Published online: 2019-10-04

Review Article

How to develop and validate a questionnaire for orthodontic research

Elbe Peter[1], R. M. Baiju[2], N. O. Varghese[3], Remadevi Sivaraman[4], David L. Streiner[5,6]

[1]Department of Orthodontics, Government Dental College, Kottayam, Kerala, India,
[2]Department of Periodontics, Government Dental College, Kottayam, Kerala, India,
[3]Department of Conservative Dentistry, PMS Dental College, University of Kerala, Thiruvananthapuram, Kerala, India,
[4]School of Health Policy and Planning Studies, Kerala University of Health Sciences, Thiruvananthapuram, Kerala, India,
[5]Department of Psychiatry and Behavioural Neurosciences, McMaster University, Hamilton, Canada,
[6]Department of Psychiatry, University of Toronto, Toronto, Canada

Correspondence: Dr. Elbe Peter
Email: elbeortho@yahoo.co.in

ABSTRACT

The use of psychometric tools to assess various psychological aspects of malocclusion and treatment is increasing in orthodontics. Mere evaluation of an orthodontic patient with normative criteria is not enough; instead, the psychological status should be assessed using a questionnaire. Many generic and few condition-specific tools are available for assessing quality of life (QoL) in orthodontics. The steps involved in the development of such tools are complex and unknown to many. This article outlines the methodology involved in the development and validation of a psychometric tool for dental and orthodontic use. It also helps the clinician to translate and cross-culturally adapt an existing QoL tool to a different setting.

Key words: Psychometric tool, quality of life, questionnaire, reliability and validity

INTRODUCTION

The broadened definition of health by the World Health Organization (WHO) in 1948 led to the development of a plethora of Patient Reported Outcome Measures (PROMs) in the health-care field to assess subjective aspects of health. There is a difference between the normative need for treatment estimated using clinical parameters or indices and self-perceived or realistic demand for treatment. Clinical measures of disease do not fully measure subjective experiences, personal values, attitudes, or beliefs. Thus, the term quality of life (QoL), initially introduced in the management sector, was adapted to health sciences. The WHO defined QoL as an individual's perception of their position in life in the context of the culture and value system, in which they live and in relation to their goals, expectations, standards, and concerns.[1] Health research concentrates on that aspect of QoL which relates specifically to an individual's health,

Access this article online

Quick Response Code:

Website: www.eurjdent.com

How to cite this article: Peter E, Baiju RM, Varghese NO, Sivaraman R, Streiner DL. How to develop and validate a questionnaire for orthodontic research. Eur J Dent 2017;11:411-6.

DOI: 10.4103/ejd.ejd_322_16

called as "Health-Related Quality of Life" (HRQoL). Locker and Allen[2] defined oral HRQoL (OHRQoL) as "the impact of oral disease and disorders on aspects of everyday life that a patient or person values, that are of sufficient magnitude, in terms of frequency, severity, or duration to affect their experience and perception of their life overall."

Two broad categories of PROMs used in clinical research are generic instruments that provide a summary of overall health and specific instruments that focus on specific conditions such as malocclusions. Generic tools such as the oral health impact profile, the child perception questionnaire, and the oral impact on daily performance were applied in orthodontics. The first condition-specific OHRQoL tool in orthodontics was introduced for orthognathic surgery patients in 2000 by Cunningham *et al*.[1] and another by Klages in 2006. Most of the OHRQoL measures are based on Locker's conceptual model,[2] adapted from International Classification of Impairment, Disability, and Handicap of the WHO. Wilson and Cleary[3] proposed another conceptual model in 199 incorporating environmental and nonmedical factors also. Recently, Masood *et al*.[4] using structural equation modeling reported a new model for patients with malocclusion.

An OHRQoL tool is essentially a psychometric test and its development is a complex process, involving various steps [Flow Chart 1]. The words tool, scale, and questionnaire are used interchangeably in the literature. An overview of tool/questionnaire development applicable in dentistry and orthodontics is outlined here.

CONCEPTUALIZATION

It is the process of taking a concept and refining it by giving a theoretical definition. Qualitative methods can be used for deriving concept. The abstract concepts are called constructs and the concrete representations as variables. Clinicians usually work with observable variables for measuring latent constructs using a tool.

DESCRIPTIVE STRUCTURE OF A TOOL

As with any research, tool development begins with a research question, and one needs to constantly refer back to it during the development process. Having a conceptual model with *a priori* hypotheses will enable the researcher to validate the hypothesis at the end of the study.

Items are the basic unit of a tool having a stem and a response format. They can be phrased either as questions or as statements. Questions can be open ended or closed. Open questions will enable the respondents to open up their minds and are useful in the initial stages of tool development; however, they are of little value as items in the final tool. Related items that define a part of the construct or domain are grouped together. To be qualified for a domain or subscale, each should have a minimum of two items under it. This step can be done initially by the researcher or experts and later by factor analysis. Apart from multiple items in a multi-item scale, there can be global rating questions, which according to some are important in assessing QoL.

Response options

The most common response format used in OHRQoL tool is Likert scale with a 5-point response option. Alternatively, a visual analog scale (VAS) format is also useful. Likert scale is bipolar with strong endorsement on one end and strong endorsement of the opposite on the other end, whereas VAS and adjectival scales are unipolar with magnitude of a feeling or belief from zero or little to very much or maximum.[5] The choice of response option depends on the nature of the question asked. For any response, the respondent needs to understand the question, recall the behavior, attitude, or belief to give a genuine answer. The responses are then converted into numerical form usually by adding the scores together to get a total score and a domain-wise subtotal for statistical analysis. Items that tap the opposite aspect of a trait need score reversal to calculate the total score.

STEPS IN TOOL DEVELOPMENT

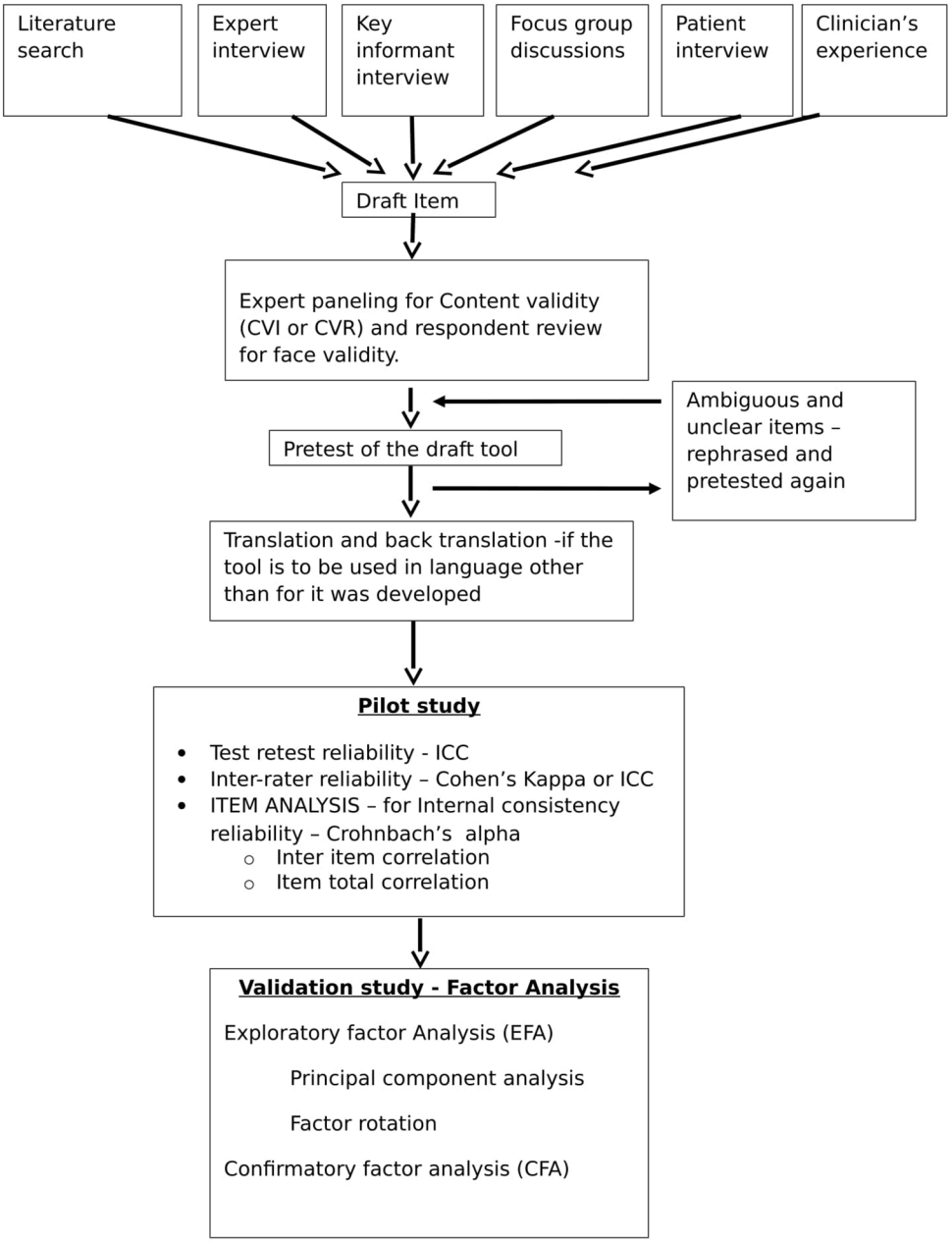
Item generation

Item source

Items are derived from various sources: existing tools, expert opinion, clinical observations, research evidence, theory, etc. The use of exiting items from a condition-specific tool is most valuable since these items have already undergone item analysis and their usefulness has been established. Patients experiencing conditions such as malocclusion and white spot lesions can be excellent source of items in a tool. Focus group discussions and key informant interviews can be utilized to acquire patient's viewpoints in a systematic manner.

Item selection

The pool of items generated in the initial phase should be in excess of what is required because item reduction



Flow Chart 1: Outline of questionnaire development

can be done at various levels to optimize the length of the tool without affecting its psychometric properties. Nothing can be done later to compensate for an item which was not included. Expert opinion is one method to ensure the content validity of the items. Draft items are administered to a panel of experts for ranking and selection is based on relevance. Raters can evaluate items on a four-point scale: 4 = highly relevant, 3 = quite relevant, 2 = somewhat relevant but needs rewording, and 1 = not relevant. A content validity index[6] or content validity ratio[7] can be calculated to decide whether to include the items. Lawshe[7] recommended a value of 0.99 for five or six raters, which is the minimum number of rates required, 0.85 for 8 raters and 0.62 for 10 raters.

Item wording, sequencing, formatting, and scoring

Items in a tool should be presented using clear and comprehensible language. It should need only basic reading skill of a 12 years old. Avoid double-barreled, ambiguous, negative, and leading questions. A mixture of items endorsed in the opposite direction may minimize the danger of acquiescent response bias, i.e., the tendency for respondents to agree with a statement or respond the same way to all items.[8] The length of each item

should be as short as possible without affecting the comprehensibility (preferably <20 words). Items should be sequenced so as to maintain a logical flow.

Translation and back translation

If the tool is developed in a different language than the intended language for clinical testing, it needs to be translated. The steps in translation are explained under cross-cultural adaptation of the tool.

Pretesting

The item pool is now pretested with the respondents to ensure that the items are comprehensible for the target group, unambiguous, and acceptable. Items that do not meet these criteria are eliminated or reframed and pretested again.

Item reduction

The draft items selected by experts and pretested in respondents can be tested for their psychometric properties. The terms "item selection" and "item reduction" are often used interchangeably. Testing of the draft items involves a minimum of two studies; one, a pilot study and another one, for factor analysis. The pilot study should be conducted in a sample of at least fifty individuals. Items should have good discriminative ability, i.e., they should discriminate between individuals having a condition and those who do not. Items where more than 90% of the sample gives the same response do not have such discriminatory power and should be deleted.

Munro[9] recommends five participants per item while others suggest ten participants per item for factor analysis, as long as there are a minimum of 100 people. The Bartlett's test of sphericity should be significant ($P < 0.05$) for factor analysis to be suitable. However, the psychometric soundness of the tool lies in the establishment of its reliability and validity;[10] hence, item reduction is *sine qua non* with establishment of reliability and validity.

Assessing the psychometric properties

Establishing reliability

Reliability refers to the repeatability, stability, and internal consistency of a tool. Before a tool can be used as a measure, it must be established that it measures "something" in a reproducible manner. Reliability is expressed as a number between 0 and 1, with 0 indicating no reliability and 1 indicating perfect reliability. Various forms of reliability are internal consistency reliability, test–retest reliability, and inter-rater reliability.

Internal consistency reliability

Internal consistency examines the inter-item correlations within an instrument and indicates how well the items fit together statistically. A low internal consistency means that the items measure different attributes or the participants' answers are inconsistent. In addition, the item-total correlation also explores a tool's internal consistency. If the items are measuring the same underlying concept, then each item should also correlate with the total score between 0.2 and 0.8. Cronbach's α, derived from Kuder–Richardson formula or split-half method, is the measure of internal consistency of a tool. Cronbach's alpha should exceed 0.70 for a new tool and 0.80 for a more established tool.[11] Alpha should be determined for separate domains of the tool which is more important than for the whole tool.

Test–retest reliability

Test–retest reliability ensures the stability of the tool over time. If the patient's condition remains same between two measurements, then the scores should be similar. To assess the stability of an instrument over time, the two measurements should be temporally close to each other. Too close an interval will result in recalling the previous answers rather than giving an independent response. The usual test–retest interval is between 10 and 14 days.

Inter-rater reliability

This form of reliability is applicable only in interviewer-administered tools. Different raters assessing the same individual should obtain similar scores. It is measured with a coefficient between 0 and 1.

The intraclass correlation coefficient is a better measure of test–retest reliability and inter-rater reliability than Pearson's *r* because it is sensitive to any bias between or among the raters or times.[12] Values more than 0.75 are considered enough to infer good agreement.[13]

Ensuring validity

Validity refers to whether one can draw accurate conclusions about the presence and degree of the attribute for an individual. In other words, validating a tool is a process by which we determine the degree of confidence, we can place on the inferences, and we make about people based on their scores from that tool.[5] Content validity and face validity are ensured in the initial stages of tool development. The classic expression of a tool's validity pertains to the three Cs of validity; content, criterion, and construct. The recent

trend is to consider all forms of validity as subsets of construct validity. As mentioned before, the validation of a tool is a dynamic process; a single study is not enough to ascertain it.[14] However, a larger sample than what is used for pilot study is essential for validation purpose.

Content validity

Content validity ensures that all aspects of the construct are represented by an adequate number of items and should not include items that are unrelated to the construct. The researcher has to ascertain the content validity of a tool by closely following the item generation steps mentioned above as there is no strict statistical procedure to ensure it. Eliminating items that have a low correlation with other item will increase the internal consistency at the expense of content validity.

Criterion validity

Criterion validity refers the correlation between new scale and some other measures of the construct. The scores on the scale are compared against clinical judgment or a gold standard. In QoL measure, there is no universally accepted gold standard tool. If both scales are administered simultaneously and results are known immediately, it is known as concurrent validation, and predictive validation when tool is administered separately and the outcome is available only at some later time.

Construct validity

Construct validation is heavily relied upon in situations when there is no criterion with which a tool can be compared. One has to generate predictions based on the hypothetical construct (hypothesis generating) which are then tested to determine construct validity. It is an ongoing process of learning more about the construct, making new predictions, and testing them. Convergent validity and divergent validity are terms used to distinguish between two aspects of construct validity. Convergent validity seeks whether the measurement is related to constructs to which it should be related if the instrument was valid, and divergent validity seeks whether the measurement is unrelated to constructs to which it should be unrelated. Discriminant validity is the ability of the tool to discriminate between different groups of participants and is quite often used interchangeably with divergent validity. Global questions such as "how do you rate your oral health today" will also help in convergent validation of the tool.

CROSS-CULTURAL ADAPTATION OF AN EXISTING TOOL

QoL being a multidimensional tool developed in one language for a population may not be appropriate for others. However, an existing tool can be translated to a different language and cross-culturally adapted with the help of linguistic and subject experts. The steps include translation of the tool (forward translation) by a team of bilingual experts, back translation, and derivation of the best version by consensus. Translation should not be done mechanically on a word-to-word basis, but cultural context should be addressed. The aim is to achieve equivalence between the two versions. Such a translated and cross-culturally adapted tool needs new studies to ensure its reliability and validity.

FUTURE OF THE TOOL

Condition-specific tools having commercial value can be copyrighted, and the use of such tool without legal transfer will incur consequences. Each application of the tool is a validation study for it reinforcing its reliability and validity. As each supportive study strengthens the construct of a tool, a well-designed experiment with negative finding can call into question the entire construct of a tool.[5]

CONCLUSION

OHRQoL is a multidimensional construct. Many studies have shown that malocclusion has an impact on individual's QoL; hence the relevance of measuring it is justified. Understanding correct tool development methodology is vital for its proper use, especially for translation and cross-cultural adaptation. An outline of QoL tool development methodology applicable in dentistry and orthodontics is presented. It is difficult to segregate the psychometric properties of a tool into watertight compartments as there is difference in the chronological order, in which they are ensured and clinically studied. No measurement technique can be valid if it is not repeatable, but it can be repeatable without being valid. Hence, reliability and validity are often studied together though explained separately. The steps involved are complicated and time consuming; use of existing tool if it fulfills your research question after cross-culturally adapting and validating it is strongly recommended. Assessment of patient-perceived change following treatment of malocclusion may become an ethical and legal prerequisite in the future.

Financial support and sponsorship
Nil.

Conflicts of interest
There are no conflicts of interest.

REFERENCES

1. Cunningham SJ, Garratt AM, Hunt NP. Development of a condition-specific quality of life measure for patients with dentofacial deformity: I. Reliability of the instrument. Community Dent Oral Epidemiol 2000;28:195-201.
2. Locker D, Allen F. What do measures of 'oral health-related quality of life' measure? Community Dent Oral Epidemiol 2007;35:401-11.
3. Wilson IB, Cleary PD. Linking clinical variables with health-related quality of life. A conceptual model of patient outcomes. JAMA 1995;273:59-65.
4. Masood M, Masood Y, Newton T, Lahti S. Development of a conceptual model of oral health for malocclusion patients. Angle Orthod 2015;85:1057-63.
5. Streiner DL, Norman GR, Cairney J. Health Measurement Scales: A Practical Guide to Their Development and Use. 5th ed. Oxford: Oxford University Press; 2015.
6. Lynn MR. Determination and quantification of content validity. Nurs Res 1986;35:382-5.
7. Lawshe CH. A quantitative approach to content validity. Pers Psychol 1975;28:563-75.
8. Rattray J, Jones MC. Essential elements of questionnaire design and development. J Clin Nurs 2007;16:234-43.
9. Munro BH. Statistical Methods for Health Care Research. Philadelphia: Lippincott Williams and Wilkins; 2005.
10. Bowling A. Research Methods in Health. Buckingham: Open University Press; 1997.
11. Everitt BS. Multivariate analysis: The need for data, and other problems. Br J Psychiatry 1975;126:237-40.
12. Keszei AP, Novak M, Streiner DL. Introduction to health measurement scales. J Psychosom Res 2010;68:319-23.
13. Indrayan A. Medical Biostatistics. 3rd ed. Boca Raton: Chapman & Hall/CRC Press; 2012.
14. Jensen MP. Questionnaire validation: A brief guide for readers of the research literature. Clin J Pain 2003;19:345-52.