COSMIN: Development and evaluation of a checklist to assess the methodological quality of studies on measurement properties

Mokkink, L.B.

2010

document version
Publisher's PDF, also known as Version of record

8

General Discussion

COSMIN

This dissertation describes the development and evaluation of the COSMIN (COnsensus-based Standards for the selection of health Measurement INstruments) checklist, i.e. the design of the Delphi study (Chapter 2), a systematic review of systematic reviews of measurement properties (Chapter 3) which is used as a preparation of the Delphi study, three articles describing the results of the Delphi study (Chapters 4 to 6), and a reliability study of the COSMIN checklist (Chapter 7).
The COSMIN checklist contains twelve boxes. Ten boxes can be used to assess whether a study meets the standard for good methodological quality. Nine of these boxes contain standards for the included measurement properties (internal consistency (box A), reliability (box B), measurement error (box C), content validity (box D), structural validity (box E), hypotheses testing (box F), cross-cultural validity (box G), criterion validity (box H) and responsiveness (box I), and one box contains standards for studies on interpretability (box J). In addition, two boxes are included in the checklist that contain general requirements for articles in which IRT methods are applied (IRT box), and general requirements for the generalisability of the results (Generalisability box), respectively. Definitions of each of the measurement properties are described in Appendix 5, and the checklist containing the standards can be found in the Appendix 7. Due to lack of time and complexity of the issues in the first two research questions (i.e. consensus on (1) terminology and definitions, and (2) standards), we could not study the third research question concerning criteria of adequacy. Therefore, these criteria were not developed. As a consequence, the COSMIN checklist cannot be used yet to evaluate the quality of a measurement instrument.

This general discussion focuses on some methodological aspects of the research described in this dissertation: the suitability of the Delphi technique in relation to the aims of the COSMIN Delphi study, the appropriateness of the COSMIN panel, and the likelihood of bias which may have influenced the outcomes of the Delphi study. Subsequently, we discuss which measurement properties of the COSMIN checklist should additionally be evaluated. Finally, we put forward recommendations for using the COSMIN checklist and suggestions for future research, and we close with some final reflections.

Methodological considerations

Was the Delphi technique suitable for the objectives of the COSMIN Delphi study?

The Delphi technique is particularly suitable when there is a lack of knowledge or agreement on a particular subject, or when there is a substantial amount of information, which is contradictory. Both situations cannot be solved by new empirical evidence. However, by optimally using the knowledge, experience and opinions of many experts, consensus can be reached. The Delphi technique is a strong design to organise and structure the group communication process for reaching consensus[1-5].

In the COSMIN Delphi study the aim was to reach consensus on terminology of measurement properties, their definitions and places in the taxonomy, and standards to evaluate the methodological quality of studies on measurement properties. First, the best term for each of the measurement properties cannot be decided empirically. This is an arbitrary choice, which should be agreed upon among a large group of experts. In the COSMIN Delphi study experts were involved with a background in different relevant fields, such as psychology, epidemiology, statistics and clinical medicine. Each field has its own set of terms, which may differ from each other, but refer to the same constructs. The chosen set of terms must be acceptable to all disciplines and not lead to confusion for any of the disciplines. For example, the term "sensitivity to change" was proposed for the measurement property which we decided to label as responsiveness. One of the reasons not to choose "sensitivity to change" was the similarity with the term sensitivity in the context of diagnostic testing. This could lead to confusion, and the term was therefore rejected.

The second topic concerned the definitions of each of the measurement properties and their places in the taxonomy. In round 1 we started by presenting several definitions found in the literature for each measurement property. Some of these differ only slightly, while others differ a lot. For example, two of the definitions we proposed for internal consistency are quite similar, i.e. 'the ability to measure a single underlying concept', and 'the degree to which all test items measure the same trait'. However, we also considered the definition 'the repeatability of the same instrument over two versions of this instrument'. We came to consensus by asking the panel members' preferred definitions, and their arguments. Based on arguments of the panel members, we understood that some of the definitions proposed reflected uni-dimensionality (also called homogeneity). This was considered to be a different construct than internal consistency, and to be a prerequisite for a good interpretation of the internal consistency coefficient. As suggested by the panel members, in round 3 we proposed the definition given by Cortina, i.e. 'the degree of the interrelatedness among the

items'[6]. Eighty percent of the panel members agreed to this definition. The distinction with homogeneity was reflected in the items of the box of internal consistency, i.e. the items A5 to A7. These items ask if uni-dimensionality is checked (A5), in a study with an appropriate sample size (A6), and whether internal consistency is assessed for each uni-dimensional scale (A7).

The construct of internal consistency, as it is now defined, does not tell anything about whether the instrument measures the construct it purports to measure, but only that the items together measure something. Therefore, internal consistency is considered part of the domain reliability, and not validity.

This example reflects that the choice of the best definition is a matter of agreement between experts, based on arguments. It cannot empirically be investigated. A Delphi study is a suitable design to reach consensus on e.g. a definition that best suits to the opinion of the majority of the group. The place of each measurement property in the taxonomy is a logical consequence of the chosen definition. Arguments for the best place of each measurement property can also be identified by a Delphi procedure.

The third topic concerned the development of the standards for studies on measurement properties. We were looking for the best study design and for the preferred statistical methods to investigate the measurement properties. The decision which standards are the best is less a matter of opinion, but should be based on strong arguments and logical reasoning, and should be in line with the definition chosen. The Delphi technique is a useful design to identify the appropriate arguments.

Instead of a Delphi study containing written rounds, we could have organized face-to-face (nominal group) meetings. However, a written Delphi technique is able to handle large groups of participants – larger than a procedure in which participants meet face-to-face[7]. Since the COSMIN panel members came from all over the world, it was not feasible to meet with all these people at the same time. Furthermore, the chance that some participants would be highly dominant, which would have influenced the results, is also less likely when written communication is used, and anonymity is guaranteed[7].

Several disadvantages of the Delphi technique should be acknowledged. Although the Delphi technique can be considered as a relative efficient way to combine knowledge and abilities of a group of experts, the method needs an extensive time commitment from the panel members[2]. This can lead to a low response rate, decreasing with each new round, and selection bias. A rough estimation of the average time it took a panel member in our study to complete a questionnaire is approximately four hours for each questionnaire. Panel members did point out this issue (remarks given in the fourth round): 'much more time consuming than

anticipated', 'exhausting for everyone', 'extremely frustrating', and 'most of us hate them – while we participate'. In the COSMIN Delphi study the response rates of round 1 was 74%, and of the rounds 2 to 4 stayed approximately the same, i.e. around 50%.

The results of a Delphi study are only a reflection of opinions, it is uncertain whether the correct answer is found[3,5]. The result of a Delphi study represents a group's opinion at a given point in time[8]. This holds for the COSMIN Delphi study as well. Perhaps in the future, an alternation of opinion may occur, by which the content of the COSMIN checklist may become outdated.

Was the COSMIN panel appropriate?

The success of a Delphi study is highly dependent on the composition of the panel. In most studies, patients are randomly selected. However, the participants in a Delphi study need not to be randomly selected. Rather the so-called purposive sampling method is more appropriate[4]. This means that people are being selected for a purpose. In the COSMIN Delphi study experts were chosen because of their knowledge of and methodological expertise in measuring health.

Furthermore, two characteristics of the panel are important, i.e., heterogeneity and anonymity. Groups with heterogeneous opinions will have different perspectives on an issue[2]. This was in our study accomplished by including people having different backgrounds, and coming from various institutions and countries. The involvement of representatives of different groups involved in measuring health, will lead to a more heterogeneous group regarding their opinions. A heterogeneous group has a higher chance to produce a high quality, and widely acceptable result[2]. Anonymity of each of the panel members provides the opportunity to present and react to ideas unbiased by the identities of it advocates[8]. It prevents panel members to be intimidated or inhibited from expressing their views[7]. In our Delphi study only one person (LM) had access to all data leading to the names of the panel members. She was a member of the Steering Committee and not of the COSMIN panel. She knew who were participating, and responding, and who gave which arguments. One of the panel members remarked in the end of the Delphi study 'I would be curious to know who where selected to participate and who participated actually till the end. Only one participant (Geoff Norman) revealed his identity in one point'.

Furthermore, the panel members must be willing and able to participate and commit to the study. To get people involved and stay committed in the Delphi study, it is important that they (1) appreciate the relevance of the study, (2) have knowledge of the topics of the study, (3) recognize that their contribution is useful and has some influence on the results, and (4) are prepared to learn from each other and if necessary alter their opinion.

In our opinion the panel involved in this Delphi study was appropriate, because it was heterogeneous. Another reason to conclude that the panel was appropriate is the fact that many of those who agreed to participate maintained involved until the process was completed, despite the large burden on their time. Of the 43 panel members who participated 58% responded three or four times.

Forms of bias that may influence the outcomes of a Delphi study

The Delphi technique is sensitive to three forms of bias, i.e. selection bias, subject bias, and bias introduced by the researchers in the interpretation of the findings. Selection bias occurs when the sample is too homogeneous with regard to their opinion[2]. Subject bias occurs when panel members change their views in line with what others are saying, because they know the group's responses[4]. In contrast, this is also considered an advantage or even the goal of a Delphi study. However, it is unclear whether panel members change their opinions on the basis of convincing arguments or feel put under pressure to conform to the group's view[3]. This latter situation is subject bias. A way to avoid subject bias is by keeping participants anonymous. Quotes given by panel members at the end of the Delphi study do not reveal a major influence of subject bias: 'seeing the arguments and votes of others helped me to make decisions', and 'interesting for contributors to read what others commented'.

A major objection against the Delphi technique is that the researchers may introduce bias in the interpretation of the findings[7]. We tried to avoid this by formulating a priori criteria for reaching consensus, by being as transparent as possible, by providing all arguments of panel members, and by explaining our choices. The tendency to include all arguments may create questionnaires that are very extensive. This could put panel members off participating[3]. Therefore, we choose to split the questionnaire and the feedback report. For each round we made a questionnaire in which the proposals and questions were presented. In this document we gave those arguments that we considered relevant. To avoid bias introduced by our choices, we made a separate document, in which we provided all responses to each question of the previous round. We also tried to avoid bias introduced by explaining our choices, by asking the panel members' opinions, and by encouraging them to provide their arguments for their choices. For example, in round 3 we proposed new terms for responsiveness, i.e. construct responsiveness and criterion responsiveness. In round 2 the panel agreed that the only difference between responsiveness on the one hand and construct and criterion validity on the other hand, was that responsiveness was about change scores and validity about single scores. Therefore, the Steering Committee proposed to introduce new terms in round 3 to emphasize the similarity with validity, including the distinction between evaluating change scores when a gold standard is available, and when not. However, panel members did not want to introduce new terms, and only 48% and 52% agreed with the terms construct and criterion responsiveness, respectively.

Should the COSMIN checklist additionally be evaluated?

The COSMIN checklist measures the methodological quality of a study on measurement properties. Since the COSMIN checklist is a measurement instrument in itself, the measurement properties of the COSMIN checklist should be thoroughly investigated. Within the COSMIN Delphi study, we reached consensus on the inclusion of individual items per measurement property. Within the panel, we did not decide on how to combine the individual scores of the items per box into an overall score for each box. However, to investigate some of the properties of the COSMIN checklist, such as construct validity or interpretability, this is needed. At the moment, the COSMIN Steering Committee in cooperation with the Working Group on Clinimetrics of the EMGO institute for Health and Care Research is developing such a rating system to assess the methodological quality of the study per measurement property that is assessed as having excellent, good, fair, or poor methodological quality. In general, each item is rated as excellent, good, fair, or poor. The overall score per box is determined by the lowest scored item. For example, if all items in the box are scored with "yes" (i.e. excellent score), except for one item that received a moderate score, this latter score determines the overall score of the box, i.e. moderate methodological quality of the study on a measurement property. For more details, see the COSMIN website: www.cosmin.nl.

To evaluate the COSMIN checklist, only three measurement properties are relevant: reliability, content validity and hypotheses testing. Also interpretability is relevant. We first explain why some measurement properties are not relevant in this case, and then we discuss the relevant properties.

Internal consistency and structural validity. Internal consistency and structural validity are not relevant because the items in the COSMIN boxes are not based on a reflective model. Within the conceptual framework of the COSMIN checklist, the items determine the methodological quality, and not visa versa. Therefore, the items don't need to be strongly correlated with each other, and are based on a formative model. For example, in a study on reliability when the authors have used an inappropriate time interval, the methodological quality of the study decreases. However, when a study has low methodological quality, it may have used an adequate time interval, but used a very small sample size.

Measurement error. Measurement error cannot be assessed, because there is no parameter of measurement error for ordinal or nominal scales.

Cross-cultural validity. Cross-cultural validity is currently not relevant because the COSMIN checklist is only available in English. We have no intention to translate the checklist into other languages.

Criterion validity. Criterion validity cannot be assessed, because there is no gold standard for assessing the methodological quality of studies on measurement properties.

Responsiveness. Responsiveness is not relevant, because the studies that are being evaluated with the checklist do not change over time.

Reliability. In chapter 7 we investigated the inter-rater reliability of single users of the checklist. Fifty-nine percent of the items had poor kappa's (below 0.40), while percentage agreement between raters per item was above 80% for 64% of the items, and above 70% for 93% of the items. However, it is highly recommendable when using the COSMIN checklist in a systematic review to complete it by two trained and independent raters, and reach consensus on each item. Therefore, in future research the inter-rater reliability of consensus obtained by two couples of trained and independent raters should be investigated. Such a procedure has been investigated to assess adverse events of clinical care, and showed poor results[9]. However, this procedure may increase the inter-rater reliability of the checklist.

Content validity. By involving many experts from different fields in the development of the checklist, it is highly likely that all relevant items of all relevant measurement properties are included. This contributes to the content validity of the checklist. However, since content validity is a subjective judgement, other researchers should judge this. Ideally, a new study should be performed in which a new expert panel is involved. This panel decides whether all items are relevant for the construct (i.e. measuring the methodological quality of studies on measurement properties), the "population" (i.e. HR-PRO instruments), and the purpose (i.e. discrimination). In addition, they should decide whether all items together comprehensively measure the methodological quality of the study.

Hypotheses testing. We consider it useful, but very difficult to formulate challenging and specific hypotheses to evaluate construct validity of the COSMIN checklist by testing hypotheses. Difficulty lies in the fact that no overall score per box can be calculated and the lack of validated comparison instruments.
It seems difficult to formulate interesting hypotheses on item level. Comparing individual items of the COSMIN checklist with items from different checklists is an aspect of content validity. However, even when an overall score for each box can be calculated, it is difficult to formulate hypotheses. Other lists, such as the SAC-MOS criteria[10], and the Terwee criteria[11], focus on the quality of the measurement instruments, rather than on the quality of the study, as measured by the COSMIN checklist. Furthermore, the validity of none of these existing checklists has been evaluated. Therefore, it will be difficult to interpret results of such a study on hypotheses testing.

An alternative, more interesting approach might be to compare scores on the COSMIN checklist with expert opinions. In the future, when the rating system is in use, studies on measurement properties can be classified into having e.g. "excellent/good/fair/poor" methodological quality. These ratings could be compared with the opinion of a group of experts. The agreement between these two methods can be hypothesized to be high (e.g. kappa >0.70) and could be tested.

Interpretability. In addition to the measurement properties, it is useful when the COSMIN checklist can be easily interpreted. Therefore, an overall score for each box is recommended, so a study can be rated as excellent, good, fair or poor. At the moment, this rating system is being developed.

Recommendations for using the COSMIN checklist

The COSMIN checklist can be used for different purposes:

- it can be used in a systematic review of measurement properties in which the methodological quality of studies on measurement properties of instruments with a similar purpose are assessed. The taxonomy of COSMIN can also be used to see whether all relevant measurement properties are studied
- authors of primary studies on measurement properties can use the checklist to design and report a study on measurement properties to make sure that their study meets the standards for excellent quality.
- editors and reviewers of manuscripts may use the checklist to assess whether the quality of a study on measurement properties is high enough to justify publication of the study
- students can use the checklist when learning about measurement properties

At www.cosmin.nl the COSMIN checklist manual can be downloaded, which contains detailed instructions for how to complete the checklist. However, different purposes may require a different use of the checklist. In some cases the checklist is adequate to use as it is. For example, when using the checklist as an author of a primary study to assess the methodological quality of one's manuscript, one can check whether the right study design and statistical methods are used, and whether all items are reported in the manuscript. If some items are negatively checked, the author can add the required information to get a positive score. Reviewers or editors can use the checklist to detect issues that are not described in the manuscript. These overlooked issues can be reported back to the authors.

However, when using the checklist in a systematic review, we recommend incorporating the checklist in the data extraction form. This way, not only the answering options "yes", "no", "not applicable", or "?" can be completed, but additional relevant qualitative information can be extracted at the same time. For example, we recommend not only to decide whether the time interval in a reliability study was appropriate (by answering item 8 of Box B of reliability with e.g. "yes"), but also to describe how long the time interval was.
In a systematic review of measurement properties, it is not only relevant to evaluate the methodological quality of the included studies, but also to describe the measurement instrument (e.g. description of construct, development process, theoretical framework, information about feasibility, costs, time to administer, administration mode, language version, etc.), the characteristics of the study (e.g., number of participants, time between administrations, description of methods used, etc.), and the results of the studies on measurement properties. All these issues could be extracted using one extensive data extraction form. It is useful that a standardized extraction form will be developed and easily available, so that researchers who conduct a systematic review of measurement properties can perform the assessment of the studies in a standardised way. At the moment, such an extensive data extraction form is being developed at the Department of Epidemiology and Biostatistics of the VU University Medical Center, and will be freely available in the future.

In Chapter 3 we gave recommendations for a high quality systematic review of measurement properties. These recommendations contained the use of an appropriate search strategy, transparent reporting of the methods used to perform the review, evaluation of the methodological quality of the primary studies, and evaluation of their results.
Guidelines for several of these recommendations have been proposed. Terwee et al. developed a highly sensitive search filter for finding studies on measurement properties in PubMed[12]. This search filter could be used in combination with search terms for the construct of interest, for the population of interest, and for the type of measurement instruments of interest, to find all relevant studies. The PRISMA statement[13] can be used to transparently report the design of the systematic review. The COSMIN checklist can be used to evaluate the methodological quality of the primary studies. Guidelines for interpretation of the results (i.e. criteria of adequacy) are unfortunately not yet consensus-based developed. Until that time, researchers may use the guidelines proposed by several initiatives, such as in the SAC-MOS criteria[10], the Terwee criteria[11], and EMPRO[14].

Recommendation for future research

R1
R2
R3
R4
R5
R6
R7
R8
R9
R10
R11
R12
R13
R14
R15
R16
R17
R18
R19
R20
R21
R22
R23
R24
R25
R26
R27
R28
R29
R30
R31
R32
R33
R34
R35
R36
R37
R38
R39

Research on the COSMIN checklist

- When using the checklist in a systematic review of measurement properties – but also in other use – we recommend that two researchers should complete the checklist and come to consensus on their ratings, similar as is being recommended for data extraction in all types of systematic reviews. Although we have studied the inter-rater reliability between single raters (Chapter 7), it would also be useful to investigate the inter-rater reliability of the consensus obtained by couples of raters, because this better reflects the real use of the checklist.
- The checklist was developed as a multidisciplinary, international collaboration with all relevant expertise involved. By using this approach, it is highly likely that all relevant items of all relevant measurement properties are included, contributing to the content validity of the checklist. However, other researchers should do an unbiased judgement on the relevance and comprehensiveness of items, and therefore, a study on the content validity of the COSMIN checklist should be conducted by another group of researchers, for example, a representative sample of users.

Research to improve the quality of systematic reviews on measurement properties

- The scores on the COSMIN checklist cannot yet be combined into an overall score for each measurement property. Within the COSMIN initiative, in cooperation with the Working Group on Clinimetrics of the EMGO institute for Health and Care Research, a rating system is being developed to classify examinations of measurement properties into excellent/good/fair/poor quality. However, it is still unknown which items in particular lead to bias, and which items should therefore be heavily weighted in the overall score. In addition, the measurement properties of this rating system should be investigated, especially inter-rater reliability, and hypotheses testing. The latter may be investigated by comparing the opinion of an expert panel about the quality of selected studies to the scores of the rating system, which are expected to have high agreement.
- In order to be able to determine if the measurement instrument itself is adequate, studies on each of the measurement properties should have good results. Some of the quality indicators for a good measurement instrument are, for example, high intraclass correlation coefficients (ICCs) with small confidence intervals (investigated in the population in which the measurement instrument will be used), a small measurement error, and a clear factor structure. However, what is "high", "small", or "clear" needs to be defined to be able to give a judgement about the quality

of the instrument. Therefore, criteria of adequacy should be developed, preferably consensus-based, and if possible based on empirical evidence for these criteria.

- In systematic reviews of efficacy of treatment, results of different studies are often pooled. Results of several studies on e.g. the internal consistency or reliability of a measurement instrument may also be pooled. Therefore, requirements of homogeneity of data and statistical methods to pool the results should be developed.

Final reflections

For decades, research was performed and articles have been published on the methodology of evaluating measurement properties. For example, in this study we learned a lot by reading the articles of Cronbach written in the Fifties of the previous century[15,16]. Due to much information, and the involvement of many disciplines in this field, one would think that everything was crystal clear. However, this was anything but, as appeared by the substantial discussions among the panel members. Moreover, still new insights are obtained on measurement properties, for example on the subject of minimal important change (MIC). The willingness of researchers to participate in the Delphi study and the reliability study – which was overwhelming – can be interpreted as a need from the field for this type of research.

I would like to finish my dissertation with a call to all researchers who need to select a measurement instrument for their research: be critical when selecting an instrument, and make a well thought decision, taking into account the quality of the measurement instruments, the quality of the studies in which they are investigated, and the clearness of interpreting the scores of the instrument.

Reference List

(1) Linstone HA, Turoff M. (2002). The Delphi Method. Techniques and Applications. Available at http://is.njit.edu/pubs/delphibook/ed. Accessed September 22, 2009.

(2) Powell C. (2003). The Delphi technique: myths and realities. Journal of Advanced Nursing, 41, 376-382.

(3) Keeney S, Hasson F, McKenna HP. (2001). A critical review of the Delphi technique as a research methodology for nursing. International Journal of Nursing Studies, 38, 195-200.

(4) Hasson F, Keeney S, McKenna H. (2000). Research guidelines for the Delphi survey technique. Journal of Advanced Nursing 32, 1008-1015.

(5) Jones J, Hunter D. (1995). Consensus methods for medical and health services research. British Medical Journal, 311, 376-380.

(6) Cortina JM. (1993). What is coefficient alpha? An examination of theory and applications. Journal of Applied Psychology, 78, 98-104.

(7) Williams PL, Webb C. (1994). The Delphi technique: a methodological discussion. Journal of Advanced Nursing, 19, 180-186.

(8) Goodman CM. (1987). The Delphi technique: a critique. Journal of Advanced Nursing, 12, 729-734.

(9) Zegers M, de Bruijne M, Wagner C, Groenewegen P, van der Wal G, de Vet H. (2009). The inter-rater agreement of retrospective assessments of adverse events does not improve with two reviewers per patient record. Journal of Clinical Epidemiology, 63, 94-102.

(10) Scientific Advisory Committee of the Medical Outcomes Trust. (2002). Assessing health status and quality-of-life instruments: Attributes and review criteria. Quality of Life Research, 11, 193-205.

(11) Terwee CB, Bot SD, De Boer MR, et al. (2007). Quality criteria were proposed for measurement properties of health status questionnaires. Journal of Clinical Epidemiology, 60, 34-42.

(12) Terwee CB, Jansma EP, Riphagen II, de Vet HCW. (2009). Development of a methodological PubMed search filter for finding studies on measurement properties of measurement instruments. Quality of Life Research, 18, 1115-1123.

(13) Moher D, Liberati A, Tetzlaff J, Altman DG. (2009). Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. Annals of Internal Medicine, 151, W65-94.

(14) Valderas JM, Ferrer M, Mendivil J, et al. (2008). Development of EMPRO: A tool for the standardized assessment of patient-reported outcome measures. Value in Health, 11, 700-708.

(15) Cronbach LJ. (1951). Coefficient Alpha and the Internal Structure of Tests. Psychometrika, 16, 297-334.

(16) Cronbach LJ, Meehl PE. (1955). Construct validity in psychological tests. Psychological Bulletin, 52, 281-302.