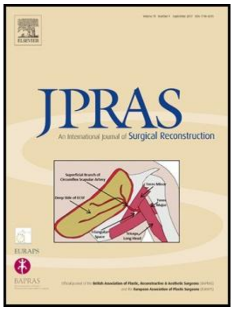
Paper:
Dobbs, T., Hughes, S., Mowbray, N., Hutchings, H. & Whitaker, I. (2018). How to decide which patient-reported outcome measure to use? A practical guide for plastic surgeons. *Journal of Plastic, Reconstructive & Aesthetic Surgery*
http://dx.doi.org/10.1016/j.bjps.2018.03.007



Title: How to decide which patient-reported outcome measure to use? A
practical guide for plastic surgeons

Author: Thomas Dobbs, Sarah Hughes, Nicholas Mowbray, Hayley A.
Hutchings, Iain S. Whitaker

PII: S1748-6815(18)30097-4
DOI: https://doi.org/10.1016/j.bjps.2018.03.007
Reference: PRAS 5623

To appear in: *Journal of Plastic, Reconstructive & Aesthetic Surgery*

Received date: 13-10-2017
Accepted date: 10-3-2018

Please cite this article as: Thomas Dobbs, Sarah Hughes, Nicholas Mowbray, Hayley A. Hutchings, Iain S. Whitaker, How to decide which patient-reported outcome measure to use? A practical guide for plastic surgeons, *Journal of Plastic, Reconstructive & Aesthetic Surgery* (2018), https://doi.org/10.1016/j.bjps.2018.03.007.

How to decide which patient-reported outcome measure to use? A practical guide for plastic surgeons

Mr Thomas Dobbs MA, BM BCh, MRCS

1 Reconstructive Surgery & Regenerative Medicine Research Group, Institute Of Life Sciences, Swansea University Medical School, Swansea, UK

2 Welsh Centre for Burns and Plastics, Morriston Hospital, Swansea, UK

Mrs Sarah Hughes, BSc, MHSc, MRCSLT

3 Health Services Research, Institute of Life Sciences, Swansea University Medical School, Swansea, UK

4 Abertawe Bro Morgannwg University Health Board, Princess of Wales Hospital, Bridgend, UK

Mr Nicholas Mowbray MBBCH, MRCS

3 Health Services Research, Institute of Life Sciences, Swansea University Medical School, Swansea, UK

Professor Hayley A Hutchings BSc, PhD

3. Health Services Research, Institute of Life Sciences, Swansea University Medical School, Swansea, UK

Professor Iain S Whitaker MA Cantab MBBChir, PhD, FRCS(Plast)

1 Reconstructive Surgery & Regenerative Medicine Research Group, Institute

Of Life Sciences, Swansea University Medical School, Swansea, UK

2 Welsh Centre for Burns and Plastics, Morriston Hospital, Swansea

Corresponding Author

Mr Thomas Dobbs MA, BM BCh, MRCS

Reconstructive Surgery & Regenerative Medicine Research Group, Institute Of Life Sciences, Swansea University Medical School, Swansea, UK

E: tomdobbs@doctors.org.uk

T: 07973986658

Meetings

This work has yet to be presented

Summary

The use of patient-reported outcome measures (PROMs) is rising across all medical specialties as their importance to patient care is validated. They are likely to play a particularly important role in plastic and reconstructive surgery where outcomes are often subjective, and the recent guidance from the Royal College of Surgeons of England advising their use in cosmetic surgery highlight this. In order to drive their routine use across our specialty it is important that clinicians are able to understand the often complex and confusing language that surrounds their design and validation. In this article we describe the process of PROM design and validation, and attempt to 'demystify' the language used in the health outcome literature. We present the important steps that a well-designed PROM must go through and suggest a straightforward guide for selecting the most appropriate PROMs for use in clinical practice. We hope that this will encourage greater use of PROM data across plastic and reconstructive surgery and ultimately help improve outcomes for our patients.

Key Words: Patient-reported outcome measures; PROM; plastic surgery; reconstructive surgery; guide

Introduction

Patient reported outcome measures (PROMs) are standardised, validated questionnaires that are completed by patients and capture one or more aspects of their health and wellbeing[1,2]. In a world where shared-decision making between clinicians' and patients' is encouraged[3], traditional measures of health outcomes have needed to change from traditional assessments conducted from the surgeon's perspective (e.g., do we as the operating surgeon think that the patient has had a "good" outcome) to encompass a more holistic and patient-centred view. Moreover, the definition of health has evolved to include outcomes such as happiness, quality of life and the ability to perform tasks of daily living. This change is so important that the World Health Organisation (WHO) defines health as 'a state of physical, mental and social well-being and not just the absence of disease or infirmity'[4]. PROMs are therefore designed to encompass and measure these aspects of health that can either not be directly observed or it is not feasible to observe[5].

Many PROMs were originally developed for assessing treatment effectiveness in the context of clinical trials[7]. They are, however, becoming more commonly used in other situations, such as routine monitoring of treatment effect and health-care service provision. NHS England has orchestrated a national PROMs programme since 2009, requiring routine collection of PROMs data for all those undergoing hip and knee replacement surgery, inguinal hernia surgery and varicose vein surgery[2,7]. More recently the Royal College of Surgeons of England advocated the routine collection of PROMs for a number of cosmetic procedures, using three prominent questionnaires, BREAST-Q[8], FACE-Q[9] and BODY-Q[10].

There are a number of benefits to incorporating PROM data into research and routine clinical practice, especially in a specialty such as plastic and reconstructive surgery where objective outcomes can be difficult to quantify. It is important that we have patient-reported data to advocate certain treatments for patients, especially in the current climate where rationing of procedures is occurring. Many regulatory bodies also demand the inclusion of patient-reported data in applications[11]. The drive for value-based healthcare requires the wider adoption of PROMs to measure health outcomes across different providers and healthcare settings[12,13] and the King's Fund report suggests that PROMs are likely to become "a key part of how health care is funded, provided and managed"[2].

Types of PROM

PROMs are typically classified as generic or disease-specific. Generic PROMs such as the EQ-5D, which is a measure of health status and SF-6D, which measures quality of life, are designed to be applied across different disease states[14]. These generic PROMs allow comparisons of quality of life across a wide range of conditions. Disease-specific (also known as condition-specific), are as the name indicates, specific to certain diseases or body areas. Unlike generic PROMs they are able to discriminate with greater sensitivity between individuals with specific conditions. A wide range of disease-specific PROMs are available in the plastic and reconstructive surgery literature (*Table 1*). PROMs are delivered in a questionnaire format, which can be administered in various ways, such as paper or computer based, or online platforms. Each question is usually scored on a Likert-type scale, with scores summed to give a total score for the underlying group of questions or 'construct'. In some instances, questions are given different weights based on their importance in contributing to the total score[15]. Typically, the PROM is applied at more than one time point during the patient pathway, allowing comparison of scores (either from the same person or pooled scores from multiple patients), pre- and post-intervention, or to evaluate changes in disease course.

Assessing the quality of a PROM

Given the ever-expanding range of PROMs it is important that clinicians and researchers are able to appraise and choose the best PROMs for their needs. In choosing which PROM to use one needs to take into account its application (clinical versus research), the condition being investigated and the validity of the PROM. The 'validity' is the extent to which a PROM measures what it intends to measure or what can be concluded about a patient's health condition based on a particular score. There have been several publications in which standards for assessing specific measurements of a PROM have been discussed, including the scientific advisory committee of the Medical Outcome Trust[16], Evaluating the Measurement of Patient-Reported Outcomes (EMPRO) tool[17], the Food and Drug Administration (FDA)[18], McDowell and Jenkinson[19], Bombardier and Tugwell[20], Andresen[21], Streiner[22], DeVon et al[23] and Terwee et al[24]. However, these publications are largely written for the health outcome specialist audience and are therefore complex and confusing for those not familiar with the literature and certain concepts are taken for granted. In order to encourage the use of PROMS in clinical practice and research, it would be beneficial for the process of design and validation to be understandable to clinicians.

This article builds on the recent publication in this journal by Wormald and Rodrigues[25], demystifying the process of PROM development and validation required for a 'good' PROM. We present a guide to choosing which PROM to use (*Figure 1*) along with a practical assessment tool for clinicians to assess PROMs, allowing them to pick the most appropriate and valid PROM for their condition or patient group of interest.

Aspects of PROM design and validation

Item selection

The first stage in any PROM development is to generate a pool of items that cover all aspects of the area of interest[27,27]. 'Items' are the questions that will be included in the questionnaire and can be derived from five main sources: literature review; patients; clinical observations; expert opinion and generic item banks.

A *literature review* is the most common way to begin developing a PROM[27]. It aims to identify PROMs that have already been developed and used in the clinical area of interest, with the questions from these PROMs possibly used in the development of a new PROM or the adaptation of an existing one. This has obvious time efficiencies when using items that have already undergone construction and psychometric evaluation.

Patients are often the most useful source of item generation and inclusion of patients is considered by the FDA to be the most important source of item generation. Many well developed PROMs in plastic surgery follow this approach[28-30].

Clinical observation is a particularly fruitful source of items, however in modern PROM development clinical observation should not be used alone.

Expert opinion is commonly employed to either generate items or comment on those that have already been suggested. However, as with patient-derived items, it is important to aim for a heterogeneous mix of individuals within the group to minimize bias when selecting questions for inclusion in a new PROM.

Finally, items can be sourced from an *item bank* such as the Patient Reported Outcome Measurement Information System (PROMIS). They are sponsored by the National Institutes of Health in the United States to develop a standardized, validated *item bank*[31,32]. (http://www.nihpromis.com).

Readability of items and cross-cultural adaptation

Once the initial pool of items has been generated using a combination of the above methods, items should be checked for complex or technical language. Each question should also follow the US Department of Health and Human Services (USDHHS) recommendation for patient-orientated health literature and be written at or below the sixth-grade level, equating to a UK reading age of 11-12 years[33]. Furthermore, it is important to note that PROMs are language specific. That is, in order that the items are understandable and have the same meaning in a different language, they require translation and cross-cultural adaptation[34], such as with the Spanish version of the Skin Cancer Index[35].

Item piloting

Following the process of item generation, a PROM should be piloted in a group of patients to determine their face and content validity and to select those items that are most relevant.

Face validity refers to whether the questions appear to be assessing the desired qualities (i.e. are they on the surface measuring what they actually are) while *content validity* is concerned with whether the whole instrument (the entire questionnaire) is measuring all that is relevant and important to the patient and their condition[15]. The results of the piloted questions can then be subject to a number of statistical methods to identify those items that are most relevant, which can then be taken forward for further psychometric evaluation. These initial items should be assessed for:

- *Frequency of endorsement:* The frequency of endorsement is a measure of the proportion of people that give a different response to each item in the questionnaire. If items (questions) have a response that either has a lot of people or very few people answer the same way it will add little to the overall questionnaire and should be eliminated. In practice only those items with a frequency of endorsement between 0.2 and 0.8 (between 20% and 80% of people answering with the same response for an item) should be retained[27].
- *Item-total correlation:* Item-total correlation (also known as item-partial total correlation) is the correlation between the individual item and the total scale score omitting that item. Generally items that correlate below 0.3 and above 0.7 should be considered for removal[36] as this indicates they are either not relevant or redundant.
- *Internal consistency:* Internal consistency is a measure of the homogeneity of a scale that contains multiple items (i.e., it assesses the extent to which each item is measuring the same concept)[15,24]. Two statistical measures for internal consistency are described, the Kuder-Richardson formula and Cronbach's alpha[37,38], although Cronbach's alpha is much more commonly used. Values for Cronbach's alpha range between 0 and 1, with a value of >0.70 suggested as the minimum requirement for internal consistency[15]. An upper limit has also been suggested of 0.95, because above this, items correlate too closely and therefore there is redundancy in the scale[24].
- *Factor analysis:* Factors analysis is a statistical method used to explain the correlation between different variables and therefore the underlying structure

of the scale (i.e., the different aspects of the “latent” or underlying trait that the scale is purporting to measure). It allows the user to determine if the scale is unidimensional (i.e., measures a single attribute) and helps identify those items that are not contributing to the scale and therefore could be removed[39]. Two basic types of factor analysis exist, exploratory factor analysis (EFA) and confirmatory factor analysis (CFA). The type of factor analysis used depends on whether the PROM developers are seeking to reduce the number of items in a scale/ instrument or confirm its underlying structure[15].

- *Principal component analysis:* Principal component analysis (PCA) is similar to factor analysis in that it is a data reduction technique used to identify items that are redundant and therefore could be removed from a scale. Many people have argued over the difference between PCA and factor analysis, however, both are used in the health outcomes literature. Data from PCA is expressed as an eigenvector and an eigenvalue. According to Kaiser’s rule those factors with an eigenvalue values of 1 and above should be retained[40]. There will usually be more than one principal component (also known as uncorrelated variables), but the total number should be less than or equal to the number of original items and have therefore reduced the quantity of data.

Reliability

Reliability refers to how consistent the results of a scale are when applied in different situations[41]. It reflects the amount of random and systematic error that is inherent in any measurement and results from the interaction of the instrument, the specific group using the instrument and the situation. Therefore it is not specifically the reliability of the instrument per se but the reliability of the results obtained when the instrument is used in that particular manner[42].

A reliability coefficient is calculated to illustrate the degree to which a PROM can differentiate between different patients. A number of different statistical methods exist for calculating the reliability coefficient, but the three most commonly seen in the literature are the Pearson correlation coefficient, the Intra-class correlation coefficient (ICC) and Kappa coefficient[15]. Scores are reported between one and zero, such that one equals perfect reliability and zero no reliability. There is little consensus on what the value of reliability should be, although it is commonly quoted that it should be greater than 0.70[43], with some authors suggesting a coefficient of greater than 0.90 for clinical tools[15]. Whilst a higher reliability coefficient indicates the test is more reliable, if it is too close to 1 then this may suggest that important items reflecting the full scope of the condition in question have been omitted, thus reducing the usefulness of the questionnaire[44].

Different forms of reliability are reported in the literature, usually dependent on the study design and type of questionnaire being assessed. Commonly used forms of reliability include:

- *Test-retest reliability* is a measure of the reliability of the instrument over the passage of time. It is assessed by getting the same group of patients to answer the questionnaire at two time points, separated by an appropriate period of time, usually 2-14 days[15]. It is important that the condition of interest in the group used for test-retest reliability is not changed during this period.
- *Inter-observer reliability* refers to the reliability of the scores between different observers (i.e., when two clinicians score the same patient at the same time the scores should be the same).

- *Intra-observer reliability* is the consistency of the instrument when the same individual is assessed by the same clinician on two separate occasions. Intra- and inter-observer reliability cannot be applied to those instruments that are self-administered.
- Finally, *parallel-forms reliability* refers to how consistent the scores are when an individual takes two or more forms of the same questionnaire. It is used when comparing questionnaires that are thought to be assessing the same domains[15].

Validity

The validity of a health outcome measure refers to whether or not it is able to measure exactly what it is intended to measure and, therefore, can accurate conclusions about the presence and degree of the attribute be deduced? It is based on inference (e.g., a person who scores highly on a measure of distress would be expected to be more distressed than someone who doesn't). Traditionally there have been three basic forms of validity[45]; content validity, criterion validity and construct validity, however further sub-types have been developed and may be reported in the literature.

Face validity is more commonly used as part of the item generation and reduction stage and has been described in more detail above.

Content validity is a judgment assessment of whether the items in a scale encompass all relevant and important areas of the concept being measured in appropriate detail[15]. This is commonly assessed by experts in the field and just reported in papers as having been carried out. More recently however there has been a drive to quantify the degree of content validity, with it being suggested that all instrument development studies should report on the content validity assessment[46]. This may be documented in the form of a content validity index (CVI), where a total scale score of 0.90 or above is considered to be excellent content validity[47].

Criterion validity compares the instrument under study to another measure of the subject of interest (ideally the 'gold standard') and assesses how well they correlate[15]. Two types of criterion validity are described: *concurrent validity* and *predictive validity*. *Concurrent validity* applies when the new scale and the 'gold standard' are administered at the same time. Correlation between the scores are typically assessed with a phi coefficient of Pearson correlation coefficient, a positive correlation being considered to be greater than 0.70. *Predictive validity* is performed when the outcome being measured occurs in the future and one is trying to determine if the new instrument is able to predict this future event and therefore give an answer earlier than the current instrument. In this case the new scale would be administered at time point 1 and then the old measure used at time point 2 in the future. Scores are then compared to see if the new measure is able to predict future outcomes[15].

Construct validity is the term used to describe the relationships between various, non-measurable factors that combine to describe something we can observe. For example, anxiety is not an 'observable' trait, but the many symptoms and signs which we attribute to anxiety can be combined into a construct to represent anxiety which can be measured. Construct validity is therefore seen as an overarching term used to encompass all forms of validity and refers to how well a measure or questionnaire is able to assess the construct that it is trying to assess[15,48,49]. It is assessed by making hypotheses as to how this measure will correlate when assessed against other measures of the same construct. The hypotheses that are generated are either a positive correlation (termed *convergent validity*) or a negative correlation (known as *divergent validity*). It is important that the hypotheses are stated in advance to avoid bias that may occur if hypotheses are developed retrospectively to fit the observed correlations between scales. Terwee et al specified that for a questionnaire to be rated as positive for construct validity, hypotheses should not only be specified in advance but that at least 75% of the results should agree with these hypotheses in a group of at least 50 patients[24].

Responsiveness and sensitivity

Responsiveness is defined as the 'ability of an instrument to measure a clinically important change' while sensitivity is the 'ability of an instrument to measure *any* change regardless of whether it is clinically meaningful'[50,51]. Many variations on these definitions are described[52-54] in the literature. Assessment of responsiveness should be based on hypothesis testing, in a similar manner to construct validity, with hypotheses made regarding the expected differences in change between 'known' groups[24].

Two broad approaches exist to test responsiveness, either anchor-based or distribution-based[55]. In an anchor-based approach the relationship between the change in the instrument score and an external variable (such as a patient-reported change in their condition or laboratory measurements) is measured. Statistical analysis of this involves measuring the area under curve (AUC)[56], with a score of > 0.70 considered adequate[24]. Distribution-based methods are based on statistical characteristics of the sample. A wide range of statistical methods exist, but the most commonly reported include Cohen's effect size[15], Guyatt's responsiveness ratio[57] and the standardized response mean[58].

Interpretability

Interpretability is closely related to responsiveness. In clinical practice a PROM that has all the required attributes above but does not have clinical meaning is potentially useless. Therefore, interpretability refers to the degree to which one can assign qualitative meaning to the quantitative score of the instrument[24,59]. To establish the interpretability of a PROM the minimal important difference (MID) also known as the minimal important change (MIC)[60], standard error of measurement (SEM), and smallest detectable change (SDC) should be calculated. The MID is defined by Jaeschke et al as 'the smallest difference in score in the domain of interest which patients perceive as beneficial and which would mandate in the absence of trouble-some side-effects and excessive cost, a change in the patient's management'[61]. It is recommended that the MID/MIC is defined by the study team and provide information as to what change in score would be considered to be clinically meaningful.

Assessing floor and ceiling effects can also be useful in helping to understand interpretability. The instrument is considered to have a floor or ceiling effect when 15% of respondents achieve either the lowest or highest possible score respectively[24]. If a floor or ceiling effect exists it will leave those patients who score at the extremes with only one direction in which they can move on the scale and thus both the responsiveness and interpretability in these groups is diminished.

A glossary of terms used in PROM development and validation is presented in *Table 2*.

Classical Test Theory versus Modern Test Theory

So far we have presented the psychometric properties of classical test theory (CTT), the traditional and most commonly used technique for the development and validation of PROMs to date[62,63]. However, another more modern psychometric technique used in PROM development and validation is item response theory (IRT)[64]. This is being increasingly used to validate plastic surgery related PROMs, such as the 'Q-series'[30]. In CTT the underlying assumption is that the observed score is a combination of the true score plus a degree of random error and that because the random error is normally distributed the expected value of all random errors equals zero[65]. This leads to a number of problems, such as the established psychometric properties only relate to the specific population and situation in which the questionnaire was developed, the assumption that all items in the scale contribute equally to the final score and difficulty with equating scores that someone achieves on different tests[15,66]. IRT aims to overcome these issues by focusing on individual items in the questionnaire rather than the overall or test-level score. It assumes that all items are measuring the same underlying construct, but that individual items have different weights and therefore do not contribute equally to the final score. IRT uses the principle of latent traits (as discussed above in *factor analysis*) and log odds units (Logits) to allow the creation of an interval scale. This allows the final scale to be truly used to determine if a patients' condition has changed and by what degree[15].

IRT is an over-riding term given to a number of different statistical methods, one of which being Rasch measurement theory (RMT)[67]. For more information on the difference between IRT and Rasch please see Cano and Hobart, 2011[66]. Many of the concepts described above for the development and validation of a PROM are the same whether CTT or IRT is used. Items still need to be developed and reduced using techniques such as exploratory and confirmatory factor analysis, which at the same time confirms whether there is unidimensionality in the items (meaning all items are measuring the same underlying construct and an assumption that is key to IRT). Further statistical testing to determine local independence (items are independent of one another and the answer to one does not depend on the answer to another) and item fit provides evidence that the instrument is both valid and reliable. When using Rasch analysis the Person Separation Index should be calculated to aid in the determination of reliability[15,68].

As a result of IRT assessing item-level psychometrics, questionnaires developed using IRT can be used in a process called adaptive testing. In CTT the questionnaire is generally only deemed valid and reliable if all items are administered, however with adaptive testing different subsets of items are given to different patients based on their answers to preceding items, usually facilitated by a computer programme and termed computerized adaptive testing (CAT). Overall questionnaire scores can still be compared between individuals and this approach has the advantage of 'tailoring' the questionnaire to the patient, therefore reducing the responder burden[25,68].

Many consider IRT to now be the 'gold standard' technique for developing and validating a PROM. Despite this there are drawbacks to its use, such as requiring larger sample sizes, added expertise in the study team and consequently greater development costs[68]. Furthermore, strict assumptions in the model can mean that items may be rejected even when they have good content validity if they do not fit the IRT model. CTT should therefore not be disregarded and many argue that it has a role to play in the validation process alongside IRT. Furthermore, it is likely that when reviewing the current literature clinicians will more commonly encounter the principles of CTT, given that many currently used PROMs were developed a number of years ago. The quality checklist proposed below therefore incorporates both CTT and elements of IRT.

Quality checklist

We propose the use of a simple checklist (*Table 3*) when appraising a PROM, which covers the five most important aspects of its design and validity testing: 1) item generation, 2) reliability, 3) validity, 4) responsiveness and 5) interpretability. This is an adaption from Alrubaiy et al.[69].

Discussion

With the increasing use of PROMs in research and clinical practice it is important that clinicians understand their development, validation and use. Without this understanding they will be at a loss when involved in clinical studies, appraising research papers and asked to collect patient-reported outcomes data in their routine clinical practice. This paper has been written to help the practicing plastic and reconstructive surgeon understand the main components that make up the design and validation of a good quality PROM. We have also included a simplified assessment checklist, which can be used when appraising different PROMs and aid in decision making as to which one to use.

Choosing the right PROM to use is very important, particularly given their increased use in routine clinical practice[70]. They must be psychometrically valid[18] and clinically meaningful. The fact that many questionnaires are too long and cumbersome, disincentivising patients from completing them[71] lends further weight to the importance of understanding the criteria of a good-quality PROM. Future efforts in the design of new PROMs (or the adaptation of old ones) need to focus on making simple, but still clinically meaningful and discriminatory PROMs. Further advances will be made in PROM development and validation by improving the ease and speed of completion and data collection and synthesis through the use of web or tablet-based platforms[72]. Integration of PROM data with other outcome measures will increase the power of big data outputs in plastic surgery, driving innovation and improving patient care.

This paper is not meant to be a detailed description of all aspects of PROM design and validation as there are many other excellent resources that cover this. We hope that through highlighting the important areas of a validated PROM, plastic surgeons will feel more comfortable in appraising a PROM for its appropriateness for their needs and the quality of its development and validation. We hope that this practical guidance will not only increase the quality of PROMs used, but will also increase uptake of their use in routine clinical practice. We all want what is best for our patients' and by asking for their opinion on their condition and treatment outcomes it is hoped that plastic and reconstructive surgery will continue to improve those aspects of patients' lives that matter most to them.

Conflicts of Interest:

We have no conflicts of interest to declare.

Acknowledgements of Funding:

Mr Thomas Dobbs is supported by the Welsh Clinical Academic Training Pathway and Professor Iain Whitaker currently holds a Royal College of Surgeons Pump Priming Grant (2017-18).

References

1. McGrail K, Bryan S, Davis J. Let's all go to the PROM: the case for routine patient-reported outcome measurement in Canadian healthcare. Healthc Pap. 2011;11(4):8–18–discussion55–8.

2. Devlin NJ, Appleby J. Getting the most out of PROMS. Putting health outcomes at the heart of NHS Decision-Making. 2010.

3. Lansley A. Equity and excellence: liberating the NHS. Department of Health; 2010.

4. Grad FP. The Preamble of the Constitution of the World Health Organization. Bull World Health Organ. World Health Organization; 2002;80(12):981–4.

5. Davidson M, Keating J. Patient-reported outcome measures (PROMs): how should I interpret reports of measurement properties? A practical guide for clinicians and researchers who are not biostatisticians. Br J Sports Med. BMJ Publishing Group Ltd and British Association of Sport and Exercise Medicine; 2014 May 1;48(9):792–6.

6. Fitzpatrick R, Davey C, Buxton MJ, Jones DR. Evaluating patient-based outcome measures for use in clinical trials. HEALTH TECHNOL ASSESS. 1998;2(14):i–iv–1–74.

7. Health DO. Guidance on the routine collection of Patient Reported Outcome Measures (PROMs). 2008 Dec 23;:1–28.

8. Cohen WA, Mundy LR, Ballard TNS, et al. The BREAST-Q in surgical research: A review of the literature 2009-2015. J Plast Reconstr Aesthet Surg. 2016 Feb;69(2):149–62.

9. Klassen AF, Cano SJ, Scott A, Snell L, Pusic AL. Measuring patient-reported outcomes in facial aesthetic patients: development of the FACE-Q. Facial plast Surg. © Thieme Medical Publishers; 2010 Aug;26(4):303–9.

10. Klassen AF, Cano SJ, Alderman A, et al. The BODY-Q: A Patient-Reported Outcome Instrument for Weight Loss and Body Contouring Treatments. Plast Reconstr Surg Glob Open. 2016 Apr;4(4):e679.

11. Revicki DA. FDA draft guidance and health-outcomes research. The Lancet. Elsevier; 2007 Feb 17;369(9561):540–2.

12. Porter ME, Larsson S, Lee TH. Standardizing Patient Outcomes Measurement. N Engl J Med. Massachusetts Medical Society; 2016 Feb 11;374(6):504–6.

13. Porter ME, Lee TH. The strategy that will fix health care. Harv Bus Rev. 2013.

14. Brazier J, Roberts J, Tsuchiya A, Busschbach J. A comparison of the EQ-5D and SF-6D across seven patient groups. Health Economics. John Wiley & Sons, Ltd; 2004 Sep 1;13(9):873–84.

15. Streiner DL, Norman GR, Cairney J. Health Measurement Scales. Fifth Edition. 2015 Jan 1.

16. Aaronson N, Alonso J, Burnam A, et al. Assessing health status and quality-of-life instruments: attributes and review criteria. Qual Life Res. 2002

May;11(3):193–205.

17. Valderas JM, Ferrer M, Mendívil J, et al. Development of EMPRO: a tool for the standardized assessment of patient-reported outcome measures. Value Health. 2008 Jul;11(4):700–8.

18. Health USDO, Human Services FDA Center for Drug Evaluation, Research USDOH, Human Services FDA Center for Biologics Evaluation, Research USDOH, Human Services FDA Center for Devices, et al. Guidance for industry: patient-reported outcome measures: use in medical product development to support labeling claims: draft guidance. Health Qual Life Outcomes. BioMed Central; 2006;4(1):79.

19. McDowell I, Jenkinson C. Development Standards for Health Measures. Journal of Health Services Research. SAGE PublicationsSage UK: London, England; 1996 Oct 1.

20. Bombardier C, Tugwell P. Methodological considerations in functional assessment. J Rheumatol Suppl. 1987 Aug;14 Suppl 15:6–10.

21. Andresen EM. Criteria for assessing the tools of disability outcomes research. Arch Phys Med Rehabil. 2000 Dec;81(12 Suppl 2):S15–20.

22. Streiner DL. A checklist for evaluating the usefulness of rating scales. Can J Psychiatry. 1993 Mar;38(2):140–8.

23. DeVon HA, Block ME, Moyle-Wright P, et al. A psychometric toolbox for testing validity and reliability. J Nurs Scholarsh. 2nd ed. Blackwell Publishing Inc; 2007;39(2):155–64.

24. Terwee CB, Bot SDM, de Boer MR, et al. Quality criteria were proposed for measurement properties of health status questionnaires. Journal of Clinical Epidemiology. 2007 Jan;60(1):34–42.

25. Wormald JCR, Rodrigues JN. Outcome measurement in plastic surgery. J Plast Reconstr Aesthet Surg. 2017;doi: 10.1016/j.bjps.2017.11.015

26. Keszei AP, Novak M, Streiner DL. Introduction to health measurement scales. Journal of Psychosomatic Research. Elsevier; 2010 Apr 1;68(4):319–23.

27. Streiner DL, Norman GR. Health measurement scales: a practical guide to their development and use 4 edition Oxford University Press. New York; 2008.

28. Cano SJ, Browne JP, Lamping DL, Roberts AHN, McGrouther DA, Black NA. The Patient Outcomes of Surgery-Head/Neck (POS-head/neck): a new patient-based outcome measure. J Plast Reconstr Aesthet Surg. 2006;59(1):65–73.

29. Matthews BA, Rhee JS, Neuburg M, Burzynski ML, Nattinger AB. Development of the facial skin care index: a health-related outcomes index for skin cancer patients. Dermatologic Surgery. Blackwell Publishing Inc; 2006 Jul;32(7):924–34–discussion934.

30. Klassen AF, Cano SJ, Schwitzer JA, et al. Development and Psychometric Validation of the FACE-Q Skin, Lips, and Facial Rhytids Appearance Scales and Adverse Effects Checklists for Cosmetic Procedures. JAMA Dermatol. 2016 Mar 2.

31. Cella D, Yount S, Rothrock N, et al. The Patient-Reported Outcomes Measurement Information System (PROMIS): progress of an NIH Roadmap

cooperative group during its first two years. Medical care. NIH Public Access; 2007 May;45(5 Suppl 1):S3–S11.

32. Cella D, Riley W, Stone A, et al. The Patient-Reported Outcomes Measurement Information System (PROMIS) developed and tested its first wave of adult self-reported health outcome item banks: 2005–2008. Journal of Clinical Epidemiology. Elsevier; 2010 Nov 1;63(11):1179–94.

33. Services UDOHAH. National action plan to improve health literacy. Washington; 2010.

34. Wild D, Grove A, Martin M, et al. Principles of Good Practice for the Translation and Cultural Adaptation Process for Patient-Reported Outcomes (PRO) Measures: Report of the ISPOR Task Force for Translation and Cultural Adaptation. Value in Health. 2005 Mar;8(2):94–104.

35. de Troya-Martín M, Rivas-Ruiz F, Blázquez-Sánchez N, et al. A Spanish version of the Skin Cancer Index: a questionnaire for measuring quality of life in patients with cervicofacial nonmelanoma skin cancer. British Journal of Dermatology. 2015 Jan;172(1):160–8.

36. Kline P. A Handbook of Test Construction. Routledge, 2015.

37. Kuder GF, Richardson MW. The theory of the estimation of test reliability. Psychometrika. Springer-Verlag; 1937;2(3):151–60.

38. Cronbach LJ. Coefficient alpha and the internal structure of tests. Psychometrika. Second ed. Springer-Verlag; 1951;16(3):297–334.

39. Floyd FJ, Widaman KF. Factor analysis in the development and refinement of

clinical assessment instruments. Psychological Assessment. American Psychological Association; 1995 Sep 1;7(3):286–99.

40. Kaiser HF. An index of factorial simplicity. Psychometrika. Springer-Verlag; 1974;39(1):31–6.

41. Testa MA, Simonson DC. Assessment of Quality-of-Life Outcomes. http://dxdoiorg/101056/NEJM199603283341306. Massachusetts Medical Society; 1996 Mar 28;334(13):835–40.

42. Gronlund NE, Linn RL. Measurement and evaluation in teaching. 6 ed. Macmillan Publication Company, 1990; 1990.

43. Terwee CB, Dekker FW, Wiersinga WM, Prummel MF, Bossuyt PMM. On assessing responsiveness of health-related quality of life instruments: Guidelines for instrument evaluation. Qual Life Res. Kluwer Academic Publishers; 2003;12(4):349–62.

44. Donovan JL, Frankel SJ, Eyles JD. Assessing the need for health status measures. J Epidemiol Community Health. BMJ Publishing Group; 1993 Apr;47(2):158–62.

45. Kaplan RM, Bush JW, Berry CC. Health status: types of validity and the index of well-being. Health Serv Res. Health Research & Educational Trust; 1976;11(4):478–507.

46. Froman RD, Schmitt MH. Thinking both inside and outside the box on measurement articles. Research in Nursing &amp; Health. Wiley Subscription Services, Inc., A Wiley Company; 2003 Oct 1;26(5):335–6.

47. Polit DF, Beck CT, Owen SV. Is the CVI an acceptable indicator of content validity? Appraisal and recommendations. Research in Nursing &amp; Health. Wiley Subscription Services, Inc., A Wiley Company; 2007 Aug 1;30(4):459–67.

48. Nunnally JC, Bernstein IH, Berge J. Psychometric theory. 1967.

49. Jenney ME, Campbell S. Measuring quality of life. Arch Dis Child. BMJ Publishing Group; 1997 Oct;77(4):347–50.

50. Guyatt GH, Deyo RA, Charlson M, Levine MN, Mitchell A. Responsiveness and validity in health status measurement: A clarification. Journal of Clinical Epidemiology. 1989 Jan;42(5):403–8.

51. Liang MH. Longitudinal construct validity: establishment of clinical meaning in patient evaluative instruments. Medical care. 2000 Sep;38(9 Suppl):II84–90.

52. Testa MA, Nackley JF. Methods for quality-of-life studies. Annual review of public health. 1994.

53. Guyatt GH, Naylor CD, Juniper E, Heyland DK, Jaeschke R, Cook DJ. Users' Guides to the Medical Literature: XII. How to Use Articles About Health-Related Quality of Life. JAMA. American Medical Association; 1997 Apr 16;277(15):1232–7.

54. Anderson JJ, Chernoff MC. Sensitivity to change of rheumatoid arthritis clinical trial outcome measures. J Rheumatol. 1993 Mar;20(3):535–7.

55. Wyrwich KW, Wolinsky FD. Identifying meaningful intra-individual change standards for health-related quality of life measures. J Eval Clin Pract. 7 ed.

Blackwell Science Ltd; 2000 Feb 1;6(1):39–49.

56. Deyo RA, Centor RM. Assessing the responsiveness of functional scales to clinical change: an analogy to diagnostic test performance. J Chronic Dis. 1986;39(11):897–906.

57. Guyatt G, Walter S, Norman G. Measuring change over time: assessing the usefulness of evaluative instruments. J Chronic Dis. 1987;40(2):171–8.

58. Revicki D, Hays RD, Cella D, Sloan J. Recommended methods for determining responsiveness and minimally important differences for patient-reported outcomes. Journal of Clinical Epidemiology. 2008 Feb;61(2):102–9.

59. Lohr KN, Aaronson NK, Alonso J, et al. Evaluating quality-of-life and health status instruments: development of scientific review criteria. Clinical Therapeutics. 1996 Sep;18(5):979–92.

60. King MT. A point of minimal important difference (MID): a critique of terminology and methods. Expert Rev Pharmacoecon Outcomes Res. 2011 Apr;11(2):171–84.

61. Jaeschke R, Singer J, Guyatt GH. Measurement of health status. Controlled Clinical Trials. 1989 Dec;10(4):407–15.

62. DeVellis RF. Classical Test Theory. Medical care. 2006 Nov 1;44(11):S50–9.

63. Crocker LAJ. Introduction to Classical and Modern Test Theory. Holt, Rinehart and Winston, 1986.

64. Hays RD, Morales LS, Reise SP. Item response theory and health outcomes

measurement in the 21st century. Medical care. NIH Public Access; 2000 Sep;38(9 Suppl):II28–42.

65. Hobart J, Cano S. Improving the evaluation of therapeutic intervention in MS: The role of new psychometric methods. Health Technol Assess. 2009;13:1-200.

66. Cano SJ, Hobart SC. The problem with health measurement. Patient Prefer Adherence. 2011;5:279-290.

67. Smith EV, Conrad KM, Chang K, Piazza J. An introduction to Rasch measurement for scale development and person assessment. J Nurs Meas. 2002;10(3):189–206.

68. Baylor C, Hula W, Donovan NJ et al. An introduction to item response theory and rasch models for speech-language pathologists. Am J Speech Lang Pathol. 2011;20(3):243-259.

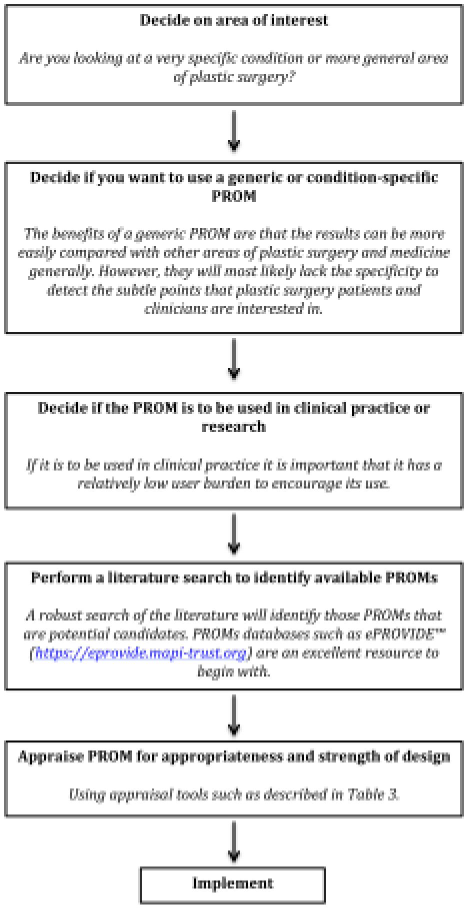
69. Alrubaiy L, Hutchings HA, Williams JG. Assessing patient reported outcome measures: A practical guide for gastroenterologists. United European Gastroenterol J. 1st ed. SAGE PublicationsSage UK: London, England; 2014 Dec;2(6):463–70.

70. Black N. Patient reported outcome measures could help transform healthcare. BMJ. 2013 Jan 28;346(jan28 1):f167–7.

71. Hutchings HA, Alrubaiy L. Patient-Reported Outcome Measures in Routine Clinical Care: The PROMise of a Better Future? Dig Dis Sci. Springer US; 2017 Aug;62(8):1841–3.

72. Miedany El Y, Gaafary El M, Youssef S, et al. Toward Electronic Health Recording: Evaluation of Electronic Patient-reported Outcome Measures System for Remote Monitoring of Early Rheumatoid Arthritis. J Rheumatol. The Journal of Rheumatology; 2016 Dec 1;43(12):2106–12.

Figure 1: A step-by-step guide demonstrating the steps to be carried out in deciding on and implementing a patient-reported outcome measure (PROM) into ones clinical or research practice.



Decide on area of interest

Are you looking at a very specific condition or more general area of plastic surgery?

Decide if you want to use a generic or condition-specific PROM

The benefits of a generic PROM are that the results can be more easily compared with other areas of plastic surgery and medicine generally. However, they will most likely lack the specificity to detect the subtle points that plastic surgery patients and clinicians are interested in.

Decide if the PROM is to be used in clinical practice or research

If it is to be used in clinical practice it is important that it has a relatively low user burden to encourage its use.

Perform a literature search to identify available PROMs

A robust search of the literature will identify those PROMs that are potential candidates. PROMs databases such as ePROVIDE™ (https://eprovide.mapi-trust.org) are an excellent resource to begin with.

Appraise PROM for appropriateness and strength of design

Using appraisal tools such as described in Table 3.

Implement

Table 1: A selection of condition-specific patient-reported outcome measures (PROMs) designed for use in the plastic and reconstructive surgery community. This is not an exhaustive list, but designed to indicate the broad spectrum of PROMs available in our specialty.

Sub-specialty	**Example of condition-specific PROM**
Burns	*CBOQ*: Children Burn Outcome Questionnaire
Breast	*BREAST-Q*™
Cleft	*CLEFT-Q*™
Cosmetic	*FACE-Q*™ *BODY-Q*™
Hand	*DASH*: Disabilities of the Arm, Shoulder and Hand score *Michigan Hand Outcome Questionnaire*
Head and Neck	*NOSE*: Nasal Obstruction and Septoplasty Effectiveness scale *EORTC QLQ-H&N43*: European Organisation for Research and Treatment of Cancer Head and Neck Module
Lower Limb	*TESS*: Toronto Extremity Salvage Score
Skin Cancer	*SCI*: Skin Cancer Index

Table 2: Glossary of terms commonly used in patient-reported outcome measure (PROM) development and psychometric validation.

Term	Definition
Classical Test Theory	The traditional method of assessing the scientific robustness of a PROM.
Content validity	Refers to whether the whole instrument is measuring all that is relevant and important to the patient and their condition.
Criterion validity	Assessment of how well the instrument being studied correlates with another instrument (ideally considered to be the gold-standard).
Face validity	A subjective measure of whether the questions are actually measuring what they are meant to be.
Instrument	A method of capturing data. In the case of patient-reported outcome measures an instrument usually refers to a questionnaire.
Items	An item is an individual question. Multiple items make up an instrument.
Interpretability	The degree to which one can assign clinical meaning to the quantitative score given by an instrument.
Modern Test Theory	Rasch measurement theory and item response theory and two methods encompassed by the term 'modern test theory'. These are newer methods of statistical analysis, designed to address some of the flaws of

	classical test theory.
Patient-reported outcome measures	Standardised and validated questionnaires that are designed to capture one or more aspect of a person's health and wellbeing.
Reliability	Refers to how consistent the results are when the instrument is applied in different situations.
Responsiveness	Refers to the ability of an instrument to measure a clinically important change.
Sensitivity	Refers to the ability of an instrument to measure any change.

Table 3: A simplified checklist for evaluating a patient-reported outcome measure (PROM).

Area of assessment	**Individual component**	**Was it performed?**
In those studies that applied Item Response Theory (IRT) models	- Was the IRT model used appropriately described? - Was the statistics package used adequately described e.g. RUMM2020, WINSTEP etc? - Was an adequate method of estimation used? - Were the assumptions of unidimensionality, local independence and item fit checked?	
Item generation	- Were the items sourced appropriately? - Was the target population included in item generation? - Was face validity assessed? - Frequency of endorsement calculated (0.2-0.8) - Item-total correlation calculated (0.3-0.7) - Internal consistency calculated (Cronbach's Alpha 0.7-0.95)	
Reliability	- Was a reliability co-efficient calculated and was it >0.7? - Was a measure of test-retest/inter-observer/intra-observer/parallel-forms reliability calculated?	

	* Person Separation Index > 0.7, to be able to differentiate between 2 groups of people	
Responsiveness	Was responsiveness assessed? If so was an appropriate method used (e.g. area under curve >0.7, Cohen's effect size >0.8, Guyatt's responsiveness ratio >0.5)	
Construct Validity	- Was construct or criterion validity assessed? - Were *a priori* hypotheses stated? *Was a test determining the unidimensionality of the scale performed? *Chi-square values summarizing the difference between observed and expected responses	
Interpretability	- Are the results clinically relevant? - Was a floor-to-ceiling effect calculated? - Was a minimally important difference (MID), standard error of measurement (SEM) and smallest detectable change (SDC) calculated?	
Burden	- Was there some assessment of the degree of burden placed on the patient completing the PROM?	

The first box is used to determine if the paper uses Item Reponses Theory (IRT) or Classical Test Theory (CTT). If CTT is used none of the first 4 questions can be answered.

* denotes areas in which IRT papers will quote differing test statistics to CTT.