Journal of Plastic, Reconstructive & Aesthetic Surgery (2019) 72, 2049–2055

ELSEVIER

BAPRAS
British Association of Plastic Reconstructive and Aesthetic Surgeons

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

Further construct validation of the CLEFT-Q: Ability to detect differences in outcome for four cleft-specific surgeries

Conrad J. Harrison[a], Charlene Rae[b], Elena Tsangaris[c], Karen W.Y. Wong Riff[d], Marc C. Swan[a,e], Tim E.E. Goodacre[e], Stefan Cano[f], Anne F. Klassen[g,*]

[a] *Department of Plastic Surgery, John Radcliffe Hospital, Headley Way, Oxford, OX3 9DU United Kingdom*
[b] *Department of Clinical Epidemiology and Biostatistics, McMaster University, 1280 Main Street West, Hamilton, Ontario, Canada*
[c] *Patient Reported Outcomes Values and Experience Center, Brigham and Women's Hospital, Harvard Medical School, 75 Francis Street, Boston, MA 02115, United States*
[d] *Department of Plastic Surgery, Hospital for Sick Children, 555 University Ave, Toronto, Ontario, M5G 1X8, Canada*
[e] *Nuffield Department of Surgery, University of Oxford, Room 6607, Level 6, John Radcliffe Hospital, Headley Way, Oxford, OX3 9DU, United Kingdom*
[f] *Modus Outcomes, Suite 210b, Spirella Building, Letchworth Garden City, SG6 4ET, United Kingdom*
[g] *Department of Pediatrics, McMaster University, Room HSC 3N27, 1280 Main Street West, Hamilton, Ontario L8S 4L8, Canada*

Received 6 May 2019; accepted 31 July 2019

KEYWORDS
CLEFT-Q;
Cleft lip and/or palate;
Patient reported outcome measure;
PROM;
Psychometrics

Summary *Background:* The CLEFT-Q is a patient-reported outcome measure developed for use in patients with cleft lip and/or palate. A significant indicator of the CLEFT-Q's validity relates to its ability to detect differences between the impact of specific aspects of clefting before and after surgery. This study compares relevant sub-scale scores of the CLEFT-Q for patients requiring four specific surgical treatments against those who either have had surgery or never needed surgery.
Methods: CLEFT-Q scores and clinical information regarding the past and future need for jaw surgery, lip revision, rhinoplasty, and speech surgery were obtained from the CLEFT-Q field-test data. Eight one-way analysis of variance (ANOVA) models were developed to compare mean scores of relevant CLEFT-Q scales between those who needed surgery, those who have had

Part of this article has been presented at BAPRAS Winter Scientific Meeting, London, 29th November 2018.

* Corresponding author at: Department of Pediatrics, McMaster University, Room HSC 3N27, 1280 Main Street West, Hamilton, Ontario L8S 4L8, Canada.
E-mail address: aklass@mcmaster.ca (A.F. Klassen).

https://doi.org/10.1016/j.bjps.2019.07.029

surgery, and those who never needed surgery. Only patients from high-income countries were included to minimize the impact of any economic confounders that could result in treatment variation. In the rhinoplasly and lip revision models, patients without a cleft lip were excluded. In the jaw surgery and speech surgery models, patients without a cleft palate or alveolus were excluded.

Results: The CLEFT-Q field test included 1938 participants from high-income countries. Participants who needed surgery scored significantly lower (worse) than those who have had surgery in each of the eight relevant CLEFT-Q scales ($p < 0.001$ in each ANOVA).

Conclusion: The ability of the CLEFT-Q to detect differences between groups based on surgical status further supports its validity.

Introduction

Cleft lip and/or palate (CL/P) is a congenital anomaly that occurs in 7.94 per 10,000 live births annually.[1] There is little consensus on the optimal management protocol for CL/P, but patients frequently require numerous surgical interventions throughout their lives for both functional and esthetic reasons.[2] These include (but are not limited to) operations to improve the form and function of the face (nose, lips, jaws) and to address problems related to speech and hearing.

Outcomes of CL/P treatment have typically been evaluated using clinician-derived measurements rather than measurements reported by the patients themselves.[3] Our team recently developed a patient-reported outcome measure (PROM) called the CLEFT-Q to provide a means of incorporating the patient perspective into outcome assessments.[4]

The consensus-based standards for the selection of health status measurement instruments (COSMIN) have outlined a checklist for assessing the methodological quality of studies on health status measurement instruments. The checklist considers three types of validity: content validity (whether the items of a scale are relevant to the disease in question, and cover all important elements of that disease comprehensively), construct validity (how well an instrument measures the trait(s) it has been designed to measure), and criterion validity (a comparison of the instrument with other measures used in that field).[5,6]

Construct validity is further divided into structural validity, cross-cultural validity, and hypothesis testing. Structural validity refers to the detailed psychometric properties of items and scales, and has been demonstrated, along with cross-cultural validity, by the results of the CLEFT-Q field test.[4] Hypothesis testing (which tests an outcome measure's ability to distinguish between groups that are hypothesised to differ) is required as part of a comprehensive assessment of construct validity, as described by COSMIN.[5]

We aimed to assess the construct validity of the CLEFT-Q through hypothesis testing. Specifically, we aimed to compare CLEFT-Q scores of participants who either needed, have had, or never needed four cleft-specific surgeries (jaw surgery, lip revision, rhinoplasty, and speech surgery).

Methods

The CLEFT-Q

The CLEFT-Q was developed in accordance with internationally recommended guidelines for PROM development.[7-10] Items (questions) were generated following a systematic review[11] and qualitative interviewing of 138 participants from six countries (Canada, Kenya, India, Philippines, UK, and USA).[12] Scales (groups of items measuring the same trait) were refined in three further rounds of patient interviews and with input from experts in the field of CL/P.[12] The international field-test sample of 2434 children from 12 different countries demonstrated reliability and validity of the CLEFT-Q based on the Rasch measurement theory (RMT) analysis.[4] In the field test dataset, 12 of the 13 CLEFT-Q subscales met the requirements of the Rasch model. This means that the scores for these 12 scales, which are transformed from 0 to 100 (with a higher number indicating a better outcome), provide continuous measurement (as opposed to nominal or ordinal measurement).

Each scale is independently functioning and designed to measure a unidimensional concept. In this manner, clinicians and researchers can pick and choose from the set of scales to ensure only those that are relevant to a specific context are used. In contrast, legacy scales, such as the child oral health impact profile (COHIP)[13] or pediatric quality of life inventory (PedsQL),[14] designed using the classical test theory approach, provide ordinal level measurement. Many legacy scales provide total scores that summate scores for separate scales. This approach to measurement can be problematic in studies that aim to compare treatments as it can mask effects, for example, when patients improve on some scales but not others. This situation might occur especially when the surgical objectives are for a subtle change.

The 12 scales that met the requirements of the Rasch model assess appearance (of the face, nose, nostrils, teeth, lips, jaws, and cleft lip scar), health-related quality of life (i.e. psychological, social, school, and speech-related distress), and facial function (speech). The 13th scale, designed to assess eating and drinking, did not fit the Rasch model, and instead represents a problem checklist. The scales were designed so that they could evaluate cleft-specific interventions across a broad range of age groups, from 8 to 29 years.

Table 1 Patient age, gender, and cleft types (unilateral or bilateral) stratified by surgical status.

Surgery	Surgical status	*N*	Age (years) Mean (SD)	Gender *N* (%) Male	Cleft-type *N* (%) Bilateral
Jaw surgery	Need	375	14 (4)	225 (60.0)	115 (35.8)
	Have had	118	20 (4)	54 (45.8)	44 (42.3)
	Never needed	1216	13 (4)	663 (54.6)	201 (25.9)
Lip revision	Need	280	14 (4)	164 (59.0)	92 (33.0)
	Have had	355	15 (5)	198 (55.8)	124 (35.0)
	Never needed	766	13 (4)	497 (64.9)	169 (22.2)
Rhinoplasty	Need	431	14(4)	254 (58.9)	116 (27.0)
	Have had	227	17 (5)	132 (58.1)	78 (34.4)
	Never needed	749	13 (4)	481 (64.4)	195 (26.2)
Speech surgery	Need	105	15 (5)	57 (54.3)	26 (36.6)
	Have had	450	14 (4)	266 (59.1)	122 (39.1)
	Never needed	1191	14 (4)	637 (53.5)	225 (26.3)

SD, standard deviation.

Items on the appearance scales begin with "How much do you like…" and have four response options (not at all, a little bit, quite a bit, very much). For example, the nostril scale includes the item "How much do you like how your nostrils look in photos?" and the lips scale includes the item "How much do you like the shape of your lips?" Each appearance scale contains between eight and 12 items. The speech distress scale contains ten items such as "I get upset when I need to repeat myself" and has three response options (never, sometimes, always). The speech function scale contains 12 items such as "I have trouble reading out loud" with the same three response options.

Study participants

The CLEFT-Q field test was conducted between October 2014 and November 2016 and involved 2434 participants in 12 countries. Data were collected either face-to-face or by post using paper booklets or electronic tablets. A clinical form asking about cleft type and past and future treatment needs was completed by site staff. All data were entered into research electronic data capture (REDCap)[15] hosted at McMaster University (Canada).

Participants included patients with cleft lip (CL), cleft palate (CP), cleft lip and palate (CLP), or cleft lip and alveolus (CLA) between the ages of eight and 29 years. Patients with a cognitive delay that would prevent them from self-reporting were excluded from the study. Research ethics committee approval was obtained from each of the hospitals participating in the CLEFT-Q field test. A more in-depth description of the field-test study and its participants has been published elsewhere.[4] In this analysis, only patients from high-income countries as classified by the World Bank 2016 were included to minimize the impact of any economic confounders that could result in treatment variation.[16] Patients who had completed the CLEFT-Q perioperatively were also excluded.

Data analysis

Data were analysed using SPSS (IBM SPSS Statistics, Version 25, IBM Corp). For each type of surgery (jaw surgery, lip revision, rhinoplasty, and speech surgery) patients were classified as follows: need surgery, have had surgery, and never needed surgery. Those who had undergone an operation but required additional surgery (i.e. a revision) were classified as needing surgery.

We performed eight separate one-way analysis of variance (ANOVA) analyses using scores from the following CLEFT-Q scales as the dependent variables: jaws and teeth scales for jaw surgery; lips and cleft lip scar scales for lip revision surgery; nose and nostrils scales for rhinoplasty; and speech distress and speech function scales for speech surgery. Surgical status was used as the independent variable in each analysis. Assumptions of normality and equal variance were tested with the Shapiro-Wilks test and the Brown-Forsythe test, respectively.

Patients who underwent multiple types of surgery were included in each ANOVA relevant to their treatment (for example, if a patient underwent lip revision and rhinoplasty, they would be included in the four ANOVAs comparing scores on the following scales: lip, cleft lip scar, nose, and nostril).

Patients with an isolated cleft palate were excluded from the rhinoplasty and lip revision analyses and patients with an isolated cleft lip were excluded from the jaw surgery and speech surgery analyses. Where responses were missing, patients were excluded from the corresponding ANOVA in a listwise fashion.

We hypothesised that patients who have had surgery would achieve higher CLEFT-Q scores than patients who needed surgery, in addition we hypothesised that patients who have had surgery would have similar scores to those who never needed surgery.[7,8]

Results

The CLEFT-Q field test included 1938 participants from high-income countries. The number of participants included in each ANOVA varied. Sample sizes and patient age, gender, and cleft types (unilateral or bilateral) are displayed in Table 1, stratified by surgical status.

The results of the assumption testing underlying each ANOVA are displayed in Table 2. All models failed the assumption of normality ($p \leq 0.02$), but normality plots

Table 2 Assumption testing of each one-way ANOVA model.

Scale	Surgery status	Normality		Equal variance	
		Shapiro-Wilks test result	*p*-value	Brown-Forsythe test result	*p*-value
Teeth	Need	0.976	<0.001	1.44	0.237
	Have had	0.958	0.003		
	Never needed	0.969	<0.001		
Jaw	Need	0.959	<0.001	1.899	0.15
	Have had	0.907	<0.001		
	Never needed	0.880	<0.001		
Lip	Need	0.975	<0.001	0.408	0.665
	Have had	0.959	<0.001		
	Never needed	0.955	<0.001		
Cleft lip scar	Need	0.960	<0.001	0.062	0.94
	Have had	0.945	<0.001		
	Never needed	0.935	<0.001		
Nose	Need	0.975	<0.001	1.849	0.158
	Have had	0.976	0.001		
	Never needed	0.965	<0.001		
Nostrils	Need	0.964	<0.001	1.713	0.181
	Have had	0.971	<0.001		
	Never needed	0.954	<0.001		
Speech distress	Need	0.951	0.001	1.44	0.21
	Have had	0.950	<0.001		
	Never needed	0.936	<0.001		
Speech function	Need	0.968	0.020	2.306	0.1
	Have had	0.960	<0.001		
	Never needed	0.933	<0.001		

showed no sign of platykurtosis or leptokurtosis. We proceeded to perform one-way ANOVA analyses as this test is still considered robust even after violating the normality assumption, particularly when sample sizes are large.[17] All models met the assumption of homogeneity of variance ($p \geq 0.10$).

The results of each one-way ANOVA and post hoc testing (Tukey honest significant difference tests) are displayed in Table 3. There was a statistically significant difference in mean score for all eight CLEFT-Q scales between groups defined by surgical status ($p < 0.001$). Post hoc tests showed that those who have had surgery had significantly higher (better) scores than those who needed surgery ($p \leq 0.003$). Those who never needed surgery had significantly greater mean scores than those who needed surgery for all eight scales ($p < 0.001$). There were no statistically significant differences in mean scores between those who have had or who never needed rhinoplasty, jaw surgery, or lip revision ($p \geq 0.10$). While patients who have had speech surgery reported higher (better) mean speech distress ($p = 0.003$) and speech function ($p < 0.001$) scores than those who needed speech surgery, this group had slightly lower mean scores than those who never needed surgery (speech distress $p = 0.004$; speech function $p < 0.001$).

Discussion

The hypothesis testing in this paper provide additional evidence of the construct validity of the CLEFT-Q scales. The eight scales examined could detect expected differences between subgroups of patients who either need, have had, or never needed four cleft-specific surgeries: rhinoplasty, jaw surgery, lip revision, and speech surgery. Participants who needed a specific surgical intervention attained significantly lower (poorer) scores than those who have had surgery in each of the relevant scales.

The results in this paper are in keeping with the results of a multi-center study of 1200 youth with CL/P that showed surgical candidates reporting lower oral health-related quality of life (poorer COHIP scores) than nonsurgical candidates.[18] Similar findings were also reported by Ranganathan et al. who examined the relationship between the desire for revision surgery and HRQOL in 71 children with cleft lip with or without palate (CLCP). This study found that children who scored poorer on the cleft evaluation profile (CEP), and PedsQL (psychosocial develop and communication skills) were more likely to desire revision surgery for some aspect of their cleft.[19] Ranganathan et al. also found general body image scores on the satisfaction with appearance questionnaire (SWAP) did not relate to the desire for surgical revision in children with CLCP. Although this current study examined "need" rather than desire for revision, the appearance scales on the CLEFT-Q could detect differences between the "need" group and both the "have had surgery" and "never needed surgery" groups. It is likely that targeting scales for particular areas of the face would provide a better indication of change or differences between groups in this population than a generalised appearance measure.

Table 3 Results of each one-way ANOVA model and post-hoc Tukey honest significant difference (HSD) tests.

Surgery	Scale	Surgery status	Valid n	ANOVA							Mean difference (need – not required)	*p*-value (Tukey HSD)
				Mean	SD	ANOVA *p*-value	Mean difference (need – had)	*p*-value (Tukey HSD)	Mean difference (had – not required)	*p*-value (Tukey HSD)		
Jaw Surgery	Teeth	Need	238	46.47	21.53	< 0.001	−15.60	< 0.001	4.76	0.099	−10.85	< 0.001
		Have had	101	64.95	21.35							
		Never needed	685	58.87	24.33							
	Jaw	Need	238	55.71	27.34	< 0.001	−18.44	< 0.001	1.05	0.922	−17.39	< 0.001
		Have had	101	74.15	22.32							
		Never needed	685	73.14	25.54							
Lip revision	Lip	Need	265	55.38	23.44	< 0.001	−6.83	0.001	−1.82	0.478	−8.66	< 0.001
		Have had	329	62.56	23.14							
		Never needed	672	63.98	24.36							
	Cleft lip scar	Need	265	55.91	27.45	< 0.001	−7.39	0.002	−2.26	0.422	−9.65	< 0.001
		Have had	329	63.21	26.58							
		Never needed	672	65.59	26.69							
Rhinoplasty	Nose	Need	407	49.10	20.91	< 0.001	−8.44	< 0.001	−2.97	0.182	−11.41	< 0.001
		Have had	213	57.46	21.37							
		Never needed	711	60.46	22.21							
	Nostrils	Need	407	45.29	27.06	< 0.001	−9.52	< 0.001	−2.76	0.406	−12.27	< 0.001
		Have had	213	54.81	26.48							
		Never needed	711	57.56	28.17							
Speech surgery	Speech distress	Need	95	62.38	23.35	< 0.001	−7.70	0.003	−3.92	0.004	−11.62	< 0.001
		Have had	390	69.99	20.85							
		Never needed	987	74.03	20.16							
	Speech function	Need	95	56.81	19.84	< 0.001	−12.35	< 0.001	−5.67	< 0.001	−18.02	< 0.001
		Have had	390	69.16	21.03							
		Never needed	987	74.70	20.42							

SD, standard deviation.

The ability of the CLEFT-Q to distinguish between different surgical groups is important and suggests that in future studies, it is likely CLEFT-Q will display responsiveness. Responsiveness refers to an outcome measure's ability to measure clinically meaningful change. There are various techniques to establish the minimal important difference (MID) in a scale's score that represents change that is meaningful to patients,[20] although these are contested in the literature and beyond the scope of this article.[5] Further work is required, with a variety of methods, to define the responsiveness of the CLEFT-Q and to calculate the MID in each subscales' scores.

Our study included a large international sample that represents the CLEFT-Q's target population. However, our large sample size means that small differences in mean scores became statistically significant. This may not be reproducible in smaller studies. While we controlled for economic confounding, treatment protocols are likely to have varied between and within high-income countries and this could not be controlled for. Furthermore, because surgical status was subjectively determined by clinicians, this could have varied between centres and led to bias.

Conclusion

The CLEFT-Q is a PROM for use in the evaluation of treatment outcomes for patients with CL/P aged eight to 29 years of age. The analyses in this paper demonstrate the CLEFT-Q scales' ability to distinguish between subgroups of patients who need, have had, or never needed four cleft-specific surgeries providing early evidence of its construct validity. The CLEFT-Q scales could soon be used to monitor clinical progress in everyday practice, and incorporate the patient perspective into future trials studying disease severity, treatment effectiveness, comparative treatment effectiveness, and treatment cost-effectiveness.

Declaration of Competing Interest and Financial Disclosure

The CLEFT-Q is owned by McMaster University and the Hospital for Sick Children. Anne F Klassen and Karen W.Y. Wong Riff are co-developers of the CLEFT-Q and, as such, could potentially receive a share of any license revenues based on their institutions inventor sharing policy. The other authors have no conflicts of interest to declare in relation to the content of this article.

CRediT authorship contribution statement

Conrad J. Harrison: Data curation, Formal analysis, Writing - original draft, Writing - review & editing. **Charlene Rae:** Data curation, Formal analysis, Writing - review & editing. **Elena Tsangaris:** Data curation, Formal analysis, Writing - review & editing. **Karen W.Y. Wong Riff:** Data curation, Formal analysis, Conceptualization, Writing - review & editing. **Marc C. Swan:** Writing - review & editing. **Tim E.E. Goodacre:** Conceptualization, Writing - review & editing. **Stefan Cano:** Writing - review & editing. **Anne F. Klassen:** Data curation, Formal analysis, Conceptualization, Writing - original draft, Writing - review & editing, Supervision.

Acknowledgments

The research described in this paper was supported by a grant from the Canadian Institutes of Health Research.

Financial disclosure

The CLEFT-Q is owned by McMaster University and the Hospital for Sick Children. Anne F. Klassen and Karen W.Y. Wong Riff are co-developers of the CLEFT-Q and, as such, could potentially receive a share of any license revenues based on their institutions inventor sharing policy.

References

1. Tanaka SA, Mahabir RC, Jupiter DC, Menezes JM. Updating the epidemiology of cleft lip with or without cleft palate. *Plast Reconstr Surg* 2012. doi:10.1097/PRS.0b013e3182402dd1.
2. Mossey PA, Little J, Munger RG, Dixon MJ, Shaw WC. Cleft lip and palate. *Lancet* 2009. doi:10.1016/S0140-6736(09)60695-4.
3. Jones T, Al-Ghatam R, Atack N, et al. A review of outcome measures used in cleft care. *J Orthod* 2014. doi:10.1179/1465313313Y.0000000086.
4. Klassen AF, Riff KWYW, Longmire NM, et al. Psychometric findings and normative values for the CLEFT-q based on 2434 children and young adult patients with cleft lip and/or palate from 12 countries. *CMAJ* 2018. doi:10.1503/cmaj.170289.
5. Mokkink LB, Terwee CB, Patrick DL, et al. The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: an international Delphi study. *Qual Life Res* 2010. doi:10.1007/s11136-010-9606-8.
6. Dobbs TD, Hughes S, Mowbray N, Hutchings HA, Whitaker IS. How to decide which patient-reported outcome measure to use? A practical guide for plastic surgeons. *J Plast Reconstr Aesthetic Surg* 2018. doi:10.1016/j.bjps.2018.03.007.
7. Cano SJ, Klassen A, Pusic AL. The science behind quality-of-life measurement: a primer for plastic surgeons. *Plast Reconstr Surg* 2009. doi:10.1097/PRS.0b013e31819565c1.
8. U.S. Department of Health and Human Services. Guidance for Industry. Patint-reported Outcome Measures: Use in Medical Product Development to support Labeling Claims. https://www.fda.gov/media/77832/download. Published 2009. Accessed 31 August 2019.
9. 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 Heal* 2011. doi:10.1016/j.jval.2011.06.013.
10. Lohr KN. Assessing health status and quality-of-life instruments: attributes and review criteria. *Qual Life Res* 2002. doi:10.1023/A:1015291021312.
11. Klassen AF, Tsangaris E, Forrest CR, et al. Quality of life of children treated for cleft lip and/or palate: a systematic review. *J Plast Reconstr Aesthetic Surg* 2012. doi:10.1016/j.bjps.2011.11.004.
12. Tsangaris E, Wong Riff KWY, Goodacre T, et al. Establishing content validity of the CLEFT-q. *Plast Reconstr Surg – Glob Open* 2017. doi:10.1097/GOX.0000000000001305.
13. Broder HL, McGrath C, Cisneros GJ. Questionnaire development: face validity and item impact testing of the child oral health impact profile. *Communi Dent Oral Epidemiol* 2007. doi:10.1111/j.1600-0528.2007.00401.x.

14. Varni JW, Seid M, Kurtin PS. PedsQL™ 4.0: reliability and validity of the pediatric quality of life inventory™ version 4.0 generic core scales in healthy and patient populations. *Med Care* 2001. doi:10.1097/00005650-200108000-00006.
15. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap) – a metadata-driven methodology and workflow process for providing translational research informatics support. *J Biomed Inform* 2009. doi:10.1016/j.jbi.2008.08.010.
16. The World Bank. World Bank Country and Lending Groups. https://datahelpdesk.worldbank.org/knowledgebase/articles/906519-world-bank-country-and-lending-groups. Published 2016. Accessed 12 December 2018.
17. Laerd Statistics. One-Way ANOVA. https://statistics.laerd.com/statistical-guides/one-way-anova-statistical-guide-3.php. Published 2018. Accessed 31 August 2019.
18. Broder HL, Wilson-Genderson M, Sischo L, Norman RG. Examining factors associated with oral health-related quality of life for youth with cleft. *Plast Reconstr Surg* 2014. doi:10.1097/PRS.0000000000000221.
19. Ranganathan K, Shapiro D, Aliu O, et al. Health-related quality of life and the desire for revision surgery among children with cleft lip and palate. *J Craniofac Surg* 2016. doi:10.1097/SCS.0000000000002924.
20. Revicki D, Hays RD, Cella D, Sloan J. Recommended methods for determining responsiveness and minimally important differences for patient-reported outcomes. *J Clin Epidemiol* 2008. doi:10.1016/j.jclinepi.2007.03.012.