COSMIN reviews: the need to consider measurement theory, modern measurement and a prospective rather than retrospective approach to evaluating patient-based measures

We are grateful for the letter from Mokkink and colleagues concerning our article, “Setting and maintaining standards for patient-reported outcome measures: Can we rely on the COSMIN checklists?” We appreciate their acknowledgment of our purpose and acceptance of many of the points we raised. Rather than reiterate our concerns, we would like to focus on specific issues raised by the letter. We still feel that there are gaps in the COSMIN process. Mokkink et al. point out that they cannot be held responsible for how people apply their recommended procedures and that reviewers should be experts in patient-reported outcome measures (PROMs). This is of course true, but the lack of expertise apparent in many of the COSMIN reviews will always occur, with many unqualified people continuing to undertake reviews, claiming that they apply the COSMIN procedures. The COSMIN standards required to evaluate the PROMs are still vague, leaving reviewers to make subjective decisions. Consequently, readers must check for themselves the articles reviewed to judge their quality.

Mokkink and colleagues report that the COSMIN Risk of Bias checklist uses a worst-score rating per study and argue that it is difficult to meet the stated standards. However, Terwee and colleagues use the example of “Reliability” (Box B) to provide a general description of the scoring system used. It should be noted that reproducibility (reliability) is a crucial statistic for evaluating the quality of instruments. Each issue in the risk of bias boxes is given a score on a four-point rating system running from “excellent” to “poor”. Descriptions of the four ratings follow:

- Items should be scored excellent if the evidence is adequate.
- They should be scored good if relevant information is not reported, but it can be assumed that they are adequate.
- Items should be rated fair if it is doubtful whether they are adequate.
- In some cases, the worst possible response option is limited to good or fair instead of poor because it is not desirable for the issue to have much impact on the instrument’s overall score.

The application of these ratings must be misleading for both reviewers and readers.

Taking a recent “COSMIN” based review at random it is possible to see whether the COSMIN intentions are achieved. However, it is likely that the COSMIN group did not have any direct influence on the quality of the review. Climent-Sanz et al. reviewed instruments designed to assess sleep problems. Five instruments were identified that were suitable for review, although these were only covered by seven publications. The reviewers were hoping to find the best instrument for use with fibromyalgia patients but only one of the five instruments was designed for such a population, while the other four were generic measures of sleep quality. One of the instruments reviewed, the Pittsburgh Sleep Quality Index was reported to have seven subscales but that a total score is generated by adding scores on the subscales together. The Jenkins Sleep Scale (JSS) consists of four items. The Sleep Quality-Numeric Rating Scale is a single item. The Medical Outcomes Study-Sleep Scale was based on questions used since the 1990s that were written for an average population. It is composed of 12 items evaluating six sleep domains. These are added to give a single score. The Fibromyalgia Sleep Diary has eight items.

The researchers concluded that all five instruments had very good quality. All were reported to be valid and reliable. Little information was provided in the review concerning the conceptual models underlying the instruments and virtually no mention was made of a type of scale or construct validity. Some information was provided on reproducibility, but it was confusing, used different methodologies and indicated that reproducibility was poor. Furthermore, no consideration was given to unidimensionality and it was obvious that the authors had no concerns about adding together different constructs to give a total score. The review tells the reader little about the measures and does not provide evidence of their psychometric properties. Consequently, the review would not be of help in selecting an appropriate instrument – though it is clear that all five measures are inadequate in several ways. Unfortunately, several such “COSMIN” related reviews are equally problematic.

Mokkink and colleagues argue that a systematic review will always be restricted to existing instruments and studies, that are predominately developed using Classical Test Theory (CTT). This is true for structured reviews. But improvements in the quality of instrument development will not result from such reviews, especially as poor outcome measures are consistently rated good by systematic reviewers. Surely, it would be better to advocate the development of high-quality PROMS using Rasch Measurement Theory (RMT). If data collected with a measure fit the Rasch model it ensures that the measure is unidimensional and provides interval level measurement. These two qualities are fundamental to good measurement but hardly addressed in the COSMIN checklists. Unfortunately, the development of PROMS using RMT is a rare skill – which explains why there is a reluctance to apply
modern measurement. Where RMT has been used it is uniformly inappropriately applied and peer reviewed[9,10]. We would expect COSMIN to insist that modern measurement techniques are applied[11]. RMT has clearly defined standards that should be met and reported in all articles describing measure development[12].

It is also time for creating measures that meet the requirements of measurement theory. Virtually all PROMs available today are ordinal scales. With such scales, it is not valid to add together item scores to give a total score[13]. Furthermore, it is not possible to calculate means or standard deviations and non-parametric statistical tests must be employed with ordinal data. Consequently, very few PROMs are reliable or valid. The lack of adherence to measurement theory also explains why there are few (if any) examples of PROMs detecting differences between two active interventions in a clinical trial. Such trials now commonly include PROMs, but the trial results generated by these instruments are rarely reported. It is for these reasons that we cannot concur that CTT and RMT provide different kinds of complementary information on the quality of measures.

As most patient-reported outcome measures are clearly outdated and invalid, their review is not the best way forward at this time. First, it is necessary to develop methodologies and practical tools that produce high-quality outcome measurement. We feel that the COSMIN group could lead the way in setting standards for instrument development using modern measurement techniques that meet the requirements of measurement theory.

Transparency

Declaration of funding

No funding was received to produce this article.

Declaration of financial/other relationships

The authors are employees of Galen Research Ltd., which develops patient-reported outcome measures.

Acknowledgements

None reported.

ORCID

Stephen P. McKenna http://orcid.org/0000-0003-4238-8333
Alice Heaney http://orcid.org/0000-0002-4534-6705

References

[1] Mokkink LB, Terwee CB, Bouter LM, et al. Reply to the concerns raised by McKenna and Heaney about COSMIN. J Med Econ. 2021. doi:10.1080/13696998.2021.1948231

[2] McKenna SP, Heaney A. Setting and maintaining standards for patient-reported outcome measures: can we rely on the COSMIN checklists? J Med Econ. 2021;24(1):502–511.

[3] Terwee CB, Mokkink LB, Knol DL, et al. Rating the methodological quality in systematic reviews of studies on measurement properties: a scoring system for the COSMIN checklist. Qual Life Res. 2012;21(4):651–657.

[4] Climent-Sanz C, Marco-Mitjavila A, Pastells-Perf R, et al. Patient reported outcome measures of sleep quality in fibromyalgia: a COSMIN systematic review. Int J Environ Res Public Health. 2020;17(9):2992.

[5] Buyssse DJ, Reynolds CF, Monk TH, et al. The Pittsburgh Sleep Quality Index: a new instrument for psychiatric practice and research. Psychiatry Res. 1989;28(2):193–213.

[6] Jenkins CD, Stanton BA, Niemczyk SJ, et al. A scale for the estimation of sleep problems in clinical research. J Clin Epidemiol. 1988;41(4):313–321.

[7] Martin S, Chandran A, Zografos L, et al. Evaluation of the impact of fibromyalgia on patients’ sleep and the content validity of two sleep scales. Health Qual Life Outcomes. 2009;7:64.

[8] Cappelleri JC, Bushmakin AG, McDermott AM, et al. Measurement properties of the medical outcomes study sleep scale in patients with fibromyalgia. Sleep Med. 2009;10(7):766–770.

[9] Yorke J, Corris P, Gaine S, et al. emPHasis-10: development of a health-related quality of life measure in pulmonary hypertension. Eur Respir J. 2014;43(4):1106–1113.

[10] Mestre TA, Carlozzi NE, Ho AK, et al. Quality of life in Huntington’s disease: critique and recommendations for measures assessing patient health-related quality of life and caregiver quality of life. Mov Disord. 2018;33(5):742–749.

[11] Al Zoubi F, Mayo N, Rochette A, et al. Applying modern measurement approaches to constructs relevant to evidence-based practice among Canadian physical and occupational therapists. Implement Sci. 2018;13(1):152.

[12] Tennant A, Conaghan PG. The Rasch measurement model in rheumatology: what is it and why use it? When should it be applied, and what should one look for in a Rasch paper? Arthritis Rheum. 2007;57(8):1358–1362.

[13] Grimby G, Tennant A, Tesio L. The use of raw scores from ordinal scales: time to end malpractice? J Rehabil Med. 2012;44(2):97–98.