Qual Life Res (2012) 21:1607–1617
DOI 10.1007/s11136-011-0082-6

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 / Published online: 21 December 2011

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

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

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

FA	Fatigue
HRQoL	Health-related quality of life
MIMIC	Multiple indicator, multiple cause
NV	Nausea and vomiting
PA	Pain
PF	Physical function
PROMIS	Patient-reported outcomes measurement information system
QLQ-C30	Quality of Life Questionnaire core 30 items
RF	Role function
RMSEA	Root mean square error of approximation
SL	Insomnia
SF	Social function
WHO	World Health Organization
WLSMV	Weighted least squares estimator with adjustment for means and variance

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 covary 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.

Fig. 1 Seven hypothesized models[a]: **a** standard model, **b** physical health, mental health and QL, **c** physical burden, mental function and QL, **d** symptom burden, function and QL, **e** HRQL and QL, **f** formative symptom burden (free weights), function and QL, **g** formative symptom burden (fixed weights) function and QL. [a]Models are described in text. Item thresholds, means, (error) variances, and correlations between first-order latent variables (in the standard model) are not represented, for clarity's sake

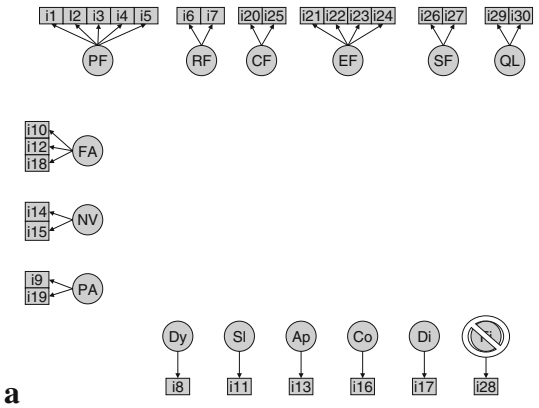
Estimators

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.

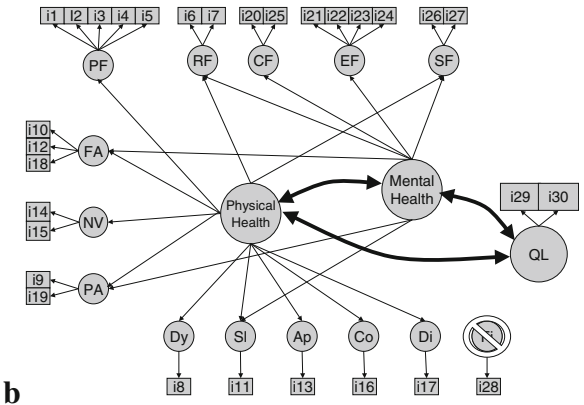
Tests of model fit and other fit indices

The χ^2 test of model fit was examined. The χ^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 ≥ 0.01 between (pairs of) TLIs/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.

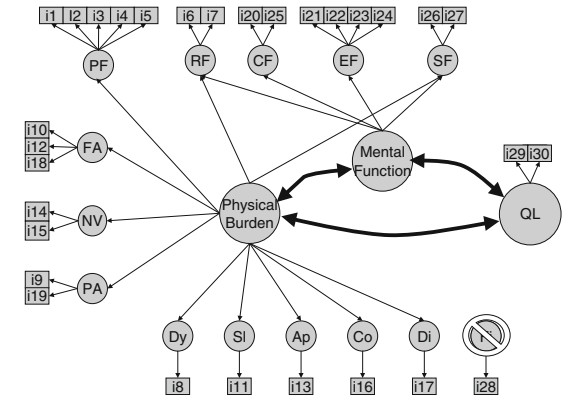
Direct comparisons of models by computing the differences between their respective chi-squares are not appropriate when using WLSMV estimators, and requires some


i1 i2 i3 i4 i5
i6 i7
i20 i25
i21 i22 i23 i24
i26 i27
i29 i30
PF
RF
CF
EF
SF
QL
i10 i12 i18
FA
i14 i15
NV
i9 i19
PA
Dy
Sl
Ap
Co
Di
i8
i11
i13
i16
i17
i28

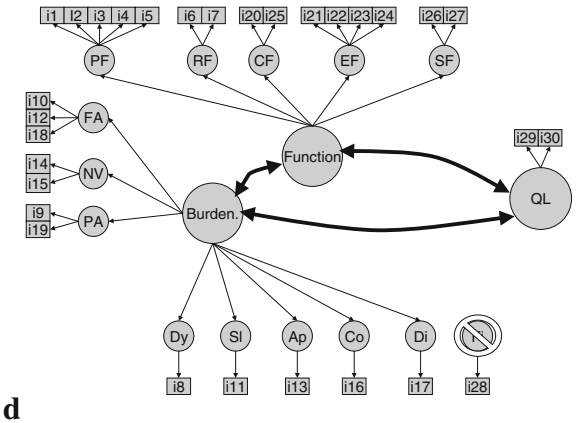
a


i1 i2 i3 i4 i5
i6 i7
i20 i25
i21 i22 i23 i24
i26 i27
PF
RF
CF
EF
SF
i10 i12 i18
FA
i14 i15
NV
i9 i19
PA
Physical Health
Mental Health
i29 i30
QL
Dy
Sl
Ap
Co
Di
i8
i11
i13
i16
i17
i28

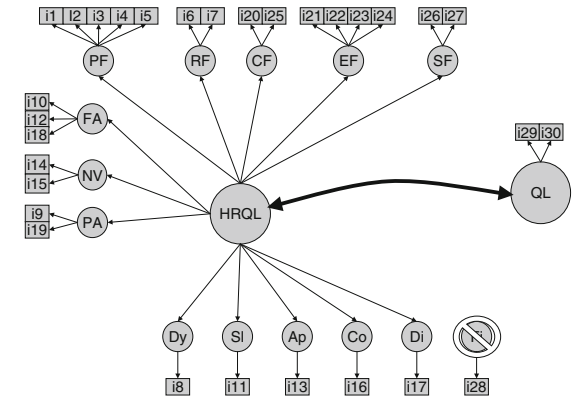
b


i1 i2 i3 i4 i5
i6 i7
i20 i25
i21 i22 i23 i24
i26 i27
PF
RF
CF
EF
SF
i10 i12 i18
FA
i14 i15
NV
i9 i19
PA
Physical Burden
Mental Function
i29 i30
QL
Dy
Sl
Ap
Co
Di
i8
i11
i13
i16
i17
i28

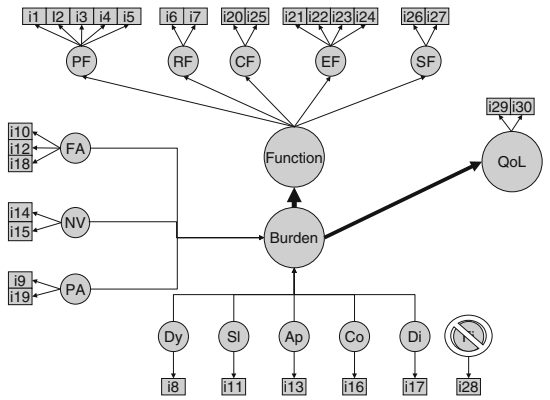
c


i1 i2 i3 i4 i5
i6 i7
i20 i25
i21 i22 i23 i24
i26 i27
PF
RF
CF
EF
SF
i10 i12 i18
FA
i14 i15
NV
i9 i19
PA
Burden.
Function
i29 i30
QL
Dy
Sl
Ap
Co
Di
i8
i11
i13
i16
i17
i28

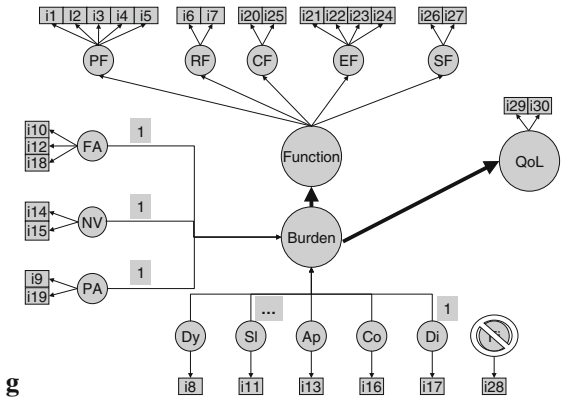
d


i1 i2 i3 i4 i5
i6 i7
i20 i25
i21 i22 i23 i24
i26 i27
PF
RF
CF
EF
SF
i10 i12 i18
FA
i14 i15
NV
i9 i19
PA
HRQL
i29 i30
QL
Dy
Sl
Ap
Co
Di
i8
i11
i13
i16
i17
i28

e


i1 i2 i3 i4 i5
i6 i7
i20 i25
i21 i22 i23 i24
i26 i27
PF
RF
CF
EF
SF
i10 i12 i18
FA
i14 i15
NV
i9 i19
PA
Function
Burden
i29 i30
QoL
Dy
Sl
Ap
Co
Di
i8
i11
i13
i16
i17
i28

f


i1 i2 i3 i4 i5
i6 i7
i20 i25
i21 i22 i23 i24
i26 i27
PF
RF
CF
EF
SF
i10 i12 i18
FA
1
i14 i15
NV
1
i9 i19
PA
1
Function
Burden
i29 i30
QoL
...
1
Dy
Sl
Ap
Co
Di
i8
i11
i13
i16
i17
i28

g

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 χ^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 ($N = 4{,}541$)

	Mean (SD)	% Missing
Age	59.6 (12.6)	9.9
	N	%
Gender		
Male	2,511	55.3
Female	1,906	42.0
Unknown	124	2.7
Stage		
I–III	1,846	40.7
IV–recurrent/metastatic	1,765	38.9
Unknown	930	20.5
Site		
Breast	663	14.6
Colorectal	245	5.4
Gynecological	375	8.3
Head and neck	801	17.6
Lung	610	13.4
Esophagus/stomach	822	18.1
Prostate	405	8.9
Other	620	13.7
Study type		
RCT	1,561	34.4
Non-RCT	1,455	32.0
Field study	1,386	30.5
Unknown	139	3.1
Country		
Belgium	193	4.3
Canada	120	2.6
France	266	5.9
Germany	477	10.5
Netherlands	228	5.0
Norway	498	11.0
Spain	402	8.9
Sri Lanka	438	9.6
Sweden	202	4.4
UK	722	15.9
USA	157	3.5
Other	838	18.5

Table 2 Tests[a] and approximate goodness-of-fit indices for various models

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

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

[a] WLSMV estimator on matrix of polychoric correlations, assuming ordinal items, with adjustment for cluster sampling. All latent error variances were free, with exception of single-item scales. Only one item loading was fixed for each scale, with the exception of the QL scale (in which both item loadings were fixed, equal to each other)

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 3 χ^2 Difference testing between 3 branches of nested models

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

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

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

Table 4 (Standardized) Regression weights for first-order factors and percentage variance explained by best fitting higher order model for each of three branches of (nested) models

First-order factors	Physical/mental health (model # 2)			Burden/function (model #4)			(free wgt.) Formative burden/function (model #6)		
	Physical	Mental	R^2	Burden	Function	R^2	(free) Formative burden	Function	R^2
PF	0.80[a]		0.64		0.76*	0.58		0.76[a]	0.59
RF	0.89*	0.04	0.84		0.89[a]	0.79		0.89*	0.80
EF		0.72[a]	0.52		0.62*	0.38		0.62*	0.38
CF		0.90*	0.82		0.80*	0.63		0.80*	0.62
SF	0.42*	0.46*	0.68		0.82*	0.67		0.82*	0.67
FA	0.82*	0.19*	0.93	0.97[a]		0.95	0.83[a]		NA
NV	0.66*		0.43	0.65*		0.42	0.04		NA
PA	0.60*	0.23*	0.62	0.79*		0.63	0.16*		NA
DY	0.80*		0.65	0.80*		0.64	0.03		NA
SL	0.05	0.77*	0.64	0.77*		0.59	0.08*		NA
AP	0.85*		0.72	0.84*		0.71	−0.08		NA
CO	0.75*		0.56	0.73*		0.54	0.04		NA
DI	0.62*		0.39	0.62*		0.38	−0.02		NA

PF physical function, *RF* role function, *CF* cognitive function, *EF* emotional function, *SF* social function, *FA* fatigue, *NV* nausea and vomiting, *PA* pain, *DY* dyspnea, *SL* insomnia, *AP* appetite loss, *CO* constipation, *DI* diarrhea

* $P < 0.01$

[a] Unstandardized weights were fixed to a value of 1.0, for purposes of model identification

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 χ^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, χ^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 *Mental Health/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.

Appendix

Members of the EORTC Quality of Life Group Cross-Cultural Meta-Analysis Project

Australia: M. King, S. Leutenegger, N. Spry; Austria: E. Greimel, B. Holzner; Belgium: A. Bottomley, C. Coens, K. West; Brazil: G. de Castro, C. de Souza; Canada: A. Bezjak, M. Whitehead; Denmark: M. Groenvold, M. Klee, M. Petersen; France: A. Bredart, T. Conroy, C. Rodary; Germany: M. Berend, B. Bestmann, M. Koller, O. Krauß, T. Kuchler, B. Malchow, R. Schwarz; Greece: K. Mystakidou; Iran: A. Montazeri; Italy: C. Brunelli, M. Tamburini; Japan: T. Matsuoka, H. Zhao; Netherlands: N. Aaronson, A. de Graeff, C. Gundy, R. de Leeuw, M.

Muller, M. Sprangers; Norway: K. Bjordal, E. Brenne, M. Hjermstad, M. Jordhøy, P. Klepstad, S. Sundstrøm, F. Wisløff; Singapore: Y. B. Cheung, S.B. Tan, J. Thumboo, H. B.Wong; South Korea: Y. H. Yun; Spain: J. Arraras; Sri Lanka: H. Jayasekara, L. Rajapakse; Sweden: M. Ahlner-Elmqvist; Switzerland: P. Ballabeni, J. Bernhard; Taiwan:W.-C. Chie; Turkey: U. Abacioglu; UK: J. Blazeby, J. Bruce, A. Davies, P. Fayers, L. Friend, Z. Krukowski, T. Massett, J. Nicklin, J. Ramage, N. Scott, A. Smyth-Cull, T. Young; USA: D. Cella, D.-L. Esseltine, C. Gotay, I. Pagano.

Contributing groups

European Organization for Research and Treatment of Cancer (EORTC) Brain Cancer Group, EORTC Breast Cancer Group, EORTC Chronotherapy 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 of Canada (NCIC) Clinical Trials Group, Swiss Group for Clinical Cancer Research (SAKK).

References

1. Aaronson, N. K., Ahmedzai, S., Bergman, B., Bullinger, M., Cull, A., Duez, N. J., et al. (1993). The European Organization for Research and Treatment of Cancer QLQ-C30: A quality-of-life instrument for use in international clinical trials in oncology. *Journal of the National Cancer Institute, 85*, 365–376.
2. Osoba, D., Aaronson, N. K., Zee, B., et al. (1997). Modification of the EORTC QLQ-C30 (version 2.0) based on content validity and reliability testing in large samples of patients with cancer. *Quality of Life Research, 6*, 103–108.
3. Aaronson, N. K., Cull, A., Kaasa, S., & Sprangers, M. A. G. (1996). The European Organization for Research and Treatment of Cancer (EORTC) modular approach to quality of life assessment in oncology: An update. In B. Spilker (Ed.), *Quality of life and pharmacoeconomics in clinical trials* (pp. 179–189). Philadelphia: Lippincott-Raven Publishers.
4. Fayers, P. M., Aaronson, N., Bjordal, K., Groenvold, M., Curran, D., Bottomley, A., et al. (2001). *EORTC QLQ-C30 scoring manual* (3rd ed.). Brussels: European Organization for Research and Treatment of Cancer.
5. Raykov, T. (2004). Behavioral scale reliability and measurement invariance evaluation using latent variable modeling. *Behavior Therapy, 35*(2), 299–331.
6. Ware, J. E., Kosinski, M., Gandek, B. G., Aaronson, N., Apolone, G., Bech, P., et al. (1998). The factor structure of the SF-36® Health Survey in 10 countries: Results from the International Quality of Life Assessment (IQOLA) Project. *Journal of Clinical Epidemiology, 51*(11), 1159–1165.
7. Guethlin, C., & Walach, H. (2007). MOS-SF 36: Structural equation modeling to test the construct validity of the second-order factor structure. *European Journal of Psychological Assessment, 23*(1), 15–23.
8. Horner-Johnson, W., Suzuki, S., Krahn, J. L., Andresen, J. M., & Drum, C. E. (2010). Structure of health-related quality of life among people with and without functional limitations. *Quality of Life Research, 19*, 977–984.
9. Boehmer, S., & Luszczynska, A. (2006). Two kinds of items in quality of life instruments: Indicator and causal variables in the EORTC QLQ-C30. *Quality of Life Research, 15*(1), 131–141.
10. McLachlan, S. A., Devins, G. M., & Goodwin, P. J. (1999). Factor analysis of the psychosocial items of the EORTC QLQ-C30 in metastatic breast cancer patients participating in a psychosocial intervention study. *Quality of Life Research, 8*(4), 311–317.
11. Van Steen, K., Curran, D., Kramer, J., Molenberghs, G., Van Vreckem, A., Bottomley, A., et al. (2002). Multicollinearity in prognostic factor analyses using the EORTC QLQ-C30: Identification and impact on model selection. *Statistics in Medicine, 21*(24), 3865–3884.
12. Ringdal, K. (1999). Assessing the consistency of psychometric properties of the HRQoL scales within the EORTC QLQ-C30 across populations by means of the Mokken Scaling Model. *Quality of Life Research, 8*, 25–43.
13. Gotay, C., Blaine, D., Haynes, S., Holup, J., & Pagano, I. (2002). Assessment of quality of life in a multicultural cancer patient population. *Psychological Assessment, 14*(4), 439–450.
14. Pagano, I., & Gotay, C. (2006). Modeling quality of life in cancer patients as a unidimensional construct. *Hawaii Medical Journal, 65*, 74–82.
15. Osoba, D., Zee, B., Pater, J., Warr, D., Kaizer, L., & Latreille, J. (1994). Psychometric properties and responsiveness of the EORTC quality of Life Questionnaire (QLQ-C30) in patients with breast, ovarian and lung cancer. *Quality of Life Research, 3*(5), 353–364.
16. Joreskog, K. (1993). Testing structural equation models. In K. Bollen & J. Long (Eds.), *Testing structural equation models* (pp. 294–316). Newbury Park, CA: Sage.
17. Scott, N. W., Fayers, P., Bottomley, A., Aaronson, N., de Graeff, A., Groenvold, M., et al. (2006). Comparing translations of the EORTC QLC-C30 using differential items functioning analyses. *Quality of Life Research, 15*, 1103–1115.
18. Scott, N. W., Fayers, P. M., Aaronson, N. K., Bottomley, A., de Graeff, A., Groenvold, M., et al. (2007). The use of differential item functioning analyses to identify cultural differences in responses to the EORTC QLQ-C30. *Quality of Life Research, 16*(1), 115–129.
19. Hjermstad, M. J., Fossa, S. D., Bjordal, K., & Kaasa, S. (1995). Test/retest study of the European Organization for Research and Treatment of Cancer Core Quality-of-Life Questionnaire. *Journal of Clinical Oncology, 13*, 1249–1254.
20. Rindskopf, D., & Rose, T. (1988). Some theory and application of confirmatory second order factor analysis. *Multivariate Behavioral Research, 23*, 51–67.
21. Koufteros, X., Babbarb, S., & Kaighobadi, M. (2009). A paradigm for examining second-order factor models employing structural equation modeling. *International Journal of Production Economics, 120*, 633–652.
22. Arnau, R., & Thompson, B. (2000). *Second order confirmatory factor analysis of the WAIS-III*. Assessment.
23. Edwards, J. R., & Bagozzi, R. P. (2000). On the nature and direction of relationships between constructs and measures. *Psychological Methods, 5*, 155–174.
24. Cella, D., Chang, C., Wright, B., Von Roenn, J., & Skeel, R. (2005). Defining higher order dimensions of self-reported health:

Further evidence for a two-dimensional structure. *Evaluation and The Health Professions, 28*(2), 122–141.
25. World Health Organization (WHO). (1946). *Constitution of the World Health Organization*. Geneva: WHO.
26. Hays, R. D., Bjorner, J., Revicki, D., Spritzer, K., & Cella, D. (2009). Development of physical and mental health summary scores from the patient-reported outcomes measurement information system (PROMIS) global items. *Quality of Life Research, 18*, 873–880.
27. Cella, D., Yount, S., Rothrock, N., Gershon, R., Cook, K., Reeve, B., et al. (2007). The Patient-Reported Outcomes Measurement Information System (PROMIS): Progress of an NIH Roadmap cooperative group during its first two years. *Medical Care, 45*(5 Suppl 1), S3–S11.
28. Wilson, I., & Cleary, P. (1995). Linking clinical variables with health related quality of life. A conceptual model of patient outcomes. *JAMA: Journal of the American Medical Association, 273*, 59–65.
29. Sousa, K. H., & Kwok, O. M. (2006). Putting Wilson and Cleary to the test: Analysis of a HRQOL conceptual model using structural equation modeling. *Quality of Life Research, 15*(4), 725–737.
30. Fayers, P. M., & Hand, D. J. (1997). Factor analysis, causal indicators and quality of life. *Quality of Life Research, 6*(2), 139–150.
31. Fayers, P. M., & Machin, D. (2000). *Quality of life: Assessment, analysis, and interpretation*. New York: Wiley.
32. Fayers, P. M., Hand, D. J., Bjordal, K., & Groenvold, M. (1997). Causal indicators in quality of life research. *Quality of Life Research, 6*(5), 393–406.
33. Fayers, P., & Hand, D. J. (2002). Causal variables, indicator variables, and measurement scales: An example from quality of life. *Journal of the Royal Statistical Society Series A, 165*(2), 233–261.
34. Howell, R., Breivik, E., & Wilcox, J. (2007). Reconsidering formative measurement. *Psychological Methods, 12*(2), 205–218.
35. Howell, R., Breivik, E., & Wilcox, J. (2007). Is formative measurement really measurement? Reply to Bollen (2007) and Bagozzi (2007). *Psychological Methods, 12*(2), 238–245.
36. Diamantopoulos, A., Riefler, P., & Roth, K. (2008). Advancing formative measurement models. *Journal of Business Research, 61*, 1203–1218.
37. MacKenzie, S., Podsakoff, P., & Jarvis, C. (2005). The problem of measurement model misspecification in behavioural and organizational research and some recommended solutions. *Journal of Applied Psychology, 90*(4), 710–730.
38. McDonald, R. (1999). *Test theory: A unified treatment*. London: Lawrence Erlbaum.
39. Hayduk, L. (1996). *LISREL: Issues, debates, and strategies*. Baltimore: Johns Hopkins Press.
40. Muthen, B., du Toit, S., & Spisic, D. (1997). *Robust inference using weighted least squares and quadratic estimating equation in latent variable modeling with categorical and continuous outcomes*. UCLA. http://www.gseis.ucla.edu/faculty/muthen/articles/Article_075.pdf. Accessed February 25, 2009.
41. Flora, D., & Curran, P. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods, 9*(4), 466–491.
42. Asparouhov, T., & Muthen, B. (2007). Computationally efficient estimation of multilevel high-dimensional latent variable models. In *Proceedings of the 2007 JSM meeting in Salt Lake City, Utah, Section on Statistics in Epidemiology* (pp. 2531–2535).
43. Barrett, P. (2007). Structural equation modelling: Adjudging model fit. *Personality and Individual Differences, 42*(5), 815–824.
44. Hayduk, L., Cummings, G., Boadu, K., Pazderka-Robinson, H., & Boulianne, S. (2007). Testing! Testing! One, two, three—testing the theory in structural equation models! *Personality and Individual Differences, 42*(5), 841–850.
45. McIntosh, C. (2007). Rethinking fit assessment in structural equation modelling: A commentary and elaboration on Barrett (2007). *Personality and Individual Differences, 42*(5), 859–867.
46. Bentler, P. (2007). On tests and indices for evaluating structural models. *Personality and Individual Differences, 42*(5), 825–829.
47. Mulaik, S. (2007). There is a place for approximate fit in structural equation modeling. *Personality and Individual Differences, 42*(5), 883–891.
48. Millsap, R. (2007). Structural equation modeling made difficult. *Personality and Individual Differences, 42*(5), 875–881.
49. Jackson, D., Gillaspy, J., & Purc-Stephenson, R. (2009). Reporting practices in confirmatory factor analysis: An overview and some recommendations. *Psychological Methods, 14*, 6–23.
50. Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*(1), 1–55.
51. Browne, M., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. Bollen & J. Long (Eds.), *Testing structural equation models* (pp. 136–162). Newbury Park, CA: Sage.
52. Cheung, G., & Rensvold, R. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling, 9*(2), 233–255.
53. Asparouhov, T., & Muthen, B. (2006). *Robust Chi square difference testing with mean and variance adjusted test statistics*. Mplus Web Notes: No. 10. http://www.statmodel.com/download/webnotes/webnote10.pdf. Accessed December 8, 2009.
54. Satora, A., & Bentler, P. M. (2001). A scaled difference Chi-square test statistic for moment structure analysis. *Psychometrika, 66*, 507–514.
55. Asparouhov, T., & Muthen, B. (4 A.D.). *Stratification in multivariate modeling*. Mplus Web Notes No. 9. Muthen & Muthen. http://www.statmodel.com/download/webnotes/MplusNote921.pdf. Accessed December 4, 2009.
56. Asparouhov, T., & Muthen, B. (2009). *Comparison of estimation methods for complex survey data analysis*. MPlus Technical Appendices. http://www.statmodel.com/download/SurveyComp21.pdf. Accessed December 3, 2009.
57. Asparouhov, T. (2006). General multilevel modeling with sampling weights. Communications in statistics. *Theory and Methods, 35*(3), 439–460.
58. Stapelton, L. (2006). An assessment of practical solutions for structural equation modeling with complex sample data. *Structural Equation Modeling, 13*(28), 58.
59. Muthen, L., & Muthen, B. (2007). *Mplus user's guide* (5th ed.). Los Angeles, CA: Muthén & Muthén.
60. Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika, 52*, 345–370.
61. Landsheer, J. A. (2010). The specification of causal models with Tetrad IV: A review. *Structural Equation Modeling, 17*(4), 703–711.
62. Liu, L. (2009). Technology acceptance model: A replicated test using TETRAD. *International Journal of Intelligent Systems, 24*(12), 1230–1242.