Author’s response to reviews

Title: Adaptation and psychometric properties of the Norwegian version of the Heart Continuity of Care Questionnaire (HCCQ)

Authors:

Irene Valaker (irene.valaker@hvl.no)

Bengt Fridlund (beppe.fridlund@gmail.com)

Tore Wentzel-Larsen (tore.wentzellarsen@gmail.com)

Heather Hadjistavropoulos (heather.hadjistavropoulos@uregina.ca)

Jan Erik Nordrehaug (jan.erik.nordrehaug@sus.no)

Svein Rotevatn (svein.rotevatn@helse-bergen.no)

Maj-Britt Råholm (maj-britt.raholm@hvl.no)

Tone M. Norekvål (tone.merete.norekval@helse-bergen.no)

Version: 1 Date: 31 Jan 2019

Author’s response to reviews:

BMRM-D-18-00507

Adaptation and psychometric properties of the Norwegian version of the Heart Continuity of Care Questionnaire (HCCQ)

Authors’ responses to referees’ comments

Dear editor-in-chief,

We greatly appreciate the comments from you and the two reviewers on our manuscript. We have made several changes in order to improve the manuscript. Below, please find the answers to the reviewers and a detailed description of the changes. All page numbers refer to locations in the revised manuscript. The text marked in red indicates the changes made.

Editor Comments:

1. Please include the role of the funder in your declaration section.
2. Please include the figure legend after the reference list.

3. Please amend the statement on "Availability of data and material". You clearly have data, so it is applicable. It is important that you state whether the data will be made available or not. Many people choose one of the following:

- Data is available from corresponding author upon reasonable request.

- Data has been made freely available at XYZ repository.

- All data is contained within this manuscript.

- Data can not be made available because of... (e.g. patient confidentiality reasons).

Authors’ response:

1. We have added the following in the declaration section: “The funding bodies had no role in the design of the study, data collection, analyses, interpretation of data or in writing the manuscript”.

2. We have included the figure legend after the reference list.

3. Data can not be made available due to patient confidentiality reasons.

Reviewer 1 (Walter Sermeus)

1. Referee’s comment:

There seems to be a problem with Q14 and Q15. I see it well presented in the analysis and in the CFA in figure 2. It is unclear however why this is and what the recommendation is to included in the Norwegian version. It would be recommendable to rerun the CFA without the 2 variables to see if the psychometric properties are improving.

Authors’ response:

The purpose of this article was to determine if the instrument, as recommended by the developers, was suitable in a Norwegian setting. However, we have run SEM without Q14 and Q15, which loaded on two factors, from the material from Canada as an additional analysis. We have revised the text (Page 15, line 349-350) (Page 18, line 407-410). The additional analysis, in which item 14 and 15 was removed did not improve the fit appreciably: Chi-square ($\chi^2$)/ Degree of freedom (df) =$2712/431$; RMSEA = 0.11; CFI =0.91 and TLI = 0.90. Standardized factor loadings ranged from 0.58 to 0.97 (p < 0.001).
2. Referee’s comment:


Authors’ response:

Thank you for emphasizing that we should have reported more accurately how we dealt with cross-cultural equivalency. We used the five levels of cross-cultural equivalence of Polit and Yang (2016): conceptual, content, semantic, technical equivalence and measurement equivalence. We have explained this in the manuscript (Page 8, line 159-170) (Page 9, line 199-200) (Page 10, line 222). When the HCCQ was selected, the conceptual and technical equivalence was discussed in the expert group and with the developer of the instrument. Familiarity with prior translation work and culture was useful. Our previous qualitative study, Valaker et al. (2017), showed that the construct being measured was meaningful in the Norwegian context. The back translation was essential for semantic equivalence to make sure that the translation reflected the same item content as the original. Additionally, the expert group discussed, on an item-by-item basis, the adequacy of the target language translation. After adaptation, a pilot test was performed and patients judged comprehensibility, relevance and completeness. Once the translated version was finalized, measurement equivalence was evaluated. In particular, the authors compared sample means and standard deviations on the two versions and the scale’s internal consistency. Additionally, CFA was used to further evaluate the measurement equivalence of the adapted and original measure. However, other analyses could have been relevant to perform if we had access to comparative data from the original instrument. We have underlined this in the limitation of the study (Page 26, line 612-615). This study has mainly focused on cross-cultural approaches for translating instruments. Therefore, more research is needed to compare the HCCQ across healthcare systems and organizations. However, we have added sentences on these matters in the discussion chapter (Page 21, line 486-493).

3. Referee’s comment:

An important concept in the cross-cultural comparability is measurement invariance evaluation: can we compare the Norwegian data with the original data. A difference is made in Configural invariance (identical factor structures across groups), Metric invariance (equality of factor loadings across groups) and Scalar invariance (equality in factor scores and intercepts). See also Bruyneel L et al., Bayesian Multilevel MIMIC Modeling for Studying Measurement Invariance in Cross-group Comparisons. Med Care. 2017 Apr;55(4):e25-e35.

Please report what level of measurement invariance is obtained and if cross-cultural comparability is obtained.

Authors’ response:
We appreciate the important input from the reviewer. According to Polit and Yang (2016) five types of factorial invariance can be assessed. Dimensional invariance requires that a multi-item instrument represents the same number for latent traits in the original and adapted instruments. Next is configural invariance, which concerns whether the same items are configured on the same factors in the original and adapted version. We have tested dimensional and configural invariance using data from one sample i.e., the sample for field testing the adapted/translated scale. The goal was to assess whether the CFA model for the adapted instrument can replicate the model supported in previous explorative factor analysis of the instrument. We did not have the original Canadian data available to evaluate metric invariance, scale and factorial invariance. Metric invariance requires the comparability of factor loadings of items onto common factors across groups. Scale invariance requires equality of the item intercepts and factorial equivalence concerns equality of the residuals. Such tests require raw data from both groups being compared to perform statistical tests. The study from Canada was published in 2004 and according to the instrument developer the original data have been destroyed as per ethical requirements. However, this kind of analysis is a good proposal. We have revised the text to inform about this in the analysis chapter (Page 15, line 328-332) as well as in the limitation chapter (Page 26, line 615-617).

Reviewer 2 (Awat Feizi)

4. Referee’s comment:

Introduction: please update the literature reviews in the introduction section. Please add some justification on why a tool should be adopted into another language.

Authors’ response:

We have conducted a new literature review in Ovid-database and PubMed with search terms including ‘continuity of care’ and its linguistic variations. Medical subject headings (MeSH), were used with AND/OR to combine search terms. We included articles describing the development and/or evaluation of the psychometric properties of instruments. Additionally, we also looked for recent articles and articles in press in relevant journals. We also hand-searched reference lists in relevant articles and included search terms related to continuity of care. A few articles were not available in English and thus were not included.

Many continuity measures have been created. However, quite a few instruments with similar and/or related constructs exist. The most commonly used measures based on healthcare services utilization patterns are the Usual Provider Continuity (UPC) and Continuity of Care (COC) indices. In particular, Napolitano et. al (2016) conducted a study among patients with multiple chronic conditions in Italy including patients with cardiac diseases. We decided not to include these articles in the introduction. Although their view is still important, continuity of care is considered to be a multidimensional concept. A few instruments have been developed to evaluate the three types of continuity of care, including Continuity of care between Care Levels Questionnaire (CCAENA) the Nijmegen Continuity Questionnaire (NQQ), and Chao Perceptions of Continuity (Chao PC). These instruments are applicable across care levels. However, they are
aimed at the general population. A systematic review by Van Melle et al. (2018) identified instruments in transitional patient safety and also used the term “continuity of care” in their search strategy but only generic instruments were identified. Another resource for development of instruments of continuity of care is the Care Coordination Measures by the U.S. Agency for Health Care Research and Quality. However, the overview over available instruments tools are mostly for patients with multiple chronic conditions.

The decision on which instrument to use depends on the characteristics of the study population, the ability to measure all three dimensions of continuity and the quality of the psychometric properties. According to a systematic review by Uijen (2012), the HCCQ seems to be a proper questionnaire for patients with heart disease. Additionally, our qualitative study (2017) found that the multi-faceted concept of continuity of care as defined by Haggerty et al. (2003) was relevant to patients undergoing percutaneous coronary intervention. We also considered an instrument developed by Uijen et al. (2012) used on patients with heart failure. However, psychometric properties were not adequately reported in their article. A generic version is translated and evaluated by Hetlevik et al. (2017) in a Norwegian context. However, as pointed out, we wanted to use a disease-specific instrument in this study. Although we did not find more work on HCCQ validation, we have added four articles in the introduction:


Furthermore, we have pointed out in the manuscript why a tool should be adapted into another language (Page 6, line 116-117)

5. Referee’s comment:

Methods section: the method used for translation and cultural adaptation must be referenced. The title of "measurement" should be changed to "study instruments" in this regards for all presented instrument the data regarding their validity and reliability, also it is not clear why two form of QoL questionnaires were used (RAND and WHO?!)
We have entered the references De Vet et al. (2001) and Beaton et al. (2000) (Page 8, line 158) used for translation process. The title of “measurement” is changed to “study instruments” (Page 10, line 224).

We used a single question from WHOQOL-BREF, which is a global quality of life scale applicable cross-culturally, and developed through the World Health Organization. The instrument allows comparisons with the general population. On the other hand, RAND-12 is a generic form with more questions and two domain scores. The instrument is designed to measure very broad aspects of health and is therefore more suitable for a wide range of patient groups and the general population. We found it useful to use instruments on different levels, global and generic, to evaluate the construct validity. Cronbach’s alpha is not computed for RAND-12 since the two summary scores there were not ordinary scale scores, but the result of more involved computations involving a norm data set. However, we have added Cronbach’s alpha for the other presented instrument.

6. Referee’s comment:

It should be declare which types of validity have been evaluated, I see only construct validity by CFA, although you did not say about the type of validity as construct validity, you should first do exploratory factor analysis then CFA, some important form of validity such criterion and discriminant validity are missed in this study, some form of items validity such item discriminant validity, item-scale correlation have not been evaluated,

Discriminant validity could be assessed internally based on subgroups of the patients.

I see three types of validity "content validity, criterion validity and construct validity. Criterion validity” in page 13 but in next page convergent validity also could be seen; please deal with the ambiguity.

Authors’ response:

We appreciate the reviewer’s perspective. We have mainly used the book “Measurement in Medicine” by De Vet et al. (2011) and Measurement and the measurement of change” by Polit and Yang (2016), which cover measurement theories, methods and criteria for evaluating instruments. Three different types of validity have been evaluated. Firstly, face validity was evaluated. Face validity is the first impression and a subjective assessment of the instrument. Thereafter, the content validity was evaluated, which focused on whether the instrument corresponds to the construct “continuity of care” with regard to relevance and comprehensiveness. Finally, three aspects of construct validity were evaluated; construct validity, hypotheses testing and cross-cultural validity. This is terminology used in the framework by De Vet et al (2011). However, we have changed the title to “Expected relationships and subgroup means” to clarify (Page 13, line 285).

Based on theories about the construct and previous research, hypotheses were formulated with regard to expected relationships with instruments measuring related construct (convergent
validity) and hypothesis that the instrument should have no or small correlations (discriminant validity). For example, that the HCCQ would have only slight correlation with RAND-12 as these were thought to assess two different constructs. Another type of hypothesis concerns expected differences between subgroups of patients (De Vet et al. 2011). For example, that patients’ cardiac rehabilitation participation is positively related to HCCQ scores, based on previous research. We have added a statement initially to clarify what we have done (Page 13, line 286-288).

De Vet et al. (2011) further emphasize that criterion validity can only be assessed when a gold standard is available. Patient-reported instruments such as the HCCQ, which often focus on subjective perceptions, usually lack a gold standard and in these situations, construct validity is applicable (Page 14, 318 – 324). As suggested we have calculated item-scale correlations between 0.45 and 0.70 (Page 19, line 441-442).

According to De Vet et al. (2011) exploratory factor analysis should be assessed if there are no clear ideas about the number and types of dimensions and CFA should be used if a priori hypotheses about dimensions of the construct are available (based on theory or previous analysis). An exploratory factor analysis was conducted on the original dataset in a Canadian population by cardiac patients. Based on those findings a confirmatory factor analysis was used in this study. We added a sentence in the Discussion section to clarify this view (Page 22, line 512 – 513). The HCCQ is developed from a well-known theory of continuity of care developed by Haggerty et al. (2003). Furthermore, Riley et. al (2017), who evaluated the HCCQ among patients with acute coronary syndrome, warranted more psychometric testing including CFA.

7. Referee’s comment:

Type of ICC used must be reported. SEM and SDC must be calculated and reported.

Authors’ response:

There are several ICC formulas, all of which are variations on the basic formula for a reliability parameter. We used the ICCagreement which is the extent to which scores across a test and a retest are identical (Page 16, line 358 -369). In that case, the error variance consists of residual variance plus variance due to systematic differences (Polit and Yang, 2016). It has been argued that in a clinical situation, absolute agreement is often more important than consistency (De Vet et al. 2011).

We agree with the reviewer that SEM and SDC should be calculated and reported. A reliability coefficient is a relative index that varies from sample to sample and across populations, SEMs by contrast, are in measurement units of the measure (Polit and Yang, 2016). Different methods can be used to obtain the SEM value. We estimated SEM from the computation of ICCs: $SEM = SD \times \sqrt{(1 - ICC)} = 0.28$. Reliable change for continuous data is often estimated using the index smallest detectable change (SDC). Operationally, the SDC has been defined as a change score that falls outside the limits of agreement on a Bland-Altman plot (De Vet et al. 2011). We estimated SDC from this formula: $SDC = d \pm 1.96 \times (\sqrt{2 \times SEM}) = \pm 1.05$. We have added a
sentence in the statistical analyses (Page 16, line 360-361) and in the results chapter (Page 22, line 512-513).

8. Referee’s comment:

The Cronbach alpha also should be evaluated for each extracted domains. it would also add ceiling and floor effects.

Authors’ response:

The Cronbach’s alpha for the three domains of informational, relational, and management continuity were 0.93, 0.87, 0.89, respectively; for the total scale Cronbach’s alpha was 0.95.

(Page 19, line 437-438). We have also highlighted in the text that alpha is assessed for each domain in statistical analyses (Page 14, line 354-355).

The HCCQ total floor effect was 0% and ceiling was 1.7%, (information continuity floor 0% and ceiling 4%, relational continuity 0.5% and ceiling 6.8%, management continuity floor 4% and ceiling 4.5%). We added this information in the analysis chapter (Page 14, line 310 – 311) and in the results (Page 17, line 393 – 396). When a large proportion of the patients is found at either the higher and the lower end of the scale, then more items are needed to discriminate between these patients. According to De Vet et al (2011) substantial floor and ceiling effects can occur if more than 15% of the patients achieve the lowest or highest possible score, respectively.

9. Referee’s comment:

Other important point: this sentence at first paragraph of statistical analysis does not make sense "Item means, standard deviations,...." also I am not really agree with the presented sentence regarding cut points from previous study "Hadjistavropoulos", during a validation and adaptation it is possible to obtain new ones!!

Authors’ response:

We agree that it is problematic to calculate average and standard deviations on ordinal data. However, Hadjistavropoulos et al. (2004) have done so in their original articles and we wanted to be able to compare with the Canadian population. We underlined this in the article (Page 14, line 307-308).

The sentence regarding cut points. The purpose of the rating scale was to allow patients to express both the direction and strength of their opinion about a topic. However, there are different approaches to compute cut points in a distribution of scores to classify or divide patients into different groups. For example, use of ROC curve which show the specificity and sensitivity for any given cut-off point. Another approach is to use the cut off score for a 5-point Likert scale = (Maximum – Minimum) / Group = (5-1)/5 = 0.8. -> Likert scale 1 to 1.08
(Strongly disagree), 1.81 to 2.60 (Somewhat Disagree), 2.61 to 3.40 (Hard to decide) 3.41 to 4.20 (Somewhat Agree), 4.21 to 5 (Strongly Agree) (Jaafar et. al 2018). Nevertheless, the instrument developer of the HCCQ, has clarified that the cut points in this case were based on clinical judgment, therefore we are not able to reproduce the procedure in our setting. Items were rated on a 5-point Likert-type scale from 1 (strongly disagree) to 5 (strongly agree), as well as the option to choose ‘not applicable’. To be able to compare with the Canadian population the cut-off value from Canada was used. However, we have underlined in in the Limitation section that this should be considered in further evaluation of the instrument (Page 26, line 617-618).

10. Referee’s comment:

This sentence "Non-parametric tests were used for ordinal variables and parametric tests for continuous variables." is not applicable I do not see the relevant results except Spearman rank and Pearson correlation coefficients! also I do not find any application for this sentence in results section?!!"and groups were compared with linear regression and analysis of variance."!!

Authors’ response:

We appreciate this observation. We have removed the two sentences “non-parametric tests were used for ordinal variables and parametric tests for continuous variables." is not applicable” and “and groups were compared with linear regression and analysis of variance”. We have sought to make the text more clear (Page 14, line 312 -314).

11. Referee’s comment:

The proposed statistical test in this sentence "The association between HCCQ and gender was .... using Pearson correlation" are not correct, first it not clear why these evaluation were conducted? then for the first you should use independent t-test and for the next one Pearson correlation should be used.

Authors’ response:

Following the comments of the reviewer, we have split the sentence in two in order to clarify which analyses have been done. “The association between HCCQ and gender was evaluated by using an independent t-test. Furthermore, the association between HCCQ and age by using Pearson correlation” (Page 14, line 325-327).

The rationale for expected differences between genders was the study of Riley et al. (2007). They found mean differences between genders even if the difference was not significant (Male 3.92 ±0.55, Female 3.82± 0.62). This study prospectively evaluated continuity of care in a large-site sample of patients hospitalized for an acute coronary event or procedure. We have added the reference to Riley et al. (2007) and gender as one of the four grouping variables we expected differences (Page 13, line 294) (Page 18, line 415)
12. Referee’s comment:

The title "Hypotheses" and the matters under it is not common in validation studies particularly the heading! majority of the matter does not make sense, for example the last sentence under heading "Hypotheses" if continuous variables have been reported as mean (SD) how about the categorical variable?

Authors’ response:

A clinometric working group in the Netherlands has sought to address problems relating to measurement terminology. Using a Delphi-type approach, they created COSMIN, the Consensus-based Standards for the selection of health Measurement Instruments (Mokkink et al., 2010a, 2010b; Terwee et al. 2012). According to De Vet et. al (2011) the basic principle of construct validation is that hypotheses are formulated about the relationship of scores on the instrument under study with scores on other instruments measuring similar (convergent validity) or dissimilar construct (discriminant validity). Additionally, another type of hypothesis concerns expected differences between subgroup of patients (known group or discriminant validity). Polit and Yang (2016) also use the terminology “hypothesis testing” when they present a taxonomy of measurement properties. However, following the comments of the reviewer we changed the heading of this section to “Expected relationships and subgroup means” (Page 13, line 285). See also response to comment 6.

13. Referee’s comment:

Discussion: In first or second para must report the main findings.

Authors’ response:

As the reviewer recommended, we have reported the main results in the first para in the discussion chapter (Page 19-20, line 450-453).

On behalf of the CONCARD Investigators,

Sincerely yours,

Irene Valaker