The St George's Respiratory Questionnaire revisited

Paap, Muirne C. S.; Brouwer, Danny; Glas, Cees A. W.; Monninkhof, Evelyn M.; Forstreuter, Benjamin; Pieterse, Marcel E.; van der Palen, Job

Published in:
Quality of Life Research

DOI:
10.1007/s11136-013-0570-y

Qual Life Res (2015) 24:67–79
DOI 10.1007/s11136-013-0570-y

QUANTITATIVE METHODS SPECIAL SECTION

The St George's Respiratory Questionnaire revisited: a psychometric evaluation

Muirne C. S. Paap · Danny Brouwer · Cees A. W. Glas · Evelyn M. Monninkhof · Benjamin Forstreuter · Marcel E. Pieterse · Job van der Palen

Accepted: 1 November 2013 / Published online: 16 November 2013

Abstract
Purpose The St George's Respiratory Questionnaire (SGRQ) has clearly acquired the status of legacy questionnaire for measuring health-related quality of life in patients with chronic obstructive pulmonary disease (COPD). The main aim of this study was to assess the underlying dimensionality of the SGRQ and to investigate the added value of the empirical weights used to calculate total scores.
Methods The official Dutch translation of the SGRQ was completed by 444 COPD patients participating in two clinical studies. These data were used for secondary data analysis in this study. Three complementary statistical methods were used to assess dimensionality: Mokken scale analysis (MSA), parametric multidimensional item response theory (IRT) and bifactor analysis. Additionally, the original SGRQ weighting procedure was compared to IRT-based weighting.
Results The results of the MSA and multidimensional item response theory (MIRT) pointed toward a unidimensional structure. The bifactor analyses indicated that there was a strong general factor, but the group factors did have additional value. Nineteen items performed poorly in the MSA, MIRT analysis or both. Shortening the scale from 50 to 31 items did not negatively impact measurement precision. SGRQ total score and IRT-derived scores correlated strongly, 0.90 for the one-parameter model and 0.99 for the two-parameter model.
Conclusion The SGRQ contains some multidimensionality, but an abbreviated version can be used as a unidimensional tool in patients with COPD. Subscale scores should be used with care. SGRQ total scores correlated highly with IRT-based scores, and thus, the weighting methods may be used interchangeably to calculate total scores.

Keywords Multidimensional item response theory · Bifactor analysis · SGRQ · Mokken scale analysis · COPD

Abbreviations

1PL	One-parameter logistic
2PL	Two-parameter logistic
CFI	Comparative fit index
COPD	Chronic obstructive pulmonary disease
ECV	Explained common variance
FA	Factor analysis

Electronic supplementary material The online version of this article (doi:10.1007/s11136-013-0570-y) contains supplementary material, which is available to authorized users.

M. C. S. Paap (✉) · C. A. W. Glas · J. van der Palen
Department of Research Methodology, Measurement, and Data-Analysis, Behavioral Sciences, University of Twente, P.O. Box 217, 7500, AE, Enschede, The Netherlands
e-mail: m.c.s.paap@utwente.nl

D. Brouwer
Department of Psychometrics and Statistics, Behavioral Sciences, University of Groningen, Groningen, The Netherlands

E. M. Monninkhof
Julius Center for Health and Primary Care, University Medical Center Utrecht, Utrecht, The Netherlands

B. Forstreuter
University of Twente, Enschede, The Netherlands

M. E. Pieterse
Department of Psychology, Health and Technology, Behavioral Sciences, University of Twente, Enschede, The Netherlands

J. van der Palen
Medical School Twente, Medisch Spectrum Twente, Enschede, The Netherlands

GA	Genetic algorithm
GPCM	Generalized partial credit model
HRQoL	Health-related quality of life
HS	Health status
IRT	Item response theory
MIRT	Multidimensional item response theory
MSA	Mokken scale analysis
PCA	Principal component analysis
SGRQ	St George's Respiratory Questionnaire
SGRQ-C	COPD-specific version of the St George's Respiratory Questionnaire
TLI	Tucker-Lewis index
QoL	Quality of life
RSMEA	Root mean square error of approximation
VAS	Visual analogue scale

Introduction

Since its introduction over two decades ago, the St George's Respiratory Questionnaire (SGRQ) has become one of the most widely used disease-specific quality of life (QoL) questionnaires in chronic obstructive pulmonary disease (COPD); it has clearly acquired the status of legacy questionnaire [1–3]. The SGRQ consists of a combination of yes/no and Likert type questions, which are grouped in subscales.

Both the SGRQ total score and subscale scores (Symptoms, Activity, Impacts) are widely used. However, the number of papers published on the factorial structure of the SGRQ is few in number. In one of the first publications about the SGRQ, Jones et al. [2] state that the SGRQ contained 76 items, and a principal component analysis (PCA) had been used to partition the questionnaire into three sections. Unfortunately, no details are provided regarding the PCA or how the authors came to the three-factorial solution. A little over a decade ago, the American translation of the SGRQ was introduced [4]; several aspects of validity were documented, and some important changes were made. For example, the recall period was modified from 1 year to 4 weeks. However, the factorial structure was not investigated. In a more recent study [5], the COPD-specific version (SGRQ-C) was introduced. To our knowledge, this was the first paper that employed a form of item response theory (IRT) [6], to validate, improve and shorten the SGRQ. The authors used a Rasch model [7], a one-parameter IRT model. They examined the fit of each item in relation to the subscale it belonged to. Hence, also this validation study did not investigate the factorial structure. All in all, there is little evidence to support the proposed three-factor structure.

Over the years, the SGRQ has been referred to as a QoL, health status (HS) or health-related QoL (HRQoL) instrument, among other things. Upon inspection of the questionnaire, it becomes clear that the items in the SGRQ can be considered a symptom/state checklist. The QoL aspect is contained by the weights. When calculating subscale scores or the total score, each answering category is multiplied by a specific "empirically derived" weight, which can be found in the manual [8]. The first papers published on the SGRQ provide some information regarding the development of the weights. Patients with asthma [3, 9] or COPD [2] were presented with each question/category in the SGRQ and asked to rate the distress they would experience if that state would be applicable to them. They rated this hypothetical distress on a 10-cm visual analogue scale (VAS), with end points "no distress" and "maximum imaginable distress." Henceforth, the average distress scores obtained from these patients were used as weights. Jones et al. [1] stated that "…the development of the SGRQ has shown that it is possible to produce a standardized measure of impaired health covering a range of disturbances to health and perceived well-being in patients…."

Although their approach was firmly based in empirical research, resulting in weights that were based on patients' perspectives, it is assumed (by using the weights) that the distress associated with a symptom/state is the same for every patient. The authors reported large individual differences in distress scores associated with a given symptom/state, so the assumption clearly does not hold [3]. On the other hand, using the same weights for all patients allows for a straightforward comparison of scores. From a practical point of view, this is a very compelling argument. However, standardized weights could also be obtained by using factor analysis or item response theory (IRT) models. It remains unclear at this point, whether the empirical weights should be preferred over psychometrically derived weights.

Our motivation to evaluate the psychometric properties of the SGRG is twofold. Firstly, we are interested in the SGRQ as a stand-alone questionnaire, since it is often used in this way in research and clinical practice. Secondly, we want to establish its psychometric properties on item level, since we are considering using the items in a computerized adaptive test (CAT) to measure QoL in patients with COPD. Our main aim is to investigate two important aspects of construct validity. Our research questions are:

1. Is the underlying dimensionality best described by a one-factor, a multiple-factor solution or are both equally viable?
2. Do the empirical weights have added value over psychometrically derived weights?

Table 1 Descriptive statistics for the two samples used in this study

	COPE	SMOKE
N	242	202
Male (%)	206 (85.1)	110 (54.5)
Age in years (mean ± SD)	65.6 ± 7.2	58.2 ± 8.6
FEV1 % predicted in l (mean ± SD)[a]	56.3 ± 15.2	65.5 ± 27.0
SGRQ total score (mean ± SD)	37.9 ± 17.1	41.5 ± 18.1
SGRQ Symptoms (mean ± SD)	48.0 ± 22.1	51.5 ± 22.7
SGRQ Activity (mean ± SD)	51.4 ± 23.1	54.9 ± 23.0
SGRQ Impact	26.8 ± 16.7	30.6 ± 18.5
Correlations SGRQ subscales		
Symptoms, activity	0.52	0.50
Symptoms, impact	0.59	0.59
Activity, impact	0.73	0.73

[a] Based on 239 cases for the COPE study and 181 for the SMOKE study *SGRQ* St George's Respiratory Questionnaire, *FEV1%* forced expiratory volume in 1 s as percent predicted for age, gender and height

Methods

Participants

We used data from two prior studies (the COPE I study and SMOKE study) involving Dutch patients with COPD for secondary data analysis. The second phase of the COPE I study [10] was a randomized, open, parallel-group single center study that compared the effect of a self-management program for patients with COPD with regular care. Follow-up was 12 months. At baseline of phase 2, 242 patients filled out the SGRQ (see Table 1). The main focus of the SMOKE study [11] was to compare the effectiveness of a newly developed smoking cessation intervention, SmokeStop-Therapy (SST), with the "Minimal Intervention Strategy for Lung Patients" (LMIS). SMOKE was a randomized controlled multicenter trial with 1-year follow-up. At baseline, all 202 patients filled out the SGRQ (see Table 1). Note that the data of these studies were pooled prior to analysis.

Measures: the SGRQ

The SGRQ consists of 50 items, of which 11 are scored on a Likert scale and 39 dichotomously. Both a total score and three subscale scores (Symptoms, 8 items; Activity, 16 items; Impacts, 26 items) are usually calculated. The Symptoms subscale contains a recall period, which differs among versions (countries). For this study, the official Dutch translation of the SGRQ was used, which has a recall period of 4 weeks. The total and subscale scores are derived by using "empirical weights." The score is then calculated by dividing the sum of the weights for all positive answers in the questionnaire/subscale by the sum of all weights for all items in the questionnaire/subscale and multiplying this number by 100.

Statistics

The dimensionality of the SGRQ was assessed using three complementary statistical methods: Mokken scale analysis (MSA), (multidimensional) IRT and bifactor analysis. A more detailed description of these methods, along with their respective strengths and weaknesses, can be found in the online supplement. Note that the analytic strategy was defined prior to viewing the data set.

Mokken scale analysis [12, 13] was applied using the *R* [14] package *Mokken* [15]. MSA is a non-parametric type of IRT analysis. In recent years, MSA has been increasing in popularity in QoL, psychiatric and medical research [e.g., 16–24]. MSA can be used to investigate the dimensionality (factorial structure) of the data and at the same time identifies scales that allow an ordering of individuals on an underlying one-dimensional scale using the unweighted sum of item scores. In order to determine which items cluster and form a scale, scalability coefficients are calculated. Similar to the item-rest correlation, the scalability coefficient (H) expresses the degree to which an item is related to other items in the scale. The scalability coefficient can be seen as a "corrected" correlation: The correlation between items is divided by the maximum expected correlation given the items' marginal score-frequency distributions. A scale is considered acceptable if $0.3 \leq H < 0.4$, good if $0.4 \leq H < 0.5$ and strong if $H \geq 0.5$ [12, 13].

We started with a confirmatory analysis. First, the total scale was analyzed, and subsequently, the three subscales were analyzed separately. Then, exploratory analyses were performed using the newly developed genetic algorithm (GA) that aims to find the optimal partitioning into Mokken scales by maximizing an objective function [25]. This function closely follows Mokken's intention that the first selected cluster contains the maximum number of items, followed by the second cluster and so on. Hence, the function reflects that an extra item in the first scale is more important than the number of items in the subsequent shorter scales. Following Sijtsma and Molenaar [13], we ran the analysis several times in a row, each time increasing the lower bound scalability coefficient (also known as the user-specified constant, c). The resulting sequence of outcomes indicates whether the data set are one-dimensional or multidimensional [13].

Parametric IRT models have the same basic assumptions as Mokken models: unidimensionality, monotonicity and local independence [26]. The main difference is that

parametric IRT models are more restrictive than Mokken models with respect to the shape of the item characteristic curve, but have the advantage of allowing the item locations and estimated trait levels to be placed on an interval scale. Furthermore, multidimensional models can be estimated when using parametric IRT; this is not possible with MSA. In this study, the specific model used was the generalized partial credit model [GPCM; 27] for polytomous items and the extended multidimensional version of the GPCM. We investigated whether a model based on the three subscales which also took into account the correlation among the scales was to be preferred over a unidimensional solution, based on model fit statistics. Marginal maximum likelihood estimation was used. Model fit was ascertained by computing absolute differences between expected and observed item scores for high, average and low scoring individuals. An absolute difference smaller than 0.10 was interpreted as sufficient item fit [cf. 28, 29]. The parametric IRT analyses were applied using the software package MIRT [30].

Reise et al. [31] demonstrated that for quality of life questionnaires, bifactor analysis can provide additional evaluations that complement traditional dimensionality investigation [see also 32, 33]. Gustafsson and Åberg-Bengtsson [34] discuss the differences between a bifactor model and a correlated trait or higher-order model. The main distinguishing feature of the bifactor model is that the items load on both the general factor and so-called group factors. It can therefore be used to investigate to what degree item variance is due to a general factor or to specific group factors. When the loadings on the group factors are large as compared to the factor loadings on the general factor, this means that there are non-ignorable sources of variance that can be attributed to different constructs other than the general construct. If this is the case, subscales for specific symptom groups need to be considered. In this study, we used both exploratory and confirmatory bifactor analysis to evaluate the extent to which items loaded onto specific (group) factors when their relationship with the main factor was accounted for. For each model, we calculated the percentage of explained common variance (ECV) that was attributable to the general factor and to group factors [33]. For each factor, the ECV is the sum of squared factor loadings for that factor divided by the sum of all squared factor loadings (the common variance) for the model. Reise et al. [35] demonstrated that when the ECV for the general factor in a bifactor model is larger than 60 %, the factor loading estimates for a unidimensional model are close to the true loadings on the general factor in the bifactor model and can be interpreted as one construct.

We performed exploratory bifactor analysis for a two- and three-factor solution using the Schmid-Leiman procedure [for an explanation of this procedure see 33, 36]. We used a polychoric correlation matrix with the *Schmid* routine included in the *psych* package [37] of the *R* software program [14]. The confirmatory one- and bifactor models were estimated using MPLUS 4.1 [38], and the mean and variance-adjusted weighted least squares estimation was used for all calibrations. We used the following fit indices and rules-of-thumb: the comparative fit index (CFI), good fit if CFI $\geq$0.95 and acceptable fit if CFI is between 0.90 and 0.95; the Tucker-Lewis index (TLI), good fit if TLI $\geq$0.90, and the root mean square error of approximation (RSMEA), good fit if RSMEA $\leq$0.06, acceptable fit if RMSEA is between 0.06 and 0.08 [39, 40].

To investigate whether a weighted score was to be preferred over an unweighted score, one-parameter IRT models were compared to two-parameter IRT models. To gain more insight into the construct being measured by the SGRQ, the relationships among the SGRQ total score and theta scores based on a one-parameter (i.e., unweighted) versus a two-parameter (i.e., weighted) model were evaluated.

Results

Descriptive statistics for both samples can be found in Table 1. The patients in the SMOKE study were younger, had better lung function (higher FEV1 %) and were mostly women. The mean differences in SGRQ scores between two samples did not exceed the minimal important difference (MID) of 4 [41, 42], and the pattern of correlations among the subscales was highly comparable. Note that 2 % of the cells in the total data set were missing. Most of these cells pertained to questions that could be skipped if they were not applicable to the patient. The missings were coded as 0 prior to the analyses.

Dimensionality analyses

MSA

The *H* values for the total scale, as well as the Activity and Symptoms subscales were acceptable to good (0.307, 0.635 and 0.341, respectively). However, the Impact scale had an *H*-value of 0.268, which is considered too low. Running exploratory analyses for increasing values of *c* indicated a unidimensional pattern: Most items were placed in the first scale, some in a second scale and an increasing number was discarded. To obtain the "optimal" partitioning, we set the *c*-value to 0.3 and ran the GA 10 times. We selected the solution (see Table 2) with the highest value for the objective function (0.6612). This led to 17 items being

Table 2 Results of IRT analyses

Item no.	Description	SGRQ scale	MSA scale	H_i	LM 2PLM	Dif. 2PLM	LM multi	Dif. multi	Problems in MSA or IRT[a]?
Problematic items									
1	Cough	S	2	0.26	35.50	0.14	0.03	0.01	Both
2	Phlegm	S	2	0.27	22.71	0.13	0.11	0.02	Both
4	Wheezing	S	1	0.27	20.65	0.10	0.29	0.03	IRT
5	Chest trouble	S	1	0.34	20.86	0.14	13.01	0.05	IRT
6	Worst attack	S	0	0.25	16.15	0.15	–	0.11	Both
8	Wheeze morning	S	0	0.16	9.19	0.03	0.14	0.01	MSA
10	Employment	I	0	0.18	2.34	0.03	1.63	0.03	MSA
11	Sitting, lying still	A	0	0.23	–	0	–	0.01	MSA
16	Breathless hills	A	0	0.48	2.95	0.01	1.45	0.01	MSA
20	Breathless talk	I	0	0.25	0.00	0	–	0.01	MSA
22	Sleep disturbed	I	2	0.25	7.53	0.03	–	0.03	MSA
24	Embarrassing	I	0	0.25	0.19	0.01	0.89	0.01	MSA
26	Panic	I	0	0.24	0.14	0	0.80	0.01	MSA
28	No hope	I	0	0.14	2.88	0.03	2.37	0.03	MSA
32	Meds 1	I	0	0.19	1.76	0.01	0.80	0.01	MSA
33	Meds 2	I	0	0.12	1.43	0.01	2.42	0.01	MSA
34	Meds 3	I	0	0.10	4.98	0.02	2.06	0.02	MSA
45	Cannot sports	I	0	0.31	3.48	0.03	1.24	0.02	MSA
46	Cannot recreate	I	0	0.21	6.38	0.02	1.37	0.02	MSA
Unproblematic items									
3	Short of breath	S	1	0.40	19.28	0.09	9.72	0.13	
7	Good days	S	1	0.27	5.65	0.06	–	0.06	
9	Chest condition	I	1	0.38	8.97	0.04	2.14	0.05	
12	Breathless wash	A	1	0.33	0.30	0	6.53	0.01	
13	Breathless walk 1	A	1	0.40	–	0.01	17.77	0.02	
14	Breathless walk 2	A	1	0.36	1.47	0.01	108.81	0.03	
15	Breathless stairs	A	1	0.40	6.95	0.02	–	0	
17	Breathless sports	A	1	0.43	0.35	0	0.26	0.01	
18	Cough hurts	I	1	0.30	–	0.01	0.31	0	
19	Cough tired	I	1	0.35	15.04	0.03	3.33	0.03	
21	Breathless bend	I	1	0.28	9.05	0.04	2.82	0.03	
23	Exhausted	I	1	0.45	2.74	0.02	–	0.03	
25	Nuisance family	I	1	0.30	6.56	0.02	3.42	0.01	
27	Not in control	I	1	0.29	2.29	0.02	1.54	0.01	
29	Invalid	I	1	0.36	–	0.01	14.23	0.02	
30	Not safe	I	1	0.26	5.13	0.03	9.80	0.03	
31	Effort	I	1	0.37	–	0.03	24.25	0.04	
35	Meds 4	I	1	0.32	–	0.01	0.36	0	
36	Slow wash	A	1	0.36	8.30	0.02	0.98	0.01	
37	Slow bath	A	1	0.32	4.83	0.02	1.27	0.02	
38	Slow walk	A	1	0.43	0.49	0.01	8.45	0.03	
39	Slow housework	A	1	0.39	0.93	0.01	3.09	0.02	
40	Slow stairs	A	1	0.37	21.91	0.03	–	0.03	
41	Slow hurry	A	1	0.47	21.87	0.04	–	0.03	
42	Activities 1	A	1	0.45	44.01	0.04	2.17	0.02	
43	Activities 2	A	1	0.57	11.68	0.02	1.18	0.01	
44	Activities 3	A	1	0.66	–	0	10.10	0.01	

Table 2 continued

Item no.	Description	SGRQ scale	MSA scale	H_i	LM 2PLM	Dif. 2PLM	LM multi	Dif. multi	Problems in MSA or IRT[a]?
47	Cannot shop	I	1	0.48	–	0	64.63	0.01	
48	Cannot housework	I	1	0.35	–	0.01	6.43	0.02	
49	Cannot move	I	1	0.56	–	0	52.17	0	
50	Doing things	–	1	0.29	1.35	0.02	–	0.01	

For the Mokken Scale Analysis (MSA) the final scale solution is shown, along with item scalability coefficients. For the parametric IRT analyses, item fit statistics for the unidimensional 2PL IRT model and the multidimensional model are shown. The last column indicates in which analysis the item showed bad item fit

The H_i values are based on a confirmatory MSA; $H = 0.31$; the numbers in the "MSA scale" column refer to the number of the scale the item was designated to (0 indicates that the item was excluded from any of the scales)

SGRQ St George's Respiratory Questionnaire, *MSA* Mokken scale analysis, H_i item scalability coefficients, *LM* Lagrange multiplier statistics, *Dif.* average absolute difference between observed and expected scores, *2PLM* unidimensional two-parameter logistic model, *multi* multidimensional 2PLM

[a] Based on unidimensional IRT analysis (columns "LM 2PLM" and "Dif. 2PLM")

excluded from the scale. The *H*-value of the resulting unidimensional scale equaled 0.407.

MIRT

The fit on item level was acceptable for most items under both the multidimensional and the unidimensional model; however, items 1, 2, 4 and 5 showed better fit under the multidimensional model compared to the unidimensional one (see Table 2). Correlations among the three dimensions were moderate to high: 0.538, 0.686 and 0.797 for Symptoms and Impact, Symptoms and Activity, and Impact and Activity, respectively. Furthermore, correlations between theta-estimates on the subscales for the model taking into account the relationship among the subscales (the multidimensional model) with the respective theta-estimates based on a model, where this relationship was not taken into account, were very high (>0.90). For five items, the absolute difference between expected and observed item scores was larger than 0.10. Upon closer inspection, it was found that the item difficulty thresholds for these items were not logically ordered.

Bifactor

The confirmatory factor analyses showed that the bifactor model had a good fit (CFI = 0.916, TLI = 0.954, RMSEA = 0.060), whereas the one-factor model had a poor fit (CFI = 0.802, TLI = 0.885, RMSEA = 0.095). The general factor showed a high ECV (68 %). The three subscales Activity, Impact and Symptoms explained 10, 10 and 12 of the common variance, respectively. Items 43, 44 and 47–49 had to be removed from the analysis since they showed very high correlations with other items and caused computational difficulties. Of the 19 items that performed badly in the MSA/(M)IRT analyses, nine items exhibited group factor loadings that exceeded their loading on the general factor (see Table 3).

Exploratory bifactor analyses resulted in a general factor with ECV = 48 % accompanied by 2 or 3 rather strong group factors. Note that the ECV for the general factor in the exploratory bifactor model was lower because items were allowed to cross-load on multiple group factors. In the exploratory bifactor model with two group factors (see Table 3), the items in the first group factor mostly related to impact of breathing symptoms on moderate to heavy activities (ECV = 32 %) and the second group factor to impact of breathing symptoms on light activities (ECV = 20 %). The explanatory three-factor bifactor solution (data not shown) more or less resembled the two-factor solution, with the exception that the second factor was now divided into two factors: The third factor (ECV = 11 %) seems to represent more specific pain or symptoms (due to the treatment).

Dimensionality and measurement precision of the "good" items

The 31 items that showed good psychometric properties in both the MSA and IRT analyses could be used as a unidimensional scale. This shorter scale contained only slightly less information than the full scale (see Fig. 1). Moreover, when the confirmatory factor analyses were repeated for this subset of items, it was found that both the one-factor and the bifactor models showed good fit, although the bifactor model showed superior fit (ECV general factor = 75 %). Exploratory bifactor analyses resulted in similar group factors as the analyses performed on the full item set.

Table 3 Results of the factor analyses (loadings)

Item no.	One-factor model	Confirmatory bifactor model				Exploratory bifactor model		
		GF	F1	F2	F3	GF	F1	F2
1	0.51	0.23	**0.78**			0.22		
2	0.53	0.27	**0.76**			0.26		
3	**0.70**	**0.65**	0.34			0.48	0.46	
4	0.49	0.37	**0.51**			0.29		0.23
5	**0.64**	0.51	**0.58**			0.44	0.29	0.22
6	0.51	0.39	**0.51**			0.31	0.24	
7	0.48	0.46	0.19			0.37	0.27	
8	0.29	0.22	0.33			0.21		0.28
9	**0.67**	**0.65**			0.26	0.47	0.46	
10	0.36	0.39			−0.02	0.20	0.28	
11	0.37	0.29		**−0.59**		0.26		0.44
12	**0.72**	**0.68**		−0.44		0.57	0.22	0.44
13	**0.77**	**0.69**		**−0.64**		**0.62**		**0.64**
14	**0.75**	**0.75**		−0.29		0.58	0.34	0.33
15	**0.74**	**0.78**		0.25		0.47	0.53	
16	**0.79**	**0.80**		0.57		0.40	**0.68**	0.22
17	**0.68**	**0.69**		0.40		0.41	0.55	
18	0.50	0.32			**0.72**	0.32		0.48
19	**0.63**	0.52			**0.57**	0.47		0.40
20	0.51	0.46			0.31	0.34	0.23	
21	0.56	**0.60**			0.00	0.43	0.41	
22	0.48	0.32			**0.67**	0.33		0.31
23	**0.78**	**0.77**			0.20	0.52	0.54	
24	0.51	0.37			**0.61**	0.37		0.24
25	0.57	0.50			0.42	0.41	0.27	0.20
26	0.52	0.47			0.33	0.41	0.24	0.24
27	0.55	0.47			0.43	0.39	0.23	0.22
28	0.28	0.36			−0.22	0.21	0.37	
29	**0.72**	**0.74**			0.06	0.52	0.43	
30	**0.60**	**0.68**			−0.23	0.38	0.49	
31	**0.78**	**0.81**			0.06	**0.62**	0.36	0.36
32	0.35	0.33			0.19	0.26		0.23
33	0.27	0.23			0.22	0.27		
34	0.19	0.12			**0.31**			0.35
35	0.55	0.49			0.35	0.52		0.56
36	**0.81**	**0.76**		−0.48		**0.68**	0.39	0.39
37	**0.73**	**0.67**		−0.51		0.59	0.27	0.41
38	**0.85**	**0.88**		−0.11		**0.69**	**0.62**	
39	**0.80**	**0.81**		−0.21		**0.64**	0.44	0.30
40	**0.72**	**0.76**		0.17		0.51	0.55	
41	**0.83**	**0.86**		0.13		0.60	**0.64**	
42	**0.79**	**0.82**		0.23		0.54	**0.63**	
43						0.60	**0.70**	
44						0.55	**0.78**	
45	**0.62**	**0.69**			−0.17	0.48	0.39	
46	0.43	0.47			−0.01	0.36		0.36
47						0.49		0.58

Table 3 continued

Item no.	One-factor model	Confirmatory bifactor model				Exploratory bifactor model		
		GF	F1	F2	F3	GF	F1	F2
48						0.52	0.21	0.40
49						0.41		**0.66**
50	0.56	0.59				0.45	0.48	

Loadings that are higher than 0.60 are printed in bold, to indicate which items showed best discrimination on the general factor; to aid interpretation, item's loadings on group factors that were larger than that item's loading on the general factor are also printed bold

GF general factor, *F1-F3* subgroup factors

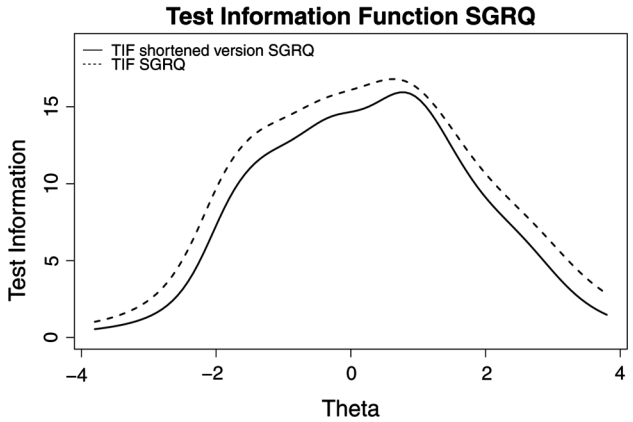

Fig. 1 Test information functions (TIFs) based on unidimensional 2PL models, with information on the *y*-axis and estimated theta values on the *x*-axis. The *black line* shows the TIF for the total scale (50 items) and the *dashed line* for the shortened scale (31 items)

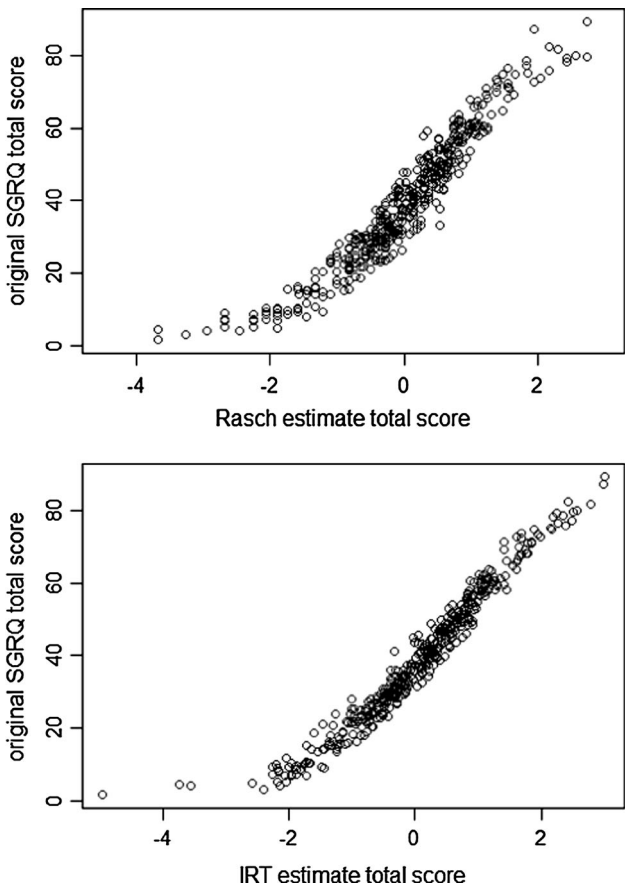

Fig. 2 Scatterplots of IRT theta-estimates (*x*-axis) by the original SGRQ total score (*y*-axis). The *upper panel* shows the Rasch (1PL) scores on the *x*-axis and the *lower panel* the 2PL scores

Weights

The fit on item level was better for the 2PL model than for the 1PL model, both for the total scale (50 items) as the shorter one (31 items).[1] Correlations between theta-estimates of the 1PL and 2PL model, the full scale and the shorter one, and theta-estimates and original SGRQ-scores were all above 0.90. As illustrated in Fig. 2, the correlation between the original SGRQ-scores and the 1PL estimates was slightly weaker (0.90) than the correlation between the SGRQ-scores and the 2PL estimates (0.99). When inspecting the correlations among the discrimination parameters, difficulty parameters and the SGRQ category weights, we found correlations of the item parameters with the SGRQ weights of 0.45 and 0.40 for discriminations[2] and difficulty parameters, respectively, and a correlation between the two item parameters of -0.15. The values of the item parameters estimated under the unidimensional 2PL IRT model, as well as the SGRQ weights can be found in Table 4.

Discussion

We investigated the factorial structure of the SGRQ using three complementary statistical methods. The findings of the MSA and IRT/MIRT analyses provided clear support for unidimensionality. The bifactor analyses indicated that

[1] Total scale: nine items showed poor fit under the 1PL model, five under the 2PL model. Shorter scale: 3 items showed poor fit under the 1PL model, 0 under the 2PL model.

[2] To obtain the discrimination parameter per category, the parameter estimate was multiplied by category number for polytomous items. This value was then used in calculating the correlations.

Table 4 Item parameters under unidimensional 2PL IRT model and SGRQ weights

Item no.	Response category	Parameter		SGRQ weight
		A	B	
1	1	0.27	−4.31	28.1
1	2		0.11	29.3
1	3		0.46	63.2
1	4		−2.23	80.6
2	1	0.28	−0.93	30.2
2	2		0.58	34.0
2	3		−0.13	60.0
2	4		−1.61	76.8
3	1	0.86	−2.13	35.7
3	2		−0.12	43.7
3	3		−0.02	71.4
3	4		−0.09	87.2
4	1	0.37	−0.02	36.4
4	2		1.32	45.6
4	3		1.54	71.0
4	4		0.06	86.2
5	1	0.49	1.59	44.2
5	2		0.14	60.3
5	3		1.96	73.5
5	4		−1.03	86.7
6	1	0.31	1.28	41.9
6	2		2.13	58.8
6	3		0.94	73.5
6	4		0.54	89.7
7	1	0.44	−3.64	15.4
7	2		1.16	61.5
7	3		0.42	76.6
7	4		1.22	93.3
8	1	0.44	1.44	62.0
9	1	1.04	−1.73	34.6
9	2		0.86	82.5
9	3		0.87	83.2
10	1	0.38	4.92	77.6
10	2		−0.50	88.9
11	1	0.76	3.93	90.6
12	1	1.75	0.73	82.8
13	1	2.23	1.47	80.2
14	1	2.07	0.70	81.4
15	1	1.78	−0.72	76.1
16	1	1.98	−1.45	75.1
17	1	1.50	−1.58	72.1
18	1	1.02	2.72	81.1
19	1	1.27	0.21	79.1
20	1	0.98	0.87	84.5
21	1	1.15	0.04	76.8
22	1	0.84	1.05	87.9
23	1	2.22	−0.45	84.0
24	1	0.90	1.00	74.1
25	1	1.15	1.23	79.1
26	1	1.02	1.21	87.7
27	1	1.17	1.18	90.1
28	1	0.48	−0.42	82.3
29	1	1.84	1.04	89.9
30	1	1.22	0.99	75.7
31	1	2.32	0.71	84.5
32	1	0.67	2.53	88.2
33	1	0.45	2.97	53.9
34	1	0.30	7.57	81.1
35	1	1.31	2.30	70.3
36	1	2.28	1.04	74.2
37	1	1.79	1.41	81.0
38	1	2.71	−0.30	71.7
39	1	2.40	0.06	70.6
40	1	1.78	−0.68	71.6
41	1	2.54	−0.86	72.3
42	1	2.32	−1.02	74.5
43	1	3.30	−1.34	71.4
44	1	3.26	−1.80	63.5
45	1	1.31	−0.05	64.8
46	1	0.91	2.52	79.8
47	1	2.07	2.50	81.0
48	1	1.75	1.88	79.1
49	1	2.40	2.92	94.0
50	1	0.91	0.14	42.0
50	2		1.16	84.2
50	3		2.61	96.7

A discrimination parameter, *B* threshold parameter

there was a strong general factor, but the SGRQ did contain some multidimensionality; however, the existing subscales do not seem to capture it very well. Nineteen items showed inferior psychometric properties, based on the MSA and IRT results. Removing these items from the analyses did not negatively impact measurement precision. Users may consider using an abbreviated version instead of the full-length instrument, either based on this study or a recent study by Meguro et al. [5]. Furthermore, the "empirically derived" weights used to calculate SGRQ-scores showed a strong relationship to the IRT item parameters, so we suggest SGRQ total scores and theta-estimates based on a 2PL model can be used interchangeably.

The SGRQ has a long-standing reputation of being a valid and reliable measurement tool in many different

languages [43–51], but hardly any evidence regarding its factorial structure could be found. No details regarding the PCA of the SGRQ subscales were found [2], and only three studies aimed to replicate the proposed three-factor solution [52–54]. Rutten-van Moken et al. and Yu et al. focused solely on PCA, a method that has many drawbacks when used for rating scale/dichotomous data [e.g., 54–57]. The sample by Rutten-van Molken and colleagues consisted of 133 Dutch COPD patients and the sample used by Yu and colleagues of 54 COPD patients from Hong Kong. In both studies, more than 10 components were found based on the Eigenvalue > 1 criterion. The first three components explained 52 % of the variance in the study by Yu and colleagues and only 28 % in the study by Rutten-van Molken and colleagues. Yu and colleagues found these results rather surprising and speculated that the sample sizes of aforementioned studies may have been too small to obtain reliable results. A PCA on our data resulted in similar findings to those reported by Rutten-van Molken et al. and Yu et al. Therefore, it seems unlikely that the sample size is to blame. In his doctoral thesis, Karpinski [54] investigated the factorial structure of the SGRQ by means of a scale- and item-analysis that closely resembles the multiple group method [see 58], using data from 429 German patients with either Asthma or COPD. He inspected corrected item-total correlations ("trennschärfe") within the subscales and the correlation of each item with the total scores of the other subscales ("kreuz-trennschärfe"). The pattern of correlations he found did not offer support for the multidimensionality of the SGRQ.

By using a combination of appropriate exploratory and confirmatory, parametric and non-parametric, IRT and factor analytic models, as suggested by several researchers [e.g., 16, 18, 29, 59–62], we managed to obtain a comprehensive picture of the psychometric properties of the SGRQ. The results from the MSA, MIRT and confirmatory bifactor analysis suggest that (a shorter version of) the SGRQ can be used to calculate a total (unidimensional) score. The existing subscales performed quite poorly. However, the data did contain some multidimensionality, and ignoring this completely may lead to a decrease in reliability since the total score does not capture this multidimensionality. More research is needed to see how this multidimensionality could best be modeled and reflected in scores. The subscales suggested by our exploratory bifactor analyses could be tentatively labeled "complaints hampering strong exertion" and "complaints hampering light exertion." A high score on the latter may be an indicator for (perceived) disease severity. It should be explored in future studies whether these subscales can be reproduced in other (larger) samples and whether these subscales make sense to patients and are useful for clinicians. In the meantime, the cautious user may prefer to refrain from using subscale scores.

Several items showed poor performance. Of the ten items omitted in the new version of the SGRQ [see 5], seven also performed badly in our MSA and IRT analyses. Strikingly, as many as six of the eight Symptom had low item discrimination parameters and poor item fit; this was most likely due to the breach of logical ordering in the threshold parameters. Interestingly, Meguro et al. [5] also reported this problem for six of the seven polytomous Symptoms items. They suggested collapsing response categories to circumvent the problem. Whereas this may be a statistically viable solution, the occurrence of the problem is still unexpected from a clinical standpoint. One would assume that a patient with higher HRQoL is more likely to answer in the lowest item category, in this case "not at all." For a somewhat lower HRQoL, the most likely category would be the next-lowest one and so on. However, this was not the case for the aforementioned Symptoms items. For example, for the second item, "Over the last 4 weeks, I have brought up phlegm (sputum)," the most extreme category had a higher probability of being endorsed than the lower categories, even for patients that had high HRQoL. The middle category "a few days a month" never had the highest probability to be preferred over the other item categories, for any value of the latent trait. It should be noted that this psychometric finding could *not* simply be explained by a low count in some categories compared to others, which is often a reason to collapse answering categories. If the middle categories had barely been chosen, collapsing them would have been a straightforward solution. However, in this case, we argue that it would be more useful to conduct interviews with patients in a future study, to investigate what may be the explanation for the illogical ordering of the threshold parameters. An additional problem pertaining to the Symptoms items is that the recall period is not the same for all countries; so when making cross-cultural comparisons, or analyzing SGRQ data in multinational trials, it would be highly difficult to compare these items across countries. The recall period was actually omitted in the SGRQ-C; it is unclear whether this, by itself, would have resolved the item threshold issue.

Among the remaining items that performed badly was an item about employment and three items pertaining to medication. These items also caused trouble in the study by Meguro et al. [5] and were therefore not included in the SGRQ-C. Many patients with COPD are retired, so this may be a reason why the employment item performs badly. One may also argue that items pertaining to medication are measuring something quite separate from the rest of the items in the SGRQ, so it does not seem unexpected that these items had poor psychometric properties. Another item that performed badly was "Walking up hills" from Part 2/Sect. 2. This could possibly be a cultural finding; the Netherlands is a very flat country, with hardly any hills.

Therefore, this item may have a different meaning to a Dutch patient than to a patient from another part of the world.

In this study, we used three sophisticated psychometric methods to investigate the psychometric properties of the SGRQ. It should be noted, however, that the sample size was on the small side when it comes to performing parametric IRT. Since it is often difficult to obtain samples of 500–1,000 patients in clinical studies, we think that this study adequately reflects what researchers in this field typically have to work with. Given these circumstances, it is all the more important to use complementary methods as was done in this study. Another limitation worth mentioning is that the analyses were based on pooled data. The two data sets were too small to run the analyses separately; hence, we were not able to assess whether the results might have turned out differently if only performed in one subgroup. However, the mean difference in SGRQ scores between the two samples did not exceed the MID of 4; moreover, the correlations among the subscales was highly similar in the two subsamples. Therefore, we felt confident to proceed with the analyses using the pooled data set.

The authors of the SGRQ took great pains to obtain weights based on empirical information [1–3, 9]. In their most recent revision of the SGRQ, the items themselves were scrutinized, but not the weights. Meguro and colleagues used Rasch analysis to analyze the items and subsequently recalculated the empirical weights. However, a logical integration of these two is possible in a two-parameter IRT model. Our results indicated that the estimates of the latent trait based on our 2PL IRT model showed a very high correlation with the SGRQ total score. In our opinion, this implies that a 2PL IRT model is highly suitable to "capture" the relative importance people assign to QoL items and could thus safely replace older, more cumbersome methods. However, it should be noted that the latent trait estimates based on the 1PL IRT (Rasch) model already showed a strong correlation with the SGRQ total scores.

In conclusion, we found psychometric support for the use of a unidimensional total score of the SGRQ. However, nineteen of the 50 items had poor psychometric properties. Importantly, omitting these items from the analyses did not lead to a substantive decrease in measurement precision. More research is needed to investigate how to best take into account the multidimensionality in this instrument; based on our findings, it is not advised to use the original subscales. Users who want to use the SGRQ as a stand-alone instrument may consider using the SGRQ-C instead or leave out a number of items based on the findings of this study. Before deciding which SGRQ items to use in our CAT, we will first conduct a cognitive interview study with COPD patients, which may lead to rewording certain items before they are added to our item bank. Such a study could possibly also generate useful qualitative information that may help explain why certain items performed poorly in this study. Finally, we found a very strong correlation between the SGRQ total score calculated using the original "empirical" weights and the IRT-based total score as calculated in this study. If our findings are generalizable to other QoL instruments, researchers need no longer invest time and money in acquiring "empirical" weights, but could use IRT to obtain comparable results.

Acknowledgments We thank Dr. Straat for advising us on the use of the GA algorithm. This study was supported by Grant No. 3.4.11.004 from Lung Foundation Netherlands.

References

1. Jones, P. W., Quirk, F. H., & Baveystock, C. M. (1991). The St George's Respiratory Questionnaire. *Respiratory Medicine, 85*(Supplement 2), 25–31.
2. Jones, P. W., Quirk, F. H., Baveystock, C. M., & Littlejohns, P. (1992). A self-complete measure of health status for chronic airflow limitation. The St. George's Respiratory Questionnaire. *American Review of Respiratory Disease, 145*(6), 1321–1327.
3. Quirk, F. H., & Jones, P. W. (1990). Patients' perception of distress due to symptoms and effects of asthma on daily living and an investigation of possible influential factors. *Clinical Science (London), 79*(1), 17–21.
4. Barr, J. T., Schumacher, G. E., Freeman, S., LeMoine, M., Bakst, A. W., & Jones, P. W. (2000). American translation, modification, and validation of the St. George's Respiratory Questionnaire. *Clinical Therapeutics, 22*(9), 1121–1145.
5. Meguro, M., Barley, E. A., Spencer, S., & Jones, P. W. (2007). Development and validation of an Improved, COPD-specific version of the St. George Respiratory Questionnaire. *Chest, 132*(2), 456–463.
6. Embretson, S. E., & Reise, S. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum.
7. Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Chicalo, IL: University of Chicago Press.
8. Jones, P. W. (2009). *St George's Respiratory Questionnaire: Manual*. London: Division of Cardiac and Vascular Science, St George's, University of London.
9. Quirk, F. H., Baveystock, C. M., Wilson, R., & Jones, P. W. (1991). Influence of demographic and disease related factors on the degree of distress associated with symptoms and restrictions on daily living due to asthma in six countries. *European Respiratory Journal, 4*(2), 167–171.
10. Monninkhof, E., van der Valk, P., van der Palen, J., van Herwaarden, C., & Zielhuis, G. (2003). Effects of a comprehensive self-management programme in patients with chronic obstructive pulmonary disease. *European Respiratory Journal, 22*(5), 815–820.
11. Christenhusz, L. C., Prenger, R., Pieterse, M. E., Seydel, E. R., & van der Palen, J. (2012). Cost-effectiveness of an intensive smoking cessation intervention for COPD outpatients. *Nicotine & Tobacco Research, 14*(6), 657–663.
12. Mokken, R. J. (1971). *A theory and procedure of scale analysis*. The Hague: Mouton.
13. Sijtsma, K., & Molenaar, I. W. (2002). *Introduction to nonparametric item response theory* (Vol. 5). Thousand Oaks: Sage Publications.

14. R Development Core Team. (2012). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing.
15. van der Ark, L. A. (2007). Mokken scale analysis in R. *Journal of Statistical Software, 20*(11), 1–19.
16. Paap, M. C. S., Kreukels, B. P. C., Cohen-Kettenis, P. T., Richter-Appelt, H., de Cuypere, G., & Haraldsen, I. R. (2011). Assessing the utility of diagnostic criteria: a multisite study on gender identity disorder. *Journal of Sexual Medicine, 8*(1), 180–190.
17. Sijtsma, K., Emons, W. H., Bouwmeester, S., Nyklicek, I., & Roorda, L. D. (2008). Nonparametric IRT analysis of Quality-of-Life Scales and its application to the World Health Organization Quality-of-Life Scale (WHOQOL-Bref). *Quality of Life Research, 17*(2), 275–290.
18. Wismeijer, A. A. J. (2012). Dimensionality analysis of the thought suppression inventory: Combining EFA, MSA, and CFA. *Journal of Psychopathology and Behavioral Assessment*, *34*(1), 116–125.
19. Watson, R., Deary, I. J., & Shipley, B. (2008). A hierarchy of distress: Mokken scaling of the GHQ-30. *Psychological Medicine, 38*(04), 575–579.
20. Watson, R., van der Ark, L. A., Lin, L.-C., Fieo, R., Deary, I. J., & Meijer, R. R. (2012). Item response theory: How Mokken scaling can be used in clinical practice. *Journal of Clinical Nursing*, *21*(19–20), 2736–2746.
21. Paap, M. C. S., Meijer, R. R., Cohen-Kettenis, P. T., Richter-Appelt, H., de Cuypere, G., Kreukels, B. P. C., et al. (2012). Why the factorial structure of the SCL-90-R is unstable: Comparing patient groups with different levels of psychological distress using Mokken Scale Analysis. *Psychiatry Research, 200*(2–3), 819–826.
22. Beukers, F., Houtzager, B. A., Paap, M. C. S., Middelburg, K. J., Hadders-Algra, M., Bos, A. F., et al. (2012). Parental psychological distress and anxiety after a successful IVF/ICSI procedure with and without preimplantation genetic screening: Follow-up of a randomised controlled trial. *Early Human Development, 88*(9), 725–730.
23. Blom, E. H., Bech, P., Hogberg, G., Larsson, J. O., & Serlachius, E. (2012). Screening for depressed mood in an adolescent psychiatric context by brief self-assessment scales—testing psychometric validity of WHO-5 and BDI-6 indices by latent trait analyses. *Health Qual Life Outcomes, 10*(1), 149.
24. Roorda, L. D., Green, J. R., Houwink, A., Bagley, P. J., Smith, J., Molenaar, I. W., et al. (2012). Item hierarchy-based analysis of the Rivermead Mobility Index resulted in improved interpretation and enabled faster scoring in patients undergoing rehabilitation after stroke. *Archives of Physical Medicine and Rehabilitation, 93*(6), 1091–1096.
25. Straat, J. H., van der Ark, L. A., & Sijtsma, K. (2013). Comparing optimization algorithms for item selection in Mokken scale analysis. *Journal of Classification*, *30*(1), 75–99.
26. Reise, S. P., & Waller, N. G. (2009). Item response theory and clinical measurement. *Annual Review of Clinical Psychology, 5*(1), 27–48.
27. Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16*(2), 159–176.
28. van den Berg, S. M., Heuven, H. C. M., van den Berg, L., Duffy, D. L., & Serpell, J. A. (2010). Evaluation of the C-BARQ as a measure of stranger-directed aggression in three common dog breeds. *Applied Animal Behaviour Science, 124*(3–4), 141–146.
29. van den Berg, S. M., Paap, M. C. S., Derks, E. M., et al. (2013). Using multidimensional modeling to combine self-report symptoms with clinical judgment of schizotypy. *Psychiatry Research*, *206*(1), 75–80.
30. Glas, C. A. W. (2010). *Preliminary manual of the software program multidimensional item response theory (MIRT)*. Enschede: Department of Research Methodology, Measurement and Data-Analysis, University of Twente.
31. Reise, S. P., Morizot, J., & Hays, R. D. (2007). The role of the bifactor model in resolving dimensionality issues in health outcomes measures. *Quality of Life Research, 16*, 19–31.
32. Reise, S. P. (2012). The rediscovery of bifactor measurement models. *Multivariate Behavioral Research, 47*(5), 667–696.
33. Reise, S. P., Moore, T. M., & Haviland, M. G. (2010). Bifactor models and rotations: Exploring the extent to which multidimensional data yield univocal scale scores. *Journal of Personality Assessment, 92*(6), 544–559.
34. Gustafsson, J.-E., & Åberg-Bengtsson, L. (2010). Unidimensionality and interpretability of psychological instruments. In S. E. Embretson (Ed.), *Measuring psychological constructs: Advances in model-based approaches* (pp. 97–121). Washington, DC: American Psychological Association.
35. Reise, S. P., Scheines, R., Widaman, K. F., & Haviland, M. G. (2013). Multidimensionality and structural coefficient bias in structural equation modeling: A bifactor perspective. *Educational and Psychological Measurement, 73*(1), 5–26.
36. Schmid, J., & Leiman, J. M. (1957). The development of hierarchical factor solutions. *Psychometrika, 22*, 53–61.
37. Revelle, W. (2012). Psych: Procedures for psychological, psychometric, and personality research. R package version 1.1–10. Retrieved from http://personality-project.org/r/psych.manual.pdf.
38. Muthén, L. K., & Muthén, B. O. (2006). *Mplus user's guide*, 4th, 6th edns. Los Angeles, CA: Muthén & Muthén.
39. Cook, K. F., Kallen, M. A., & Amtmann, D. (2009). Having a fit: impact of number of items and distribution of data on traditional criteria for assessing IRT's unidimensionality assumption. *Quality Life Research, 18*(4), 447–460.
40. Hu, L., & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to unparameterized model misspecification. *Psychological Methods, 3*, 424–453.
41. Jones, P. W. (2002). Interpreting thresholds for a clinically significant change in health status in asthma and COPD. *European Respiratory Journal, 19*(3), 398–404.
42. Jones, P. W. (2005). St. George's Respiratory Questionnaire: MCID. *COPD: Journal of Chronic Obstructive Pulmonary Disease, 2*(1), 75–79.
43. Al-Shair, K., Atherton, G. T., Kennedy, D., Powell, G., Denning, D. W., & Caress, A. (2013). Validity and reliability of the St. George's Respiratory Questionnaire in assessing health status in patients with chronic pulmonary aspergillosis. *Chest*, *144*(2), 623–631.
44. Weldam, S. W., Schuurmans, M. J., Liu, R., & Lammers, J. W. (2013). Evaluation of Quality of Life instruments for use in COPD care and research: A systematic review. *International Journal of Nursing Studies*, *50*(5), 688–707.
45. Bae, Y. J., Kim, Y. S., Park, C. S., Lee, Y. S., Chang, Y. S., Cho, Y. S., et al. (2011). Reliability and validity of the St George's Respiratory Questionnaire for asthma. *The International Journal of Tuberculosis and Lung Disease, 15*(7), 966–971.
46. Tafti, S. F., Cheraghvandi, A., Mokri, B., & Talischi, F. (2011). Validity and specificity of the Persian version of the Saint George Respiratory Questionnaire. *Journal of Asthma, 48*(6), 589–592.
47. Liang, W.-M., Chen, J.-J., Chang, C.-H., Chen, H.-W., Chen, S.-L., Hang, L.-W., et al. (2008). An empirical comparison of the WHOQOL-BREF and the SGRQ among patients with COPD. *Quality of Life Research, 17*(5), 793–800.
48. El Rhazi, K., Nejjari, C., Benjelloun, M. C., Bourkadi, J., Afif, H., Serhier, Z., et al. (2006). Validation of the St. George's Respiratory Questionnaire in patients with COPD or asthma in Morocco. *The International Journal of Tuberculosis and Lung Disease, 10*(11), 1273–1278.

49. Sanjuas, C., Alonso, J., Prieto, L., Ferrer, M., Broquetas, J. M., & Anto, J. M. (2002). Health-related quality of life in asthma: A comparison between the St George's Respiratory Questionnaire and the Asthma Quality of Life Questionnaire. *Quality of Life Research, 11*(8), 729–738.
50. Engstrom, C. P., Persson, L. O., Larsson, S., & Sullivan, M. (1998). Reliability and validity of a Swedish version of the St George's Respiratory Questionnaire. *European Respiratory Journal, 11*(1), 61–66.
51. Ferrer, M., Alonso, J., Prieto, L., Plaza, V., Monso, E., Marrades, R., et al. (1996). Validity and reliability of the St George's Respiratory Questionnaire after adaptation to a different language and culture: the Spanish example. *European Respiratory Journal, 9*(6), 1160–1166.
52. Yu, D. T. W., Scudds, R. J., & Scudds, R. A. (2004). Reliability and validity of a Hong Kong Chinese Version of the St George's Respiratory Questionnaire in Patients with COPD. *Hong Kong Physiotherapy Journal, 22*(1), 33–39.
53. Rutten-van Molken, M., Roos, B., & Van Noord, J. A. (1999). An empirical comparison of the St George's Respiratory Questionnaire (SGRQ) and the Chronic Respiratory Disease Questionnaire (CRQ) in a clinical trial setting. *Thorax, 54*(11), 995–1003.
54. Karpinski, N. (2005). *Validierung von Lebensqualitäts-Assessments bei chronisch–obstruktiven Atemwegserkrankungen (validation of quality of life-assessments in patients with chronic obstructive pulmonary disease)*. Bremen: University of Bremen. Retrieved from http://d-nb.info/991362772/34.
55. Paap, M. C. S. (2011). *Examining the validity of the assessment of gender identity disorder: Diagnosis, self-reported psychological distress and strategy adjustment*. Oslo: University of Oslo. Retrieved from https://www.duo.uio.no/bitstream/handle/10852/27940/dravhandling-paap.pdf.
56. Wismeijer, A. A. J., Sijtsma, K., van Assen, M. A. L. M., & Vingerhoets, A. J. J. M. (2008). A comparative study of the dimensionality of the self-concealment scale using principal components analysis and Mokken scale analysis. *Journal of Personality Assessment, 90*(4), 323–334.
57. Nunnally, J. C. (1978). *Psychometric theory*. New York: McGraw-Hill.
58. Timmerman, M. E. (2005). Factor analysis. Retrieved March 7, 2013, from http://www.ppsw.rug.nl/~metimmer/.
59. Emons, W. H. M., Sijtsma, K., & Pedersen, S. S. (2012). Dimensionality of the Hospital Anxiety and Depression Scale (HADS) in cardiac patients: Comparison of Mokken Scale analysis and factor analysis. *Assessment*, *19*(3), 337–353.
60. Salaffi, F., Franchignoni, F., Giordano, A., Ciapetti, A., Gasparini, S., & Ottonello, M. (2013). Classical test theory and Rasch analysis validation of the recent-onset arthritis disability questionnaire in rheumatoid arthritis patients. *Clinical Rheumatology*, *32*(2), 211–217.
61. Brouwer, D., Meijer, R. R., Weekers, A. M., & Baneke, J. J. (2008). On the dimensionality of the Dispositional Hope Scale. *Psychological Assessment, 20*(3), 310–315.
62. Sousa, R. M., Dewey, M. E., Acosta, D., Jotheeswaran, A. T., Castro-Costa, E., Ferri, C. P., et al. (2010). Measuring disability across cultures–the psychometric properties of the WHODAS II in older people from seven low- and middle-income countries. The 10/66 Dementia Research Group population-based survey. *International Journal of Methods In Psychiatric Research, 19*(1), 1–17.