The final publication is available at IOS Press through https://dx.doi.org/10.3233/EFI-180216

Education for Information: Interdisciplinary Journal of Information Studies, 2019, 35(1): 7-20. DOI : 10.3233/EFI-180216

Towards a comprehensive Questionnaire Origin and Development Appraisal tool: A literature review and a modified nominal group

Joshua Hamzeh[a], Navdeep Kaur[b], Paula Bush[a], Catherine Hudon[c], Tibor Schuster[b], Isabelle Vedel[b], Quan Nha Hong[d] and Pierre Pluye[a,b]

a. Method Development, Quebec SPOR Support Unit, Montréal, QC, Canada
b. Department of Family Medicine, McGill University, Montréal, QC, Canada
c. Department of Family Medicine and Emergency Medicine, Université de Sherbrooke, Sherbrooke, QC, Canada
d. EPPI-Centre, Social Science Research Unit, UCL Institute of Education, University College London, London, UK

ABSTRACT

The questionnaires' origin (sources from which elements of the questionnaire are derived) and initial development (process of making a questionnaire from elements) should be assessed before their measurement properties. There is no Critical Appraisal Tool (CAT) that comprehensively assesses the origin and initial development of questionnaires therefore, our objective was to develop one. To develop the Questionnaire Origin and Development Appraisal (QODA) tool, we first developed QODA version-one from psychometric tool development guidelines, and pilot tested it, resulting in QODA version-two. Second, we performed a review that identified six CATs that evaluate origin and initial development. A pool of items was derived from these six CATs and QODA version-two. Using a nominal group method, this item pool was reduced and QODA version-three (20 items) was developed. The QODA may be useful to aca- demic librarians who assist researchers and students from various disciplines. Information professionals including research trainees can learn how to use QODA and ensure sound questionnaire origin and initial development. This will improve the accuracy of data collected in research and professional practice (e.g., information service evaluation). Future research will study the QODA's measurement properties.

Keywords: Critical appraisal tools, questionnaires, questionnaire origin and development

1. INTRODUCTION

In Library and Information Science (LIS), questionnaires are among the most commonly used research methods (Chu, 2015). Numerous Critical Appraisal Tools (CATs) assess measurement properties of questionnaires (e.g., validity). Rosenkoetter and Tate (2018) identified six CATs that evaluate questionnaire measurement properties. However, none of these CATs comprehensively assesses the quality of origin and initial development of evaluation questionnaires, i.e., they do not evaluate all dimensions (sub-concepts to be measured) of origin and initial development con- structs (main concepts to be measured). To avoid potentially compromised data, the origin and initial development of a questionnaire need to be well established before assessing questionnaire measurement properties. Indeed, the first step of the validation process requires that questionnaires have been developed such that they accurately measure their targeted construct(s) (Hinkin, 1995). Questionnaires' origin and initial development are included in the broader validation process. They constitute two initial and interdependent validation phases, including the search for relevant and representative questionnaire items.

According to Haynes et al., (1995), the origin of a questionnaire is based on "sources from which the questionnaire items and responses are derived" (Step 4). Adding to this, we suggest origin is the sources from which the questionnaire's con- struct, dimensions, items and response options are derived. Constructs, dimensions, items and response options can derive from various sources such as empirical literature, previous questionnaires, theoretical frameworks and suggestions from con- tent experts and intended users of the questionnaire. For example, constructs and dimensions within the health partnership assessment questionnaire by King et al., 2009 originated from a research outcome theory and items within the questionnaire originated from a literature review and interviews with community members and researchers affiliated with participatory research. The initial development of a questionnaire is defined as "the specification of the construct, dimensions, assessment function, matching of items to dimensions, response parameters, instructions to respondents and stimuli, e.g., social scenarios" (Steps 1–3 & 5–9) (Haynes et al., 1995).

The evaluation of questionnaire origin and initial development concerns the conceptual and methodological quality, respectively, of a questionnaire. This differs from the assessment of content validity, which concerns solely the methodological quality of a questionnaire. Content validation is the next step (second step) of the validation process. For example, assessing content validity consists of an evaluation and explanation of the relevance and representativeness of the questionnaire dimensions and items. Content validation serves to measure (quantitative content validation) and justify (qualitative content validation) whether items fit within the confines of the chosen construct (relevance), i.e., main

concept to be measured, and tap all dimensions of the construct (representativeness) (Haynes et al., 1995).

The objective of this study was to describe the conceptualization and construction of the comprehensive Questionnaire Origin and Development Appraisal (QODA) tool. Librarians may use QODA to ensure sound questionnaire origin and initial development. This will improve the accuracy of data collected in research and professional practice (e.g., information service evaluation).

2. METHODS

Unlike researchers focusing on the theory of questionnaire origin and development, Haynes et al. (1995) provide practical guidelines for ensuring that questionnaires derive from appropriate sources and are appropriately developed (Appendix 1). These guidelines are foundational as they align with current conceptual international standards for education and psychology (American Psychological As- sociation, American Educational Research Association, & National Council on Measurement in Education, 2014). Therefore, we used Haynes et al. (1995) guidelines to develop the QODA.

In phase 1, we transformed the Haynes et al. (1995) guidelines into a list of items, created a QODA version-one (V1), and pilot tested it; this led to create QODA version-two (V2). In phase 2, we performed a review that identified six CATs including at least one item appraising the origin and initial development of questionnaires. A pool of items was derived from these six CATs and QODA V2. Using a nominal group method, this item pool was reduced, and QODA version-three (V3) was developed. Each phase is described in detail below.

2.1. Phase 1: Development and pilot testing of preliminary versions of the QODA

Using Steps 1–2 and 4–8 on origin and development from Haynes et al. (1995) guidelines, a team member (PP) developed a 25-item V1 of the QODA tool (Appendix 1). Specifically, he reworded steps from the guidelines as questionnaire items, and added instructions and response options. Guidelines' Steps 3 and 9–13 were not used for the following reasons: Step 3 pertains to all types of assessment methods, and was not used because the QODA applies exclusively to questionnaires; Step 9 was not used as it pertained to stimuli, which is a minor component of questionnaire development; Steps 10–13 were not used because they pertain to subsequent validation steps (content validation and testing of measurement properties). QODA V1 was tested by four team members (CH, JH, PB and PP) by appraising a health partnership assessment questionnaire (King et al., 2009). Pilot test results led to clarify the definition of the main construct (origin and initial development) including 15 origin-related and 13 development-related items. Results also led to the revision of definitions (e.g., the definition of questionnaire 'domain') and items. Moreover, response options were modified from yes/no to Likert

scales, to reflect the degree to which each origin and development item was accomplished. This 28-item QODA V2 was pilot tested using five questionnaires included in the lead author's (JH) systematic review of partnership questionnaires (Hamzeh, 2018).

2.2. Phase 2: Development of QODA V3

2.2.1. Literature review

A literature review was performed to (a) identify CATs that evaluate the quality of the origin and development of questionnaires, and (b) determine whether these CATs were validated and/or reliability tested. The search strategy was developed with assistance from a specialized librarian (Appendix 2). A search was conducted in the Scopus database, covering content indexed from inception to June 2018. An article was eligible for inclusion if it reported (a) in English or French, (b) the development and/or validation and/or reliability testing of any CATs that evaluate the quality of origin and initial development of a questionnaire, and (c) empirical research (original quantitative, qualitative or mixed methods study). Questionnaires were included when they pertained to clinical, educational, psychometric and service/policy assessment. Documents from all countries were included. All Scopus records were imported into EndnoteX8 for screening.

The lead author (JH) screened titles and abstracts of all records retrieved from Scopus, and titles of publications associated to the CATs included in the review by Rosenkoetter and Tate (2018), then read the full texts corresponding to relevant abstracts. If it was unclear whether full-texts satisfied inclusion criteria, the last author (PP) made the final decision. In addition, citations of the publications associated to the CATs in the Rosenkoetter and Tate (2018) review were searched (published from 2016 onwards as the latest CAT of the review was developed in 2016).

From the selected CATs, items evaluating the origin and the initial development were extracted by the lead author (JH) and redundancies removed. For example, items such as "The concept to be measured is clearly stated" and "The main questionnaire construct is clearly defined", were considered to have covered the same subject matter. The last author (PP) reviewed this step, and an item pool was generated. All items were classified according to the Haynes et al. (1995) guidelines and organized into dimensions.

2.2.2. Nominal group

The item pool was reduced by using a modified nominal group technique (McMillan et al., 2016). The traditional nominal group technique, which seeks to build collective agreement on the best ideas (e.g., questionnaire items) of a given topic, is comprised of four steps: silent generation of ideas by

panelists, individual presentation of ideas, clarification of ideas and voting on ideas by panelists. The modified nominal group places some variation on this tradition; for example, as in the case of this study, ideas were derived from a literature review, rather than panelists (McMillan et al., 2016). The nominal group included seven subject experts, i.e., panelists, clinician-researchers and researchers with quantitative, qualitative and mixed methods research backgrounds. The goal of the nominal group was to identify the 'most relevant' items, i.e., those that best match the 'origin and initial development' construct. The panelists collectively (a) discussed the clarity of items, (b) deleted any unclear items, (c) merged similar items, (d) modified the wording of items (if necessary), and (e) suggested new items. The suggestions provided by the panelists were incorporated into a revised item pool, which the panelists then used to independently rate the relevance of each item according to four criteria (1 = not at all relevant, 2 = somewhat relevant, 3 = relevant, 4 = very relevant). Finally, the lead and last authors discussed the results, and included items that met panel consensus, i.e., received average scores of > 3. This led to the QODA V3.

3. RESULTS

3.1. Origin of QODA: Literature review

Using the above-mentioned search strategy, 61 records were retrieved. After screening the titles and abstracts, no record satisfying the eligibility criteria was identified; however, the review by Rosenkoetter and Tate (2018) was identified, including five relevant CATs. After screening the references of the articles associated to the CATs within the review by Rosenkoetter and Tate (2018), we screened the references citing these papers, which led to the identification of one relevant record (Mokkink et al., 2018). One CAT in the review by Rosenkoetter and Tate (2018), the COSMIN tool by (Mokkink et al., 2010) was excluded; while it evaluated content validity, it did not have items for the evaluation of origin and initial development. In the end, the five CATs within the review by Rosenkoetter and Tate (2018) and the CAT within Mokkink et al. (2018) were included. A pool of items was derived from these six CATs and QODA V2. The list of dimensions and items are presented in Table 1.

3.2. Development of QODA: Nominal group

The calculated means of the ratings of relevance for each item and the consensus are presented in Table 1. Of these, nine items did not meet consensus of the panelists and were removed from the item pool. Although item 19 met consensus for being relevant, it was removed from the item pool because all panelists found it was redundant.

As a result of the nominal group, the QODA V3 has 20 items (Table 2).

4. DISCUSSION

The QODA V3 aims to comprehensively appraise the quality of origin and initial development of questionnaires. QODA V3 improves upon the other CATs identified within the literature review in two main ways (Table 3 and Appendix 3). While the QODA V3 contains 20 items evaluating origin and initial development, the other CATs only have a few. Moreover, the QODA V3 may be used to evaluate questionnaires from all domains, not only specific health and information science domains. Thus, the QODA V3 is more comprehensive and generic compared to previous CATs. It is available online in a free public academic wiki (McGill Family Medicine Studies Online, 2019, 13:e06).

The three main strengths of this work are as follows: the literature review search strategy was developed by a specialized librarian; the nominal group panelists had diverse research backgrounds; and the panelists contributed to refine and finalize the QODA V3 with their valuable suggestions based on their expertise and experience. The limits are that the literature review included only one database, one reviewer and had no quality appraisal. While literature reviews are typically performed prior to tool development, our study first developed a tool, and then refined the tool using literature findings. While the nominal group was small, the review by McMillan et al. (2016) found that nominal group techniques ranged from 2 to 14 panelists.

Overall, the QODA V3 is of interest to academic librarians who support re- searchers and students in various disciplines. The QODA V3 may also be valuable for evaluators and researchers involved in the training and continuing education of information professionals, e.g., researchers, graduate students, and professionals. For example, the QODA V3 can be used in two main ways: (a) to perform quality appraisal of relevant questionnaires in literature reviews of questionnaires (in preparation for designing an evaluation or a study), and (b) to plan, conduct and assess the creation and development of a questionnaire. The importance of evaluating the quality of origin and initial development of questionnaires cannot be understated. The theoretical foundation of questionnaires should be well established before the testing of measurement properties (Costello & Osborne, 2005; DeVellis, 2016; Hinkin, 1995). To this end, the QODA V3 can provide information evaluators, professionals and researchers a means to ensure the questionnaires they are using are derived from appropriate sources and are well-constructed.

5. CONCLUSION

In conclusion, the QODA V3 is a practical tool that can comprehensively assess the origin and initial development of questionnaires. Given that the review was con- ducted in Scopus (all sciences), the QODA V3 might actually be useful for educators, evaluators, professionals and researchers from all disciplines to assess whether questionnaires they are using are derived from appropriate sources and have been appropriately developed. Future research is needed to study the measurement properties of QODA V3 (content and construct validity and reliability testing and more as needed). For example, we plan to validate the QODA V3 with the help of experts in qualitative research (ecological content validation with QODA users) and statistics (Bayesian rapid construct validation).

ACKNOWLEDGEMENTS

The authors thankfully acknowledge the support of the Method Development plat- form of the Quebec SPOR SUPPORT Unit, Department of Family Medicine, McGill University, Montréal, Canada. We sincerely thank Dr. Joel Ankri, who dedicated his valuable time to participate in the nominal group.

CONFLICT OF INTEREST

No conflict of interests were declared.

REFERENCES

American Psychological Association, American Educational Research Association, & National Council on Measurement in Education (2014). Standards for Educational & Psychological Tests: American Psychological Association.

Andresen, E. M. (2000). Criteria for assessing the tools of disability outcomes research. *Archives of Physical Medicine and Rehabilitation*, *81*, S15-S20.

Chu, H. (2015). Research methods in library and information science: A content analysis. *Library & Information Science Research*, *37*(1), 36-41.

Costello, A. B., & Osborne, J. W. (2005). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical Assessment, Research & Evaluation*, *10*(7), 1-9.

DeVellis, R. F. (2016). Scale Development: Theory and Applications (Vol. 26). Thousand Oaks, California: Sage publications.

Francis, D. O., McPheeters, M. L., Noud, M., Penson, D. F., & Feurer, I. D. (2016). Checklist to operationalize measurement characteristics of patient-reported outcome measures. *Systematic Reviews*, *5*(1), 129.

Hamzeh, J. (2018). Processes and Outcomes of Organizational Participatory Research Partnerships in Health: A Systematic Mixed Studies Review with Framework Synthesis. (MSc), McGill University, Montreal. Retrieved from http://digitool.library.mcgill.ca/webclient/StreamGate?folder_id=0&dvs=1538688945716~139.

Haynes, S. N., Richard, D. C. S., & Kubany, E. S. (1995). Content validity in psychological assessment: A functional approach to concepts and methods. *Psychological Assessment*, *7*(3), 238-247.

Hinkin, T. R. (1995). A Review of Scale Development Practices in the Study of Organizations. *Journal of Management*, *21*(5), 967-988.

King, G., Servais, M., Kertoy, M., Specht, J., Currie, M., Rosenbaum, P. et al., (2009). A measure of com- munity members' perceptions of the impacts of research partnerships in health and social services. *Evaluation and Program Planning*, *32*(3), 289-299.

Lohr, K. N., Aaronson, N. K., Alonso, J., Burnam, M. A., Patrick, D. L., Perrin, E. B., & Roberts, J. S. (1996). Evaluating quality-of-life and health status instruments: development of scientific review criteria. *Clinical Therapeutics*, *18*(5), 979-992.

McMillan, S. S., King, M., & Tully, M. P. (2016). How to use the nominal group and Delphi techniques. *International Journal of Clinical Pharmacy*, *38*(3), 655-662.

Mokkink, L. B., de Vet, H. C. W., Prinsen, C. A. C., Patrick, D. L., Alonso, J., Bouter, L. M., & Terwee, C. B. (2018). COSMIN Risk of Bias checklist for systematic reviews of Patient-Reported Outcome Measures. *Quality of Life Research*, *27*(5), 1171-1179.

Mokkink, L. B., Terwee, C. B., Patrick, D. L., Alonso, J., Stratford, P. W., Knol, D. L., et al., (2010). The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: an international Delphi study. *Quality of Life Research*, *19*(4), 539-549.

Rosenkoetter, U., & Tate, R. L. (2018). Assessing features of psychometric assessment instruments: A comparison of the COSMIN checklist with other critical appraisal tools. *Brain Impairment*, *19*(1), 103-118.

Scientific Advisory Committee of the Medical Outcomes, T. (2002). Assessing Health Status and Quality- of-Life Instruments: Attributes and Review Criteria. *Quality of Life Research*, *11*(3), 193-205.

Terwee, C. B., Bot, S. D. M., de Boer, M. R., van der Windt, D. A. W. M., Knol, D. L., Dekker, J. et al., (2007). Quality criteria were proposed for measurement properties of health status questionnaires. *Journal of Clinical Epidemiology*, *60*(1), 34-42.

Terwee, C. B., Prinsen, C. A. C., Chiarotto, A., Westerman, M. J., Patrick, D. L., Alonso, J. et al., (2018).

COSMIN methodology for evaluating the content validity of patient-reported outcome measures: a Delphi study. *Quality of Life Research*, *27*(5), 1159-1170.

Valderas, J. M., Ferrer, M., Mendívil, J., Garin, O., Rajmil, L., Herdman, M., & Alonso, J. (2008). Development of EMPRO: A Tool for the Standardized Assessment of Patient-Reported Outcome Measures. *Value in Health, 11*(4), 700-708.

TABLES AND FIGURES

Table 1 Consensus of nominal group on QODA V2 (28- items)

# Item	Mean*	Consensus (mean≥3)**
Initial development (construct & item specification)		
1. The definition of the construct(s) measured by the questionnaire is/are appropriate	3.9	√
2. The context of use is appropriate.	2.6	No
3. The target population is appropriate.	3.6	√
4. The purpose of the questionnaire is appropriate.	3.1	√
5. The response scale(s) is/are appropriate.	3.3	√
6. The dimension(s) measured by the questionnaire is/are appropriate.	2.9	No
7. The dimension(s) measured by the questionnaire is/are appropriate.	3.6	√
8. There is coherence between dimensions and the construct.	3.3	√
9. There is coherence between items and response scale(s).	3.4	√
Origin (construct & item& function sources)		
10. The rationale for combining multiple items is supported by theoretical and/or empirical work(s).	2.7	No
11. Dimensions and items are derived from the input of content experts who are not the intended users of the questionnaire.	3.2	√
12. Dimensions and items are derived from the practical experience of the intended users of the questionnaire.	3.3	√
13. Dimensions and items are derived from other questionnaires relevant to the construct.	3.5	√
14. Dimensions and items are derived from a conceptual framework or theory relevant to the construct.	3.4	√
15. Dimensions and items are derived from empirical studies relevant to the construct.	3.4	√
16. The purpose of the questionnaire is supported by theoretical and/or empirical work(s).	2.3	No

17. The response scale of questionnaire measurement is supported by theoretical and/or empirical work(s).	1.9	No
Origin (methodological quality of studies on origin of construct of items)		
18. The involvement of the target population in generating items for the questionnaire was appropriate.	2.1	No
19. The questionnaire development study was performed in a sample representing the target population.	3.3***	No
Initial development (clarity of construct & items; specification of responses & scales & instructions)		
20. Information on the questionnaire development phase is provided (Screening Q)	3.4	√
21. Items were edited by questionnaire developers.	1.9	No
22 The rationale for any modification of dimensions was appropriate.	3.5	√
23 All dimensions were appropriately worded.	2.3	No
24 The rationale for any modification of items was appropriate.	3.2	√
25 Methods for deriving scores for the questionnaire and/or its dimensions are appropriate.	2.7	No
26 The rationale for transforming data (such as weighting and standardization) is appropriate.	3.3	√
27 The instructions for administering and scoring the questionnaire are clear and complete.	3.3	√
28 The instructions for completing the questionnaire are clear and complete	4.0	√

*Mean > 3 = relevant or very relevant. **Consensus: √ = Yes. **Panelists found this item was redundant therefore, it was removed from the item pool.

Table 2 Final version of the QODA V3 (20 items)

#	Item
Initial development (construct & item specification)	
1	The definition of the construct(s) measured by the questionnaire is/are appropriate.
2	The target population is appropriate.
3	The purpose of the questionnaire is appropriate.
4	The response scale(s) is/are appropriate.
5	There is coherence between dimensions and the construct.
6	There is coherence between dimensions and items.
7	There is coherence between items and response scale(s)
Origin (construct & item & function sources)	
8	Dimensions and items are derived from the input of content experts who are not the intended users of the questionnaire.
9	Dimensions and items are derived from the practical experience of the intended users of the questionnaire.
10	Dimensions and items are derived from other questionnaires relevant to the construct.
11	Dimensions and items are derived from a conceptual framework or theory relevant to the construct.
12	Dimensions and items are derived from empirical studies relevant to the construct.
13	The purpose of the questionnaire is supported by theoretical and/or empirical work(s).
Origin (methodological quality of studies on origin of construct of items)	
14	The questionnaire development study was performed in a sample representing the target population.
Initial development (clarity of construct & items; specification of responses & scales & instructions)	
15	Information on the questionnaire development phase is provided (Screening Q).
16	The rationale for any modification of dimensions was appropriate.
17	The rationale for any modification of items was appropriate.
18	The rationale for transforming data (such as weighting and standardization) is appropriate.
19	The instructions for administering and scoring the questionnaire are clear and complete.
20	The instructions for completing the questionnaire are clear and complete.

Table 3 Comparison of QODA V3 with Other CAT's

	Applies to all questionnaires within health sciences	Made from theoretical foundation	Made from literature	Made from stakeholder experience	Compr-ehensive
LOHR 1996	No (only applies to health outcome measures)	Unclear	Unclear	Unclear (perhaps those within committee are instrument development experts)	No
Andersen 2000	No (only applies to questionnaire within disability outcomes research)	Unclear (theoretical articles may have been found within the literature search)	Yes (literature, e.g., textbooks, articles, and guidelines, were used)	Yes (disability and research experts con-tributed)	No
Terwee 2007	No (designed to assess health status questionnaires)	No	Yes (made from previous criteria that evaluate questionnaires)	Yes (experts adapted the criteria based on their experiences from pilot testing)	No
Valderas 2008	No (applies to patient Reported Outcome measures)	No	Yes (adapted from previous criteria on evaluating questionnaires)	Yes (Experts were involved in adapting previous criteria into a new questionnaire)	No
Francis 2016	No (applies to patient Reported Outcome measures)	No	Yes (Items were derived from a literature	Yes (Items were pilot-tested. and reliability tested with	No

			review that included empirical studies, textbooks, guidelines and reports related to evaluating PROMs)	clinicians and researchers)	
Mokkink 2018	No (applies to patient Reported Outcome measures)	Unclear (theoretical articles may have been found within the literature search)	Yes (three literature searches were made)	Yes (e-Delphi with experts)	No
Hamzeh 2018	Yes	Yes	Yes	Yes	Yes

*The features in the above table are ordered according to steps 1–5 from (Haynes et al., 1995) (Appendix 1) checklist for content validation of psychometric instruments: (1) Specification of construct(s), domain, dimensions and facets, (2) Specification of intended functions of the instrument, (3) Initial selection and generation of items (deduction, clinical experience, theories relevant to construct, empirical literature relevant to construct, other assessment instruments, expert suggestion, target population suggestion), and (4) Matching item to facets and dimensions (use table of facets to insure coverage, generate multiple items for each facet, insure proportional representation of items across facets). **Theory that gives explanation of phenomenon without testable hypotheses (Gregor, 2006). ***Includes nearly all dimensions of construct (i.e., no missing constructs when compared with related frameworks).

APPENDICES

Appendix 1: Haynes et al. 1995 Guidelines

Procedures and sequence of content validation according to Haynes et al. guidelines (Haynes et al., 1995)

1. Specify the construct(s) targeted by the instrument
    a. Specify the domain of the construct
        i. what is to be included
        ii. what is to be excluded
    b. Specify the facets and dimensions of the construct
        i. factors of construct to be covered
        ii. dimensions (e.g., rate, duration, and magnitude)
        iii. mode (e.g., thoughts and behavior)
        iv. temporal parameters (response interval and duration of time-sampling) situations
2. Specify the intended functions of the instrument (e.g., brief screening, functional analysis, and diagnosis)
3. Select assessment method to match targeted construct and function of assessment
4. Initial selection and generation of items (e.g., questionnaire items, behavior codes, psychophysiological measures, and behaviors monitored)
    a. from rational deduction
    b. from clinical experience
    c. from theories relevant to the construct
    d. from empirical literature relevant to the construct (e.g., studies on construct validity of potential items)
    e. from other assessment instruments (i.e., borrowing items from other instruments that have demonstrated validity)
    f. from suggestions by experts
    g. from suggestions by target population
5. Match items to facets and dimensions
    a. use table of facets to insure coverage (include all relevant dimensions, modes, temporal parameters, and situations)
    b. generate multiple items for each facet
    c. insure proportional representation of items across facets (i.e., the relative number of items

in each facet should match the importance of that facet in the targeted construct)

6. Examine structure, form, topography, and content of each item
    a. appropriateness of item for facet of construct
    b. consistency and accuracy, specificity and clarity of wording, and definitions
    c. remove redundant items
7. Establish quantitative parameters
    a. response formats and scales
    b. time-sampling parameters (sampling intervals and durations)
8. Construct instructions to participants
    a. match with domain and function of assessment instrument
    b. clarify; strive for specificity and appropriate grammatical structure
9. Establish stimuli used in assessment (e.g., social scenarios, and audio and video presentations) to match construct and function
10. Have experts review the results of methods 1–3 and 5–9
    a. quantitative evaluations of construct definition, domain, facets, mode, and dimensions
    b. quantitative evaluation of relevance and representativeness of items and stimuli
    c. quantitative evaluation of response formats, scales, stimuli, situations, time-sampling parameters, data reduction, and aggregation
    d. match of an instrument attributes to its function
    e. qualitative evaluation – suggested additions, deletions, and modifications
11. Have target population sample the results – review quantitative and qualitative evaluation of items, stimuli, and situations
12. Have experts and target population sample review the modified assessment instrument
13. Perform psychometric evaluation and contingent instrument refinement – criterion-related and construct validity, and factor analysis

Appendix 2: Literature Review Search Strategy (Scopus)

(TITLE (appraisal*) OR TITLE (“critical assessment*”) OR TITLE (“critical evaluation*”) OR TITLE (“quality evaluation*”) OR TITLE (“quality review*”) OR TITLE ({risk of bias}) OR TITLE (“quality assessment*”) OR TITLE ({methodological quality}) OR TITLE (assess* AND methodolog*) AND TITLE (questionnaire*) OR KEY ({surveys and questionnaires}) OR TITLE (American Psychological Association et al.) OR TITLE (instrument) OR TITLE (instruments) OR TITLE ({psychometric properties}) AND KEY (checklist*) OR KEY (instrument*) OR KEY (tool*) OR KEY (guideline*) OR KEY (scale*) AND KEY (improv*) OR KEY (refin*) OR KEY (outset) OR KEY (design*) OR KEY (preliminary) OR KEY (develop*) OR KEY (origin) OR KEY (deriv*)).

Appendix 3: Purpose, audience, dimensions and measurement properties of included CATs

The COSMIN Risk of Bias checklist for systematic reviews of Patient Re- ported Outcome Measures (Mokkink et al., 2018) assesses the methodological quality of studies on measurement properties of Patient-Reported Outcome measures (PROMs), specifically for use in systematic reviews. This CAT was developed by making updates to a previous COSMIN Risk of Bias Tool (Mokkink et al., 2010). The Content Validity section was updated through an international e-Delphi with 159 experts (experts in qualitative research, instrument development and validation and systematic reviews) from 21 countries, whereby experts rated agreement to items derived from previous guidelines, task-force articles, and methods-articles, amongst others (Terwee et al., 2018). The Content Validity section was updated to include PROM development, and previous items pertaining to content validity were also re- vised. The tool was pilot tested on two systematic reviews on PROMs. Only the PROM development dimension has items relating to origin and initial development.

The checklist created by Francis et al., (2016) assesses the quality of measurement properties of PROMs, and whether a PROM is appropriate for use under given circumstances. The tool is an 18-item checklist, with 6 dimensions. The tool can be used by systematic reviewers, researchers, and clinicians; users do not require a minimal level of expertise with instrument development. Items were derived from a literature review that included empirical studies, textbooks, guidelines and reports related to evaluating PROMs. The items were pilot-tested, and reliability tested (interrater reliability) with clinicians and researchers. While the tool is an efficient way to evaluate measurement properties of PROMs, there are few items pertaining to origin and development; the conceptual model and content validity dimensions each have 3 items, which pertain to origin and initial development. Furthermore, lacking an overall score makes it difficult for users to determine whether a given PROM has undergone adequate development prior to testing of measurement properties.

Andresen (2000) provides evaluation items for assessing questionnaires within disability outcomes research. The items were derived from instrument developmental and evaluation literature (e.g., guidelines for instrument selection), stemming largely from the field of Quality of Life measurement, and disability and research experts. Of the eleven dimensions, the Conceptual Model and Alternate/accessible forms dimensions each contain an item related to evaluation of origin and/or initial development. Furthermore, these evaluation items were developed specifically for measuring rehabilitation and disability outcomes research measures and cannot generalize easily to other questionnaires.

The Scientific Advisory Committee of the Medical Outcomes (2002) provides updated evaluation items for health status and QoL instruments, including questionnaire and interview guides. These items

are an update to previously developed items by Lohr et al. (1996). The Conceptual and Measurement Model dimension within the criteria by Lohr et al. (1996) included items related to origin and initial development. These previous criteria were developed by a committee to evaluate health outcome measures, to be retained within a repository and distributed to interested users. The update was made by changes the definitions and items to reflect modern test theory, and by better differentiating between definitions of dimensions and items within a given dimensions.

The criteria were then reviewed by six international researchers. The criteria by Terwee et al. (2007) was created to assess the quality of studies on the development and testing of health status questionnaires. The criteria originated from previous guidelines that evaluate questionnaires: Scientific Advisory Committee of the Medical Outcomes (2002) and Andresen (2000), amongst others. The criteria were then pilot tested by the authors in two systematics reviews and revised accordingly. Each criterion could be rated as positive (present), negative (absent) or indeterminate, within each questionnaire development study. Of the eight properties evaluated, only Content Validity contained an item that matched origin and initial development. While Terwee et al. (2007) provide clearly worded criteria, these criteria do not allow one to rate the degree to which each criteria has been achieved.

Valderas et al. (2008) developed a tool called the Evaluating the Measurement of Patient-Reported Outcomes (EMPRO). The EMPRO has 39-items and eight dimensions. Four experts in the development and testing of PROMs adapted the evaluation items by Scientific Advisory Committee of the Medical Outcomes (2002) and included a response scale for each item, to make the EMPRO. The EMPRO was then pilot tested on six PROM questionnaires. EMPRO demonstrated high internal consistency (alpha = 0.95 and interrater concordance was 0.87–0.94) and good construct validity. The Conceptual and Measurement model dimension contained nine items related to evaluation of origin and/or initial development.