Patient Related Outcome Measures

Dovepress
open access to scientific and medical research

Open Access Full Text Article

ORIGINAL RESEARCH

Exploring how to evaluate a qualitative patient-centered outcome measure: literature review and illustrative example – a Perthes child-friendly measure

This article was published in the following Dove Press journal:
Patient Related Outcome Measures

Andrew F Long[1]
Tina Gambling[2]

[1]School of Healthcare, Faculty of Health and Social Care, University of Leeds, Leeds, UK; [2]School of Health Sciences, College of Biomedical and Life Sciences, Cardiff University, Cardiff, Wales, UK

Purpose: To explore the question of 'how to evaluate a qualitative patient-centred outcome measure', comprising predominantly open-ended items, including perhaps emojis, story writing and/or pictures, in a way that does not compromise the strictures of the qualitative paradigm, doing so in a credible and authoritative manner. The paper aims to promote debate and discussion in the measurement validation community.

Methods: Comprehensive literature review of three electronic databases (PubMed; SCOPUS; Web of Science/Knowledge) and searches of three outcome-focused journals.

Results: The vast majority (>90%) of the papers only used qualitative methods in the initial, in particular, content validation of a measure and then used (quantitative) psychometric validation procedures. The remaining papers comprised articles that were either methodologically or methods focused and the role of qualitative research. A number of key issues are raised, inter alia: giving primacy to the patient's perspective; exploring the meaning and interpretation respondents place on the concept and possible items in a measure; prioritising maximising meaningful discrimination from the respondent's perspective; ensuring face and content validity and relevance of items in the item content pool; and using appropriate qualitative methods, for example, concept elicitation, "think-aloud" and cognitive interviews and expert respondent panels/judges. This approach is applied to validate a child-friendly outcome measure for children with Perthes disease, a paediatric hip condition presenting primarily amongst male children aged 5-8 years.

Conclusions: The core messages are to: (i) not force validation of a qualitative outcome measure into psychometric validation; but (ii) retain full adherence to the principles of the qualitative paradigm and employ procedures drawn from that paradigm. In this manner, primary emphasis would lie on issues of meaningfulness, face and content validity, the meaning of item and measure scores to respondents and, for a child-friendly measure, the child-friendliness of the measure.

Keywords: qualitative validation, outcome measure, psychometric validation, qualitative paradigm, Perthes disease, child-friendly measure

A Muse:
I wonder how we might evaluate a measure which comprises a majority of open-ended questions, including perhaps use of emojis, story writing and pictures. How best, and in a credible and authoritative way, can this be done and thus demonstrate its validity, reliability and responsiveness to change?

Correspondence: Andrew F Long
School of Healthcare, Health Systems Research, University of Leeds, Baines Wing, Leeds LS2 9UT, UK
Email a.f.long@btinternet.com

submit your manuscript | www.dovepress.com
DovePress
http://doi.org/10.2147/PROM.S215425

Introduction

There are established, tried and tested approaches to the design and testing of a quantitative outcome measure.[1–4] A common step-by-step approach embraces the following:

- Decide on the aim, purpose, general scope and breadth of the proposed measure, for example: measuring what concept or phenomenon?; used for what purpose (to discriminate between people at one point in time or to evaluate change over time for individuals or a group)?;[5,6] what aspect of the concept or concept itself is the aim of measurement?; and, what should be the extent of patient-centredness[7] and grounding in patient perspectives and phraseology?
- Develop possible measure content/the item content pool, via, inter alia: reviewing existing measures or those in a related field; open-ended interviews and focus group discussions with the target patient/condition group; and using patient experts and/or expert judges.
- Draw up a pilot measure and evaluate its content and face validity, ease of completion, question phrasing (ease of understanding; unambiguous phrasing, no double questions, sufficiency of response levels for level of discrimination required, relevance of "don't know" or "not applicable" response option); time taken to complete the measure; potential and patient-perceived burden of measurement. Make use of, for example: focus group discussions with the target patient group and as appropriate clinicians; cognitive interviewing using the "think-aloud" approach.
- Explore the practicality and feasibility of use in clinical practice and clinical/patient utility,[8] via interviews with the target groups, for example, clinicians and patients.
- Refine the measure and re-evaluate as above, continuing as necessary until a prototype measure has been developed ready for psychometric testing.
- Undertake psychometric testing, exploring: the measure's internal reliability (internal consistency); test–retest reliability; inter-rater reliability, if relevant; criterion validity; and construct validity. For all, use established psychometric approaches, including item reduction, factor analysis and correlation analysis.
- Refine the prototype to maximize its measurement properties.
- Repeat as necessary the psychometric approaches leading to a final measure ready for use in the target area(s).
- Assess responsiveness to change, if the measure is intended to be used as an evaluative measure.

While these approaches make sense for a measure that predominantly comprises fixed-choice questions, and thus potentially quantifiable responses (for example, using a 5-point Likert scale), it is not self-evident how relevant they are for a measure that comprises predominantly open-ended, and thus non-quantifiable, questions. To the best of our knowledge, however, and confirmed from discussion with colleagues with expertise in outcome measurement, there is a dearth of discussion of or literature on this topic, save for research exploring the meaning and interpretations that potential respondents place on items in an outcome measure[9] and the role of qualitative research in ensuring attention lies on the patient perspective.[10] In part, such a lack of literature on validation for a qualitative measure could be accounted for on the argument that a thematic coding scheme could be developed that allocates a particular type of response into a code/number. The resultant set of codes would then take on the form of quantitative measurement, if only at a nominal level, and could then be subject to the psychometric validation process. However, this option fails to directly address the core issue in a way that preserves the principles of the qualitative paradigm.[10–12]

It is to address the Muse that this paper is directed, with the aim of stimulating debate and adding to methodological understanding in the field of measurement validation. Following a comprehensive literature search, possible ways to evaluate a qualitative measure which comprises predominantly open-ended questions and, moreover, in a manner that honors the principles of qualitative research, are explored. To aid insight into the potential issues involved, the discussion and approach are situated against one newly-designed qualitative outcome measure, developed for young children (here, aged 5–8 years old) with the pediatric hip condition of Perthes.[13]

Literature searches

An initial literature search on Google Scholar was undertaken in January 2019 to locate methodologically oriented literature and/or discussion of ways to evaluate an instrument that comprises predominantly open-ended questions. The keywords of "qualitative validation," "qualitative measure," "qualitative outcome measure" were used. This uncovered only a small number of potentially relevant articles.

A comprehensive search was then undertaken on three electronic databases in June 2019, with no data restrictions or other search limitations: PubMed (for bio-medical and health care-related literature); SCOPUS; and Web of Knowledge/Science (for social science-oriented literature). MESH search terms were derived from PubMed for the PubMed searches. For SCOPUS and Web of Science/Knowledge, keywords were used in combination (using "AND"). The search and keywords terms and search yields are summarized in Table 1. The abstracts of the papers were first assessed, and full papers obtained for papers of potential relevance. The reference lists of the full papers were then explored for additional references.

Finally, to supplement these searches, an electronic search was conducted, using the keywords "qualitative validation," "qualitative measure," "qualitative outcome measure," within three outcome measure focused journals, Patient Related Outcome Measures, Quality of Life Research and Health and Quality of Life Outcomes. In order to get as close as possible to the issue raised in the Muse, a self-titled "qualitative validation" of a fixed-choice, Likert-style outcome measure was also identified[14] to examine how it set out to evaluate the measure, while paying heed to the principles of the qualitative paradigm.

Findings

The paper yield generated from the set of searches is summarized in Table 2. The searches of the three databases overall generated a similar set of papers, and thus numerous duplicate papers. Two major groupings were evident (see Table 2).

The first grouping, representing the overwhelming majority of articles (>90%) were those that used qualitative methods in the initial, in particular, content validation of a measure, and then, for all subsequent validation, used (quantitative) psychometric validation procedures. This approach was entirely appropriate as the measures themselves were most commonly of a Likert-type or fixed-scale response variety. This was also the case for the searches undertaken of the three journals. For example, a search of Health and Quality of Life Outcomes identified 12 articles. Typical examples were an article exploring the FACIT fatigue scale[35] or an article exploring the Patient Uncertainty Questionnaire for rheumatology, the PUG-R.[36]

The second grouping comprised articles that were either methodologically or methods focused and centered on the use of qualitative research. This second grouping was subsequently divided into six thematic areas (Table 2):

(i) Methodologically/method-oriented and/or broader theoretical/philosophical discussions of validity;
(ii) Guides to best practice for measure development;
(iii) Use and importance of qualitative research in the development of a patient-reported outcome measure (PROM);
(iv) Use of qualitative approaches in constructing a measure and generating an item pool;
(v) Use of qualitative approaches in establishing construct validity and, in particular, content validity of a PROM;
(vi) Use of qualitative approach to explore the validity or reliability of a qualitative measure.

The most notable finding from the literature review is the lack of focus on the topic, or issues surrounding, validation of a qualitative outcome measure or how this might be accomplished. If at all, attention centered on the use of concept elicitation interviewing, with experts or potential respondents to the measure (to elaborate the nature of the concept and

Table 1 Overview of databases, search terms and yield

Database	Search terms	Yield (number of papers)
PUBMED (MESH Terms)	Psychometrics; outcome assessment; health care	162
	Outcome assessment; health care; psychometrics; qualitative research	146
	Patient reported outcome measures; outcome assessment; health care; methods. Subheading: methods.	155
SCOPUS (Word Search Terms)	Qualitative, outcome measure, scale, development, validation	125
	Qualitative outcome measure, scale development. Subheading: psychometrics	88
Web of Science (Word Search Terms)	Qualitative research, outcome measure, psychometric validation	236

Table 2 Illustrative examples of group two articles by theme

Thematic area	Illustrative articles and content
Methodologically/Method Oriented	*A: Discussion of "What is Validity and Reliability" in Qualitative Research*: • Winter[15]: for example, points to. a "realist" approach – an account is valid if it reflects the perspectives of the actors in that situation (p. 7) • Creswell & Miller[16]: stress importance of exploring validity from the lens of participants, for example, via member checking and peer (and cognitive) debriefing • Golafshani[17]: inter alia, argues (p601) that validity, and reliability, in qualitative paradigms are assessable in terms of Credibility, Neutrality or Confirmability, Consistency or Dependability and Applicability or Transferability[18] and of the importance of triangulation from multiple perspectives • Frost et al[19]: focus on what are the psychometric properties to generate "sufficient evidence" for the validity and reliability of a PROM; argues for importance of establishing content validity as primary task, and use of qualitative research to do this, in particular, focus group and cognitive interviews; and then use of psychometric validation approaches
Theoretical Discussions of Validity	*B. Theoretical/Philosophical Discussions* • Zumbo[20]: what is validity, and particularly construct validity, and its implications for process of validation? Presents a "contextualised and pragmatic explanation" of validity; construct validity "should provide an explanation for the test scores for the observed variation in test scores" (p. 69); validity as "establish (ing) the "why" and "how…" (p. 70) and as "support (ing) inferences … (made) from test scores…" (p. 70); validation is a "higher order integrative process…involving…concept formation…." (p. 69); argues for "multilevel testing and measurement" for a multilevel construct (p. 78). • Gadermann et al[21]: importance of asking, "what are the underlying cognitive processes that result in respondents providing responses to self-report questions" (p. 39) in the way they do; use cognitive interviewing to do this. • Hubley & Zumbo[22]: explore meaning of 'response processes'; argue that response processes should be considered as "mechanisms that underlie what people, do, think or feel...when...responding to an item" (p. 2); research in this area should "become more explanation-based" (p. 8) and explore "the broader context (i.e. purpose of testing, setting, culture)" (p. 8) when the response to an item is completed.
Guides to Best Practice for Measure Development	• Wild et al[23]: exploring best practice for cultural adaptation and translation of a PROM; includes use of persons in new cultural context with experience in qualitative interviewing and/or cognitive interviewing to explore translation and cultural adaptation • Brod et al[24] present best practice guide in use of qualitative research and exploration of content validity • Luyt[25]: drawing on Adcock and Collier[26] presents a framework for measure development comprising three interconnected stages: (i) measure development (background concept; developing concept definition; devising indicators); (ii) measure validation and (iii) measure revision. Advocating use of qualitative (in stage 1) and quantitative (for stage 2).

(*Continued*)

Table 2 (Continued).

Thematic area	Illustrative articles and content
Use and Importance of Qualitative Research in Development of a PROM	• Lasch et al[27]: qualitative research as providing sound and rigorous basis for PROM development; role of theoretical saturation (in coding categories) and triangulation, to explore from multiple perspectives • Cheung and Clark[28]: major role of qualitative research in PROM development and also cultural adaptation
Use of Qualitative Approaches in Constructing a Measure and Generating the Item Pool	• Mallinson[29]: importance of exploring the meaning and interpretation that respondents place on items in a PROM, focusing here on the SF-36, a fixed-choice measure; item interpretation and meanings attached may interact with a range of social and cultural factors affecting the respondent; use of face-to-face (cognitive, debriefing) interviews while respondent completes the measure • Viswanathan et al[30]: explore measurement implications of scale responses, depending on whether the primary concern is maximizing discrimination between scale responses, whilst retaining reliability) and meaningful discrimination from the perspective of the respondent; argues for greater emphasis to be placed on meaningful discrimination from the perspective of the potential respondent, and not measure developer/researcher. • Luyt[25]: explores measure development phase, in particular, the "constellation of meanings and understandings associated with a given concept" (p. 4), using focus groups • Cheung and Clark[28]: significance in ensuring explicit focus on patient perspectives; critical role lies in both item generation and establishing content validity (eg, concept elicitation, cognitive debriefing) • Breyer et al[31]: develop a patient-grounded measure on the symptoms, functions and impacts of urethral stricture disease; use of concept elicitation and cognitive interviews, followed by patients prioritizing items in terms of their bothersome-ness
Use of Qualitative Approaches in Establishing Construct Validity and, In particular, Content Validity of a Patient Reported Outcome Measure (PROM)	• Hardesty and Bearden[32]: explore the use of expert judges in assessing the face and content validity of items • Cremenns et al[33]: present a literature review of health self-report measures for children aged 3–8 years; range of measures found, using formats of Likert scales, graphical (pictorial), facial (cartoon) or visual analog; in 40% of measures children involved in item development (researcher talking with child) and in 47% pilot testing with children, where authors reported on content validity, in 40% of children themselves informed this; argues that measure developers should draw on the child's perspective from the child, and not just rely on researcher/expert panel

(*Continued*)

Table 2 (Continued).

Thematic area	Illustrative articles and content
Use of Qualitative Approach to Explore the Validity or Reliability of a Qualitative Measure	• Golafshani[17]: advocates the use of quality criteria drawn from Lincoln & Guba[18] – credibility, confirmability, consistency transferability; and exploring from multiple perspectives • Cremenns et al[34]: explore the development of a generic quality of life measure for school-age (6–9 year old) children; advocate and use "think aloud"/cognitive interviews; develop coding categories to select 30 items for the measure; and strategy used by children to answer items for the measure.; use of two independent raters, leading to exploration of intra- and inter-rater reliability; use of both qualitative measures (measure development) and quantitative (for reliability testing and comparison of strategy use) • Gadermann et al[21]: develop coding and sub-coding categories, guided by research purposes (in the paper, strategies employed by children to respond to measure's items); present in tree diagram format; illustrate category content; add frequency counts; compare tree diagrams with counts attached (here by strategy categories in item response – absolute, relative, general positive, unclear) • Luyt[25]: suggests the use of multiple (two or more) coders for qualitative data analysis, then exploring of inter- and intra-rater reliability quantitatively

develop a conceptual model for the measure) and/or cognitive and/or debriefing interviewing (to explore the meaning and interpretation placed on items in a measure), important and central issues for both a quantitative and qualitative outcome measure.

Key issues arising from thematic areas depicted in Table 2, primarily those centered on the use of qualitative research in constructing a measure, generating the item pool and establishing construct validity and, in particular, content validity of a PROM, are summarized below. Focus lies on implications and/or suggestions for how best to evaluate a qualitative outcome measure that comprises predominantly open-ended questions.

The importance and value of qualitative research in constructing a PROM has been widely advocated[27] and most especially in assessing and ensuring content validity.[24] Most recently, Cheung and Clark[28] in an editorial highlight the major role that qualitative research should play in the development, and any subsequent cultural adaptation, of a PROM. In particular, they point to its bringing patient perspectives to the fore and its value in generating an item pool and establishing content validity. Luyt[25] also suggests the use of focus groups, in order to gain insight into the meanings that potential respondents to a measure associate with the underlying concept.

Winter[15] and Creswell and Miller[16] both point to the importance and value of exploring validity from the perspectives of those completing a measure. Mallinson[29] addresses this issue directly for one (then) highly popular and widely advocated PROM, the SF-36,[37,38] focusing on the meaning and interpretation of fixed-choice questions. She draws attention to the fact that:

> Standardisation of the survey text does not automatically lead to standardisation of meaning. (p. 12)

Moreover,

> …The meanings of words does not inhere in the words themselves but is a product of the situation and the relationship between those interacting and can be affected by a range of social and cultural factors… (p. 12)

To explore the core issue of the meanings and interpretations potential respondents place on the questions, and their phrasing, she suggests use of:

- "Think-aloud" protocols;
- Face-to-face interviews; and,
- Use of "experts," in particular, expert patients or patient panels.

Whichever of these methods are chosen, primary interest centers on exploring the face and content validity of the questions from the perspective of the potential respondents, and, critically, to gain deeper insight into where problems over intended meaning and interpretation arise. Findings can then be used either to temper the interpretations placed on the results of the measure, here ratings on the SF-36, and/or to assist the scale developer and/or researchers' measure to further refine the measure to enhance its validity for the target group.

Again, drawing patient perspectives to the fore, Viswanathan et al[30] explore the measurement implications of scale responses, depending on whether the primary concern is maximizing discrimination between scale responses (for example, where the difference between a 4 and a 5 on a five-point Likert scale is important, particularly for the researcher/measurer, whilst retaining reliability) and meaningful discrimination from the perspective of the respondent/consumer. Commonly, they comment, scale developers focus on maximizing discrimination, as long as reliability is not compromised. Indeed, some researchers would argue that a scale with too few categories (for example, 3 or 5) does not enable sufficient discrimination and, furthermore, that a larger number of scale levels often leads to a more reliable scale.

In contrast, Viswanathan et al[30] argue in favor of maximizing meaningful discrimination, that is, ensuring a scale has an appropriate number of response categories to facilitate this. For example, use of a seven-point Likert scale asks the consumer to indicate a rating of an item on a scale from 1 to 7. One consumer may rate the item as a 5, another as a 6 and another as a 7. However, the consumers in their judgment may be re-interpreting/translating the scale values into more meaningful values, such as "low" (for example, 1, 2 and perhaps 3), "medium" (perhaps 4 and 5) and "high" (6 and 7; and maybe 5). For these three consumers, the 6 and the 7 would thus mean and be meaningful as "high," and the 5 as "medium" or perhaps even "high." To address this issue, Viswanathan et al[30] recommend that a scale item should comprise the number of categories that the consumer finds meaningful. This will result in a more valid scale, and one that is able to "(generate) valuable diagnostic information about consumer attitudes and behaviours" (p. 123) and "validly measure differences in products" (p. 123–124). The challenge for the measure developer is then to clarify how many rating levels are meaningful to the target group, for example, through the use of expert consumer panels or "think-aloud" interviews as consumers complete a selection of items from the content pool.

A number of papers explore the use of concept elicitation, "think-aloud"/cognitive and debriefing interviewing, again to ensure the grounding of a PROM in the patient perspectives. Indeed, Gadermann et al[21] build on Zumbo's[20] extended concept of validity and construct validity, using the process of cognitive interviewing in their empirical study. This provides a way to explore the understanding, meanings and interpretation that potential respondents to a measure ascribe to items in a measure and may help in understanding what an overall measure and associated item scores mean to them, for example, in their cultural context.[23] A useful example is provided by Breyer et al's study.[31] They demonstrate how they developed a "patient-grounded" measure on the symptoms, functioning and impacts of urethral stricture disease in their everyday lives, using concept elicitation and cognitive interviews, followed by patient prioritization of items in terms of their impact on their quality of life. Cremenns et al[34] similarly used think aloud/cognitive interviews in their development of a generic quality of life measure for children aged 6–9 years.

In contrast, Hardesty and Bearden[32] focus on issues surrounding the use of expert judges in the development of a scale or measurement tool. In the first part of their paper, in a similar manner to Mallinson,[29] their emphasis lies on the concepts of face and content validity, which they argue are often confused or used seemingly interchangeably. To illustrate the differences in the two concepts, they draw an analogy to a dartboard. To establish content validity, darts must land all over the dartboard, and not to just one side or adjacent segments. In contrast, to establish face validity, the darts have just to hit the dartboard; items in the item/content pool must therefore all "hit the dartboard," and so reflect the desired construct. Moreover, all the items in the final content pool and resultant measure must have face validity. But, as they appropriately comment, face validity is just one part of construct validity, to ensure that the measure reflects what it is intended to measure. Other aspects, they point out, embrace content validity (items then representing a "proper" sample of the domain(s) of the concept being measured) and aspects such as discriminant, convergent and predictive validity. The second part of the paper reviews a number of "expert judging" decisions rules, to make sense of the findings from a panel of expert judges. They conclude by advocating the "sum-score" rule (that is, calculating the total score for an item

across all the judges, and then selecting the highest valued items, above a pre-defined score threshold). They end on a note of caution, commenting:

> …Simply judging items may (sic, does) not guarantee the selection of the most appropriate items for a scale. (p. 106)

Other approaches of potential significance for the validation of a qualitative measure arise from three other papers.[17,21,25] The former points to the relevance of classic qualitative quality criteria,[18] in particular, the measure's and contents' credibility and confirmability (for example, from others' perspectives or with other data). Gadermann et al[21] point to the importance of developing coding and sub-coding categories, built on patient perspectives, guided by the research's and/or measure's purposes (in this instances, strategies used by patients to respond to the measure's items). In a similar vein, and taking the discussion a quality assessment[17] a step further, Luyt[25] suggests use of at least two coders and then to explore intra- and inter-rater reliability.

Exploration of this literature suggests a number of issues and potential ways to address the guiding question to which this paper is directed: "how best and in an authoritative and credible manner can a qualitative measure/outcome measure be validated paying heed to the principles of the qualitative paradigm?" Nine key points are extracted.

1. The important role of qualitative research in bringing patient perspective to the forefront in the development of a PROM.
2. The need for and clarity of the underlying conceptual model of the proposed measure, basing this on patient perspectives through, for example, concept elicitation interviews and/or focus groups.
3. A need to ensure face and content validity in the measure's item content pool by exploring this with potential respondents.
4. The importance of exploring and elucidating the meaning and interpretation potential respondents place on the scale's items/questions and their phraseology. This should be extended to include any guide and/or instructions provided to respondents in relation to how to fill in the measure, item completion and meanings of the rating procedure (that is, the meaning the scale designer gives to a 1 to 5 for a five-point Likert scale).
5. The value of maximizing meaningful discrimination from the perspective of the potential respondent, and thus using the appropriate number of (rating) levels that they can manage and use, rather than prioritizing maximum discrimination from the perspective of the scale developer.
6. Subsequent further exploration of meaningful discrimination for the trimmed items in the measure's content pool.
7. Retention of items for theoretically informed reasons or because of their respondent-related importance/significance, notwithstanding their psychometric features.
8. Use of appropriate qualitative methods to clarify and explore these issues including, for example: "think-aloud" protocols; cognitive interviews; expert respondent and/or other expert panels/judges.
9. Potential of drawing on the quality assessment criteria commonly employed in qualitative research, in particular, the measure's and contents' credibility and confirmability, along with the use of multiple coders of the qualitative data, to explore the validity and reliability of a measure.

To cast further light on the guiding Muse conundrum, an article with the term "qualitative validation" of a measure in its title was selected from a key outcome-focused journal, Quality of Life Research. This article[14] focused on one established, widely used and psychometrically validated, self-administered scale, the Minnesota Living with Heart Failure (MLHF) questionnaire.[39–41] Indeed, the paper's authors partly justify their choice of this measure because "it is the most widely used QoL instrument in clinical trials in heart failure" (p. 418). For their validation study, they conducted two to three semi-structured interviews (76 in total), guided by a checklist, with a small sample (n=31) of patients recruited from two settings (a hospital with a nurse-led clinic, and one without) and selected from a large 2-year prospective observational study. Their validation approach used "simple qualitative pre-testing techniques from the field of questionnaire design" (p. 420) aimed at exploring the feasibility of the instrument, particularly its possible respondent burden (physical and mental), practical and interpretative problems respondents experienced and perceived face validity. For example, they observed respondents while they were completing the MLHF measure (using "think-aloud"), talked with them about the process of completing

the measure (respondent debriefing, using retrospective probes about what they were doing or thinking for a particular item) and sought comments on problems experienced with the questionnaire items, their interpretability and item relevance (face validity). A number of problems areas were evident.

Firstly, Hak et al[14] found that respondents did not read or not read fully the instructions on how to complete the questionnaire and, thus in consequence, were answering the questions in other ways than those intended by the scale developer and researchers. Notably, the instructions for the MLHF explicitly draw attention to its core focus as: "did your heart failure prevent you from living as you wanted during the last month…" Questions should thus be answered for the time frame of "last month." focus on whether the respondent was "prevented from living as they wanted," and refer only to symptoms or handicaps "caused by their heart failure" and not for other reasons/causes. Respondents' spontaneous comments showed that these instructions were however not being followed and/or not fully understood. Most commonly, a different time frame was used (for example, the previous week), responses provided in relation to things they found difficult to do (but not necessarily were "prevented" from doing) and/or relating to symptoms or handicaps other than due to their heart failure (for example, old age) or symptoms that varied a lot (items such as swollen ankles or shortage of breath, with some answering from an "at present" time perspective). A further implication was that this might compromise test–retest validity). Overall, Hak et al[14] comment: "the 'true' validity of the MLHF is low, in the sense that items are not read (or completed) as intended" (p. 421).

Other sets of problems their study identified related to respondents' understanding of items, lack of a "not applicable" option and responding to items separated by an "or." For example, respondents were unsure how to interpret the meaning of "loss of grasp"; they then made sense of it themselves, as it were, as the authors put it, "inventing" a meaning "on the spot." Similarly, respondents did not know how to respond if they perceived an item as "not applicable," as items in the MLFH did not provide this as a possible response. Finally, respondents were unsure how to respond to a question which separated two issues by an "or," especially if it was not considered applicable to their current situation.

Findings from Hak et al's[14] study reinforce the arguments drawn from the literature review concerning the development of a credible and authoritative approach to validating a qualitative measure, that is, one comprising predominantly open-ended items, where the translation of open-ended responses is deemed inappropriate or as violating the qualitative paradigm. In summary, they point toward the need in a validation of a qualitative measure to prioritize the following:

1. Importance of clarity about, and basing the measure upon, an underlying conceptual model of the measure (and thus the concept it is aiming to measure).
2. Primary focus on face and content validity, and maximizing meaningful discrimination from the perspective of potential respondents.
3. The importance of exploring and elucidating the meaning and interpretation potential respondents place on the scale's items/questions.
4. Exploring areas of difficulty and problems experienced, if any when completing the measure.
5. Exploring item relevance and interpretability from the perspectives of potential respondents.
6. Retention of items for theoretically informed reasons or because of their respondent-related importance/significance, irrespective of their psychometric features.
7. Use of appropriate qualitative methods to clarify and explore these issues: for example, cognitive interviews; observing respondents while completing the measure, combining this with "think-alouds" or cognitive interviewing and/or respondent debriefing using retrospective probes; expert respondents and/or other expert panels/judges.

Attention now turns to apply the points raised in the literature review to the development of a protocol for a qualitative outcome measure designed by the authors, in collaboration with colleagues at the University of Liverpool, to explore the impact of Perthes disease on the affected child and their family.

Developing a protocol to validate a child-friendly outcome measure for Perthes disease

Need for a measure and the development process

Perthes' disease is a condition that affects predominantly young male children presenting between 4 and

7 years of age.[42] Commonly reported outcomes are radiographic, focusing on the shape and congruency of the femoral head.[43] Patient-centered outcomes, in particular, the potential major psycho-social, emotional and quality of life (QoL) impact of Perthes on the lives of affected children and their families have not been explored in the literature.

Following an approach by a leading Perthes surgeon from Alder Hey Children's Hospital, Liverpool, UK, we, together with colleagues in Liverpool, developed a child-friendly measure for the child to complete either on their own or with the help of their parents.[12] The measure was grounded in two tape-recorded open-ended interviews with members of two families (a mother, a father, respectively). We designed a topic guide for the interview, beginning by asking the parent to tell the story of their child with a hip condition, and its subsequent identification as Perthes, from their initial concern that something was wrong and its impact on the child, themselves and other children in the family and ending at its impact at the present time and stage of disease management. Follow-up questions and prompts were used to ensure full coverage of a range of potential impacts, for example, the impact on siblings, schooling, playing and socializing with friends, pain following activities, psycho-social effects limitations related to what the child could do, and wider influences on daily life activities. The interviews were thematically analyzed [TG, AFL], leading to the development of a prototype measure. This was centered around uncovering Perthes' impact on the child on a "typical good day" and a "typical bad day." It explored the social, emotional and QoL impact of Perthes on the child. Each item was accompanied by emojis/"smiley faces." These were used as they are child-friendly, easy to interpret and fun to complete. In addition, the child was encouraged to write a brief story of a typical good and a typical bad day.

The measure was piloted with the same two families and their child affected by Perthes (one aged 5 and pre-surgery; one aged 8 and post-Perthes surgery). If necessary, and the case for the 5-year old, the child could seek parental help in completing the measure. Finally, the measure was revised based on parental comments and further methodological advice from a research colleague highly experienced in collecting data from young and teenage children. This led to rephrasing of some items to ease interpretation and to ensure the items were as meaningful as possible to the child. Extracts of the child booklet/measure are presented in Box S1 (the opening page), Box S2 (examples of items to rate by an emoji) and Box S3 (story writing) to illustrate the type and form of a qualitative measure that asks either for emoji responses and/or open-ended comments, stories and pictures. A copy of the full measure can be found in Leo et al.[12]

Consultation with the Health Research Authority confirmed that ethical approval was not required for the research; it was deemed to be service development, aiming to determine important outcomes related to standard care (reference 60/89/81). A signed consent form was collected from parents, in particular, to seek their permission for the interview to be recorded, analyzed and subsequently disseminated, if appropriate, in an anonymized format.

Developing a possible validation protocol

Returning to the question posed in the guiding Muse, the starting point is to reflect on what parts of the standard measure validation methodology are appropriate to utilize. Looking overall, the first two stages in this methodology appear fitting, albeit with some modifications to ensure full adherence to the qualitative paradigm. However, other stages seem more problematic.

Stage 1 is the process of scale/measure development. Common foci, and appropriate in this qualitative context, are features including: patient base and/or patient-centeredness; primary concern with face and content validity; focus on user domain-specific utility (in a health context, patients and clinicians); and practicality and feasibility to use, in both research and, in a health care context, routine clinical practice. The initial item pool may also draw on previous measures and experts' views as long as the content pool also draws on and is grounded in patient views.

Further requirements must be addressed for a child-friendly measure. Examples include: interviews that involve children of the relevant age range and gender; ensuring that, and then exploring if, the measure captures aspects that are important, relevant and meaningful to children and engage their participation in completing the measure.

Stage 2 involves detailed exploration of the measure's face and content validity and its meaningfulness to potential respondents. Techniques would include:

- Cognitive interviewing;
- Interpretability and understanding of the instructions short completing the measure;
- Exploration of the measure's content in terms of its meaningfulness to the potential respondent, be it adult or child; appropriateness of specific measure content,

for example, use of emojis, story writing and pictures, and generation of meaningful discrimination;
- Exploring respondent's (adult or child) engagement and ways to enhance this, if necessary, and respondent burden (for example, time to complete, level of enjoyability and concentration level; and respondents' ability to express themselves verbally, orally and/or pictorially).

However, Stage 3, the process of standard psychometric validation, does not seem appropriate for a qualitative measure, due to its quantitative nature. The only way that psychometric validation methods can be applied would be to transpose appropriate parts of the initial qualitative measure into a quantitative form. The most obvious example for the Perthes measure relates to the emojis; these can straightforwardly be translated into 5-point Likert-type items. For textual data, a coding scheme could perhaps be drawn up based on thematic analysis. Each code would then be allocated a numerical value and a new "scale" developed for each open-ended question, for example, counting the number of allocated codes to one person's textual comments as a proportion of the maximum number of codes arising from all respondents. For other qualitative data, for example, in the form of pictures, perhaps, but somewhat problematical, some sort of marking scheme could be developed. However, these approaches seem somewhat dubious and contrary to the principles and spirit of the qualitative paradigm.

In order to adhere to the qualitative paradigm, a different strategy is called for. So, what approaches can be utilized over and above those outlined in Stages 1 and 2 delineated above? The simplest answer is to say "none," building on the approaches and arguments discussed above in the literature review. In contrast, primary concern must lie on face and content validity, meaningfulness to respondents, respondent-friendliness (be it adult or child), ease of phraseology and understanding by respondents.

In other words, the answer to the Muse is perhaps quite simple:

1. Accept the "qualitative" nature of the measure;
2. Recognize and give primacy to the strictures of the qualitative paradigm, including its emphasis on multiple perspectives, potential of concept saturation and the quality assessment criteria commonly employed in qualitative research (for example, credibility and confirmability);
3. Employ approaches that adhere to the principles and practice of the qualitative paradigm, for example, those used by Mallinson[29] and Hak et al[14]
4. Ensure the measure is grounded in users' perspectives and experiences;
5. Explore in depth the measure's face and content validity, interpretability, its meaningfulness and utility to target groups.

The above approach would seem to provide an acceptable, authoritative and credible approach, and potential gold standard way to validate a qualitative measure and one that adheres to the spirit and principle of the qualitative paradigm.

Discussion and conclusion

This paper set out to draw attention within the PROM field to the issue of how to evaluate a qualitative measure which comprises predominantly open-ended questions. The issue is of special significance in light of the increasing policy and practice of user (for example, patient, adult, child) related outcome measures, along with user-centredness and measures grounded in users' views. Furthermore, within the health field, there is heightened interest in a focus on patient perspectives and the potential and power of collaborative, patient-practitioner decision making.[44–46]

An approach to validate a qualitative measure, and thus address the Muse posed at the beginning of the paper, has been presented. In essence, the outlined approach gives priority to, firstly, using methods that adhere to the principles and practice of the qualitative paradigm, and, secondly, focus on face and content validity, interpretability, meaningfulness to users and utility inter alia in a health context, discussions with patients.

Whether or not such an approach would be perceived as credible and potentially authoritative by advocates of psychometric validation remains to be seen. Notwithstanding, it is evident that psychometric validation is fitting only to a quantitative (outcome) measure. However, it is not appropriate to use psychometric validation procedures where a measure comprises predominantly open-ended questions and where translation of respondents' views into numerical (nominal level) codes, is inappropriate and/or is seen as violating the principles and practice of the qualitative paradigm.

In conclusion, the response to the Muse conundrum is:

> Do not force validation of a qualitative measure, itself comprising predominantly open-ended questions, into a

quantitatively based, psychometric validation 'straight jacket'/procedure.

Retain and continue adherence to the principles of the qualitative paradigm and employ procedures drawn solely from that.

This means in practice placing the emphasis of meaningfulness, face validity and content validity, the meaning of item and measure scores to potential respondents, and, in the context of a child-friendly measure, a focus on the child's views on the above and child-friendly-ness features of the measure. It is hoped that this paper promotes a debate and discussion and ultimately leads to the development of an authoritative and credible approach to qualitative measure validation and one that is recognized within the psychosocial and health research and practice community.

Disclosure

The authors report no conflicts of interest in this work.

References

1. Nunnally JC, Bernstein IR. *Psychometric Theory*. New York: McGraw-Hill; 1994.
2. Scientific Advisory Committee of the Medical Outcomes Trust. Assessing health status and quality-of-life instruments: attributes and review criteria. *Qual Life Res*. 2002;11;193–205.
3. Smith SC, Lamping DL, Banerjee S, et al. Measurement of health-related quality of life for people with dementia: development of a new instrument (DEMQOL) and an evaluation of current methodology. *Health Technol Assess (Rockv)*. 2005;9(10):1–112.
4. Farnik M, Pierzchała W. Instrument development and evaluation for patient-related outcome assessments. *Patient Relat Outcome Meas*. 2012;3:1–7. doi:10.2147/PROM.S14405
5. Guyatt GH, Kirshner B, Jaeschke R. Measuring health status: what are the necessary measurement properties? *J Clin Epidemiol*. 1992;45(12):1341–1345.
6. Guyatt GH. Measuring health-related quality of life: general issues. *Can Respir J*. 1997;4(3):123–130. doi:10.1155/1997/271269
7. Long AF, Dixon P. Monitoring outcomes in routine practice: defining appropriate measurement criteria. *J Eval Clin Pract*. 1996;2:71–78.
8. Greenhalgh J, Long AF, Brettle AJ, Grant MJ. Reviewing and selecting outcome measures for use in routine practice. *J Eval Clin Pract*. 1998;4:339–350.
9. Donovan JL, Frankel S, Eyels JD. Assessing the need for health status measures. *J Epidemiol Community Health*. 1993;47:158–162. doi:10.1136/jech.47.2.158
10. Wolcott HF. *Ethnography: A Way of Seeing*. Lanham, MD: Altamira Press; 1999.
11. Charmaz K. *Constructing Grounded Theory*. 2nd ed. London: Sage; 2014.
12. Corbin J, Strauss A. *Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory*. 4th ed. London: Sage; 2015.
13. Leo DG, Murphy R, Gambling T, Long AF, Jones H, Perry DC. Perspectives on the social, physical and emotional impact of living with Perthes' disease in children and their family: a mixed methods study. *Glob Pediatr Health*. 2019;6:1–10. doi:10.1177/2333794X19835235
14. Hak A, Willems R, van der Wal G, Visser F. A qualitative validation of the Minnesota living with heart failure questionnaire. *Qual Life Res*. 2004;13:417–426. doi:10.1023/B:QURE.0000018487.35591.6e
15. Winter G. A comparative discussion of the notion of 'validity' in qualitative and quantitative research. *Qual Rep*. 2000;4(3/4):1–14.
16. Creswell JW, Miller DL. Determining validity in qualitative inquiry. *Theory Pract*. 2000;39(3):124–131. doi:10.1207/s15430421tip3903_2
17. Golafshani N. Understanding reliability and validity in qualitative research. *Qual Rep*. 2003;8(4):597–607.
18. Lincoln YS, Guba EG. *Naturalistic Inquiry*. Beverly Hills, CA: Sage; 1985.
19. Frost MH, Reeve BB, Liepa AM, Stauffer JW, Hays RD. The Mayo/FDA patient-reported outcomes consensus meeting group. What is sufficient evidence for the reliability and validity of patient-reported outcome measures? *Value Health*. 2007;10(Suppl 2):S94–S105. doi:10.1111/j.1524-4733.2007.00272.x
20. Zumbo BD. Validity as contextualized and pragmatic explanation, and its implications for validation practice. In: Lissetz RW, editor. *The Concept of Validity: Revisions, New Directions and Applications*. Charlotte: Information Age Publishing, Inc.; 2009:65–82.
21. Gadermann AM, Guhn M, Zumbo BD. Investigating the substantive aspect of construct validity for the satisfaction with life scale adapted for children: a focus on cognitive processes. *Soc Indic Res*. 2011;100(1):37–60. doi:10.1007/s11205-010-9603-x
22. Hubley AM, Zumbo BD. Response processes in the context of validity: setting the scene. In Zumbo BD, Hubley AM, editors. *Understanding and Investigating Response Processes in Validation Research*. Social Indicators Research Series 69, Springer International Publishing AG; 2017:1-12.
23. 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 Health*. 2005;8(2):94–104. doi:10.1111/j.1524-4733.2005.04054.x
24. Brod M, Tesler LE, Christensen TL. Qualitative research and content validity: developing best practices based on science and experience. *Qual Life Res*. 2009;18(9):1263–1278. doi:10.1007/s11136-009-9540-9
25. Luyt R. A framework for mixing methods in quantitative measurement development, validation, and revision: a case study. *J Mix Methods Res*. 2011;6(4):294–316. doi:10.1177/1558689811427912
26. Adcock R, Collier D. Measurement validity: a shared standard for qualitative and quantitative research. *Am Polit Sci Rev*. 2001;95:529–546. doi:10.1017/S0003055401003100
27. Lasch KE, Vigneux M, Abetz L, et al. PRO development: rigorous qualitative research as the crucial foundation. *Qual Life Res*. 2010;19:1087–1096. doi:10.1007/s11136-010-9655-z
28. Cheung KKF, Clark AM. Qualitative methods and patient-reported outcomes: measures development and adaptation. *Int J Qual Methods*. 2017;16:1–3.
29. Mallinson S. Listening to respondents: a qualitative assessment of the short-form 36 health status questionnaire. *Soc Sci Med*. 2002;54:11–21. doi:10.1016/s0277-9536(01)00003-x
30. Viswanathan M, Sudman S, Johnson M. Maximum versus meaningful discrimination in scale response. Implications for validity of measurement of consumer perceptions about products. *J Bus Res*. 2004;57:108–124. doi:10.1016/S0148-2963(01)00296-X
31. Breyer BN, Edwards TC, Patrick DL, Voelzke BB. Comprehensive qualitative assessment of urethral stricture disease: toward the development of a patient centered outcome. *J Urol*. 2007;198:1113–1118. doi:10.1016/j.juro.2017.05.077
32. Hardesty DM, Bearden WO. The use of expert judges in scale development: implications for improving face validity of measures of unobservable constructs. *J Bus Res*. 2004;57:98–107. doi:10.1016/S0148-2963(01)00295-8
33. Cremeens J, Eiser C, Blades M. Characteristics of health-related self-report measures for children aged three to eight years: a review of the literature. *Qual Life Res*. 2006;15(4):739–754. doi:10.1007/s11136-005-4184-x

34. Cremeens J, Eiser C, Blades M. A qualitative investigation of school-aged children's answers to items from a generic quality of life measure. *Child Care Health Dev*. 2007;33(1):83–89. doi:10.1111/j.1365-2214.2006.00665.x
35. Acaster S, Dickerhoof R, DeBusk K, Bernard K, Strauss W, Allen LF. Qualitative and quantitative validation of the FACIT-fatigue scale in iron deficiency anemia. *Health Qual Life Outcomes*. 2015;13:60.
36. Cleanthous S, Isenberg DA, Newman SP, Cano SJ. Patient Uncertainty Questionnaire Rheumatology (PUQ-R): development and validation of a new patient-reported outcome instrument for systemic lupus erythematosus (SLE) and rheumatoid arthritis (RA) in a mixed methods study. *Health Qual Life Outcomes*. 2016;14(33):1–14. doi:10.1186/s12955-015-0404-4
37. Ware JE, Sherbourne CD. The MOS 36-item short-form health survey (SF-36) 1. Conceptual framework and item selection. *Med Care*. 1992;30:473–483. doi:10.1097/00005650-199206000-00002
38. Brazier JE, Harper R, Jones NM, et al. Validating the SF-36 health survey questionnaire: new outcome measure for primary care. *BMJ*. 1992;305(6846):160–164. doi:10.1136/bmj.305.6846.160
39. Rector TS, Kubo SH, Cohn JN. Patient's self-assessment of their congestive heart failure: content, reliability, and validity of a new measure: the Minnesota living with heart failure questionnaire. *Heart Fail*. 1987;3:198–219.
40. Morcillo C, Aguado O, Delas J, Rosell F. Utility of the Minnesota living with heart failure questionnaire for assessing quality of life in heart failure patients. *Rev Esp Cardiol*. 2007;60(10):1093–1096. doi:10.1157/13111242
41. Billbao A, Escobar A, Garcia-Perez L, Navarr G, Quiros R. The Minnesota living with heart failure questionnaire: comparison of different factor structures. *Health Qual Life Outcomes*. 2016;14:23. doi:10.1186/s12955-016-0426-6
42. Perry DC. Unravelling the enigma of Perthes disease. *Ann R Coll Surg Engl*. 2013;95:311–316. doi:10.1308/003588413X13629960046192
43. Hailer YD, Haag AC, Nilsson O. Legg-Calvé Perthes' disease: quality of life, physical activity, and behaviour pattern. *J Pediatr Orthop*. 2014;34:514–521. doi:10.1097/BPO.0000000000000157
44. Charles C, Gafni A, Whelan T. Shared decision making in the medical encounter: what does it mean (or it takes two to tango). *Soc Sci Med*. 1997;44:681–692. doi:10.1016/S0277-9536(96)00221-3
45. Edwards A, Elwyn G. Inside the black box of shared decision-making: distinguishing between the process of involvement and who makes the decision. *Health Exp*. 2006;9:307–320.
46. Long AF, Gambling T. Enhancing health literacy and behavioural change within a tele-care education and support intervention for people with type 2 diabetes. *Health Exp*. 2011;15:267–282.

Supplementary materials

For the Child with Perthes

This is new booklet to help us find out how your Perthes affects what you do and the way you feel. We have tried to make these questions as easy as possible to answer. Please try to answer all the questions. If you need your mum or dad to help you, that is fine.

In the tables on the next two pages, we are asking you what you can and cannot do and how you feel on a typical *good* day and a typical *bad* day. Please choose the smiley face that best matches how you feel or what you are able to do.

Example:

On a *good* day, I was able to do everything I wanted to, play outside and with my friends.

Your answer might look like this.

On a *bad* day, I was unable to do what I wanted. The pain was too much. I could not go to school and just hoped some of my friends or my brother or sister might come and sit with me.

Your answer might look like this.

Box S1 Opening page of measures.

Items	On a typical good day…	On a typical bad day…
My hip is painful.		
When my hip does not hurt, I feel happy.		
I can see and play with my friends.		
I am able to go to pre-school or school.		
I am able to sleep well at night.		

Box S2 Example of items to rate with an Emoji.

Can you write us a short story…?

It would be really helpful if you would write down how your hip affects you. Can you do this for a recent good day and then again for a recent bad day?

On a recent good day…

Can you tell us about a recent *good* day, when you have been able to do lots of thing you would normally do and had little pain? Tell us, if you can what you did from when you woke up and got up to when you went to bed. Maybe write it as a short story and draw us a picture too. Do ask your mum or dad to help if you want.

Please write what you want here and on another sheet of paper if you like.

When I woke up and got up in the morning….

When I went to school …

When I came back from school until I went to bed …

Box S3 Story writing.