Comparing higher order models for the EORTC QLQ-C30

Gundy, C.M.; Fayers, P.M.; Groenvold, M.; Petersen, M.A.; Scott, N.W.; Sprangers, M.A.G.; Velikova, G.; Aaronson, N.K.

Published in:
Quality of Life Research

DOI:
10.1007/s11136-011-0082-6

Link to publication

Citation for published version (APA):

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: http://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.
Comparing higher order models for the EORTC QLQ-C30

Chad M. Gundy · Peter M. Fayers · Mogens Groenvold · Morten Aa. Petersen · Neil W. Scott · Mirjam A. G. Sprangers · Galina Velikova · Neil K. Aaronson

Accepted: 28 November 2011
© The Author(s) 2011. This article is published with open access at Springerlink.com

Abstract
Purpose To investigate the statistical fit of alternative higher order models for summarizing the health-related quality of life profile generated by the EORTC QLQ-C30 questionnaire.
Methods A 50% random sample was drawn from a dataset of more than 9,000 pre-treatment QLQ-C30 v 3.0 questionnaires completed by cancer patients from 48 countries, differing in primary tumor site and disease stage. Building on a “standard” 14-dimensional QLQ-C30 model, confirmatory factor analysis was used to compare 6 higher order models, including a 1-dimensional (1D) model, a 2D “symptom burden and function” model, two 2D “mental/physical” models, and two models with a “formative” (or “causal”) formulation of “symptom burden,” and “function.”
Results All of the models considered had at least an “adequate” fit to the data: the less restricted the model, the better the fit. The RMSEA fit indices for the various models ranged from 0.042 to 0.061, CFI’s 0.90–0.96, and TLI’s from 0.96 to 0.98. All chi-square tests were significant. One of the Physical/Mental models had fit indices superior to the other models considered.
Conclusions The Physical/Mental health model had the best fit of the higher order models considered, and enjoys empirical and theoretical support in comparable instruments and applications.

Keywords Health-related quality of life · Confirmatory factor analysis · Higher order factor · EORTC QLQ-C30

Abbreviations
AIC Akaike Information Criterion
AP Appetite loss
CFI/TLI Comparative Fit Index/Tucker–Lewis Index
CF Cognitive function
CO Constipation
DF Degrees of freedom
DI Diarrhea
DY Dyspnea
EF Emotional function
EORTC European Organization for Research and Treatment of Cancer

Disclaimer: The contents of this publication and methods used are solely the responsibility of the authors and do not necessarily represent the official views of the EORTC.

C. M. Gundy · N. K. Aaronson
Division of Psychosocial Research and Epidemiology, The Netherlands Cancer Institute, Plesmanlaan 121, 1066 CX Amsterdam, The Netherlands
e-mail: n.aaronson@nki.nl

P. M. Fayers · N. W. Scott
Section of Population Health, Institute of Applied Health Sciences, University of Aberdeen, Aberdeen, UK

P. M. Fayers
Department of Cancer Research and Molecular Medicine, Faculty of Medicine, Norwegian University of Science and Technology, Trondheim, Norway

M. Groenvold · M. Aa. Petersen
The Research Unit, Department of Palliative Medicine, Bispebjerg Hospital, Copenhagen, Denmark

M. Groenvold
Institute of Public Health, University of Copenhagen, Copenhagen, Denmark

M. A. G. Sprangers
Department of Medical Psychology, Academic Medical Centre, University of Amsterdam, Amsterdam, The Netherlands

G. Velikova
St James’s Institute of Oncology, Leeds, UK

Published online: 21 December 2011

DOI 10.1007/s11136-011-0082-6
Introduction

Since its release in 1993, the EORTC QLQ-C30 has become a widely used “core” instrument for the study of cancer-specific health-related quality of life (HRQoL) [1–4]. It comprises 9 multi-item scales and 6 single-item measures. While the multidimensional profile generated by the QLQ-C30 is invaluable in providing a detailed picture of the impact of cancer and its treatment on patients’ HRQoL, there is also interest in developing “summary” scores that can simplify analyses and minimize the chance of Type I errors due to multiple comparisons. In addition, it might sometimes be more useful, particularly in clinical trials, to employ a composite variable measured with greater precision [5], as opposed to many variables, each measured with less precision. This interest in summarizing data generated from multidimensional HRQoL profiles is reflected in the development of so-called “higher order models,” such as those available for the SF-36 Health Survey and other instruments [6–8].

To date, there have been a limited number of analyses of the structure of the QLQ-C30, all of which relied on either relatively small sample sizes (e.g., N < 200), a subset of the QLQ-C30 items, and/or exploratory techniques [9–15]. The aim of the present study was to fill this gap, by examining empirically and comparing the statistical “fit” of a number of alternative “higher order” measurement models for the QLQ-C30, using confirmatory factor analysis in a large sample of patients [16]. The results of this study may be used to identify one or more, higher order measurement models that could be used for the computation of simpler, summary scores for this questionnaire. The results are also of interest from a theoretical perspective, hopefully allowing us to place the pragmatically oriented QLQ-C30 in the context of a number of established, theoretical HRQoL models.

Methods

Data source

The data used in this study were originally collated for the Cross-Cultural Assessment Project of the EORTC Quality of Life Group, and have been described elsewhere [17, 18]. Briefly, a total of 124 individual datasets were received: 54 from the EORTC Data Center, with permission from the relevant EORTC Clinical Groups, and an additional 70 datasets from other individuals and organizations from around the world. Included were datasets from 48 countries and for 33 translations of the QLQ-C30. The resulting dataset consisted of 38,000 respondents, of whom more than 30,000 completed baseline (pre-treatment) questionnaires. Of these 30,000 respondents, 9,044 completed the most recent version (3.0) of the QLQ-C30. We selected a 50% random sample for the present investigation. The remaining observations were retained for future analyses.

Relevant information from each dataset was extracted, recoded into a standard format, and combined into one large project database. In addition to the QLQ-C30, other data collected included age, gender, country, language of administration, primary disease site, and stage of disease.

The QLQ-C30

The EORTC QLQ-C30 version 3.0 [1–4] includes 30 items comprising 5 multi-item functional scales (physical (PF), role (RF), cognitive (CF), emotional (EF), and social (SF)), 3 multi-item symptom scales (fatigue (FA), nausea and vomiting (NV), and pain (PA)), 6 single-item symptom scales (dyspnea (DY), insomnia (SL), appetite loss (AP), constipation (CO), diarrhea (DI), and financial difficulties (FI)), and a two-item global quality of life scale (QL). The FI item was excluded from all of the present analyses, as it may be considered peripheral to the other scales in the instrument, and often is left unreported in the literature. The questionnaire uses a 1-week time frame and 4-point Likert-type response scales (“not at all,” “a little,” “quite a bit,” and “very much”), with the exception of the two items of the overall QL scale which use a 7-point response scale.

The QLQ-C30 has been shown to be reliable and valid in a range of patient populations and treatment settings. Across a number of studies, internal consistency estimates (Cronbach’s coefficient α) for the scores of the multi-item scales exceeded 0.70 [3]. Test–retest reliability coefficients range between 0.80 and 0.90 for most multi-item scales and single items [19]. Tests of validity have shown the...
QLQ-C30 to be responsive to meaningful between-group differences (e.g., local vs. metastatic disease, active treatment vs. follow-up) and changes in clinical status over time [1, 3].

Measurement models

Seven HRQoL measurement models [20–22] were fit to the data. The models were chosen on the basis of a review of recent HRQoL literature, general knowledge of psychometric literature, discussions among the co-authors, and suggestions made by external experts. Analyses were conducted by means of confirmatory factor analysis. The fit of each model was considered separately, and in relationship to the other models when possible.

The models to be compared in this study were organized in 3 branches of nested models, each branch beginning with the same Standard model in the root node. The first branch consists of the Standard model, followed by a two-dimensional Physical health–Mental health model, a two-dimensional Physical burden–Mental function model, and culminating in a one-dimensional HRQL model. The second branch begins with the Standard model, followed by a two-dimensional Burden and Function model, and again culminating in the same one-dimensional HRQL model just mentioned. Finally, the third group of models utilizes a different, so-called “formative”—or “causal”—approach to measurement. Two variants, a fixed weight and a free weight, of these formative models are included in this branch. These two models are nested within a third “branch” emanating from the Standard model mentioned above.

These 7 models are described in more detail below. (See Fig. 1 for a graphical representation of the models. (Straight lines, with one-sided arrows, represent regression coefficients; arced lines, with two-sided arrows, represent correlation coefficients.)

(1) The Standard 14-dimensional QLQ-C30 model corresponding to the original 13 QLQ-C30 scales and one overall QL scale, with each scale modeled as a first-order latent variable. All first-order factors were allowed to correlate with each other. Here we also assumed that the single-item symptom scales were manifestations of latent variables (the so-called “spurious” model [23]). This Standard model formed a fundamental “building block,” used as the basis for all of the other models described here.

(2) The two-dimensional, Physical health and Mental health model, which has been used for the SF-36 [6, 7], has been considered in a large, multi-instrument study [24] and is consistent with the PROMIS domain mapping project and the WHO framework [25–27]. Unfortunately, it is difficult to map the QLQ-C30 a priori to the physical-mental distinction in only one, unambiguous manner (see, e.g., [24] for an alternative mapping). In the current case, implementation of the Physical–Mental model requires that some symptom-related first-order latent variables map to the Mental as well as the Physical higher order factors. Specifically, PF, NV, DY, AP, CO, and DI were allowed to load only on the Physical higher order factor; EF and CF were allowed to load only on the Mental factor; while RF and SF, and the symptoms FA, PA, and SL were allowed to load on both the Mental health and Physical health factors. We assumed that QL was not subsumed by either higher order component.

(3) This variant of the previous model, labeled the Physical burden and Mental function model, requires all symptom first-order latent variables to load onto only one higher order factor. Thus, PF, FA, NV, PA, DY, SL, AP, CO, and DI were allowed to load only on the Physical burden factor; EF and CF were allowed to load only on the Mental function factor; and RF and SF were allowed to load on both factors. Again, QL was not subsumed by either higher order component.

(4) The Wilson and Cleary model [28] describes HRQoL as consisting of (a sequence of causal effects between) 5 groups of latent variables: physiological states, symptom status, functional status, general health perception, and overall HRQoL. This model was recently tested in a structural equation model [29], using a number of different instruments in a sample of HIV/AIDS patients. This model also seems to have a natural correspondence with the content of the QLQ-C30, which emphasizes symptom burden, functional health, and overall QoL. Thus, paralleling this approach, PF, SF, RF, CF, and EF were only allowed to load on Function; and FA, NV, PA, DY, SL, AP, CO, and DI were only allowed to load on Burden. Again, QL was not subsumed by either higher order component.

(5) The parsimonious, and highly restrictive, one-dimensional HRQL model has recently been considered using the QLQ-C30 in a multicultural sample of cancer patients [13, 14]. It assumes that all first-order latent variables (with the exception of QL) load on only one underlying HRQL dimension. Again, QL remained unsubsumed.

(6) & (7) Boehmer and Luszczynska [9] published a study of a model inspired by the work of Fayers et al. (e.g., [30, 31]). It is somewhat similar to the Burden–Function model presented above, yet allows the symptom items
to simultaneously play the role of reflective indicators for Burden (or “symptomatology”) and formative indicators for Function. This model illustrates the potentially important distinction between formative and reflective scales, and the on-going controversy concerning their use and interpretability [23, 32–36]. Formative scales, when mis-specified as being reflective, will generally lead to bias and poorer model fit [37]. We therefore include a formative variant of Burden, to be used in the Burden-Function model mentioned above. As formative scales can have either equal, fixed weights, or freely estimated weights for their components, we consider both types of weighting, forming models (6) and (7). This model architecture is also closely related to the “multiple indicator, multiple cause” (MIMIC) model [38].

Statistical analysis

The 7 models described above were fitted to the QLQ-C30 version 3.0 item scores. All of these models were fitted under the following assumptions and methods:

**Basic model architecture**

The original QLQ-C30 multi- and single-item measures were modeled as first-order latent variables. The QL scale was also included in the models as a latent variable, and was allowed to co-vary with all other (higher order) latent variables, yet remained distinct from other higher order latent variables. Only those items originally associated with a specific scale were associated with the corresponding latent variable. All items were treated as being ordinal.

In order to identify latent variable models, it is customary to fix one of the item loadings to a value of 1.0. (Both loadings of items corresponding to the QL latent variable were also fixed.) This problem of model identification is especially critical for latent variables having only one item/indicator, and requires one to also fix the error variances for the five latent variables with only a single indicator. We therefore estimated the reliability of the one item latent variables on the basis of test–retest correlations reported elsewhere [19], and accordingly fixed the latent error variances to be equal, at 20% of the total variance for these latent variables [39]. This assumption is tantamount to assuming that the single-item scales perform satisfactorily, even though they are not perfect. Preliminary analyses indicated that model-fit statistics were only slightly affected by varying this assumption within reasonable bounds. This architecture corresponds to the Standard model mentioned above. It may also be viewed as a liberalization of the original QLQ-C30 scales, for it allows unequal item weights, assumes an ordinal measurement level for each item, and estimates error variances where possible.

**Estimators**

As all items were treated as being ordinal, polychoric correlations were estimated and a (robust) weighted least squares estimator with adjustment for means and variance (WLSMV)—with MPLUS’ default “delta” parameterization—was used [40]. This estimator is robust for small sample sizes and deviations from normality [41] and is nearly optimal for multi-level models [42]. The WLSMV estimator utilizes pair-wise deletion of missing observations as default. Alternative, (robust) maximum likelihood estimators would have required numerical integration—or Monte Carlo simulations—in more than 14 dimensions, which would present a computational burden straining the capacity of modern, desktop computers.

**Tests of model fit and other fit indices**

The \( \chi^2 \) test of model fit was examined. The \( \chi^2 \) test is sensitive to sample size, leading easily to rejection of the null hypothesis in models with a large number of observations. Approximate goodness-of-fit indices (AGFI) are less sensitive to sample size: the CFI/TLI (Comparative Fit Index/Tucker–Lewis Index) and the RMSEA (Root Mean Square Error of Approximation). There is a great deal of controversy concerning the proper use of the chi-square and AGFI (e.g., [43–48]), and since we foresee no consensus on this matter in the near future, we will report both [49, 50]. A commonly used rule of thumb is that a RMSEA < 0.05 indicates close approximate fit, while values between 0.05 and 0.08 indicate acceptable fit, and values >0.10 indicate poor approximate fit [51]. Another rule of thumb is that a value of CFI or TLI > 0.95 indicates good fit and a value > 0.90 indicates acceptable fit [50]. Differences \( \geq 0.01 \) between (pairs of) TLI/CFIs and RMSEAs are considered to be substantial enough to merit attention [52]. In the case of inadequate model fit, modification indices and residuals were examined, in order to detect possible causes.

Direct comparisons of models by computing the differences between their respective chi-squares are not appropriate when using WLSMV estimators, and requires some
additional computations [53, 54]. Direct comparisons between model chi-squares can only be made when one model is nested within the other model.

**Correction for cluster sampling**

The dataset was composed of data collected from dozens of different studies of various populations. It was suspected that this heterogeneity in populations and procedures could lead to biased parameter estimates and fit statistics. For this reason, a correction was made to the estimation procedure to take cluster sampling into account, and to adjust the standard errors and chi-square statistics [42, 55, 56]. Additional techniques, such as utilizing sampling weights [57, 58], other (i.e., maximum likelihood) estimators, or attempting to explicitly model the sampling process, may also have added value, but were not utilized in the current analysis. In the present case, a cluster was defined as a dataset from a source with a unique study identifier code, possibly extended with the treatment group as coded in the original dataset.

**Software**

Analyses were conducted using the Mplus v.5.2 program [59].

**Statistical significance**

Unless otherwise indicated, a significant result is defined as $P < 0.01$.

**Results**

The characteristics of the patients included in the study are presented in Table 1. The average age of the patients was 60 years, with slightly more males than females, and more early than advanced cancer. A number of study types (clinical trials, non-randomized comparative studies, and observational studies), a wide variety of (primarily European) countries, and a range of disease sites were also represented.

No item had more than 2.6% missing observations; for most items this was less than 1%. However, all items, with the exception of the two items of the QL scale, were highly skewed; approximately half of the items had 50% or more of the responses in the lowest category (data not shown). The polychoric correlations between the 29 items were generally moderate (i.e., >0.30) to strong (>0.50) (data not shown).

The fit indices for the various models are presented in Table 2. As might be anticipated given the large sample size, no model passed the stringent $\chi^2$ test of model fit. However, all models were deemed to be at least "adequate" approximations to the data, as determined by the previously noted rules of thumb applied to the CFI/TLI and RMSEA indices. As expected [20], the less restricted the model, the better the model fit, with the *Standard* model even achieving a "good" fit. The *Mental–Physical* models had approximate fit indices slightly superior to all of the other higher order models. The correlations between higher order factors (in the multi-factor models) were generally

---

**Table 1** Respondent characteristics ($N = 4,541$)

<table>
<thead>
<tr>
<th></th>
<th>Mean (SD)</th>
<th>% Missing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>59.6 (12.6)</td>
<td>9.9</td>
</tr>
<tr>
<td>Gender</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>2,511</td>
<td>55.3</td>
</tr>
<tr>
<td>Female</td>
<td>1,906</td>
<td>42.0</td>
</tr>
<tr>
<td>Unknown</td>
<td>124</td>
<td>2.7</td>
</tr>
<tr>
<td>Stage</td>
<td></td>
<td></td>
</tr>
<tr>
<td>I–III</td>
<td>1,846</td>
<td>40.7</td>
</tr>
<tr>
<td>IV–recurrent/metastatic</td>
<td>1,765</td>
<td>38.9</td>
</tr>
<tr>
<td>Unknown</td>
<td>930</td>
<td>20.5</td>
</tr>
<tr>
<td>Site</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Breast</td>
<td>663</td>
<td>14.6</td>
</tr>
<tr>
<td>Colorectal</td>
<td>245</td>
<td>5.4</td>
</tr>
<tr>
<td>Gynecological</td>
<td>375</td>
<td>8.3</td>
</tr>
<tr>
<td>Head and neck</td>
<td>801</td>
<td>17.6</td>
</tr>
<tr>
<td>Lung</td>
<td>610</td>
<td>13.4</td>
</tr>
<tr>
<td>Esophagus/stomach</td>
<td>822</td>
<td>18.1</td>
</tr>
<tr>
<td>Prostate</td>
<td>405</td>
<td>8.9</td>
</tr>
<tr>
<td>Other</td>
<td>620</td>
<td>13.7</td>
</tr>
<tr>
<td>Study type</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RCT</td>
<td>1,561</td>
<td>34.4</td>
</tr>
<tr>
<td>Non-RCT</td>
<td>1,455</td>
<td>32.0</td>
</tr>
<tr>
<td>Field study</td>
<td>1,386</td>
<td>30.5</td>
</tr>
<tr>
<td>Unknown</td>
<td>139</td>
<td>3.1</td>
</tr>
<tr>
<td>Country</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Belgium</td>
<td>193</td>
<td>4.3</td>
</tr>
<tr>
<td>Canada</td>
<td>120</td>
<td>2.6</td>
</tr>
<tr>
<td>France</td>
<td>266</td>
<td>5.9</td>
</tr>
<tr>
<td>Germany</td>
<td>477</td>
<td>10.5</td>
</tr>
<tr>
<td>Netherlands</td>
<td>228</td>
<td>5.0</td>
</tr>
<tr>
<td>Norway</td>
<td>498</td>
<td>11.0</td>
</tr>
<tr>
<td>Spain</td>
<td>402</td>
<td>8.9</td>
</tr>
<tr>
<td>Sri Lanka</td>
<td>438</td>
<td>9.6</td>
</tr>
<tr>
<td>Sweden</td>
<td>202</td>
<td>4.4</td>
</tr>
<tr>
<td>UK</td>
<td>722</td>
<td>15.9</td>
</tr>
<tr>
<td>USA</td>
<td>157</td>
<td>3.5</td>
</tr>
<tr>
<td>Other</td>
<td>838</td>
<td>18.5</td>
</tr>
</tbody>
</table>
quite high, often exceeding 0.95 (see Table 2). This indicates that these higher order factors were virtually indistinguishable, thus implying that additional factors were of limited explanatory value. Exceptions are the models positing Mental and Physical factors, which have lower correlations between these higher order factors.

The results of (corrected) chi-squared difference tests between pairs of models within each branch of nested models [53] are presented in Table 3. Differences between each successive pair of nested models in each branch were significant, indicating that each successive tightening of restrictions resulted in a significant decrement in model fit.

The standardized regression weights (for the first-order factors on the higher order factors) for the best fitting models for each of the three branches are presented in Table 4. The percentage of variance for each first-order factor explained by their corresponding higher order factor is presented as well. All postulated factor regression weights for the Burden/Function and the Mental health/Physical health model were significant, with the exception of SL on the Physical health factor. However, the percentages of explained variance for PF, EF, CF, and SL are markedly inferior for the Burden/Function model.

Only the hypothesized regression weights for the FA, SL, and PA symptom scales for the formative Burden/Function model (in the third branch of nested models) were statistically significant. FA was the only symptom with a substantial loading on the formative Burden variable, which more or less ignored the other symptoms. The amount of explained variance was again inferior for the PF, SF, and CF scales, as compared to the Mental health/Physical health model.

Examination of the modification indices and residuals indicated that item q22 (“worry”) was a source of ill-fit for all models. There also appeared to be some relationships between EF and the other scales not fully captured by the higher order factors (data not shown).

Table 2 Tests* and approximate goodness-of-fit indices for various models

<table>
<thead>
<tr>
<th>Model</th>
<th>$\chi^2$</th>
<th>df</th>
<th>CFI/TLI</th>
<th>RMSEA</th>
<th>Remarks</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. “Standard” model</td>
<td>134</td>
<td>15</td>
<td>0.96/0.98</td>
<td>0.042</td>
<td>14 Latent variables, excluding FI</td>
</tr>
<tr>
<td>2. Physical health, mental health and QL</td>
<td>234</td>
<td>19</td>
<td>0.92/0.98</td>
<td>0.050</td>
<td>Correlation physical health and mental health = 0.74</td>
</tr>
<tr>
<td>3. Physical burden, mental function and QL</td>
<td>248</td>
<td>18</td>
<td>0.92/0.97</td>
<td>0.053</td>
<td>Correlation physical burden and mental function = 0.81</td>
</tr>
<tr>
<td>4. Symptom burden, function and QL</td>
<td>294</td>
<td>18</td>
<td>0.90/0.97</td>
<td>0.058</td>
<td>Correlation burden and function = 0.97</td>
</tr>
<tr>
<td>5. HRQL and QL</td>
<td>297</td>
<td>18</td>
<td>0.90/0.97</td>
<td>0.058</td>
<td></td>
</tr>
<tr>
<td>6. Formative symptom burden (free weights), function and QL</td>
<td>277</td>
<td>17</td>
<td>0.91/0.97</td>
<td>0.058</td>
<td>Correlation formative burden and function = 0.96</td>
</tr>
<tr>
<td>7. Formative symptom burden (fixed weights), function and QL</td>
<td>300</td>
<td>17</td>
<td>0.90/0.96</td>
<td>0.061</td>
<td>Correlation formative burden and function = 0.95</td>
</tr>
</tbody>
</table>

* All $\chi^2$ tests of model fit were significant at $P < 0.001$

The standardized regression weights for the best fitting models for each of the three branches are presented in Table 4. The percentage of variance for each first-order factor explained by their corresponding higher order factor is presented as well. All postulated factor regression weights for the Burden/Function and the Mental health/Physical health model were significant, with the exception of SL on the Physical health factor. However, the percentages of explained variance for PF, EF, CF, and SL are markedly inferior for the Burden/Function model.

Examination of the modification indices and residuals indicated that item q22 (“worry”) was a source of ill-fit for all models. There also appeared to be some relationships between EF and the other scales not fully captured by the higher order factors (data not shown).

Table 3 $\chi^2$ Difference testing between 3 branches of nested models

<table>
<thead>
<tr>
<th>Model</th>
<th>$\Delta\chi^2$ wrt previous model in Branch 1</th>
<th>df</th>
<th>$\Delta\chi^2$ wrt previous model in Branch 2</th>
<th>df</th>
<th>$\Delta\chi^2$ wrt previous model in Branch 3</th>
<th>df</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Standard model (14 latents), incl. QL</td>
<td>Root node</td>
<td>17</td>
<td>Root node</td>
<td>15</td>
<td>Root node</td>
<td>12</td>
</tr>
<tr>
<td>2. Physical health, mental health and QL</td>
<td>293</td>
<td>2</td>
<td>377</td>
<td>47</td>
<td>336</td>
<td>5</td>
</tr>
<tr>
<td>3. Physical burden, mental function and QL</td>
<td>77</td>
<td>2</td>
<td>15</td>
<td>51</td>
<td>12</td>
<td></td>
</tr>
<tr>
<td>4. Symptom burden, function and QL</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td></td>
</tr>
<tr>
<td>5. HRQL and QL</td>
<td>241</td>
<td>3</td>
<td>2</td>
<td>4</td>
<td>–</td>
<td></td>
</tr>
<tr>
<td>6. Formative symptom burden (free weights), function, and QL</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>336</td>
<td>12</td>
</tr>
<tr>
<td>7. Formative symptom burden (fixed weights), function, and QL</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>241</td>
<td>5</td>
</tr>
</tbody>
</table>

All $\chi^2$ difference tests of model comparisons were significant at $P < 0.01$

$\chi^2$ difference testing—when using the WLSMV estimator—is not a simple difference between two model $\chi^2$'s. In addition, a model can only be directly compared—using $\chi^2$ difference testing—with other models in the same branch of (nested) models.
Discussion and conclusions

The present study tested the statistical fit of seven alternative measurement models for the QLQ-C30. This was done by using confirmatory factor analysis to compare empirically their adequacy in representing the EORTC QLQ-C30 in a sample of 4,541 cancer patients. The point of reference was the Standard model, a latent variable model which employed the architecture of the standard, 14-dimensional QLQ-C30 model (excluding the FI item).

As mentioned previously, the models studied here were organized into three independent branches of nested models: three models in the so-called Mental–Physical branch, two in the Burden-Function branch, and two in the “formative” Burden-Function branch. The Standard model stands at the apex of each of the three branches.

None of the models examined passed the stringent \(\chi^2\) test of model fit, indicating that none of these models captured all of the systematic variation in the data. It should be noted, however, that with 4541 observations, a chi-square test is quite sensitive to detecting small deviations. Importantly, all models demonstrated at least an “adequate” approximation to the data [50]. The Standard QLQ-C30 model actually demonstrated a “good” fit to the data. Moreover, \(\chi^2\) “difference testing” demonstrated that each addition of restrictions in each of the successively nested models in each branch led to a statistically significant deterioration in model fit.

The MentalHealth/ PhysicalHealth model, the least restricted higher order model in the first branch studied, is significantly better than its nested alternatives, and gives an adequate, albeit imperfect, approximation to the data. The Burden/Function model was the best approximation to the Standard model in the second branch. We note that the Burden/Function model is only slightly superior to the simpler one-dimensional HRQL model, for its two dimensions are almost indistinguishable.

Unfortunately, we cannot use the chi-square test to directly test the fit of the models nested in these two, different branches. However, we did use the approximate fit indices to compare those models, with the results indicating that the MentalHealth/PhysicalHealth model is slightly superior to the Burden/Function model. Additionally, the MentalHealth/PhysicalHealth model achieves better explanatory power for the CF, PF, EF, and SL scales than does the Burden/Function model. For these reasons, the MentalHealth/PhysicalHealth model is preferable.

A third branch of nested models, consisting of “causal” or “formative” latent variables, represents an alternative approach for the modeling of HRQL questionnaires. The model with free weights was a statistically better fit to the data than the fixed (equally weighted) model. However, the potential improvements in fit indices, which are to be expected if the formative conceptualization was more appropriate than the reflective one [37], were not observed in the current analysis. Additionally, the only symptom that
appears to strongly predict Function is fatigue, a result also reported previously [9]. This indicates that the other symptoms may be regarded as largely irrelevant as predictors of Function for this group of patients, which may be an overly zealous simplification of the Standard model. One could argue that this result disqualifies this branch of models.

It is interesting to note that question 22 (i.e., “did you worry?”) of the QLQ-C30 emotional function scale was frequently flagged as being a source of ill-fit. This may have to do with possible ambiguity in the meaning of “worry,” either as an indication of healthy concern in a difficult situation, or as an indication of psychological distress.

Several possible limitations of this study should be noted. First, the use of pair-wise deletion for (the relatively sparse) missing data in the computation of the polychoric correlations resulted in some loss of data. A second limitation concerns the possible bias introduced from the clustered sampling of data from various data sources. While we did apply a correction to the chi-square statistics and standard errors, additional corrections for parameter estimates, possibly based on sampling weights, would arguably have been even better. Third, it would have been useful to have access to Akaike Information Criterion (AIC), and other related statistics [60] in order to compare non-nested models across the various branches. The use of full information maximum likelihood estimation procedure could have provided a solution for all three problems simultaneously; however, the computational burden for such an estimation procedure is prohibitive.

A fourth limitation concerns the choice of models, which was neither exhaustive for all plausible, theoretical models, nor sufficient for capturing all of the systematic variation in the data. On the other hand, the “alternative models” approach used here is methodologically stronger than a purely exploratory approach [16]. For this reason, we refrained from “tweaking” either the standard or any of the other alternative models in order to achieve some improvement in fit, a practice frowned upon as potentially capitalizing on chance. Nevertheless, we recognize that there are other, more exploratory approaches that might be used. For example, causal discovery techniques and software (e.g., TETRAD) employ rigorous algorithms to locate all well-fitting models for a set of observed data, to which theory can then be applied to choose the most suitable or plausible model(s). While beyond the scope of the current paper, the utility of such approaches could be the subject of future studies [61, 62].

Summarizing, we believe that the PhysicalHealth/MentalHealth model is the most appropriate conceptualization for our goal of offering a simplified form of QLQ-C30 outcomes. This model was found to provide an “adequate” fit to the data, slightly superior to the alternative, higher order models examined here. We believe that it is the best of the approximations to the Standard model considered in this study. The Physical Health/MentalHealth conceptual model has also been utilized and successfully tested for other HRQoL instruments [6, 7], has been considered in a large, multi-instrument study [24], and is consistent with the PROMIS domain mapping project and the WHO framework [25–27]. For these reasons, we consider it to be the most promising of the models considered here.

Nevertheless, the “superiority” of this PhysicalHealth/MentalHealth model is modest, and it remains to be seen whether its extra complexity—as compared to e.g., the simple HRQL model—provides tangible (clinical) benefits. We therefore intend to further examine the suitability of the PhysicalHealth/MentalHealth model by testing its measurement equivalence across sub-populations and over time. We will also attempt to use this model to predict external criteria and outcomes, as well as comparing it to other instruments purporting to measure similar concepts. These efforts will culminate in an algorithm for the computation of higher order factors for the QLQ-C30.

Acknowledgments We gratefully acknowledge the many individuals who contributed datasets to this study. This work was funded by the EORTC Quality of Life Group and the Netherlands Cancer Institute, and carried out under the auspices of the EORTC Quality of Life Group. The authors thank the EORTC Headquarters and the various EORTC Clinical Cooperative Groups for permission to use the data from their trials for this research. Some of the results of this study were presented at the Annual Conference of the International Society for Quality of Life Research, New Orleans, USA, October 30, 2009.

Open Access This article is distributed under the terms of the Creative Commons Attribution Noncommercial License which permits any noncommercial use, distribution, and reproduction in any medium, provided the original author(s) and source are credited.

Appendix

Members of the EORTC Quality of Life Group

Cross-Cultural Meta-Analysis Project


Contributing groups

European Organization for Research and Treatment of Cancer (EORTC) Brain Cancer Group, EORTC Breast Cancer Group, EORTC Gastro-Intestinal Group, EORTC Genito-Urinary Group, EORTC Gynecological Group, EORTC Head and Neck Cancer Group, EORTC Leukemia Group, EORTC Lung Cancer Group, EORTC Lymphoma Group, EORTC Melanoma Group, EORTC Quality of Life Group, EORTC Radiotherapy Group, EORTC Soft Tissue Group, National Cancer Institute Grant CA60068, National Cancer Institute Radiotherapy Group, EORTC Soft Tissue Group, National Cancer Institute of Canada (NCIC) Clinical Trials Group, Swiss Group for Clinical Cancer Research (SAKK).

References

Further evidence for a two-dimensional structure. Evaluation and The World Health Professions, 28(2), 122–141.