Dear Professor Moher and Colleagues

Instruments to Measure Patient Experience of Health Care Quality in Hospitals: A Systematic Review

Thank you for considering the above manuscript for publication in Systematic Reviews. We also wish to thank the reviewers and Editor-in-Chief for their time and helpful critique of our work. Collectively, these comments have considerably strengthened the paper.

All changes have been coloured in red font within the manuscript. We have responded to individual reviewers comments in the table below:

Reviewer 1 Comments (Caroline Terwee) Authors’ Response
Major Compulsory Revisions
This is an interesting and well performed systematic review of instruments to measure patient experience of health care quality in hospitals. Some aspects, however, need further clarification.

No action required. We are encouraged by the reviewer’s remark that that the review has been well performed. By addressing the detailed comments we are hopeful the revisions provide the clarification necessary.
It is important to provide a clear definition of the construct of interest of the review. This is needed for evaluating the completeness and appropriateness of the article selection. But also for evaluating the validity of the included instruments for measuring the construct of interest. The construct of interest in this review is patient expectations. I would like to see a clear definition of what exactly is meant by ‘patient expectations’, preferably supported by relevant references in the introduction of the paper. Some guidance is provided on page 5, but a clear definition is lacking.

The construct of interest is the patients’ perspective of hospital quality of care, measured by the patients’ experience. On reflection this was unclear with too much reliance on referring to the published protocol. Two paragraphs have been added defining the current conceptualisation of quality of care and patient experience. The rationale for measuring patient experience as opposed to satisfaction has also been explained (page 5). Additional references have been added to support this content (Beattie et al 2012, Donabedian 1980 and Institute of Medicine 2001).

In the study objectives it is important to clearly state 4 key elements of the review: the construct of interest, the target population of interest, the type of measurement instruments of interest, and the (type of) measurement properties of interest. In the current objective (no 1) the target population and type of measurement instruments of interest are not clearly stated.

Study objectives have been re-written to focus on the key elements of interest; namely construct (patient experience of quality of care), population (general, adult, inpatients), type of measurement instrument (questionnaire) and type of measurement property (validity and reliability studies). An additional objective has been added to explain the need to retain papers pertaining to other aspects of utility (cost efficiency, acceptability and educational impact). Detailed on page 7.

The in- and exclusion criteria should be more clearly described with regard to the same 4 key elements mentioned above.

More detail has been provided on the inclusion and exclusion criteria, which includes the 4 key elements highlighted (page 8). The inclusion selection form was published in the protocol and has now been referenced (Beattie et al 2014).

The quality of the instruments was rated using quality criteria used in previous reviews, such as the review of Schellingerhout et al. These criteria were developed and published by Terwee et al. in 2007 (J Clin Epidemiol 2007:60:34-42). These criteria were not developed by the COSMIN group.

Please refer to the article by Terwee et al. instead of COSMIN or Schellingerhout et al.

Apologies for this important error. All references to the quality criteria used have been amended to reference Terwee et al 200724 instead of COSMIN or Schellingerhout et al. The Schellingerhout et al reference has been removed from the reference list.
5. My most important comment refers to the utility matrix. This part of the data synthesis is not clear to me. It is not clear how the results of COSMIN ratings for the quality of the studies and the results of the studies were combined to rate the utility of the instrument. For example, what if you have one study of excellent quality with a positive result and one study of fair quality and a positive result. How is this taken into account in the utility rating? Another example, what if you have one study with positive findings and another study with negative findings (both of same or different quality)? Etc. It is also unclear how some of the items are rated. For example, how was the number of assessments needed to ensure reliable data determined? The description of the utility matrix should be completely transparent so that the review could be reproduced. Currently, this would not be possible. I also suggest to look at other systematic reviews that used COSMIN (such as the review of Schellingerhout) who used a level of evidence approach, based on Cochrane guidelines. This could be considered as an alternative way of combining the results from different studies, taking the methodological quality and the (consistency of) the results into account.

Many thanks for your constructive feedback on the matrix. We synthesized the quality of results by presenting the average result. For example, if two studies of the same measurement property had fair and excellent results we presented these as good. However, we previously made a ‘judgement’ in situations where there was no average point i.e. if one study was good and another excellent we came to a consensus agreement whether this should be good or excellent overall. We now realize the limitation of this decision making, particularly if others were to replicate the study. We now propose that where study results differ and the average point would be in between two categories these are presented as in between both ranks. For example, if one study is good and the other is excellent we present the combined result as good to excellent (***/***). We have explained this on page 14.

The number of assessments required for a reliable sample was extracted from retained papers. For example, the HCAHPS instrument requires a minimum of 300 questionnaires per hospital to achieve a minimum of 0.8 reliability for all reported measures (Giordano et al 200925).

We have added a fuller description of how the additional aspects of utility were rated (see page 11-12). We have also added a table to show our responses to individual questions per instrument to make our decision making more explicit (see Table 6: Results of Additional Aspects of Utility).

The part of the matrix which rated the importance of aspects of utility has been removed. Whist useful this requires further work and explanation out with the primary purpose of this review.

6. I wondered why the Consumer Quality Index (which is often used in the Netherlands) was not included in this review? It seems to meet the inclusion criteria, but maybe not?
We did identify the Consumer Quality Index (CQI) and multiple associated papers during our search. Although a valuable instrument, the CQI was rejected as it is a measure of collaboration between general practitioners and medical specialists, therefore not meeting our inclusion criteria. Our review was interested in identifying a global patient experience measure of quality of hospital care.

7. It is stated that only three instruments had tested criterion validity. I wonder whether these studies really addressed criterion validity because I don’t think a gold standard exists for patient experiences and patient-reported outcomes in general. Maybe the authors of the included papers called it criterion validity, but according to the COSMIN definitions the studies should be considered studies on construct validity?

During our critique we had much debate over the criterion validity category. Whilst we agree that there is no gold standard measure of patient experience we categorized these three studies as criterion validity as they were all studies comparing shortened versions with their original, longer version. We used the explanation and instructions in the COSMIN Manual (2012, p38) “The criterion used should be considered as a reasonable ‘gold standard.’ The COSMIN panel reached consensus that no gold standard exists for HR-PRO instruments. The only exception to this is when a shortened instrument is compared to the original long version. In that case, the original long version can be considered the gold standard.” We have clarified this in text at page 17/18.

8. The study on the HCAHPS was given a poor score for methodological quality because structural equation modelling was used for assessing structural validity and the authors didn’t know how to deal with this in the COSMIN box for structural validity (they also made a remark about this in the discussion). However, the authors may not be aware that structural equation modelling is a form of confirmatory factor analyses and therefore it can be scored in the COSMIN box for structural validity.

We did apply Box E using COSMIN for the HCAHPS study (Keller et al 2005) for structural equation modelling and rated this as excellent in terms of methodological quality (see table 5).

We ran into difficulties when applying structural equation modelling to the Quality Criteria for Results as the quality criteria set out for structural validity was:

Positive: Factors should explain at least 50% of the variance
Indeterminate: Explained variance not mentioned
Negative: Factors explain < 50% of the variance

Variance was not mentioned in the structural equation modelling therefore the results were categorised as indeterminate (?).

This finding was incorrectly referred to as COSMIN methodological criteria as
opposed to results criteria by Terwee et al 2007. This had been corrected on page 18 and throughout the paper.

9. In the discussion the authors state that some studies may have been rated as poor or fair because only one item in the COSMIN box was rated as poor. This is indeed how the ‘worst score counts’ method works. However, one should keep in mind that the COSMIN 4-point rating scale was developed in such a way that only a fatal flaw in the methodology of the study is rated as poor (that is also why some items cannot be scores as poor because we don’t consider them fatal). We argued that fatal flaws cannot be compensated by the quality of other methodological issues. Please see the article by Terwee et al. (Qual Life Res 2012;21:651-657) for an explanation of the ‘worst score counts’ method.

We did apply the COSMIN criteria, using the worst score counts analogy, but had not explained the design feature to compensate for minor discrepancies as opposed to fatal flaws. A paragraph has been added in the discussion section to explain this feature (see page 23).

Minor Essential Flaws

10. It is stated that 10% of the records were screened by a second reviewer. Where these abstracts of full-text articles? Please specify.

Where decisions could not be made on title and abstract alone, full papers were retrieved (n=110). For the 10% sample checked by the second reviewer 17 out of 100 required full text review to apply the inclusion criteria. We have clarified this by adding a sentence on page 9 and page 14 (under Screening Results).

11. In the results section it is stated that instruments were found from the Netherlands but I cannot find them in Table 3.

This was an error and has now been corrected as Scandinavia (page 15).

Discretionary Revisions

12. I was wondering why a search in was not performed. In my experience, often additional relevant articles are found in EMBASE.

Pragmatic and cost limitations prevented us from completing an EMBASE search, as our University does not subscribe to the EMBASE platform. We did conduct a thorough search of MEDLINE, CINHAL and Psych Info, as well as specialist databases in the field of patient experience. Despite checking secondary references we found no other instruments meeting our inclusion criteria. This statement has now been acknowledged as a limitation in the discussion section (page 22/23).

13. I completely agree with the authors that by using standard criteria some instruments will be rated negatively because they fall just below the criteria. This does not mean that the instrument is poor or cannot be used, but perhaps it should be regarded as a warning. I nevertheless believe it is helpful to use standard criteria to make reviews and their conclusions transparent.

The relevant paragraph has been re-written to provide a more balanced
discussion of the strengths and limitations of using standardised criteria (see page 23).

Reviewer 2 Comments (Wieneke Mokkink)

Major Compulsory Revisions

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.

No action required. We appreciate the reviewer’s acknowledgement of the work that has gone into this review and their helpful 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.

Content has been added to provide an understanding on the construct of interest in relation to patient experience of quality of care, as opposed to satisfaction (see page 5). We have explained that there is no standard definition of quality in relation to health care, rather quality is conceptualised as multi-dimensional with definitions varying across stakeholders. We have added a description of the most commonly used dimensions of quality in the background section. This is a vast topic area, hence the word limit prevents fuller discussion so we have referred to our earlier work on defining quality in relation to health care (Beattie el at 2012). Also, we did not want to set limitations on how quality was defined to enable a wide capture of relevant instruments, therefore we included all definitions or conceptualisations of quality if they had been devised from the patients' perspective. We have added another sentence into the results of domains covered in Table 4 (page 16).

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.
We agreed that internal consistency is about the relationship between items and the construct of interest. We have explained this more fully under the section on development of the matrix (see page 13). We have also added detail under results on page 16 and included the findings of whether instruments were derived from a reflective or formative model in Table 4.

We agree that Kappa is not a measure of simple agreement. To avoid any confusion we have removed Kappa and report the percentage only in terms of level of agreement between reviewers (page 15).

We have a measurement error column within the matrix (Table 7), but have now more fully explained this in the text (page 13).

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.

During our critique we had much debate over the criterion validity category. Whilst we agree that there is no gold standard measure of patient experience we categorized three studies as criterion validity as they were all studies comparing shortened versions with their original, longer version. We used the explanation and instructions in the COSMIN Manual (2012, p38) “The criterion used should be considered as a reasonable ‘gold standard.’ The COSMIN panel reached consensus that no gold standard exists for HR-PRO instruments. The only exception to this is when a shortened instrument is compared to the original long version. In that case, the original long version can be considered the gold standard.” We have clarified this under ‘Instrument Quality and Results’ page 17/18.

We did present QPP result as negative instead of a positive result in error. We have amended this important mistake. The quality of the results is positive using Quality Criteria for Measurement Properties (Terwee et al 200724). The quality of the methods, however was rated as poor using the COSMIN criteria.

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. It could be that no study on responsiveness was found in this review, but I would suggest to reflect on this issue.

Many thanks for highlighting this important gap. We have now added the
We found no studies assessing the ability of an instrument to detect change over time in the construct to be measured, otherwise known as responsiveness. This was surprising given that the purpose of the patient experience instruments was to measure hospital care quality for national comparisons and/or evaluation of local improvement work. This review highlights the necessity and current gap in studies assessing responsiveness of these instruments.

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 generalisability. So, are these items relevant?

We considered Van der Vleuten’s five aspects of utility helpful in providing an overview for instrument selection. There are overlaps between issues raised in the additional dimensions with validity and reliability criteria, such as number of items needed in terms of reliability, as well as cost. However, we think it is important to consider these issues from alternative perspectives to aid decision making on instrument choice. For example, Cronbach’s alpha can be improved by increasing the number of questionnaire items, but this needs to be balanced with a consideration of the time is takes users to complete the questionnaire (page 6 for background and 11 for description of questions).

The number of assessments required for a reliable sample was extracted from retained papers. For example, the HCAHPS instrument requires a minimum of 300 questionnaires per hospital to achieve a minimum of 0.8 reliability for all reported measures (Giordano et al 200925).

Further detail has been added to explain how the additional aspects of utility were rated (see page 11).

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).
A table has been added, as suggested, to show responses to individual 
questions on the additional dimensions of utility, per instrument. This makes our 
decision making more explicit (see Table 6: Results of Additional Aspects of 
Utility).

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).

One of our key conclusions is that instrument selection requires a balanced 
consideration of all five aspects of utility, largely determined by the purpose of 
the measurement. We, therefore do not think it would be possible, or helpful to 
recommend one best instrument. Alternatively, we have suggested the ‘best’ 
instrument in some circumstances i.e. HCAHPS being the instrument of choice 
for high stakes purposes i.e. hospital league tables associated with financial 
incentives. We also suggest instruments of choice when the brevity of an 
instrument is the priority i.e. PPE-15 to act as an initial screening survey (this is 
discussed on page 22).

Specific content and domains appear to be ever evolving in patient experience 
instruments. We considered it to be helpful to provide readers with an overview 
of the domains covered by each instrument in Table 4. As you have highlighted, 
the ongoing testing of validity continually refines factors specific to each 
instrument.

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/maternity care were excluded.

More detail has been provided on the inclusion and exclusion criteria (page 8). 
Our first submission relied too heavily on reference to our published protocol, 
where the inclusion and exclusion criteria had been detailed. We agree that 
adding this information makes the paper more complete.

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.

All references to the quality criteria used have been amended to reference
Terwee et al 200724 instead of COSMIN or Schellingerhout et al. The Schellingerhout et al reference has been removed from the reference list.

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 studied in two different studies, how were the results combined?

As suggested we have combined our original step 1 and 2 to report application and scoring of COSMIN checklists as one step. We have now added step 4 as synthesis of results of validity and reliability and amended Figure 1 accordingly. We have added the paragraph below to explain how we combined results (page 14).

We presented ratings of study quality as star ratings; excellent (****), good (**), fair (*) and poor * and the quality of results as positive (+), negative (-) and indiscriminate (?). Where more than one study, from the same measurement category had been conducted, we determined the average point to rate the quality of the methods. For example, if structural validity scored validity scored ‘excellent’ and cross cultural validity scored ‘fair’ our overall rating would be ‘good’. If however, structural validity scored ‘excellent’ and cross cultural validity scored ‘good’ we would rate validity overall as good to excellent (represented as ***/*). Where the quality of study results varied, within the same measurement property, we presented these as mixed. For example, if structural validity results scored positive and cross cultural validity scored negative we presented these as mixed (+/-).

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.

As identified, only 10% of the application of the inclusion criteria was independently completed by a second reviewer. This was purely a resource decision as the systematic review was an unfunded piece of work conducted as part of Michelle Beattie’s PhD by publication. All other decision making steps were conducted by two reviewers independently i.e. application of COSMIN and Terwee et al results criteria. We have added a sentence on this limitation within the discussion (page 23).

We agree Kappa is beyond a measure of simple agreement. To avoid confusion we have removed the Kappa figure and retained percentage only (page 15).

12. The authors talk about ‘high stakes instruments’, what is this?
We define high stakes as a situation where the outcome of the results has important consequences for the individual or organisation, hence requiring a measurement instrument with high reliability. For example, results from patient experience instruments are used for league tables in the US and low results result in financial penalties. In this situation the relative importance of the five aspects of utility would heavily favour reliability, whilst tolerating high cost, in terms of resource. Whereas an instrument being used to measure team performance in improvement activity may tolerate lower levels of reliability to focus on educational impact and cost (see page 6).

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. As suggested, we have removed supplement 3 from the submission and provided a Summary of Exclusion Results (Table 3).

14. P5, second paragraph. First sentence: ‘there is an increasing…’, please add ‘of care’ behind ‘measuring quality’.
This paragraph and sentence was re-written, however ‘of care’ has been added to the 3rd sentence, 2nd paragraph on page 5.

15. P7, methods/design paragraph: PRIMSA should be changed into PRISMA. Typographical error amended accordingly.

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

17. P16 about half way ‘Most results of construct validity…. (see Table one): ‘one’ should be written with the number 1. Error corrected as suggested.

We trust that these revisions address the reviewers’ concerns with our paper, and thank you in advance for re-considering this manuscript for publication in Systematic Reviews.

We look forward to hearing your decision.

Yours sincerely

Michelle Beattie
Lecturer