurrent validity and predictive validity [78]. Assessment of concurrent validity tests whether the new scale is correlated with the ‘gold standard’ measure; both scales are administered at the same time [78]. Assessment of predictive validity tests the extent to which the scores of the tool predict the scores of the

gold standard. A Delphi panel reached consensus that a gold standard does not exist for PROMs that measure health-related quality of life (HR-QOL) [83]. The COSMIN (COnsensus-based Standards for the selection of health Measurement INstruments) guideline suggested that when the scores of a new instrument are compared to one or several widely used PROMs, construct validity, instead of criterion validity, is being assessed with hypotheses about the magnitude and direction of the correlation between the instruments being formulated and tested [83].

Construct validation establishes the degree to which a PROM works as it is intended to work based on prior knowledge about the constructs being studied [78]. Construct validation is a continuous process of learning about the construct. There is no one single experiment that tests construct validity, but rather each supportive study serves to strengthen the network of predictions of a theory [78,84]. Construct validity involves the following: 1) identifying the theoretical concepts and their relatedness to each other, 2) developing or identifying scales that measure these constructs, and 3) testing the correlations among these constructs [78,84]. In summary, construct validity is a framework of hypotheses testing based on the knowledge of the underlying construct. The validation process asks whether the empirical findings correspond with the theoretical expectations about the instrument [78]. Cross-sectional construct validity examines hypotheses about correlations of scores of measures with related constructs at one point in time, whereas longitudinal construct validity focuses on change scores (i.e., between two or more points in time) [85,86,87].

Construct validity of the preliminary CLEFT-Q scales was first addressed in the field-test publication [77]. In this thesis, cross-sectional construct validity of the final CLEFT-Q scales was examined by comparing the scores of the CLEFT-Q to the scores of the COHIP and CHASQ collected at baseline.

Responsiveness to Change

Responsiveness is the ability of an instrument to detect clinically important change over time [88]. There are two types of responsiveness: internal and external [89].

Internal responsiveness refers to the ability of a measure to detect statistically significant and clinically important change over a prespecified time frame, in which the construct examined changes spontaneously or due to receiving treatment [87,90]. External responsiveness refers to the extent to which the score changes in a new instrument relate to the score changes in another outcome measure examining the trait or disorder over a prespecified time frame [87]. In this thesis, internal responsiveness of the CLEFT-Q scales was examined by determining the statistically significant and clinically important change following four cleft-related surgeries.

The minimally important difference (MID) is the smallest difference in scores in the outcome of interest that patients or proxies perceive as an important deterioration or improvement [91,92,93]. There are several reasons for the usefulness of the MID including the following: 1) MIDs are easily understood by researchers and clinicians, 2) MIDs prioritize the patient's perspective, 3) MIDs can inform judgements about the successfulness of an intervention, 4) MIDs help to estimate the sample size for clinical trials and inform the design of the study, and 5) individuals achieving a score equal to or greater than the MID may be considered as the beneficiary of an intervention [92,93]. A disadvantage of MIDs is that estimates are known to vary across patients and patient groups, and therefore should be applied only to the patient population for which the estimate was calculated [92,93,94]. Establishing MIDs for each CLEFT-Q scale in this thesis may provide a basis for estimating sample sizes in future studies and assist healthcare professionals in interpreting the meaning of changes in scores obtained from the scales [95,96,97].

There is no consensus on the best method to examine responsiveness of an instrument. Several strategies have been used to achieve an understanding of the meaning of scale scores [98]. One of the strategies refers to the anchor-based approach and examines the relationship between scores on the target instrument and some independent measure called the anchor. Anchor-based methods compare changes in PROM scores to an anchor that is interpretable and requires at least moderate correlation of the change on

the anchor with the change on the target instrument. A quality tool to assess the credibility of an anchor for estimates of MIDs for PROMs has been developed [99].

Another strategy for establishing responsiveness is termed distribution-based and relies solely on the statistical characteristics of the obtained PROM scores. The distribution-based method interprets results in terms of the relationship between the magnitude of effect and some measure of variability in the scores [100]. There are three distribution-based approaches. The first approach relies on statistical significance and examines the score change in relation to the probability that this change is a result of a random variation of scores. Examples of this approach are the paired t-statistic and growth curve analysis [87;101]. The paired t-test approach has been employed in this thesis. The second approach examines the score change in relation to sample variation using either the baseline standard deviation (SD), variation of change scores or variation of change scores in a stable group [102,103,104,105]. The third approach examines the score change in relation to measurement precision. Examples of this approach are standard error of the mean and a reliable change index [106,107]. The investigators may choose between-patient variability or within-patient variability as a measure of variability.

The most commonly used distribution-based method is based on the between-person SD, often referred to as the effect size (ES) [102,103]. There are several distribution-based variants of the ES. Cohen's ES is the ratio of the mean difference to the SD of baseline scores [102]. The denominator of the Cohen's ES is the SD at baseline of the control group, referred to as the Glass's delta, or the pooled SD at baseline of the treatment and control groups, referred to as Cohen's d [108]. Cohen's d is a more stable estimate of the SD as it uses all data. Based on Cohen's criteria, a difference of 0.2 SD units represents a small change, 0.5 a moderate change and 0.8 or above a large change [102]. These criteria can be applied to the change observed in a single group from pre-test to post-test scores of treatment and control groups, or the difference between changes in scores of treatment and control groups. Guyatt's measure is specific to a pre-test/post-test two group design and is a variant of the ES [109]. Guyatt's measure is the ratio of the mean change in the treatment group to the SD of the change score in the control group

[109]. The standardized response mean (SRM), another version of the ES, is the ratio of the mean change in a single group to the SD of the change scores [110]. In this thesis, both the Cohen's d and SRM were calculated to determine the magnitude of CLEFT-Q scale score change following surgery as a gold-standard approach has not been established.

There are several conceptual problems with the responsiveness psychometric property. Any PROM is more sensitive to large treatment effects than to small ones, therefore it is difficult to untangle the characteristics of a PROM from the characteristics of the treatment effect [76]. A possible solution is to administer multiple measurement tools to the same group of patients, including the tool undergoing the validation process along with the existing validated tools [76]. The second solution is to follow a group of patients for a period of time and administer the measurement tool before and after the intervention [76]. The patients may then be asked individually if they got better, stayed the same or got worse. The average change on the instrument for those who got better may then be computed. The challenge with this latter approach is that in any cohort of subjects followed over time, some individuals will get better or worse merely by chance, due to factors that cannot be understood or controlled [111]. Both of these solution methods have been applied to the methodology of the study reported in this thesis.

The CLEFT-Q

The development of the CLEFT-Q followed a modern psychometric approach (i.e., IRT) and was conducted in three steps (see Figure 1) [66]. This approach involves collection of data and analysis to examine whether the data fit the mathematical Rasch model. The CLEFT-Q was developed to measure concerns of patients from high-, middle, and low-income countries. The intention was to engage clinician stakeholders during the entire developmental process and form a network of collaborators who would use the CLEFT-Q in their line of work. Prior to development of the CLEFT-Q scales, a systematic review of the literature was performed to determine if a condition-specific instrument for patients with CLP already existed, and to discover PROMs that have been validated and

used in patients with CLP to identify a preliminary conceptual framework with categories that included physical, psychological and social health [43].

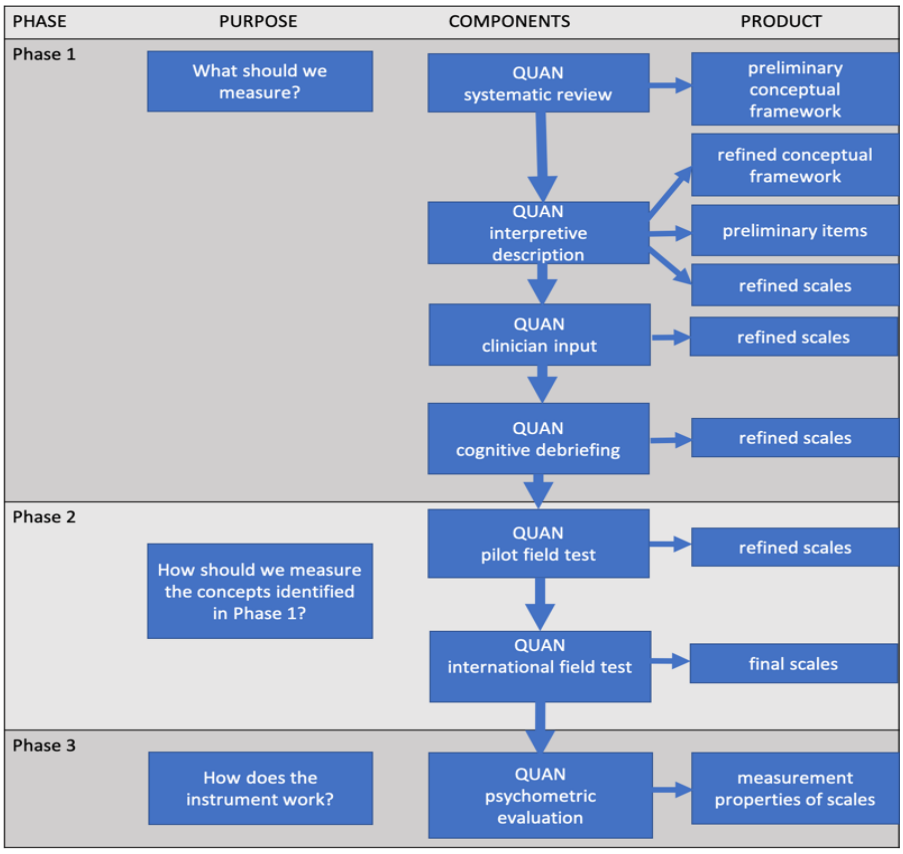

Figure 1. There are 3 main phases to developing a PROM, including item generation, item reduction and psychometric evaluation. Adopted from "International multiphase mixed methods study protocol to develop a cross-cultural patient-reported outcome instrument for children and young adults with cleft lip and/or palate (CLEFT-Q)", by Riff KWY, Tsangaris E, Goodacre T, et al., 2017, *BMJ Open*, 7: e015467.

Phase I: Pretesting the CLEFT-Q Scales

Phase I of the CLEFT-Q development consisted of identifying concepts that were important to patients with CLP from their perspective, developing a conceptual framework based on these concepts, and creating CLEFT-Q scales to measure patients' concerns [66,79]. The study followed the qualitative methodology of Interpretive Description, which assumes that theoretical knowledge, clinical knowledge and a scientific basis inform the study [66]. A total of 136 interviews with 138 participants across six countries (Canada, Kenya, India, Philippines, England and USA) were

conducted. Patient interviews were transcribed verbatim. An extensive list of potential scale items was created by coding qualitative data generated through patient interviews. Three main domains emerged from the interviews and formed the basis of the refined conceptual framework, with sub-domains within each of these top-level domains. The three domains with subdomains included: appearance (face, nose, nostrils, teeth, lips, jaws, cleft lip scar), HR-QOL (psychological, social, school, speech-related distress) and facial function (speech, eating/drinking).

Each sub-domain is measured by an independently functioning scale. The content for each scale was designed using patients' own words, while maintaining the lowest possible grade reading level. Positive or neutral language was used to minimize any potential negative impact caused by completing the scales. After the preliminary scales were formed, cognitive interviews with 69 patients were performed to ensure that patients understand the items on the scales and that relevant concepts were not missing [79]. Furthermore, 44 specialists in cleft care from multiple countries provided expert input. The refined versions were translated into multiple languages in preparation for the international field-test study [112,113].

Phase II: Field-testing and Psychometric Evaluation of the Preliminary CLEFT-Q Scales

Phase II of the CLEFT-Q development involved a field-test study that included 2,434 patients from 30 centres in 12 countries. In the field-test, participants with CLP between the ages of eight and 29 years were recruited to complete the CLEFT-Q scales. Individuals with a cognitive delay were excluded [77]. The RMT analysis was performed to determine which items in each scale were most effective in measuring the concepts of interest [77]. In phase II, the RMT analysis was used to measure performance of scale items and to determine the measurement properties of the scales.

Construct validity encompasses components of structural validity, which examines internal relationships, hypotheses testing and cross-cultural validity [114]. Structural validity and cross-cultural validity were addressed in the RMT analysis with

unidirectionality and DIF. Several hypotheses to assess construct validity were tested: 1) patients with a visible difference would report lower scores on appearance scales compared to those with an invisible difference, 2) patients undergoing speech surgery or speech therapy would report lower scores on the speech scales than those not requiring speech intervention, 3) patients requiring nose, lip or jaw surgery would report lower scores on the appearance and HR-QOL scales compared to those not requiring such treatment, 4) patients who rated their overall appearance or speech to be higher on a 4-point scale would have higher scores on the appearance or speech and HR-QOL scales, and 5) patients who were receiving psychological counseling or therapy would report lower scores on the HR-QOL scales [77].

Phase III: Psychometric Evaluation of the Final CLEFT-Q Scales

The CLEFT-Q validation process was integrated into each stage of the development. In phase I, content validity was examined. In phase II, structural and cross-cultural validity was evaluated. Phase III, the focus of this thesis, aimed to continue the psychometric evaluation of the CLEFT-Q scales by examining: 1) cross-sectional construct validity and 2) internal responsiveness (See Table 2).

Table 2. Checklist for psychometric property assessment of CLEFT-Q scales based on COSMIN recommendations. Phase III forms the basis of this thesis.

Psychometric property	**Examination Completed**	**Examination Required**
Validity		
Content validity	✓ (Phase I)	
Criterion validity	Not applicable	
Cross-sectional construct validity		**✓ (Phase III)**
Longitudinal construct validity		✓ (Not done)
Structural validity	✓ (Phase I)	
Cross-cultural validity/measurement invariance	✓ (Phase II)	
Reliability		
Internal consistency	✓ (Phase II)	
Reproducibility		✓ (Not done)
Measurement error		✓ (Not done)
Responsiveness		
Internal responsiveness		**✓ (Phase III)**
External responsiveness		✓ (Not done)

CHAPTER 2: METHODOLOGY

Objectives

The overall objective of this thesis was twofold. We aimed to perform a phase III study to examine psychometric properties of the CLEFT-Q that had not been explored to date. Further, we planned to compare the performance of the CLEFT-Q for these psychometric properties with that of two other PROMs (i.e., COHIP and CHASQ) used in research with patients with CLP. The two specific aims were to examine the following:

Aim 1: Cross-sectional construct validity: whether baseline (pre-treatment) scores collected with CLEFT-Q, COHIP and CHASQ correspond with the theoretically expected scores;

Aim 2: Internal responsiveness: whether change (pre-post treatment) scores collected with CLEFT-Q, COHIP and CHASQ were statistically significant and of clinical importance.

Data Collection

The phase III prospective study was conducted at six cleft centers in Canada, United States of America (USA) and UK including The Hospital for Sick Children, Children's Hospital of Pittsburgh, University Hospitals Birmingham (Queen Elizabeth Hospital Birmingham & Birmingham Women's and Children's Hospital), Great Ormond Street Hospital for Children, Broomfield Hospital as well as Oxford and Salisbury Cleft Centers. The research ethics approval was attained at each participating center prior to commencement of the study.

The CLEFT-Q, COHIP and CHASQ data were collected before and six months after four surgical operations: 1) rhinoplasty, 2) orthognathic, 3) cleft lip scar revision and 4) alveolar bone graft (ABG). Eligible participants were patients ages eight to 29 years at the preoperative assessment undergoing any of the four cleft-related surgeries at any of the six participating cleft centers. Individuals with a cognitive delay were excluded. Patient recruitment and follow up methodology differed at each site based on their preferences and logistics (See Table 3). Most sites collected the pre- and

postoperative data from participants at the hospital during a clinic appointment. Non-respondents were contacted three times by phone and email. If no response was received after the third contact, the non-respondent was considered lost to follow up.

Each participant completed a core set of CLEFT-Q scales that included appearance scales (i.e., face, nose, nostrils), HR-QOL scales (i.e., psychological, social and school) and function checklist (eating/drinking). Additional appearance scales were administered to individuals undergoing the orthognathic, cleft lip scar revision and ABG operations (See Table 4). Several CLEFT-Q scales in Table 4 were excluded from the analysis in this thesis. Eating/drinking checklist was excluded as no scale score can be derived. Speech function and speech distress scales were excluded as few participants with a speech problem were involved in this study. The school scale was excluded as it is only relevant to patients ages eight to 18 years and therefore not completed by the entire sample. All data were entered into a REDCap database hosted at the coordinating site at McMaster University, Canada. Data were downloaded from REDCap into IBM SPSS Statistics for Mac, Version 26.0, for analyses.

Table 3. Recruitment methodology before and after operation at each site.

Site	**Recruiter**	**Location (preop)**	**Data collection**	**Location (postop)**	**Data collection**
The Hospital for Sick Children	Researcher	Hospital	Tablet	Home	Electronic
Children's Hospital of Pittsburgh	Researcher	Clinic/ home	Paper	Clinic/ home	Paper
Broomfield Hospital	Research nurse	Clinic	Paper	Clinic	Paper
University Hospitals Birmingham	Research nurse	Clinic	Paper	Clinic	Paper
Great Ormond Street Hospital	Psycholo-gist	Clinic	Paper	Clinic	Paper
Oxford & Salisbury Cleft Centers	Research nurse	Clinic	Paper	Clinic	Paper

Table 4. CLEFT-Q scales included in assessment of each operation.

Rhinoplasty	**Orthognathic**	**Cleft lip scar revision**	**ABG**
Face*	Face	Face	Face
Nose*	Nose	Nose	Nose
Nostrils*	Nostrils	Nostrils	Nostrils
Psychological*	Teeth	Lips	Teeth
Social*	Jaws	Cleft lip scar	Lips
School*	Lips	Psychological	Psychological
Eating/Drinking*	Speech distress	Social	Social
	Speech function	School	School
	Psychological	Eating/Drinking	Eating/Drinking
	Social		
	School		
	Eating/Drinking		

* core scales

The Analysis

Scale scores for preoperative and postoperative assessments for the CLEFT-Q, COHIP and CHASQ were calculated according to the developers' instructions. Only scales within the CHASQ and COHIP that measured concepts similar to the CLEFT-Q scales examined in this thesis were included in the analysis to be able to draw comparisons. Thus, the COHIP function and oral health domains were excluded. For the CLEFT-Q scales, the raw scores were converted into Rasch transformed scores ranging from 0 to 100. For the COHIP, scores for the socio-emotional domain (ten items) were summed according to the scoring code provided by the authors [45]. The version of the COHIP provided by the developer to our research team was missing an item from the socio-emotional domain. For this item, the mean of the remaining items was imputed for completeness. For the CHASQ, total scores for the first feature (nine items) were computed using the scoring guidelines provided by the authors [115]. The normality assumption was examined through assessment of skewness and kurtosis as well as normality plots with significance tests. The data were considered normally distributed if the Skewness and Kurtosis values were below an absolute value of two [116].

Aim 1: Cross-sectional construct validity: whether baseline (pre-treatment) scores collected with the CLEFT-Q, COHIP and CHASQ correspond with the theoretically expected scores.

Assessment of cross-sectional construct validity was examined through testing of hypotheses about correlations of preoperative scores between the CLEFT-Q, COHIP and CHASQ subscale. A total of 53 hypotheses were composed that were in accordance with the COSMIN recommendations [117]. Spearman correlations between a total of 11 scale scores measuring appearance or HR-QOL were performed to test these prespecified hypotheses. Appearance scales included seven CLEFT-Q scales (face, nose, nostrils, teeth, jaws, lips, cleft lip scar) and the CHASQ subscale. HR-QOL scales included two CLEFT-Q scales (psychological and social) and the COHIP socio-emotional scale.

The cross-sectional construct validity hypotheses were formed based on the following criteria: 1) correlations between appearance scales would be strong, 2) correlations between HR-QOL scales would be strong, and 3) correlations between appearance and HR-QOL scales would be moderate (See Table 5). Correlations were interpreted as follows: <0.30 weak, 0.30-0.50 moderate and ≥0.50 strong [117].

Aim 2: Internal responsiveness: whether change (pre-post treatment) scores collected with the CLEFT-Q, COHIP and CHASQ were statistically significant and clinically important.

Assessment of internal responsiveness was performed though the distribution and anchor-based approaches for a comprehensive analysis. Both analyses were conducted separately for rhinoplasty, orthognathic, cleft lip scar revision and ABG surgeries as each operation focuses on a different aspect of appearance and, therefore, includes a unique set of CLEFT-Q scales. The ABG operation was considered a control operation in the assessment of the ability of the CLEFT-Q appearance scales to detect change, as this operation is not known to result in visible change in appearance.

To examine whether cleft-related surgery results in statically significant change in scale scores, two distribution-based approaches were used as a gold standard analysis has not been identified. Table 6 in Appendix provides a list of formulas used to calculate the distribution-based statistics used in this thesis. Paired sample t-tests were conducted to compare the pre- and postoperative scores of the CLEFT-Q scales, COHIP and CHASQ subscale. The magnitude of change was examined by calculating two parametric ES estimators: 1) Cohen's d and 2) SRM. Cohen's d was interpreted using the Cohen's criteria: 0.20-0.49 small, 0.50-0.79 moderate and ≥ 0.80 large [102,118].

The first internal responsiveness hypothesis was such that the ESs on the appearance scales would be larger than on the HR-QOL scales. The second internal responsiveness hypothesis was such that the ESs on the appearance scales examining aspects of appearance most affected by the operation would be larger than on the appearance scales examining aspects of appearance less affected by the operation. For instance, the magnitude of change on the jaws scale following the orthognathic operation would be larger than on the nose scale. The ESs on the appearance scales following the ABG operation were hypothesized to not be statistically significant as this operation is not known to result in visible change in appearance.

The clinically important change (i.e., MID) was calculated with both the distribution and anchor-based approaches as a gold standard approach has not been established. Two distribution-based approaches were conducted by calculating scale score change that pertains to 1) 0.5 ES and 2) ½ SD of the preoperative mean scores.

The anchor-based method to determine the MID was conducted by associating a PROM score with an independent measure, i.e., "anchor", that is understandable and relevant to patients. A transition-rating anchor was included in the postoperative assessment booklet for each type of surgery. For example, the rhinoplasty booklet asked: "How does your nose look now compared to before you had your nose operation?". Five response options were provided as follows: "a lot worse now than before my nose operation", "a little worse now than before my nose operation", "about the same as before

my nose operation", "a little better than before my nose operation", and "a lot better than before my nose operation". The MIDs for CLEFT-Q nose, jaws and cleft lip scar scales were determined by calculating the difference in pre- and postoperative scores for participants who answered that their nose/jaws/cleft lip scar looked a little better now than before the operation. The MIDs were hypothesized to be ½ SD as this value has been a frequently reported MID for QOL scales [119].

Table 5. Hypotheses for direction and magnitude of correlation between the CLEFT-Q, CHASQ and COHIP PROMs.

	CLEFT-Q Appearance	**CLEFT-Q HR-QOL**	**COHIP**	**CHASQ**
CLEFT-Q Appearance	Strong correlation (≥0.5)	-	-	-
CLEFT-Q HR-QOL	Moderate correlation (0.3<x<0.5)	Strong correlation (≥0.5)	-	-
COHIP	Moderate correlation (0.3<x<0.5)	Strong correlation (≥0.5)	-	-
CHASQ	Strong correlation (≥0.5)	Moderate correlation (0.3<x<0.5)	Moderate correlation (0.3<x<0.5)	-

Sample Size Calculation

The phase III study sample size calculation aimed to detect an ES of 0.50, with $p<0.05$ and power of 0.80. The aim was to recruit a sample size of 50 participants for each surgery group. A problem of fulfilling the calculated sample size was not anticipated, as the Hospital for Sick Children alone performs 45 orthognathic, 40 rhinoplasty and 30 cleft lip scar revisions annually.

CHAPTER 3: RESULTS

Descriptive Data

Sample characteristics are shown in Table 7. A total of 177 participants were included in the Phase III study. A total of 52 participants did not complete the postoperative assessment. The non-respondents to the postoperative follow up were similar to respondents in terms of mean age and type of operation but were more likely to be male ($\chi^2 = 4.39$, $p = 0.04$) and to live in England rather than Canada or the USA ($\chi^2 = 10.52$, $p = 0.005$). The time between pre- and postoperative assessments ranged between five months for rhinoplasty to seven months for ABG operation.

The type of operation for the 125 respondents who provided the pre- and postoperative data were as follows: rhinoplasty (n=31), orthognathic (n=21), cleft lip scar revision (n=18) and ABG (n=58).

Table 7. Characteristic of participants in the CLEFT-Q phase III study.

Characteristic	**No. of participants at baseline (%) n=177**	**No. of non-respondents (%) n=52**
Country		
Canada	69 (39.0%)	19 (36.5%)
England	70 (39.5%)	28 (53.8%)
USA	38 (21.5%)	5 (9.6%)
Age, yr		
8-11	74 (41.8%)	24 (41.8%)
12-15	24 (13.6%)	6 (13.6%)
16-20	52 (29.4%)	10 (29.4%)
≥ 21	27 (15.3%)	12 (15.3%)
Gender		
Female	70 (39.5%)	15 (28.8%)
Male	107 (60.5%)	37 (71.2%)
Student		
Yes	142 (80.2%)	40 (76.9%)
No	35 (19.8%)	12 (23.1%)
Cleft type		
Cleft lip only	8 (4.5%)	5 (9.6%)
Cleft palate only	3 (1.7%)	0 (0%)
Cleft lip and palate	140 (79.1%)	38 (73.1%)
Cleft lip and alveolus	24 (13.6%)	8 (15.4%)

Missing	2 (1.1%)	1 (1.9%)
Current speech problem		
No speech problem	95 (53.7%)	28 (53.8%)
Mild speech problem	63 (35.6%)	15 (28.8%)
Moderate speech problem	9 (5.1%)	5 (9.6%)
Missing	10 (5.6%)	4 (7.7%)
Syndrome or craniofacial anomaly		
Yes	10 (5.6%)	3 (5.8%)
No	161 (91.0%)	45 (86.5%)
Missing	6 (3.4%)	4 (7.7%)
Operation type		
Rhinoplasty	38 (21.5%)	7 (13.5%)
Orthognathic	27 (15.3%)	6 (11.5%)
Cleft lip scar	28 (15.8%)	10 (19.2%)
ABG	84 (47.5%)	29 (55.8%)

Cross-sectional Construct Validity

The analysis to examine the first aim of this thesis (whether baseline scores collected with the CLEFT-Q, COHIP and CHASQ correspond with theoretically expected scores) included a robust sample of 177 participants. Spearman correlations and number of participants included in each analysis are shown in Table 8. A total of 11 scale scores from the CLEFT-Q scales, COHIP and CHASQ subscale were correlated. Correlations between the cleft lip scar scale and the jaws and teeth scales were not possible, as no participant who completed the cleft lip scar scale was asked to complete either of the other two scales. Of 53 correlations, 39 (74%) aligned with the predetermined hypotheses. Below, the findings are described in more detail.

1. Correlations between appearance scales

Correlations between the appearance scales were expected to be strong ($r \geq 0.50$). A total of 26 correlations were performed to compare the eight appearance scales. Of 26 hypotheses, 20 (71%) were supported by the results. Six of the seven hypotheses to examine correlations between the CHASQ subscale and CLEFT-Q appearance scales were supported ($r \geq 0.5$, $p=0.01$). The exception was the correlation between the CHASQ

subscale and CLEFT-Q teeth scale, which was slightly weaker than predicted (r=0.46, p=0.01). Thirteen of 20 hypotheses testing correlations amongst the CLEFT-Q appearance scales were supported (r≥0.5, p=0.01). Of the remaining seven correlations, four were moderate (0.3<r<0.5, p=0.01) and one was weak (r<0.3, p=0.01).

2. *Correlations between HR-QOL scales*

Correlations between the three HR-QOL scales were expected to be strong (r≥0.50). Of three hypotheses tested, two were supported by the results. The hypotheses comparing the two CLEFT-Q scales, and between the COHIP and CLEFT-Q social scale were supported (r≥0.5, p=0.01). The correlation between the COHIP and CLEFT-Q psychological scale was lower than predicted (0.3<r<0.5, p=0.01).

3. Correlations between appearance and HR-QOL scales

Correlations between the appearance and HR-QOL scales were expected to be moderate (0.3<r<0.5). Of 24 hypotheses, 17 (71%) were supported by the results. In the analyses between the CLEFT-Q scales, 10 of 14 hypotheses to evaluate correlations between the CLEFT-Q appearance and CLEFT-Q psychological and social scales were supported by the results (0.3<r<0.5, p=0.01). The three exceptions were between the HR-QOL scales and cleft lip scar and teeth scales, which were weaker than expected (0.3<r, p=0.01), and the correlation between the psychological and face scales, which was stronger than expected (r≥0.5, p=0.01).

In the analyses between the CLEFT-Q appearance scales and COHIP, six of seven hypotheses were supported (0.3<r<0.5, p=0.01). The exception was a weaker than predicted correlation between the nostrils scale and COHIP (0.3<r, p=0.01). In the analyses between the CLEFT-Q HR-QOL scales and CHASQ subscale, both hypotheses were supported by a moderate correlation (0.3<r<0.5, p=0.01). Finally, the correlation (r≥0.5, p=0.01) between the CHASQ subscale and COHIP was stronger than expected and did not support the hypothesis.

Table 8. Correlations of preoperative scores of patients undergoing rhinoplasty, orthognathic, cleft lip scar revision and ABG operation.

		CLEFT-Q Appearance							CLEFT-Q HR-QOL		COHIP	CHASQ
		Face	**Jaws**	**Lips**	**Nose**	**Nostrils**	**Scar**	**Teeth**	**Psych**	**Social**	**SE**	**-**
Face	Spearman	1	0.75**	0.71**	0.63**	0.60**	0.32*	0.56**	0.53**	0.47**	0.47**	0.67**
	N	177	25	135	175	174	28	108	173	172	165	165
Jaws	Spearman	0.75**	1	0.67**	0.57**	0.53**	.	0.70**	0.41**	0.37*	0.39**	0.73**
	N	25	25	25	25	25	0	25	24	24	24	24
Lips	Spearman	0.71**	0.67**	1	0.60**	0.59**	0.56**	0.56**	0.44**	0.38**	0.42**	0.67**
	N	135	25	135	134	135	28	107	134	133	127	128
Nose	Spearman	0.63**	0.57**	0.60**	1	0.62**	0.38*	0.39**	0.39**	0.33**	0.35**	0.56**
	N	175	25	134	175	173	28	107	172	171	164	164
Nostrils	Spearman	0.60**	0.53**	0.59**	0.62**	1	0.24	0.45**	0.35**	0.35**	0.28**	0.52**
	N	174	25	135	173	174	28	108	173	172	165	165
Scar	Spearman	0.32*	.	0.56**	0.38*	0.24	1	.	0.18	0.11	0.34*	0.60**
	N	28	0	28	28	28	28	0	28	28	28	27
Teeth	Spearman	0.56**	0.70**	0.56**	0.39**	0.45**	.	1	0.29**	0.33**	0.39**	0.46**
	N	108	25	107	107	108	0	108	107	106	100	102
Psych	Spearman	0.53**	0.41**	0.44**	0.39**	0.35**	0.18	0.29**	1	0.66**	0.44**	0.57**
	N	173	24	134	172	173	28	107	173	172	165	165
Social	Spearman	0.47**	0.37*	0.38**	0.33**	0.35**	0.11	0.33**	0.66**	1	0.55**	0.50**
	N	172	24	133	171	172	28	106	172	172	164	164
COHIP	Spearman	0.47**	0.39**	0.42**	0.35**	0.28**	0.34*	0.39**	0.44**	0.55**	1	0.59**
	N	165	24	127	164	165	28	100	165	164	165	163
CHASQ	Spearman	0.67**	0.73**	0.67**	0.56**	0.52**	0.60**	0.46**	0.57**	0.50**	0.59**	1
	N	165	24	128	164	165	27	102	165	164	163	165

Internal Responsiveness to Change

The findings to address the second aim of this thesis (whether change scores collected with the CLEFT-Q, COHIP and CHASQ were statistically significant and clinically important) are described below for each cleft-specific operation in turn. Although the sample size was not achieved for the rhinoplasty, orthognathic and cleft lip scar operation samples, statistically significant and clinically important change in scores was identified using the distribution-based approaches for the CLEFT-Q appearance scales and the CHASQ subscale. The anchor-based approach, on the other hand, could not be performed as the distribution of responses to the category "a little better than before the operation" in anchor questions for each operation investigated in this study was too small to move forward with the analysis (see Table 9 in Appendix) .

In rhinoplasty, ESs were larger for the nose and nostrils scales than for the face scale. For orthognathic surgery, the largest ESs were observed for the jaws, teeth and face scales, with a large but slightly smaller ES for the lip scale. In cleft lip scar surgery, moderate ESs were observed for nose, lips and cleft lip scar scales. Statistically significant ESs were not found on HR-QOL scales. Statistically significant ESs following the ABG operation, serving as the control operation, were not detected as predicted. The MIDs were calculated using the ½ SD and 0.5 ES approaches. More specific results for each operation in turn are presented below.

Rhinoplasty Operation

The rhinoplasty sample consisted of 31 participants. The mean scores before and after surgery are shown in Table 10. There was statistically significant change between the pre- and postoperative scores on the CLEFT-Q face (mean difference = 7.61, SD = 18.97, $p = 0.033$), nose (mean difference = 17.10, SD = 25.95, $p = 0.001$) and nostrils (mean difference = 25.33, SD = 29.71, $p < 0.001$) scales as well as the CHASQ subscale (mean difference = 11.77, SD = 14.18, $p < 0.001$). Change scores were not statistically significant on the CLEFT-Q HR-QOL scales and the COHIP.

Cohen's d and SRM are shown in Table 10. The magnitude of change was large on the nose (Cohen's d – 0.92, SRM – 0.67) and nostrils (Cohen's d – 0.94, SRM – 0.85) scales, and moderate on the face (Cohen's d – 0.51, SRM – 0.40) scale and the CHASQ subscale (Cohen's d – 0.74, SRM – 0.83).

The first hypothesis regarding the magnitude of change was supported as the ESs were larger on the appearance scales than on the HR-QOL scales (See Table 10). The second hypothesis was similarly supported as the ESs on the appearance scales pertaining to facial areas most directly addressed by rhinoplasty (nose, nostrils) were larger than on the appearance scales pertaining to facial areas least directly addressed by rhinoplasty (face).

Table 10. Comparison of pre- and post-rhinoplasty scores using parametric methods.

		Mean	N	Sig.	Cohen's d	SRM	MID ½ SD	MID 0.5 ES
Face	Post	57.00	31	0.033	0.51	0.40	6.22	9.49
	Pre	49.39	31					
Nose	Post	66.86	29	0.001	0.92	0.66	9.70	12.98
	Pre	49.76	29					
Nostrils	Post	58.37	30	<0.001	0.94	0.85	10.43	14.86
	Pre	33.03	30					
Psych	Post	71.33	30	0.441	0.13	0.19	8.05	8.64
	Pre	68.87	30					
Social	Post	76.50	30	0.491	0.12	0.13	6.67	7.33
	Pre	74.63	30					
COHIP	Post	26.63	30	0.303	0.18	0.19	2.83	3.57
	Pre	25.27	30					
CHASQ	Post	66.47	30	<0.001	0.74	0.83	7.11	7.09
	Pre	54.70	30					

Orthognathic Operation

The orthognathic operation sample consisted of 21 participants. The mean scores before and after surgery for the sample are reported in Table 11. There was statistically significant change between the pre- and postoperative scores on the CLEFT-Q face (mean difference = 18.33, SD = 20.89, p = 0.001), nose (mean difference = 11.65, SD = 20.00,

p = 0.017), teeth (mean difference = 21.85, SD = 19.21, p < 0.001), jaws (mean difference = 37.05, SD = 23.10, p <0.001) and lips (mean difference = 22.25, SD = 23.70, p <0.001) scales as well as the CHASQ subscale (mean difference = 15.70, SD = 14.37, p < 0.001). Change scores were not significant on the CLEFT-Q nostrils and the HR-QOL scales.

Cohen's d and SRM are reported in Table 11. The magnitude of change was large on the face (Cohen's d - 1.15, SRM - 0.88), teeth (Cohen's d - 1.16, SRM - 1.14), lips (Cohen's d - 0.94, SRM - 0.94), jaws (Cohen's d - 1.80, SRM - 1.60) scales and the CHASQ subscale (Cohen's d – 1.08, SRM – 1.09), moderate on the COHIP (Cohen's d – 0.55, SRM – 0.56), and small on the nose scale (Cohen's d = 0.40, SRM = 0.58).

The first hypothesis of appearance scales having larger ESs than the HR-QOL scales was supported, with the exception of the nostrils scale, as no difference in change on this scale was detected. The ESs being larger on the appearance scale pertaining to facial areas most directly addressed by the orthognathic surgery (jaws) were larger than on the appearance scales pertaining to facial areas least directly addressed by the orthognathic surgery (teeth, lips, nose, nostrils and face), thus supporting the second hypothesis.

Table 11. Comparison of pre- and post-orthognathic operation scores using parametric methods.

		Mean	N	Sig.	Cohen's d	SRM	MID ½ SD	MID 0.5 ES
Face	Post	60.90	21	0.001	1.15	0.88	6.20	10.45
	Pre	42.57	21					
Nose	Post	47.05	20	0.017	0.40	0.58	14.38	10.00
	Pre	35.40	20					
Nostrils	Post	44.90	20	0.065	0.38	0.44	13.71	13.17
	Pre	33.35	20					
Teeth	Post	66.70	20	<0.001	1.16	1.14	5.88	9.61
	Pre	44.85	20					
Lips	Post	62.05	20	<0.001	0.94	0.94	9.24	11.85
	Pre	39.80	20					
Jaws	Post	72.75	20	<0.001	1.80	1.60	7.80	11.55
	Pre	35.70	20					
Psych	Post	69.70	20	0.311	0.31	0.23	9.96	12.34

	Pre	63.95	20					
Social	Post	80.75	20	0.408	0.20	0.19	8.70	10.16
	Pre	76.90	20					
COHIP	Post	26.55	20	0.021	0.55	0.56	3.25	2.97
	Pre	23.20	20					
CHASQ	Post	60.20	20	<0.001	1.08	1.09	7.70	7.18
	Pre	44.50	20					

Cleft Lip Scar Operation

The cleft lip scar revision sample included 18 participants. The mean scores for the sample before and after surgery are reported in Table 12. Statistically significant change between the pre- and postoperative scores on the CLEFT-Q nose (mean difference = 12.23, SD = 22.14, p = 0.035), lips (mean difference = 15.47, SD = 21.54, p = 0.009) and cleft lip scar (mean difference = 11.06, SD = 20.00, p = 0.043) scales as well as the CHASQ subscale (mean difference = 9.47, SD = 16.60, p = 0.044) was demonstrated. Change scores were not statistically significant on the CLEFT-Q face, nostrils and HR-QOL scales or the COHIP.

Cohen's d and SRM are reported in Table 12. The magnitude of change was moderate on the cleft lip scar (Cohen's d - 0.50, SRM - 0.55), lips (Cohen's d - 0.58, SRM - 0.72), nose (Cohen's d - 0.76, SRM - 0.56) scales and the CHASQ subscale (Cohen's d - 0.78, SRM - 0.57).

The ESs on the appearance scales were larger than on the HR-QOL scales supporting the first hypothesis. The pattern of the ESs for the cleft lip scar revision sample differed from the expected, with the largest ESs on the nose scale, followed by the lips and the cleft lip scar scales, thus not fully supporting the second hypothesis.

Table 12. Comparison of pre- and post-cleft lip scar revision scores using parametric methods.

		Mean	**N**	**Sig.**	**Cohen's d**	**SRM**	**MID ½ SD**	**MID 0.5 ES**
Face	Post	56.39	18	0.125	0.54	0.38	4.35	9.87
	Pre	48.89	18					
Nose	Post	57.41	17	0.035	0.76	0.56	6.60	11.07

	Pre	45.06	17					
Nostrils	Post	47.71	17	0.085	0.58	0.45	9.70	16.19
	Pre	33.29	17					
Lips	Post	54.82	17	0.009	0.58	0.72	6.44	10.77
	Pre	39.35	17					
Scar	Post	54.13	16	0.043	0.50	0.55	12.34	10.00
	Pre	43.06	16					
Psych	Post	64.88	17	1.000	0	0	6.66	6.59
	Pre	64.88	17					
Social	Post	77.76	17	0.150	0.42	0.37	7.90	9.61
	Pre	70.71	17					
COHIP	Post	26.63	16	0.155	0.51	0.37	2.40	4.09
	Pre	23.56	16					
CHASQ	Post	62.00	15	0.044	0.78	0.57	3.98	8.30
	Pre	52.53	15					

ABG Operation

The ABG sample consisted of 57 participants. The mean scores for the sample before and after surgery are reported in Table 13. Changes on the CLEFT-Q appearance and HR-QOL scales, COHIP and CHASQ subscale were not found.

The prespecified hypothesis predicting non-significant ESs on the CLEFT-Q appearance scales were supported by the results (Table 13).

Table 13. Comparison of pre- and post-ABG operation scores using parametric methods.

		Mean	**N**	**Sig.**	**Cohen's d**	**SRM**	**MID ½ SD**	**MID 0.5 ES**
Face	Post	57.50	56	0.919	0.02	0.01	9.55	11.70
	Pre	57.82	56					
Nose	Post	54.02	55	0.410	0.12	0.11	11.38	13.95
	Pre	57.15	55					
Nostrils	Post	51.65	55	0.658	0.08	0.06	12.19	16.68
	Pre	53.65	55					
Teeth	Post	42.77	53	0.209	0.19	0.18	9.17	10.79
	Pre	46.55	53					
Lips	Post	58.55	53	0.791	0.04	0.04	10.47	12.91
	Pre	59.49	53					
Psych	Post	72.02	53	0.260	0.15	0.16	9.87	10.00

	Pre	75.15	53					
Social	Post	71.74	50	0.241	0.16	0.17	8.59	9.18
	Pre	74.82	50					
COHIP	Post	28.17	41	0.753	0.06	0.05	4.06	4.93
	Pre	27.68	41					
CHASQ	Post	64.78	46	0.470	0.12	0.11	8.70	9.68
	Pre	62.70	46					

CHAPTER 4: DISCUSSION

The CLEFT-Q scales comprise a condition-specific PROM for patients with CLP that are being rapidly adopted by clinicians and academics around the globe. The fast uptake of the CLEFT-Q scales is demonstrating a considerable need for a rigorously developed PROM for patients with CLP for use in clinical practice and research. Psychometric properties of the CLEFT-Q scales have been tested in the first and second phase studies of its development. This thesis focuses on the third phase of the CLEFT-Q development, which details examination of additional psychometric properties of the CLEFT-Q scales, including cross-sectional construct validity and internal responsiveness.

The first aim of this thesis examined whether the scores collected at the study baseline corresponded with the theoretical expectations, i.e., cross-sectional construct validity. The values used for the prespecified hypotheses to examine cross-sectional construct validity between the appearance constructs and HR-QOL constructs were based on the published correlations between CLEFT-Q scales from the phase II field-test sample of 2343 individuals with CLP [77]. Correlations amongst the scales within their top-level domains (appearance and HR-QOL) were predicted to be strong. Correlations between scales in different top-level domains were predicted to be moderate. Overall, the findings from this first thesis aim provided broad support for the cross-sectional validity of the CLEFT-Q scales. A total of 53 correlations were computed to examine relationships between the CLEFT-Q, COHIP and CHASQ, and 39 (74%) aligned with the predetermined hypotheses. Of the 14 hypotheses that were not supported, 11 correlations were weaker than anticipated and three were stronger. Eight of the 14 correlations were very close to the prediction, while six were not. Of these six correlations, five were comparing the CLEFT-Q appearance (face, nose and nostrils) and HR-QOL (psych and social) with the CLEFT-Q cleft lip scar scale. The cleft lip scar scale sample size was not reached (n=28) (Table 8), which may offer a possible explanation for the correlation coefficients being lower than anticipated. The remaining correlation was between the CLEFT-Q teeth and nose scales. This correlation may not have closely reached its prediction due to the teeth scale being administered only to individuals undergoing

operations involving their jaw and gums, i.e., orthognathic and ABG, which do not directly affect the nose.

The findings on cross-sectional construct validity add to the published information about construct validity from the field-test sample. Specifically, mean scores from 1938 patients who needed, had and did not require jaw surgery, cleft lip scar revision, rhinoplasty and speech surgery were published [120]. The authors reported that participants who needed surgery scored significantly lower than those who had surgery on CLEFT-Q scales relevant to each surgery. Thus, the CLEFT-Q scales were shown to detect differences between groups cross-sectionally based on surgical status [120]. Further construct validation performed in this phase III prospective study demonstrated cross-sectional construct validity when CLEFT-Q scale scores were compared to other frequently used PROMs for patients with CLP (i.e. CHASQ and COHIP). Although some correlations were stronger than expected, most prespecified hypotheses about constructs being examined between the CLEFT-Q, COHIP and CHASQ were supported by the study results.

The second aim of this thesis examined internal responsiveness, whether the changes in scores collected before and after operation were statistically significant and clinically important. Even though the sample size was not achieved, the analysis demonstrated that the CLEFT-Q appearance scales were able to detect statistically significant change following rhinoplasty, orthognathic and cleft lip scar revision surgeries. Also, as hypothesized, statistically significant change following the ABG operation was not detected with the appearance scales. The main surgical goals of an ABG operation include closure of the oro-nasal fistula, stabilization of the maxillary arch, provision of support for roots of teeth adjacent to the cleft, and provision of support for the alar base and future prosthesis [121]. These surgical improvements do not result in visible change in appearance, therefore non-significant differences between the pre- and postoperative appearance scale scores were expected [122]. This analysis, therefore, represented a control operation.

Statistically significant differences on the CLEFT-Q HR-QOL scales (psychological and social) were not observed for any of the four operations. The HR-QOL construct was not expected to change as much as appearance, since the HR-QOL is a more distant construct than appearance in the context of cleft-related surgery outcomes. Nevertheless, postoperative scores were hypothesized to be higher (i.e., better) than preoperative on the HR-QOL scales. A sample size larger than achieved in this phase III study may be required to detect differences in aspects of HR-QOL following cleft-related surgery. Another possibility is that a longer time between assessments, i.e., several years, may be needed to detect changes in HR-QOL that are a result of a combination of surgical and non-surgical treatment modalities. For instance, unlike appearance scales that detect improvement in satisfaction with appearance of facial features, HR-QOL scales detect changes in social and psychological well-being (i.e., being accepted by friends, feeling confident and fitting in) that may require a combination of surgical and non-surgical therapeutics, i.e., psychological or counseling services, and a longer time frame to improve. For example, a study by Nichols et al demonstrated statistically significant change in HR-QOL five years after treatment using the Oral Health Impact Profile (OHIP-14) in a sample of 57 patients with CLP who underwent orthognathic surgery [123].

In this phase III study, a statistically significant difference between the pre- and postoperative scores in the orthognathic surgery sample was also detected by the COHIP socio-emotional domain. This result could be explained by the considerable improvement in facial appearance and function that orthognathic surgery (i.e., Le Fort I and II osteotomies) offers due to correction of the maxilla that permits proper alignment and positioning of the bones and teeth in relation to the base of the skull [124,125,126,127,128,129,130,131]. Furthermore, the CLEFT-Q appearance scales examining the orthognathic operation reported highest ESs (i.e., Cohen's d 1.80 on jaws scale) relative to other cleft-related operations. Thus, the drastic changes in facial appearance could in turn result in having a larger impact on HR-QOL.

Related to statistical significance, the magnitude of change on the appearance scales was larger than on the HR-QOL scales. This was a predicted result based on the assumption that the HR-QOL construct is a more distant construct than appearance with respect to cleft-related treatment. Furthermore, the magnitude of change on the appearance scales measuring facial aspects most affected by rhinoplasty and orthognathic operations was larger than on the appearance scales measuring facial aspects least affected by these surgeries. These findings were expected and demonstrate that the CLEFT-Q scales detect appropriate amounts of change on one appearance scale relative to another, depending on the specific cleft-related surgery. The pattern differed for the cleft lip scar revision group as the magnitude of change was largest on the nose scale, followed by the lips and cleft lip scar scales. The sample size for this group was particularly small, which may explain the difference between the actual ESs and the hypothesized. Furthermore, the lip revision scars tend to be more visible and take longer to mature than the rhinoplasty scars, thus offering another explanation for smaller improvement on the lip and cleft lip scar scales than on the nose scale. However, this result is not entirely surprising as the cleft lip scar revision surgery has previously demonstrated improvement in appearance of the nasolabial (i.e., nose) region [132].

Clinically important change was determined through approximation of MIDs following each surgery for each scale that demonstrated change between the pre- and postoperative scores. As there is no gold standard for calculating MIDs using the distribution-based approach, two distribution-based methods for calculating MIDs were used (i.e., ½ SD of preoperative mean scores and mean change scores corresponding to 0.5 ES) (See Tables 10, 11, 12 and 13). The MID estimates varied between methodological approaches and surgery types. For instance, MIDs for the CLEFT-Q nose scale were 9.7 (½ SD approach) and 13.0 (0.5 ES approach) in the rhinoplasty sample, but 14.4 (½ SD approach) and 10.0 (0.5 ES approach) in the orthognathic sample. The calculated MIDs using both approaches should be used as a range to approximate whether a change detected by the scales is clinically important.

The variability in MIDs between the two approaches should be interpreted with caution due to the small sample sizes for each operation and the fact that the distribution-based approaches rely solely on the statistical characteristics of the PROM scores rather than on the patients' perspective as in the anchor-based approach. The COSMIN guidelines do not suggest examining responsiveness solely through the distribution-based approach [89]. As a gold standard approach has not been established, numerous studies have attempted both methods for comparison and examination of robustness of the results [133,134,135,136,137]. In this study, in addition to the distribution-based approach, the anchor-based approach was attempted. However, this approach had to be abandoned due to an insufficient number of responses per anchor question (See Table 9 in Appendix).

The MID estimates calculated for the CLEFT-Q appearance scales in this study compare to the MIDs calculated for the BREAST-Q appearance and HR-QOL scales in another study [138]. In this CLEFT-Q study, the MIDs based on ½ SD ranged between 6.2-10.4 for rhinoplasty, 5.9-14.4 for orthognathic surgery and 6.4-12.3 for cleft lip scar revision approximately 6 months after the operation. In the BREAST-Q study that included 3052 patients, the MIDs based on 1/5 SD ranged between 3.2-5.1 for autologous reconstruction, 3.3-5.3 for alloplastic reconstruction and 3.3-5.6 for radiation therapy one year after the operation [138]. Given the variability in approaches taken to estimate the MIDs, i.e., ½ SD approach versus 1/5 SD approach, the estimates determined in this study are reasonable in relation to the estimates determined in the BREAST-Q study. This comparison further demonstrates the variability in MIDs when different distribution-based approaches are used, and highlights the need for confirmation of these MID estimates with the anchor-based approach.

The ES and MID estimates for the COHIP socio-emotional domain calculated in this study are larger than previously reported. In this study, the COHIP ESs of 0.55 (Cohen's d) and 0.56 (SRM) were calculated for the orthognathic surgery sample. The observed MIDs for the socio-emotional domain for this group were 3.3 (1/2 SD) and 3.0 (0.5 ES). In comparison, a study by Russ et al found ESs for the COHIP socio-emotional domain to be 0.22 (Cohen's d) and 0.27 (SRM) for individuals with CLP who did not

receive a surgical operation during the time between assessments [139]. The MIDs were reported to be 0.12 and 0.19 using two different anchor-based approaches. The larger ESs and MIDs in this study may likely have been the consequence of the orthognathic surgery, as analyses in the study by Russ et al did not evaluate the effects of a cleft-related surgery [139]. The varying approaches to estimating MIDs taken by both studies may have additionally contributed to the differences in MIDs between the two studies.

Establishing responsiveness of the CLEFT-Q to change is a vital component of its validation process as CLEFT-Q is an evaluative measure of appearance, function and HR-QOL related to patients' CLP. Furthermore, knowledge of responsiveness is essential for use of PROMs in clinical trials, and lack thereof represents a serious potential reason for being less certain about a trial's evidence. Assessment of score changes using the distribution-based approach demonstrates that the CLEFT-Q scales are responsive to change. Further interpretation of HR-QOL change scores and exploration of MIDs using the preferred anchor-based approach are required in a larger and more clinically and culturally diverse sample.

Internal responsiveness psychometric property was confirmed with statistically significant and appropriate in magnitude differences in pre- and postoperative scores on relevant to each surgery CLEFT-Q scales. Further examination of clinically important differences for each scale and cleft-related operation in a study of a sample similar in magnitude and diversity to the that of the second phase field-test study (and as reported in Harrison et al) would be ideal [120]. A more plausible approach may be to confirm MIDs through the benchmarking processes conducted by medical centers involved in the ICHOM's global initiative as well as sharing of data collected through research projects and clinic efforts around the world.

As mentioned above, the distribution-based methods are not derived from patients' input, thus it is suggested that the anchor-based approach should be used instead or in parallel. Nevertheless, the anchor-based approach is also limited. This approach relies on 1) the anchor and 2) the analytical methods [140]. The proper use of an anchor

requires knowing the magnitude of change on the anchor that represents a small and important change to the patient [140]. Most times this magnitude is difficult to deduce. Another limitation is that there are numerous types of anchors and no consensus on the type of anchor that best suits the process of estimation of MIDs [140]. In addition to issues related to the choice of an anchor, the analytical approaches to estimating MIDs using an anchor are vast and each with its benefits and limitations yield different results [140,141,142,143].

The preliminary MIDs determined in this study should nevertheless be verified with the anchor-based approach. A transition-rating anchor such as "how does your [e.g., nose] look now compared to before your [e.g., rhinoplasty] operation?" should be used to examine patients who have undergone a rhinoplasty operation, with the response options of "a lot worse now than before my nose operation", "a little worse now than before my nose operation", "about the same as before my nose operation", "a little better than before my nose operation", "a lot better than before my nose operation". The small and important change is estimated to be the change in scale scores corresponding to "a little better than before my nose operation". Ideally, in future studies, anchors for each CLEFT-Q scale should be adopted or generated.

Subsequent studies should additionally focus on assessment of remaining psychometric properties that have not been examined to date, which include reproducibility and measurement error as well as external responsiveness. Theoretically, reproducibility should be examined through assessment of intra-rater reliability, inter-rater reliability and test-retest reliability. The measurement error should be determined thereafter by calculating the precision of measurement of each of the three reliability tests. External responsiveness should be assessed by testing hypotheses about correlations of change scores between CLEFT-Q and other PROMs often used in cleft-related research.

Limitations

This study has a number of important limitations. The main limitation of this study was the inability to meet the intended sample size of 50 participants per operation based

on the preliminary sample size calculations and COSMIN recommendations [144]. This limitation may have impacted the ability of the collected data to precisely examine responsiveness and estimate the MIDs. The high dropout rate and variation in response rates between countries may have limited this study's generalizability. Most individuals included in the study were representative of those lost to follow up, however the highest drop our rate was in England and the lowest in the USA (See Table 7). Although all sites were contacted monthly about study progress, geographic distance made it challenging to aid in the recruitment process closely.

The process of recruitment itself was challenging. While the Hospital for Sick Children admits a large number of patients with CLP, in this study patients were consented two hours prior to their operation as this was the established protocol. At this time, participants were often worried and busy responding to their medical team. Thus, while some patients simply did not have enough time to complete the preoperative assessment, others were not permitted by their parents. Furthermore, the heightened anxiety they felt right before their operation may have affected their answers to the questionnaire. The postoperative assessment was completed by patients at home by accessing the online version of the questionnaire through a personal electronic device. Although the questionnaire is designed for self-report, it is possible that parents might have influenced or helped their children complete the assessment. Completion of the postoperative questionnaire through an online platform was the preferred method by patients, however technical difficulties or lack of designated time may have resulted in some participants failing to submit this assessment.

The variability in time between the pre- and postoperative assessments for each operation was another limitation. The time between assessments ranged between five months for the rhinoplasty to seven months for the ABG operation. This variation in time could be attributed to differences in cleft care at each site, including varying time between operation and follow-up clinic visit, and different amounts of resources allocated to hiring research staff to oversee the recruitment process. However, consensus among experts

regarding the ideal time to measure outcomes following each cleft-related surgery should be reached and followed as closely as possible.

The Americleft, Eurocleft and Scandicleft studies demonstrated that due to an abundance of existing surgical techniques, there is a substantial variation in the treatment protocols for management of patients with CLP within and between countries. As patients included in this third phase CLEFT-Q study were treated at different sites within and between countries, any deviation in management of the CLP could have introduced additional variability to the postoperative scores between sites.

Furthermore, an essential component of examining responsiveness was to ensure that each participant underwent only one cleft-related operation (i.e., the operation at study inclusion) between assessments. Of all, only one participant underwent the orthognathic surgery (i.e., Le Fort I osteotomy with horizonal advancement) and rhinoplasty (i.e., septorhinoplasty) in one operation session.

Finally, another limitation was an inability to include all CLEFT-Q scales in the pre- and postoperative assessments, as some were not relevant to the surgeries involved in this study. However, further research could be conducted to examine construct validity and responsiveness for the scales that were excluded from this study.

CHAPTER 5: CONCLUSION

The process of developing the CLEFT-Q has been a multidisciplinary and multi-site initiative with partners around the globe. The privilege of collaborating with international teams ensured that the rigorous development and validation processes account for multicultural perspectives on cleft-related care. The rapid uptake of CLEFT-Q in close to 40 countries and translations in 19 languages is evidence to its useful, comprehensive and relevant nature. Further, inclusion of the CLEFT-Q scales in the ICHOM standard set for assessment of CLP, pediatric facial palsy and craniofacial microsomia allows medical centers worldwide to adopt the scales for clinical outcome measurement and global benchmarking initiatives [64].

The opportunity to measure functional and psychosocial outcomes using the CLEFT-Q scales at the Hospital for Sick Children provided a chance to personally meet with patients with CLP. Most participants greatly appreciated being included in the study and our team's efforts to develop and validate the CLEFT-Q PROM. Although some patients expressed concerns about the instrument's length, most shared their gratitude through the comment function for 1) the opportunity to express how they feel from their perspective, 2) their highly valued medical and research teams, 3) and having felt better about their CLP condition after completing the CLEFT-Q scales.

This thesis is the first study to examine cross-sectional construct validity and internal responsiveness of the CLEFT-Q scales. The assessment of cross-sectional construct validity revealed that the correlations between the CLEFT-Q, COHIP and CHASQ at baseline supported most prespecified hypotheses. The distribution-based approach used to examine internal responsiveness demonstrated that the appearance scales were highly responsive to detecting change following cleft-specific surgeries. Furthermore, the magnitude of change on the CLEFT-Q appearance scales examining the facial feature that is being directly addressed by the operation was larger than on the scales measuring facial features not directly addressed by the operation, as predicted. The CLEFT-Q scales were responsive to change following cleft-related operation, however further examination of MIDs in a larger and more clinically and culturally diverse sample is necessary.

REFERENCES

1. Haque S, Alam MK. Common dental anomalies in cleft lip and palate patients. *Malays J Med Sci*. 2015;22(2):55-60.
2. Pavri S, Forrest CR. Demographics of orofacial clefts in Canada from 2002 to 2008. *Cleft Palate Craniofac J*. 2013;50(2):224-230.
3. Matthews JL, Oddone-Paolucci E, Harrop RA. The epidemiology of cleft lip and palate in Canada, 1998 to 2007. *Cleft Palate Craniofac J.* 2015;52(4):417-424.
4. Mossey PA, Little J, Munger RG, et al. Cleft lip and palate. *Lancet*. 2009;374(9703):1773-1785.
5. Reid J. A review of feeding interventions for infants with cleft palate. *Cleft Palate Craniofac J*. 2004;41(3):268-278.
6. Nagarajan R, Savitha VH, Subramaniyan B. Communication disorders in individuals with cleft lip and palate: An overview. *Indian J Plast Surg*. 2009;42 Suppl(Suppl):S137-S143.
7. Teele DW, Klein JO, Rosner B. Epidemiology of otitis media during the first seven years of life in children in greater Boston: a prospective, cohort study. *J Infect Dis*. 1989;160(1):83-94.
8. Nahai FR, Williams JK, Burstein FD, et al. The Management of Cleft Lip and Palate: Pathways for Treatment and Longitudinal Assessment. *Semin Plast Surg.* 2005;19(4):275-285.
9. Zreaqat MH, Hassan R, Hanoun A. Chapter 6: Cleft lip and palate management from birth to adulthood: an overview. In: Manakil J. *Insights into various aspects of oral health.* Intech Open Limited; 2017.
10. O'Brien K, Wright J, Conboy F, et al. Prospective, multi-center study of the effectiveness of orthodontic/orthognathic surgery care in the United Kingdom. *Am J Orthod Dentofacial Orthop*. 2009;135(6):709-714.
11. Grossi GB, Garagiola U, Santoro F. Measuring effectiveness of orthognathic surgery by electromyography: a restrospective clinical study. *Minerva Stomatol*. 2017;66(3):98-106.

12. Glushko A, Drobyshev A, Drobysheva N, et al. Effectiveness of the one-stage orthognathic surgery and rhinoplasty. *Int J Oral Maxillofac Surg.* 2017;46:Suppl(Suppl):S156.
13. Chaithanyaa N, Rai KK, Shivakumar HR, et al. Evaluation of the outcome of secondary rhinoplasty in cleft lip and palate patients. *J Plast Reconstr Aesthet Surg*. 2011;64(1):27-33.
14. Pinto V, Piccin O, Burgio L, et al. Effect of early correction of nasal septal deformity in unilateral cleft lip and palate on inferior turbinate hypertrophy and nasal patency. *Int J Pediatr Otorhinolaryngol*. 2018;108:190-195.
15. Wang Z, Wang P, Zhang Y, et al. Nasal Airway Evaluation After Le Fort I Osteotomy Combined With Septoplasty in Patients With Cleft Lip and Palate. *J Craniofac Surg*. 2017;28(1):207-211.
16. Trindade IE, Bertier CE, Sampaio-Teixeira AC. Objective assessment of internal nasal dimensions and speech resonance in individuals with repaired unilateral cleft lip and palate after rhinoseptoplasty. *J Craniofac Surg*. 2009;20(2):308-314.
17. Li W, Steinbacher DM. Unilateral Cleft Lip Revision with Conversion to the Modified Inferior Triangle. *Plast Reconstr Surg*. 2015;136(3):353e-361e.
18. Nadjmi N, Amadori S, Van de Casteele E. Secondary Cleft Lip Reconstruction and the Use of Pedicled, Deepithelialized Scar Tissue. *Plast Reconstr Surg Glob Open*. 2016;4(10):e1061.
19. Agrawal K. Cleft palate repair and variations. *Indian J Plast Surg*. 2009;42 Suppl(Suppl):S102-S109.
20. Semb G, Brattstrom V, Molsted K, et al. The Eurocleft study: intercenter study of treatment outcome in patients with complete cleft lip and palate. Part 4: relationship among treatment outcome, patient/parent satisfaction, and the burden of care. *Cleft Palate Craniofac J.* 2005 42:83-92.
21. Brattstrom V, Molsted K, Prahl-Andersen B, et al. The Eurocleft study: intercenter study of treatment outcome in patients with complete cleft lip and

palate. Part 2: craniofacial form and nasolabial appearance. *Cleft Palate Craniofac J.* 2005 42:69-77.

22. Molsted K, Brattstrom V, Prahl-Andersen B, et al. The Eurocleft study: intercenter study of treatment outcome in patients with complete cleft lip and palate. Part 3: dental arch relationships. *Cleft Palate Craniofac J*. 2005 42:78-82.
23. Shaw WC, Brattstrom V, Molsted K, et al. The Eurocleft study: intercenter study of treatment outcome of patients with complete cleft lip and palate. Part 5: discussion and conclusions. *Cleft Palate Craniofac J*. 2005; 42:93-8.
24. Long RE Jr, Hathaway R, Daskalogiannakis J, et al. The Americleft study: an inter-center study of treatment outcomes for patients with unilateral cleft lip and palate part 1. Principles and study design. *Cleft Palate Craniofac J.* 2011; 48:239-43.
25. Hathaway R, Daskalogiannakis J, Mercado A, et al. The Americleft study: an inter-center study of treatment outcomes for patients with unilateral cleft lip and palate part 2. Dental arch relationships. *Cleft Palate Craniofac J.* 2011; 48:244-51.
26. Daskalogiannakis J, Mercado A, Russell K, et al. The Americleft study: an inter-center study of treatment outcomes for patients with unilateral cleft lip and palate part 3. Analysis of craniofacial form. *Cleft Palate Craniofac J.* 2011; 48:239-43.
27. Mercado A, Russell K, Hathaway R, et al. The Americleft study: an inter-center study of treatment outcomes for patients with unilateral cleft lip and palate part 4. Nasolabial aesthetics. *Cleft Palate Craniofac J*. 2011; 48:259-64.
28. Russell K, Long RE Jr, Hathaway R, et al. The Americleft study: an inter-center study of treatment outcomes for patients with unilateral cleft lip and palate part 5. General discussion and conclusions. *Cleft Palate Craniofac J.* 2011; 48:265-70.

29. Semb G, Enemark H, Friede H, et al. A Scandcleft randomised trials of primary surgery for unilateral cleft lip and palate: 1. Planning and management. *J Plast Surg Hand Surg.* 2017 Feb;51(1):2-13.
30. Rautio J, Andersen M, Bolund S, et al. Scandcleft randomised trials of primary surgery for unilateral cleft lip and palate: 2. Surgical results. *J Plast Surg Hand Surg*. 2017;51(1):14-20.
31. Bannister P, Lindberg N, Jeppesen K, et al. Scandcleft randomised trials of primary surgery for unilateral cleft lip and palate: 3. Descriptive study of postoperative nursing care following first stage cleft closure. *J Plast Surg Hand Surg*. 2017;51(1):21-26.
32. Lohmander A, Persson C, Willadsen E, et al. Scandcleft randomised trials of primary surgery for unilateral cleft lip and palate: 4. Speech outcomes in 5-year-olds - velopharyngeal competency and hypernasality. *J Plast Surg Hand Surg*. 2017;51(1):27-37.
33. Willadsen E, Lohmander A, Persson C, et al. Scandcleft randomised trials of primary surgery for unilateral cleft lip and palate: 5. Speech outcomes in 5-year-olds - consonant proficiency and errors. *J Plast Surg Hand Surg*. 2017;51(1):38-51.
34. Heliovaara A, Küseler A, Skaare P, et al. Scandcleft randomised trials of primary surgery for unilateral cleft lip and palate: 6. Dental arch relationships in 5-year-olds. *J Plast Surg Hand Surg*. 2017;51(1):52-57.
35. Karsten A, Marcusson A, Hurmerinta K, et al. Scandcleft randomised trials of primary surgery for unilateral cleft lip and palate: 7. Occlusion in 5-year-olds according to the Huddart and Bodenham index. *J Plast Surg Hand Surg*. 2017;51(1):58-63.
36. Molsted K, Humerinta K, Küseler A, et al. Scandcleft randomised trials of primary surgery for unilateral cleft lip and palate: 8. Assessing naso-labial appearance in 5-year-olds - a preliminary study. *J Plast Surg Hand Surg*. 2017;51(1):64-72.

37. Feragen KB, Rumsey N, Heliövaara A, et al. Scandcleft randomised trials of primary surgery for unilateral cleft lip and Palate: 9. Parental report of social and emotional experiences related to their 5-year-old child's cleft diagnosis. *J Plast Surg Hand Surg*. 2017;51(1):73-80.
38. Feragen KB, Semb G, Heliövaara A, et al. Scandcleft randomised trials of primary surgery for unilateral cleft lip and palate: 10. Parental perceptions of appearance and treatment outcomes in their 5-year-old child. *J Plast Surg Hand Surg*. 2017;51(1):81-87.
39. Valderas JM, Kotzeva A, Espallargues M, et al. The impact of measuring patient-reported outcomes in clinical practice: a systematic review of the literature. *Qual Life Res*. 2008;17(2):179-193.
40. Weldring T, Smith SM. Patient-Reported Outcomes (PROs) and Patient-Reported Outcome Measures (PROMs). *Health Serv Insights*. 2013;6:61-68.
41. Monmouth Partners. A Guide to Patient Reported Measures – Theory, Landscape and Uses. Available from http://www.monmouthpartners.com/assets/pdf/A%20Guide%20to%20Patient%20Reported%20Measures.pdf (accessed 15 February 2020).
42. Shaye D. Update on outcomes research for cleft lip and palate. *Curr Opin Otolaryngol Head Neck Surg*. 2014;22(4):255-259.
43. Klassen A, Tsangaris E, Forrest CR, et al. Quality of life of children treated for cleft-lip and/or palate: a systematic review. *J Plast Reconstr Aesthet Surg.* 2012; 65(5):547-57.
44. Patrick DL, Topolski TD, Edwards TC, et al. Measuring the quality of life of youth with facial differences. *Cleft Palate Craniofac J.* 2007 Sep;44(5):538-547.
45. El Osta N, Pichot H, Soulier-Peigue D, et al. Validation of the child oral health impact profile (COHIP) french questionnaire among 12 years-old children in New Caledonia. *Health Qual Life Outcomes*. 2015;13:176.
46. Broder HL, McGrath C, Cisneros GJ. Questionnaire development: face validity and item impact testing of the Child Oral Health Impact Profile. *Community Dent Oral Epidemiol*. 2007;35(Suppl 1):8-19.

47. Dunlow N, Phillips C, Broder HL. Concurrent validity of the COHIP. *Community Dent Oral Epidemiol*. 2007;35(Suppl 1):41-9.
48. Wilson-Genderson M, Broder HL, Phillips C. Concordance between caregiver and child reports of children's oral health-related quality of life. *Community Dent Oral Epidemiol*. 2007;35(Suppl 1):32-40.
49. Broder HL, Wilson-Genderson M. Reliability and convergent and discriminant validity of the Child Oral Health Impact Profile (COHIP Child's version). *Community Dent Oral Epidemiol.* 2007;35(Suppl 1):20-31.
50. Broder HL, Wilson-Genderson M, Sischo L. Reliability and validity testing for the Child Oral Health Impact Profile-Reduced (COHIP-SF 19) [published correction appears in J Public Health Dent. 2013 Winter;73(1):86]. *J Public Health Dent.* 2012;72(4):302-312.
51. Ohannessian P, Berggren A, Abdiu A. The cleft lip evaluation profile (CLEP): a new approach for postoperative nasolabial assessment in patients with unilateral cleft lip and palate. *J Plast Surg Hand Surg.* 2011 Feb;45(1):8-13.
52. Network, C.P.C.E., Cleft Hearing Appearance and Speech Questionnaire (CHASQ) - user guide. 2015: Unpublished work.
53. Emerson M, Spencer-Bowdage S, Bates A. Relationships between self-esteem, social experiences and satisfaction with appearance: standardisation and construct validation of two cleft audit measures in The Craniofacial Society of Great Britain and Ireland. Annual Scientific Conference. "Setting a positive agenda in cleft and craniofacial care", 2004.
54. Mani M, Reiser E, Andlin-Sobocki A, et al. Factors related to quality of life and satisfaction with nasal appearance in patients treated for unilateral cleft lip and palate. *Cleft Palate Craniofac J.* 2013 Jul;50(4):432-9.
55. Mani MR, Semb G, Andlin-Sobocki A. Nasolabial appearance in adults with repaired unilateral cleft lip and palate: Relation between professional and lay rating and patients' satisfaction. *J Plast Surg Hand Surg.* 2010 Nov;44(4-5):191-8.

56. Feragen KB, Stock NM. Risk and protective factors at age 10: Psychological adjustment in children with a cleft lip and/or palate. *Cleft Palate Craniofac J.* 2016 Mar;53(2):161-79.
57. Crerand CE, Sarwer DB, Kazak AE, et al. Body image and quality of life in adolescents with craniofacial conditions. *Cleft Palate Craniofac J.* 2017 Jan;54(1):2-12.
58. Gibbons E, Black N, Fallowfield L, et al. Patient-reported outcome measures and the evaluation of services. In: Raine R, Fitzpatrick R, Barratt H, et al. *Challenges, solutions and future directions in the evaluation of service innovations in health care and public health.* Southampton (UK): NIHR Journals Library; 2016 May. (Health Services and Delivery Research, No. 4.16.) Essay 4. Available from: https://www.ncbi.nlm.nih.gov/books/NBK361255/
59. Vodicka E, Kimb K, Devinea E, et al. Inclusion of patient-reported outcome measures in registered clinical trials: evidence from ClinicalTrials.gov (2007–2013). *Contemp Clin Trials.* 2015;43:1–9.
60. Jeevan R, Cromwell DA, Browne JP, et al. Findings of a national comparative audit of mastectomy and breast reconstruction surgery in England. *J Plast Reconstr Aesthet Surg*. 2014;67(10):1333-1344.
61. Patrick D, Peach H. Disablement in the Community. Oxford: Oxford University Press; 1989.
62. Gotay CC, Kawamoto CT, Bottomley A, et al. The prognostic significance of patient-reported outcomes in cancer clinical trials. *J Clin Oncol.* 2008;26:1355–63.
63. Nilsson E, Orwelius L, Kristenson M. Patient-reported outcomes in the Swedish National Quality Registers. *J Int Med.* 2015;279:141–53.
64. 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):540-554.

65. Harrison CJ, Geerards D, Ottenhof MJ, et al. Computerised adaptive testing accurately predicts CLEFT-Q scores by selecting fewer, more patient-focused questions. *J Plast Reconstr Aesthet Surg.* 2019 Nov;72(11):1819-1824.
66. Wong Riff KWY, 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). *Br Med J Open*. 2017;7:e015467.
67. Streiner DL, Norman GR, Cairney J. Chapter 12: Item response theory. In: Streiner DL, Norman GR, Cairney J. *Health measurement scales: A practical guide to their development and use.* Oxford University Press; 2015.
68. Hambleton RK, Swaminathan H. *Item response theory: Principles and applications*. Kluwer Nijhoff, Boston, 1985.
69. Hambleton RK, Swaminathan H, Rogers HJ. *Fundamentals of item response theory*. Sage, Newbury Park, NJ, 1991.
70. Scott RL, Pampa WM. The MMPI-2 in Peru: A normative study. *J Pers Assess*, 2000;74:95–105.
71. Streiner DL, Norman GR, Cairney J. Chapter 5: Selecting the Items. In: Streiner DL, Norman GR, Cairney J. *Health measurement scales: A practical guide to their development and use.* Oxford University Press; 2015.
72. Bond TG, Fox CM. *Applying the Rasch model: Fundamental measurement in the human sciences*. Lawrence Erlbaum Associates, Mahwah, NJ, 2001.
73. Rasch G. *Probabilistic models for some intelligence and attainment tests*. Nielson and Lydiche, Copenhagen, 1960.
74. Terwee CB, Bot SD, de Boer MR, et al. Quality criteria were proposed for measurement properties of health status questionnaires. *J Clin Epidemiol.* 2007;60:34-42.
75. Mokkink LB, Terwee CB, Patrick DL, et al. The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. *J Clin Epidemiol.* 2010;63:737-745.

76. Streiner DL, Norman GR. Chapter 8: Reliability. In: Streiner DL, Norman GR, Cairney J. *Health measurement scales: A practical guide to their development and use.* Oxford University Press; 2015.
77. Klassen AF, Riff KWW, 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. *Can Med Assoc J*. 2018;190(15):E455-E462.
78. Streiner DL, Norman GR. Chapter 10: Validity. In: Streiner DL, Norman GR, Cairney J. *Health measurement scales: A practical guide to their development and use.* Oxford University Press; 2015.
79. 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:e1305.
80. Patrick DL, Burke LB, Gwaltney CJ, et al. Content validity-- establishing and reporting the evidence in newly developed patient-reported outcomes (PRO) instruments for medical product evaluation: ISPOR PRO good research practices task force report: part 1--eliciting concepts for a new PRO instrument. *Value Health.* 2011;14:967-977.
81. Patrick DL, Burke LB, Gwaltney CJ, et al. Content validity-- establishing and reporting the evidence in newly developed patient-reported outcomes (PRO) instruments for medical product evaluation: ISPOR PRO Good Research Practices Task Force report: part 2--assessing respondent understanding. *Value Health.* 2011;14:978-988.
82. Terwee CB, Prinsen CAC, Chiarotto A, et al. COSMIN methodology for evaluating the content validity of patient-reported outcome measures: a Delphi study. *Qual Life Res*. 2018;27(5):1159-1170.
83. Mokkink LB, Terwee CB, Knol DL, et al. The COSMIN checklist for evaluating the methodological quality of studies on measurement properties: a clarification of its content. *BMC Med Res Methodol*. 2010;10:22.

84. Cronbach LJ, Meehl PE. Construct validity in psychological tests. *Psychol Bull.* 1955 Jul;52(4):281-302.
85. Stucki G, Liang MH, Fossel AH, et al. Relative responsiveness of condition-specific and generic health status measures in degenerative lumbar spinal stenosis. *J Clin Epidemiol*. 1995;48:1369–78.
86. Kirshner B, Guyatt GG. Methodological framework for assessing health indices. *J Chron Dis* 1985;38:27–36.
87. Husted JA, Cook RJ, Farewell VT, et al: Methods for assessing responsiveness: a critical review and recommendations. *J Clin Epidemiol.* 2000, 53(5):459-468.
88. Guyatt GH, Deyo RA, Charlson M, et al. Responsiveness and validity in health status measurement: a clarification. *J Clin Epidemiol.* 1989;42(5):403–408.
89. Angst F. The new COSMIN guidelines confront traditional concepts of responsiveness. *BMC Med Res Methodol*. 2011;11:152.
90. Chiu EC, Hung TM, Huang CM, et al. Responsiveness of the Personal and Social Performance scale in patients with schizophrenia. *Psychiatry Res*. 2018;260:338-342.
91. Jaeschke R, Singer J, Guyatt GH. Measurement of health status. Ascertaining the minimal clinically important difference. *Control Clin Trials.* 1989 Dec;10(4):407-15.
92. Schünemann HJ, Guyatt GH: Commentary – goodbye M(C)ID! Hello MID, where do you come from? *Health Serv Res.* 2005, 40(2):593-597.
93. Schünemann HJ, Puhan M, Goldstein R, et al: Measurement Properties and Interpretability of the Chronic Respiratory Disease Questionnaire (CRQ). *J Chron Obstruct Pulmon Dis*. 2005, 2:81-89.
94. Santanello NC, Zhang J, Seidenberg B, et al: What are minimal important changes for asthma measures in a clinical trial? *Eur Respir J.* 1999, 14(1):23-27.
95. Locker D, Jokovic A, Clarke M. Assessing the responsiveness of measures of oral health-related quality of life. *Community Dent Oral Epidemiol*. 2004;32(1):10–18.

96. Deyo RA, Diehr P, Patrick DL. Reproducibility and responsiveness of health status measures. Statistics and strategies for evaluation. *Control Clin Trials*. 1991;12(4 Suppl):142S–158S.
97. Guyatt GH, Osoba D, Wu AW, et al. Clinical Significance Consensus Meeting Group. Methods to explain the clinical significance of health status measures. *Mayo Clin Proc*. 2002;77(4):371–383.
98. Lydick E, Epstein RS. Interpretation of quality of life changes. *Qual Life Res*. 1993;2(3):221–226.
99. Devji T, Carrasco-Labra A, Qasim A, et al. Development and inter-rater reliability of an instrument to evaluate the credibility of anchor-based minimal important difference estimates for patient reported outcomes. *Br Med J.* Dec 2018 Submitted.
100. Brozek JL, Guyatt GH, Schünemann HJ. How a well-grounded minimal important difference can enhance transparency of labelling claims and improve interpretation of a patient reported outcome measure. *Health Qual Life Outcomes*. 2006;4:69.
101. Speer DC, Greenbaum PE. Five methods for computing significant individual client change and improvement rates: support for an individual growth curve approach [published correction appears in *J Consult Clin Psychol*, 2002 Dec;70(6):1239]. *J Consult Clin Psychol.* 1995;63(6):1044-1048.
102. Cohen J. *Statistical Power Analysis for the Behavioral Sciences.* 2nd ed. Hillsdale, NJ: Lawrence Erlbaum Associates; 1988.
103. Kazis LE, Anderson JJ, Meenan RF. Effect sizes for interpreting changes in health status. *Med Care.* 1989, 27(3 Suppl):S178-189.
104. Stucki G, Liang MH, Fossel AH, et al. Relative responsiveness of condition-specific and generic health status measures in degenerative lumbar spinal stenosis. *J Clin Epidemiol.* 1995, 48(11):1369-1378.
105. Guyatt GH, Bombardier C, Tugwell PX. Measuring disease-specific quality of life in clinical trials. *Can Med Assoc J.* 1986, 134(8):889-895.

106. Jacobson NS, Truax P. Clinical significance: a statistical approach to defining meaningful change in psychotherapy research. *J Consul Clin Psychol.* 1991, 59(1):12-19.
107. Wyrwich KW, Tierney WM, Wolinsky FD. Further evidence supporting an SEM-based criterion for identifying meaningful intra-individual changes in health-related quality of life. *J Clin Epidemiol.* 1999, 52(9):861-873.
108. Rosenthal R. Science and ethics in conducting, analyzing, and reporting psychological research. *Psychol Sci.* 1994;5(3):127-134.
109. Guyatt G, Walter S, Norman G. Measuring change over time: assessing the usefulness of evaluative instruments. *J Chronic Dis.* 1987;40(2):171-178.
110. McHorney CA, Tarlov AR. Individual-patient monitoring in clinical practice: are available health status surveys adequate? *Qual Life Res.* 1995;4(4):293-307.
111. Norman GR, Regehr G, Startford PS. Bias in the retrospective calculation of responsiveness to change: The lesson of Chronbach. *J Clin Epidemiol.* 1997;8:869-879.
112. Tsangaris E, Riff KWYW, 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.
113. Tsangaris E, Riff KWYW, Dreise M, et al. Translation and cultural adaptation of the CLEFT-Q into Arabic, Dutch, Hindi, Swedish, and Turkish. *Eur J of Plast Surg.* 2018 Dec;41(6):643-652.
114. Strauss ME, Smith GT. Construct validity: advances in theory and methodology. *Annu Rev Clin Psychol.* 2009;5:1-25.
115. Nguyen VT, Persson M, Jagomägi T. Application of a new patient-reported outcome measure in orofacial clefts: An exploratory study in two countries. *Stomatologija.* 2019;21(3):72-78.
116. Kim HY. Statistical notes for clinical researchers: assessing normal distribution (2) using skewness and kurtosis. *Restor Dent Endod.* 2013;38(1): 52-54.

117. Prinsen CAC, Mokkink LB, Bouter LM, et al. COSMIN guideline for systematic reviews of patient-reported outcome measures. *Qual Life Res.* 2018;27(5):1147-1157.
118. Masood M, Masood Y, Saub R, et al. Need of minimal important difference for oral health-related quality of life measures. *J Public Health Dent.* 2014;74(1):13–20.
119. Norman GR, Sloan JA, Wyrwich KW. Interpretation of changes in health-related quality of life: the remarkable universality of half a standard deviation. *Med Care.* 2003;41(5):582-592.
120. Harrison CJ, Rae C, Tsangaris E, et al. Further construct validation of the CLEFT-Q: ability to detect differences in outcome for four cleft-specific surgeries. *J Plast Reconstr Aesthet Surg*. 2019;72(12):2049-2055.
121. Emodi O, Noy D, Hazan-Molina H, et al. Secondary bone grafting of the cleft maxilla following reverse quad-helix expansion in 103 patients. *Ann Maxillofac Surg.* 2015;5(1):32-36.
122. Gillgrass TJ, MacDonald JP, Mossey PA, et al. The impact of alveolar bone grafting on cleft lip and palate: a literature review. *South Eur J Orthod Dentofacial Res.* 2014;1:15-22.
123. Nichols GAL, Antoun JS, Fowler PV, et al. Long-term changes in oral health-related quality of life of standard, cleft, and surgery patients after orthodontic treatment: A longitudinal study. *Am J Orthod Dentofacial Orthop.* 2018 Feb;153(2):224-231.
124. Zamboni R, de Moura FRR, Brew MC, et al. Impacts of orthognathic surgery on patient satisfaction, overall quality of life, and oral health-related quality of life: a systematic literature review. *Int J Dent*. 2019;2019:2864216.
125. Murphy C, Kearns G, Sleeman D, et al. The clinical relevance of orthognathic surgery on quality of life. *Int J Oral Maxillofac Surg.* 2011;40(9):926-930.
126. Huang S, Chen W, Ni Z, et al. The changes of oral health-related quality of life and satisfaction after surgery-first orthognathic approach: a longitudinal prospective study. *Head Face Med.* 2016;12:2.

127. Berridge N, Soneji B, Heliotis M. A 5-year audit evaluating patient satisfaction following orthognathic surgery and implications for the future. *Plast Recon Surg.* 2014; 134(4S-1):49.
128. Alves e Silva AC, Carvalho RA, Santos TdeS, et al. Evaluation of life quality of patients submitted to orthognathic surgery. *Dental Press J Orthod.* 2013;18(5):107-114.
129. Silva I, Suska F, Cardemil C, et al. Stability after maxillary segmentation for correction of anterior open bite: a cohort study of 33 cases. *J Craniomaxillofac Surg.* 2013;41(7):e154-e158.
130. Rustemeyer J, Eke Z, Bremerich A. Perception of improvement after orthognathic surgery: the important variables affecting patient satisfaction. *Oral Maxillofac Surg.* 2010;14(3):155-162.
131. Cunningham SJ, Hunt NP, Feinmann C. Perceptions of outcome following orthognathic surgery. *Br J Oral Maxillofac Surg*. 1996;34(3):210-213.
132. Mercado AM, Phillips C, Vig KW, et al. The effects of lip revision surgery on nasolabial esthetics in patients with cleft lip. *Orthod Craniofac Res.* 2014;17(4):216-225.
133. Mouelhi Y, Jouve E, Castelli C, et al. How is the minimal clinically important difference established in health-related quality of life instruments? Review of anchors and methods. *Health Qual Life Outcomes.* 2020 May 12;18(1):136.
134. Jayadevappa R, Malkowicz SB, Wittink M, et al. Comparison of distribution- and anchor-based approaches to infer changes in health-related quality of life of prostate cancer survivors. *Health Serv Res.* 2012 Oct;47(5):1902-25.
135. Yost KJ, Eton DT. Combining distribution- and anchor-based approaches to determine minimally important differences: the FACIT experience. *Eval Health Prof.* 2005 Jun;28(2):172-91.
136. Cella D, Eton DT, Lai JS, et al. Combining anchor and distribution-based methods to derive minimal clinically important differences on the Functional Assessment of Cancer Therapy (FACT) anemia and fatigue scales. *J Pain Symptom Manage.* 2002 Dec;24(6):547-61.

137. Frans FA, Nieuwkerk PT, Met R, et al. Statistical or clinical improvement? Determining the minimally important difference for the vascular quality of life questionnaire in patients with critical limb ischemia. *Eur J Vasc Endovasc Surg.* 2014 Feb;47(2):180-6.
138. Voineskos SH, Klassen AF, Cano SJ, et al. Giving Meaning to Differences in BREAST-Q Scores: Minimal Important Difference for Breast Reconstruction Patients. *Plast Reconstr Surg.* 2020 Jan;145(1):11e-20e.
139. Ruff RR, Sischo L, Broder HL. Minimally important difference of the Child Oral Health Impact Profile for children with orofacial anomalies. *Health Qual Life Outcomes.* 2016 Oct 3;14(1):140.
140. Devji T. Enhancing methods for analyzing and interpreting patient-reported outcomes in clinical research and evidence-based decision making (Doctoral dissertation, McMaster University, Hamilton, Canada). Available from MacSphere McMaster University Libraries Institutional Repository http://hdl.handle.net/11375/23923.
141. King M. The interpretation of scores from the EORTC quality of life questionnaire QLQ-C30. *Qual Life Res.* 1996;5(6):555-567.
142. Mills KA, Naylor JM, Eyles JP, et al. Examining the Minimal Important Difference of Patient-reported Outcome Measures for Individuals with Knee Osteoarthritis: A Model Using the Knee Injury and Osteoarthritis Outcome Score. *J Rheumatol.* 2016;43(2):395-404.
143. Terwee CB, Roorda LD, Dekker J, et al. Mind the MIC: large variation among populations and methods. *J Clin Epidemiol.* 2010;63(5):524-534.
144. Mokkink LB, de Vet HCW, Prinsen CAC, et al. COSMIN Risk of Bias checklist for systematic reviews of Patient-Reported Outcome Measures. *Qual Life Res.* 2018;27(5):1171-1179.

APPENDIX 1

Table 1. Details about each instrument included in the analysis.

	CLEFT-Q	**COHIP**	**CHASQ**
Domains	1) Appearance 2) HR-QOL 3) Function (excluded)	1) Oral health (excluded) 2) Socio-emotional 3) Function (excluded)	1) Feature 1 2) Feature 2 (excluded)
Scales/ Checklists	1) Appearance: face, lips, nose, nostrils, jaws, teeth, cleft lip scar 2) HR-QOL: psychological, social, school (excluded), speech distress (excluded)	Single scale	Single scale
Items	Face (9 items): 1. "How much do you like…how your face looks when you look your best?" Nose (12 items): 1. "How much do you like…the length of your nose (from the top of the tip)?" Nostrils (6 items): 1. "How much do you like…how your nostrils look when you smile?" Teeth (8 items): 1. "How much do you like…the size of your teeth?" Jaws (7 items): 1. "How much do you like…the size of your jaws?" Lips (9 items): 1. "How much do you like…how your lips look when you smile?" Cleft lip scar (7 items): 1. "How much do you like…the colour of your cleft lip scar?" Psychological (10 items): 1. "I am happy with my life." Social (10 items): 1. "My friends accept me." School (10 items): 1. "I like seeing my friends at school."	Socio-emotional (10 items): 1. "been unhappy or sad" 2. "felt worried or anxious" 3. "avoided smiling or laughing with other children" 4. "felt that you look different" 5. "been worried about what other people think about your ..." 6. "been teased, bullied, or called names by other children" 7. missed school for any reason" 8. "not wanted to speak/read out loud in class" 9. "been confident" 10. "felt that you were attractive (good looking)"	Feature 1 (9 items): "How happy are you with: 1) how your face looks? 2) the whole of your appearance? 3) side view/profile? 4) how good-looking do you think you are?" "How do you feel about these parts of your face? : 5) nose 6) lips 7) chin 8) teeth 9) cheeks"

Table 6. Formulas used in the distribution-based approach.	
Variable	**Formula**
Cohen's d	$Cohen's\ d = (N1 - N2)/SDpooled$ $SDpooled = \sqrt{(((SD_1^2+SD_2^2)/2))}$ Or $Cohen's\ d = Mean/SDbaseline$
Standardizes Response Mean	$SRM = Mean/SDchange$
MID ½ SD for Parametric Data	$MID = SDchange/2$
MID ½ ES for Parametric Data	$MID = 0.5 \times SDbaseline$

Table 9. Number of participants by their answers to anchor questions.					
	A lot worse	**A little worse**	**Same**	**A little better**	**A lot better**
Nose	0	2	2	5	20
Jaws	0	0	2	3	16
Cleft lip scar	0	2	2	7	5
ABG	-	-	-	-	-