Reviewer's report

Title: Instruments to measure patient experience of health care quality in hospitals: A systematic review

Version: 2  
Date: 5 May 2015

Reviewer: Wieneke Mokkink

Reviewer's report:

In the manuscript “Instruments to measure patient experience of health care quality in hospitals: A systematic review” a review has been described, in which the measurement properties on 11 instruments has been criticized, as well as the ‘cost efficiency’, ‘acceptability’, and educational impact.

I’m aware and appreciate the amount of work that was put into this review. The authors have tried to be clear in what they did. However, I have several questions on their methods and suggestions for improvement.

1. The authors describe that ‘quality of care focusing on measuring experience’ is another construct as ‘quality of care focusing on satisfaction’. It would be nice to read more on their understanding of ‘quality of care focusing on measuring experience’. What kind of experiences are included, what is their definition or framework of the construct they are interested in. is it considered a unidimensional construct? Also, at the result section p 14 the authors describe that instruments cover the similar domains, as well as different domains. It would be nice to read what those similar domains are (in addition to the distinct domains that are described) to get a better understanding of the construct.

2. Internal consistency is about the interrelatedness among the items. To be able to interpret a Cronbach alpha, the instrument should be based on a reflective model, and the (sub)scale should be unidimensional. I think these two issues should be addressed in the review. Internal consistency is not about the structure of the instrument (see also p 16 section ‘reliability’). Furthermore, the measurement properties reliability and agreement or measurement error should not be used as synonyms. For example, a kappa of 0.9 should not be interpreted as % agreement. In Table 5 it should be nice to add a column on measurement error.

3. The measurement property ‘criterion validity’ was included in the ratings. For a construct like ‘patients’ experience about quality of care’ there is likely no reasonable gold standard available, only the longer version of a shortened questionnaire. For the QPP this is also indicated, and a correlation of 0.90 was found. However, the quality rating of the methods was given a ‘poor’ score. For the QPPS also correlations were reported (called criterion validity), and here, the quality of the methods was rated as excellent. This seems not appropriate (and not in line with COSMIN), since in most cases a significant level is not the appropriate way to investigate validity. I’m a bit confused about these ratings.
4. Authors describe two main purposes of use of the instruments (p18), to compare between hospitals (for example in a cross-sectional design), or to investigate improvement or change over time within a hospital. In the latter case, responsiveness of the instruments is also an important measurement property. In could be that no study on responsiveness was found in this review, but I would suggest to reflect on this issue.

5. The authors used five aspects of utility described by Van der Vleuten. In addition to the measurement properties, cost efficiency, acceptability and educational impact are considered as relevant aspects. It is not clear to me how the items used to assess these aspects (Table 2) were developed, and why these items were chosen. Cost efficiency is assessed using four items. I have some questions about these items: How did the authors determine what the number of observations is to ensure reliable data? Did they consider measurement error? What is considered as completion time, especially who’s time? The time needed for a patient to complete an online self-reported questionnaire; the researcher only needs to download all data once in a while? Versus the time needed for an interviewer to complete the interview-based questionnaire, and afterwards import the data? Was this equally weighted? Acceptability was assessed using three items. The first item seems to be an aspect of face or content validity. If there are many missings (second item), it may have consequences on reliability and validity, and this aspect is also taken into account in the COSMIN checklist. The latter item may be an aspect of generalizability. So, are these items relevant?

6. In addition, it would be nice to see which item for each instrument determines the overall rating (Table 5). A (supplementary) table with information on how the decision was made would be nice (for example in line with how the authors report on the measurement properties).

7. I agree with the authors that in different contexts another instrument may be preferred. However, it would strengthen the review if the authors more explicitly recommend one best instrument. For users of this review, their choice will probably also depend on the content of each instrument. In line with my suggestion about the definition, more information on content and subscales of instruments would be nice. (For example, in Table 4 two studies on structural validity were performed for the HKIEQ. One resulted in 17 factors and the other in 18 factors).

Minor issues and typo’s:

8. It would be nice to describe the eligibility criteria more explicitly. For example, which study types, which settings, and which populations were included? Also, some explanation about the choices would be nice to read. For example, I was wondering why studies focusing on nursing care or on maternity care were excluded.

9. The COSMIN checklist only assesses the methodological quality of the
included studies. The quality criteria used to assess the results of those studies (your step 3, Table 1) are not work of the COSMIN initiative. Please delete the name COSMIN in this context.

10. Furthermore, usually when the COSMIN checklist is applied, and the four-point rating system is used, and this is done in one and the same step. I suggest to combine your first two steps in Figure 1. However, a step could be added how the authors would combine scores of studies measuring the same measurement property of the same instrument (after your step 3). For example, internal consistency and structural validity of the HKIEQ was studies in two different studies, how were the results combined?

11. When performing an systematic review, I recommend to perform the whole review procedure by two independent reviewers. In this review, 90% of the article selection was performed by only one reviewer. This should be acknowledged in the discussion. Moreover, it seems that the kappa is interpreted as a measure of agreement, which it is not. It is more informative here to report actual percentages of positive and negative agreement between the two raters.

12. The authors talk about ‘high stakes instruments’, what is this?

13. Suppl 3: in my opinion this table is not relevant to publish. It would be more interesting to read a summary of the reasons of exclusion in Figure 2.

14. P5, second paragraph. First sentence: ‘there is an increasing…’, please add ‘of care’ behind ‘measuring quality’.

15. P7, methods/design paragraph: PRIMSA should be changed into PRISMA.

16. P15 instrument quality and results: ‘Table 4’ should be with a capital letter.

17. P 16 about half way ‘Most results of construct validity…. (see Table one): ‘one’ should be written with the number 1.

Level of interest: An article of importance in its field

Quality of written English: Acceptable

Statistical review: No, the manuscript does not need to be seen by a statistician.

Declaration of competing interests:

I am a member of the COSMIN steering committee, and I receive royalties for the book 'Measurements in Medicine' which I co-authored, and in which a chapter is written on systematic reviews on measurement instruments. I hand these royalties over to the Knowledge Center of Measurement Instruments, VU University Medical center, Amsterdam, the Netherlands.