Answers

1. Evaluate the methodological quality of a study with the COSMIN checklist

We follow the four steps as presented in Table 9.2.

Step 1: The following measurement properties are evaluated in the study: internal consistency (page 1063), structural validity (confirmatory factor analysis, page 1063), and construct validity (hypotheses testing, page 1063). The memorial anxiety scale for prostate cancer (MAX-PC questionnaire) was translated into Dutch (page 1062). Therefore, items 4-11 of the COSMIN box cross-cultural validity can be completed to evaluate the quality of the translation. However, the measurement property cross-cultural validity was not evaluated. To do this, a more detailed comparison between e.g. the original language version and the Dutch language version should have been performed (see Section 6.5.3.3 about measurement invariance). Therefore, the other items in the box cross-cultural validity cannot be completed.

Step 2: IRT methods were not used, therefore, the IRT box need not to be completed.

Step 3: Evaluation of the methodological quality of the study on the properties identified in step 1.

<u>Box internal consistency</u>

Box A. Internal consistency

		yes	**no**	**?**
1	Does the scale consist of effect indicators, i.e. is it based on a reflective model?	☑	□	□
Design requirements		**yes**	**no**	**?**
2	Was the percentage of missing items given?	☑	□	
3	Was there a description of how missing items were handled?	☑	□	
4	Was the sample size included in the internal consistency analysis adequate?	☑	□	□

5	Was the unidimensionality of the scale checked? i.e. was factor analysis or IRT model applied?	☑	□	
6	Was the sample size included in the unidimensionality analysis adequate?	☑	□	□
7	Was an internal consistency statistic calculated for each (unidimensional) (sub)scale separately?	☑	□	□
8	Were there other important flaws in the design or methods of the study?	□	☑	
Statistical methods		**yes**	**no**	**NA**
9	for Classical Test Theory (CTT): Was Cronbach's alpha calculated?	☑	□	□
10	for dichotomous scores: Was Cronbach's alpha or KR-20 calculated?	□	□	☑
11	for IRT: Was a goodness of fit statistic at a global level calculated? e.g. χ^2, reliability coefficient of estimated latent trait value (index of (subject or item) separation)	□	□	☑

A1. It was not formally described whether the questionnaire was based on a formative or reflective model. However, when you read the items of the MAX-PC (some examples are given in the Method sections page 1062), you may conclude that it was based on a reflective model. When the anxiety related to prostate cancer would increase, it will be likely that one agrees with all of the statements.

A2. Results section (page 1063) first paragraph: "all 129 men completed all 18 MAX-PC items". The people who completed the questionnaire did not miss any item.

A3. There were no missing items on the MAX-PC. Therefore, we answer 'yes' here. Not applicable would be most suitable, but that option is not offered here.

A4. n=129. We consider a sample size of ≥50 persons as adequate.

A5. Unidimensionality was checked with a factor analysis in this article.

A6. n=129. The scale has 18 items. For an adequate sample size, we recommend to have 7 x the number of items AND at least 100 persons. For this scale this means: 7 x 18 = 126.

A7. For each of the three sub-scales Cronbach's alpha coefficients were given (page 1063 under results). In addition, a Cronbach's alpha for the total score was given. This alpha is difficult to interpret, because it is not based on a unidimensional scale.

A8. There were no other important flaws in this study.

A9-A10. The score is considered to be a continuous score and no IRT was used, therefore, Cronbach's alpha is the preferred statistic, and the other two items (A10, A11) were not applicable.

Box E Structural validity

Box E. Structural validity				
		yes	**no**	**?**
1	Does the scale consist of effect indicators, i.e. is it based on a reflective model?	☑	□	□
Design requirements		**yes**	**no**	**?**
2	Was the percentage of missing items given?	☑	□	
3	Was there a description of how missing items were handled?	☑	□	
4	Was the sample size included in the analysis adequate?	☑	□	□
5	Were there other important flaws in the design or methods of the study?	□	☑	
Statistical methods		**yes**	**No**	**NA**
6	for CTT: Was exploratory or confirmatory factor analysis performed?	☑	□	□
7	for IRT: Were IRT tests for determining the (uni-) dimensionality of the items performed?	□	□	☑

E1. It was not formally described whether the questionnaire was based on a formative or reflective model. However, when you read the items of the MAX-PC (some examples are given in the Method sections page

1062), you may conclude that it was based on a reflective model. When the anxiety related to prostate cancer would increase, it will be likely that one agrees with all of the statements.

E2. Results section (page 1063) first paragraph: “all 129 men completed all 18 MAX-PC items”. The people who completed the questionnaire did not miss any item.

E3. There were no missing items on the MAX-PC. Therefore, we answer ‘yes’ here. Not applicable would be most suitable, but that option is not offered here.

E4. n=129. The scale has 18 items. For an adequate sample size, we recommend to have 7 x the number of items AND at least 100 persons. For this scale this means: 7 x 18 = 126.

E5. There were no other important flaws in this study.

E6-E7. The study was not based on IRT methods, therefore, an exploratory or confirmatory factor analyses (CFA) should be performed. There was information available about the structure of the questionnaire, i.e. at page 1063 third/fourth paragraph, the authors wrote: “an initial CFA model was fitted in which each item was assigned to one of the three underlying factors, similar to the original publication, to verify that the original factor structure was present in our data”. In this case a CFA is indeed preferred over the exploratory factor analysis.

Box Hypotheses testing

Box F. Hypotheses testing

Design requirements		**yes**	**no**	**?**
1	Was the percentage of missing items given?	☑	□	
2	Was there a description of how missing items were handled?	☑	□	
3	Was the sample size included in the analysis adequate?	☑	□	□
4	Were hypotheses regarding correlations or mean differences formulated a priori (i.e. before data collection)?	☑	□	□*
		yes	**no**	**NA**

5	Was the expected *direction* of correlations or mean differences included in the hypotheses?	☑	□	□
6	Was the expected absolute or relative *magnitude* of correlations or mean differences included in the hypotheses?	☑	□	□
7	for convergent validity: Was an adequate description provided of the comparator instrument(s)?	☑	□	
8	for convergent validity: Were the measurement properties of the comparator instrument(s) adequately described?	□	☑	
9	Were there other important flaws in the design or methods of the study?	□	☑	
Statistical methods		**yes**	**no**	**NA**
10	Were design and statistical methods adequate for the hypotheses to be tested?	☑	□	□

F1. Results section (page 1063) first paragraph: “All 129 men completed all 18 MAX-PC items. In 8 men, the Decisional Conflict Scale (DCS), CES-D, STAI-6 or SF-36 score were discarded, as one or more items of one of these scales were missing.”.

F2. Scores of people with one or more missings on the additional scales were discarded. In Table 2, however, scores and distributions of all questionnaires were given for 129 persons. It is likely that the results of the additional scales in this Table are based on only 121 men.

F3. Analyses based on more than 50 persons are considered to be adequate.

F4. In the section on Quality criteria (page 1063) it is written that “Correlations with r >0.3 were considered relevant. We hypothesized that higher scores on the total MAX-PC and sub-scales have to be related to higher scores on DCS, CES-D, and STAI-6, and to lower SF-12 MCS scores; that correlations were highest with STAI-6, as this measure also is anxiety specific, and lower with DCS, CES-D, and SF-12 MCS; and that correlations with the ‘PSA anxiety’ sub-scale were lower, because this is a very specific sub-scale, previously found to show lower construct validity. We tested differences between correlations with MAX-PC total for significance with a bootstrap procedure to obtain standard errors [18]. At least 75% of the results should be in accordance with a priori hypotheses.”

F5/F6. The authors considered a correlation above 0.3 as relevant. As this statement contains the direction and a magnitude, we assign a positive score in the COSMIN list. However, the hypotheses are a bit vague and unchallenging. The authors expect positive correlations (F5), but it is difficult to interpret what is meant by ‘relevant above 0.3’. Do they expect all (sub-)scales to correlate at least 0.3 even if scales do not aim to measure the same construct? (Sub-)scales that aim to measure the same construct could be expected to correlate higher with the MAX-PC subscales, and scales that aim to measure different constructs could be expected to correlate less with the MAX-PC subscales. A correlation below 0.3 can therefore be very relevant. For example, the subscale *prostate-specific antigen (PSA) anxiety* is likely to measure a different construct as the *Decisional Conflict Scale* that measures decisional conflict on the choice for active surveillance. When a correlation below 0.3 is found, this is relevant.

Words like ‘higher’ and ‘lower’ do indicate some guidance about the direction of the expected correlation, but they do not tell anything about the expected magnitude of the correlation. Therefore, the hypotheses could have been more explicit. For example, the subscales *prostate cancer anxiety* and *fear of recurrence* may measure a more general construct than the subscale *PSA anxiety*. Therefore, one could, for example, hypothesize that the expected correlation between *PSA anxiety* and the STAI-6 (which measures generic anxiety) is below 0.30 and the expected correlation between *prostate cancer anxiety* and *fear of recurrence* respectively and the STAI-6, is between 0.30 and 0.60. Note that the COSMIN checklist does not require to judge the adequacy of the hypotheses.

F7. The comparator instruments were the Decisional Conflict Scale (DCS), which construct was described as “decisional conflict on the choice for active surveillance”; the Centre for Epidemiologic Studies Depression scale (CES-D), which measures depression (this was not further defined), the abridged State Trait Anxiety Inventory (STAI-6), which measures generic anxiety (also not further defined), and the Mental Component Summary (MCS) Score of the Short Form health survey 12 (SF-12), which construct was not further defined. Although the constructs of the comparator instruments were mostly described, the information provided was very brief.

F8. None of the measurement properties of any of the instruments are described.

F9. There were no other important flaws in this study.

F10. Pearson correlations coefficients are calculated (described in Table 4), which are adequate methods to compare continuous scores when data are reasonable normally distributed.

Translation

Box G. Cross-cultural validity

	Design requirements	**yes**	**no**	**?**
1	Was the percentage of missing items given?	□	□	
2	Was there a description of how missing items were handled?	□	□	
3	Was the sample size included in the analysis adequate?	□	□	□
4	Were both the original language in which the HR-PRO instrument was developed, and the language in which the HR-PRO instrument was translated described?	☑	□	
5	Was the expertise of the people involved in the translation process adequately described? e.g. expertise in the disease(s) involved, expertise in the construct to be measured, expertise in both languages	☑	□	
6	Did the translators work independently from each other?	☑	□	□
7	Were items translated forward and backward?	☑	□	□
8	Was there an adequate description of how differences between the original and translated versions were resolved?	☑	□	
9	Was the translation reviewed by a committee (e.g. original developers)?	□	☑	

10	Was the HR-PRO instrument pre-tested (e.g. cognitive interviews) to check interpretation, cultural relevance of the translation, and ease of comprehension?	☑	□	
11	Was the sample used in the pre-test adequately described?	□	☑	
12	Were the samples similar for all characteristics except language and/or cultural background?	□	□	□
13	Were there any important flaws in the design or methods of the study?	□	□	
Statistical methods		**yes**	**no**	**NA**
14	for CTT: Was confirmatory factor analysis performed?	□	□	□
15	for IRT: Was differential item function (DIF) between language groups assessed?	□	□	□

Cross-cultural validity was not formally tested, because for cross-cultural validation data from two similar populations are needed; one population completes the original version of the questionnaire, and the other population completes the new cross-culturally adapted version (see Chapter 6.5.3). In this study a description of the translation process was provided. Therefore, G4 to G11 can be completed to rate the quality of the translation process.

G4. The MAX-PX was developed in the USA. In this study it was translated into Dutch.

G5. The people who did the forward translation were all fluent in English and had a medical background. A native English speaker fluent in Dutch, also with a medical background did the back translation. We assume that people with a medical background have some knowledge of the construct anxiety.

G6. It is not explicitly described whether all translators worked independently. But it is highly likely that they did, since the three translated versions were pooled together after a consensus meeting. In addition, they state that the guidelines by Guillemin were used, which recommend to undertake the translations at least by two independent translators.

G7. There were three forward and one backward translation.

G8. It is stated that the back translation “showed some discrepancies with the source document, but these were mainly related to the wording and not to the specific meaning of the items. Consensus was reached by discussion.”

G9. There was no formal committee who reviewed the translation.

G10-G11. The questionnaire was tested face-to-face with 5 participants, using the thinking aloud method. Unfortunately, no description was given about the characteristics of these participants.

Step 4: assess the generalisability of the results for the properties identified in step 1

When completing the COSMIN checklist, we recommend to complete the Generalisability box for each evaluated measurement property. In this paper the same sample is used for the evaluation of each measurement property. Therefore, we complete the box only once. When different samples are used for the evaluation of different measurement properties, the box should be completed several times, i.e. for each sample.

Box Generalisability box

		yes	**no**	**NA**
	Was the sample in which the HR-PRO instrument was evaluated adequately described? In terms of:			
1	median or mean age (with standard deviation or range)?	☑	□	
2	distribution of sex?	☑	□	
3	important disease characteristics (e.g. severity, status, duration) and description of treatment?	☑	□	□
4	setting(s) in which the study was conducted? e.g. general population, primary care or hospital/rehabilitation care	☑	□	

	5	countries in which the study was conducted?	☑	□	
	6	language in which the HR-PRO instrument was evaluated?	☑	□	
7	Was the method used to select patients adequately described? e.g. convenience, consecutive, or random		☑	□	
			yes	**no**	**?**
8	Was the percentage of missing responses (response rate) acceptable?		☑	□	□

1-4. Described in Table 1, page 1063.

5-6. The Dutch version of the questionnaire was evaluated by the Erasmus University Medical Centre in Rotterdam, the Netherlands (Method section *Patients* page 1062).

7. If diagnosed, men received a questionnaire (Method section *Patients* page 1062). Therefore, we assume that people were consecutively included.

8. 150 questionnaires were sent, and 129 were completed (86%) (Results section page 1063), we consider this acceptable.

2. Data synthesis of 8 studies on the reliability of one instrument

a. The studies vary widely in their characteristics. Study 7 included patients with acute low back pain, while the other studies included patients with chronic low back pain. Study 3 recruited patients from home retirement communities, study 1 recruited patients from general practice, and the other studies recruited patients from physiotherapy or outpatient clinics. The patients in study 3 are on average much older than the patients in the other studies and in study 5 the number of male patients was very small as compared to the other studies. Finally, three studies were performed in English-speaking patients, while the languages of the patients in the other studies were all different. It is difficult to imagine what possible effect these differences would have on the reliability. Therefore, we decide to consider the results of all studies together in the next steps of the analysis, but keep these differences in mind. We recommend to elaborate on these differences in the discussion of the review.

b. There are also quite some differences in the methodological quality of the studies. Two studies are of good methodological quality, four studies are of fair quality and two studies are of poor quality. Study 6 was rated as having poor quality because the time interval was too long (42 days). Such a long time interval can reduce the reliability. Study 7 was rated as having poor quality because of the small sample size (23 patients).

How should be dealt with these differences in methodological quality?

If you decide to statistically pool the results of the data, we recommend to include only those studies of at least fair methodological quality. However, when you perform statistical pooling, sample size is no longer an issue. If we would not consider the sample size, the quality of study 7 would be fair. Therefore, for statistical pooling only study 6 would be excluded. When applying levels of evidence (Table 9.18, see question d) all studies are considered, but as you will notice, the poor quality studies do not contribute to the evidence.

c. The results of the measurement properties are fairly consistent. All ICCs are above 0.70, except for the ICC found in study 7. Two remarks can be made about study 7: first, this study was rated as having poor quality because of the low sample size. This is reflected in the large confidence interval (0.20-0.78). Therefore, this study does not contribute to the evidence. Second, this was the only study which included acute low back pain patients. Perhaps the QBPDS is not reliable in acute patients. This could also be a point of discussion in the review. A recommendation could be made for a high quality reliability study in acute low back pain patients.

d. As the ICCs of the remaining 7 studies are consistently high and there are multiple studies of good methodological quality, a strong level of evidence should be applied.

e. Our overall conclusion would be that there is strong evidence that the reliability of the QBPDS is good (above 0.70) in patients with chronic low back pain. In addition we can conclude that the evidence on the reliability of the QBPDS in acute low back pain patients is (yet) unknown.