

---

ANTONIO J. ROJAS TEJADA and OSCAR M. LOZANO ROJAS

APPLICATION OF AN IRT POLYTOMOUS MODEL  
FOR MEASURING HEALTH RELATED  
QUALITY OF LIFE\*

(Accepted 8 September 2004)

**ABSTRACT.** *Background:* The Item Response Theory (IRT) has advantages for measuring Health Related Quality of Life (HRQOL) as opposed to the Classical Tests Theory (CTT). *Objectives:* To present the results of the application of a polytomous model based on IRT, specifically, the Rating Scale Model (RSM), to measure HRQOL with the EORTC QLQ-C30. *Methods:* 103 terminal cancer patients cared for by the home services of the Servicio Andaluz de Salud (Andalusian Health Service) (Andalusia, Spain) participated. These patients responded to the adapted Spanish version of the EORTC QLQ-C30. The application was carried out individually in the patients' homes. *Results:* The results show that there is an adequate global fit between the data and the IRT model applied. The analysis of the items shows that for 31 of the 33 items there is a good fit. The items which measure the general perception of health and the perception of quality of life present a lack of fit. The study of the response categories of the items (by means of Category Probability Curves) indicates that all the alternatives work extremely well. *Conclusions:* The EORTC QLQ-C30 presents good metric qualities, under the RSM, ratifying the feasibility to measure HRQOL already shown in other studies carried out with CTT.

**KEY WORDS:** cancer patients, health related quality of Life, item response theory, Rasch polytomous models, rating scale model

## INTRODUCTION

Health Related Quality of Life (HRQOL) can be understood as the state of physical, psychological and social health perceived by an individual during the course of an illness or disease (Siegrist and Junge, 1989). It is an indicator which reflects the way in which illness affects the lives of individuals, since this has negative consequences on a physical level (problems with regard to mobility and physical activities), on a psychological level (depression, stress,

anxiety, etc.) and on a social level (interaction with relatives, friends, etc.). As treatments help to improve the individual's situation, HRQOL is used as an indicator of their efficacy. This variable has been widely used to evaluate the effectiveness of treatments for a variety of illnesses, and has often been applied as a result variable for the evaluation of cancer. In some cases, the harsh treatments which patients are subjected to or the search for the control of these symptoms in other patients has generated an interest in the development of instruments to measure the HRQOL in cancer patients. Amongst these instruments we can find: Support Team Assessment Schedule (STAS) (Higginson and McCarthy, 1989), Therapy Impact Questionnaire (TIQ) (Toscani, 1996), Quality of Life Inventory (QOLI) (Spitzer et al., 1981), European Organization for Research and Treatment of Cancer Core Quality of Life Questionnaire (EORTC QLQ-C30) (Aaronson et al., 1993), etc. which incorporate specific items about alleviating treatments, consequences of this illness and specific symptoms of its patients. Therefore, the development of tests that will measure the HRQOL of these patients adequately is a priority objective for many health workers. The relevance of an accurate measure of this variable is highlighted amongst advanced cancer patients, since in many cases, the improvement of their HRQOL is one of the therapeutic aims.

As a rule, the psychometric theory employed to measure HRQOL has been the Classical Test Theory (CTT). However, the metric characteristics of this theory present a double invariance problem: (1) the measures for each patient depend on the instrument utilised (e.g. a patient will have different scores in HRQOL depending on the test used: EORTC QLQ-C30, STAS, etc.); (2) the estimations of the items and tests properties depend on the sample of individuals used for this purpose (e.g. the reliability of a test will depend on the sample of people used to calculate it). Besides, in the CTT we suppose, (hardly credible though it may be) that once the reliability of a test has been estimated for a certain population, this reliability (accuracy) remains constant for all ability levels (e.g. it will remain identical when estimating the measures in persons with high, medium or low HRQOL values). Whereas most frequently accuracy is lower when measuring the extremes of this continuum (high and low ability values).

The advances in Psychometrics have helped to displace the CTT in favour of the use of Item Response Theory (IRT)-based models. With

these models invariable measures can be obtained, regardless of the instruments used and of the individuals evaluated (Hambleton, 1985). The calibration procedure is independent of the sample to which the test is administered (it is invariant over the population), and the measures of persons are also test-free (it does not matter which selection of items is used to estimate them). Besides, it makes no sense to refer to the reliability of the test where the IRT is concerned. In the IRT the accuracy of the measure is estimated specifically (standard measure error) for each ability level in the variable.

On the other hand, IRT models are focused on the joint measure of people and items, which means they are placed in the same measure continuum with the same metrics, unlike the CTT which only locates persons within the continuum and supposes that all the items contribute equally to the measure of the construct.

However, not all IRT models are optimum to quantify the same variables. On the contrary, they must be selected according to the nature of the variable that needs to be quantified and to the way in which the tests collect data (e.g. items format).

Thus, the nature of the HRQOL variable requires the use of rating scales (e.g. Likert-type) with preference to any other method. These scales present a series of graduated response categories (or  $k$  response categories), such as finite, exhaustive and excluding. The categories represent growing amounts or increases of the variable being measured in the item (e.g. HRQOL). People with their responses graduate their agreement or disagreement with the statement expressed in the item. In order to mark these items with regard to the scores whole consecutive numbers are given to the consecutive categories, that is to say, these scales require a first identification of several ordered response levels in the items, and then, a partial score assigned to these response categories.

One model which fits the characteristics of this variable is the Rating Scale Model (RSM). This is a polytomous model derived from Rasch dichotomous model. The starting points of the RSM are the following (Andrich, 1978a): unidimensionality (all the items must measure the same construct, e.g. HRQOL), local independence of items and persons (this means that an individual's response to any of the items in the test is not affected by their response to other items, and the same applies to items) and homogenous discrimination of the items (all items have the same discrimination power).

With the RSM we estimate the a person's probability of responding to a certain category in an item, deducing this from the difference estimated between the person's ability level in the variable being measured and the intensity or ability level of the items used to measure this variable. In order to do this, the model also uses the *item steps* concept, defined as the point within the ability continuum where the transition between two adjacent response categories takes place. Thus, an item with four response categories will have three steps (the first step is the transition between categories one and two, the second step is the transition between categories two and three, and the third step is the transition between categories three and four). Generally, in an item with  $m + 1$  response categories, the steps are determined by  $m$ . The function that determines the probability of a person  $n$  of responding to a certain response category  $x$  in item  $i$  is given by the following logistic function (Andrich, 1978 a,b; Masters, 1982).

$$\pi_{nix} = \frac{\exp \sum_{j=1}^x [\beta_n - (\delta_i + \tau_j)]}{\sum_{x=1}^{m+1} \exp \sum_{j=1}^x [\beta_n - (\delta_i + \tau_j)]}, \quad x = 1, 2, \dots, m + 1.$$

where  $\beta_n$  is the location or scale value of the person  $n$ ,  $\delta_i$  the location or scale value of the item  $i$ , and  $\tau_m$  called threshold parameter between categories is the location of step  $m$  related to the scale value of the item. In the Rating Scale Model, the step  $j$  parameter of item  $i$  ( $\delta_{ij}$ ) is defined as  $\delta_{ij} = \delta_i + \tau_m$  where all the parameters  $\tau_m$  remain constant for all the items. In the model we are concerned with, the only difference between the items is due to their location  $\delta_i$  in the one-dimensional continuum of the variable which is being measured. One condition imposed by the RSM on the threshold parameters of the items ( $\tau_m$ ) is that these remain constant through all the items that constitute the scale, and it is assumed that they only depend on the response categories proposed, as in the case of the tests where all the items have identical response categories (e.g. Likert-type).

Bearing in mind the advantages that the IRT brings to the HRQOL measure and, more specifically, the appropriateness of the RSM when applied to this variable, the objective of this study is to measure HRQOL with the Spanish version of the EORTC QLQ-C30 test applying the RSM. For this purpose, we analyse the way in which

the test and each one of the items and their respective response categories work globally. Moreover, we include the result of the conversion of the test scores obtained by the persons with their corresponding measures (and standard errors).

## METHOD

### *Participants*

The sample consists of individuals included in the home care programmes for terminal cancer patients in the provinces of Córdoba and Cádiz, Andalusia, Spain.

Initially, a random sampling was carried out among all the patients included in the above mentioned programme. However, due to the high exit rates found in Cádiz, all the patients included in this programme were interviewed. 39% of the patients were male and 51% female. 44.6% of the patients were visited once a week or more often. The rest of the patients were visited at least once a fortnight. 8.5% did not know what illness they were suffering from; 19.5% only knew about it partially and 54.2% knew exactly what illness they had. 17.8% did not respond.

The final sample consisted of 103 patients. However, 14 patients who got the highest score in all items cannot be measured (the measure corresponding to an extreme -perfect or zero-score cannot be estimated by RSM), but their responses were useful to item analysis.

### *Instruments*

To measure HRQOL a adapted Spanish version of EORTC QLQ-C30 (Version 2.0) was used (Arrarás et al., 1995) (see appendix). This test consists of five functional scales (physical, role, cognitive, emotional, and social), three symptomatic scales (fatigue, pain, and nausea and vomiting), items which measure specific symptoms affecting cancer patients (dyspnoea, loss of appetite, etc.) and of two items which measure the patients' perception of their global health and quality of life.

With regard to the format, five items are measured with a dichotomic format, 26 items are measured with a Likert-type scale with four alternative responses and the two items on health and quality of life

perception are measured with a Likert-type format with five response categories.

Numerous studies have focused on analysing the reliability of these scales from CTT. To name a few, Aaronson and colls (Aaronson et al., 1993) find that the reliability measured by Cronbach's alpha fluctuates-depending on the scales- between 0.54 and 0.89. Ringdal and Ringdal (Ringdal and Ringdal, 1993) find values between 0.55 and 0.86, depending on the scale used. In a study (Godoy et al., 1999) where the Spanish version of this questionnaire was used they found Cronbach's alpha coefficient values which fluctuate between 0.92 and 0.62.

With regard to validity evidence there are studies in Spanish on this scale where we can find data supporting the fact that this scale discriminates appropriately amongst the various groups of patients and about its concurrent validity (Arrarás et al., 1995; Godoy et al., 1999).

### *Procedure*

The main carers of the patients applied the test individually to those examined. The applications took place in the homes of the main carers. Initially, the carers were interviewed and information was collected about the quality of the home care service and about the patient's health and socio-demographic variables. Later on the patients were interviewed and asked to respond to the test.

In order to apply the Rating Scale Model the programme BIG-STEPS, Version 2.82 (1998), developed by Wright and Linacre (Linacre and Wright, 1988) was used. To interpret the ability values, both in the persons' measures and in the calibration of the items a logarithmic transformation of the data was carried out, so that both remain on an interval scale, called 'logit scale' (log-odds units), with 0 mean and a standard deviation of 1 (Wright and Masters, 1982).

## RESULTS

### *Fit Between Data and Model*

In the use of the RSM one can only obtain benefits if there is a fit between data and model. With the analysis of the fit of the data to the

model we find out if the model explains the data obtained. This is a distinctive characteristic of the IRT models. The fit can be studied from a triple perspective: (1) the overall fit or total fit (which enables the fit of data to the model to be evaluated as a whole); (2) item fit (which helps identify poor items); (3) person fit (which helps identify a suspicious pattern of responses). For the objectives of this study we shall focus on total fit and item fit sections. Person fit section, although of vital importance, will not be examined in this study, since we are focusing mainly on the analysis of the metric properties of the EORTC QLQ-C30 rather than on the analysis of the response patterns of the individuals.

In order to check if there is a fit between the data and the model we follow different procedures, the most frequently used in the RSM being the residuals analysis (INFIT and OUTFIT) proposed by Wright and Masters (Wright and Masters, 1982). Fit statistics (INFIT and OUTFIT) are reported as mean-square residuals, which have approximate Chi-square distributions. These are also reported standardized,  $N(0,1)$ .

INFIT is an information-weighted fit statistic, which is more sensitive to unexpected behavior affecting responses to items close to the person's ability level. BIGSTEPS reports two INFIT statistics: MNSQ (the mean-square infit statistic with expectation 1. Values substantially below 1 indicate dependency in data, values substantially above 1 indicate noise) and ZSTD (the infit mean-square fit statistic standardized to approximate a theoretical mean 0 variance 1 distribution). OUTFIT is an outlier-sensitive fit statistic, more sensitive to unexpected behavior by persons on items far from the person's ability level. BIGSTEPS also reports two OUTFIT statistics: MNSQ (the mean-square outfit statistic, with expectation 1: values substantially less than 1 indicate dependence in data; values substantially greater than 1 indicate the presence of unexpected outliers) and ZSTD (the outfit mean-square fit statistic standardized to approximate a theoretical mean 0 and variance 1 distribution) (Wright and Masters, 1982; Linacre and Wright, 1988)).

Lunz, Wright and Linacre (Lunz et al., 1990) propose that the values of the MNSQ statistics situated in the interval between 0.6 and 1.4 would have an acceptable fit for small samples. An acceptable ZSTD statistics fit would fluctuate between values equal to or higher

than  $-2$  and equal to or lower than  $+2$  (Bond and Fox, 2001). It is necessary for the data fit to occur both for persons and for items, the above mentioned statistics being used in both cases. In Table I, first two rows, we can see the total fit values. OUTFIT (MNSQ and ZSTD) and INFIT (MNSQ and ZSTD), both for persons and for items. The results obtained with persons (MNSQ INFIT mean = 1.04; ZSTD INFIT mean = 0.0 and S.D. = 1.0; MNSQ OUTFIT mean = 1.00; ZSTD OUTFIT mean =  $-0.1$  and S.D. = 1.0), and items (MNSQ INFIT mean = 1.03; ZSTD INFIT mean =  $-0.1$  and S.D. = 1.0; MNSQ OUTFIT mean = 1.00; ZSTD OUTFIT mean =  $-0.1$  and S.D. = 1.0), support the interpretation of an adequate total fit of persons and items.

One can also appreciate the ability mean values, standard deviations, the maximum and minimum levels for persons and items in the test, as well as the corresponding measure error values of the model.

#### *Item Fit*

After fitting the items to the model we can establish which items work correctly according to the model and which do not, hence its usefulness to determine which items need to be reviewed or eliminated and which are deemed adequate (Rojas et al., 2002). That is to say, this fit can help to get information about 'the quality' of the item. It would be equivalent to the classical items analysis.

To analyse the items fit we use the statistics seen above, and the same criteria are followed to decide whether a good fit exists or not.

The RSM also provides information about the position of each item within the continuum, their measure error, and conventional point bi-serial correlation coefficients (these are reported not only for items but also for persons). In the RSM analysis, point bi-serial correlation coefficients is a useful diagnostic indicator of data mis-coding or item miskeying: negative or zero values indicate items or persons with response strings that contradict the variable (Linacre and Wright, 1998).

In Table I the analysis of each item is shown. One can observe that all the items, except number 32 (How would you rate your overall health during the past week?) and 33 (How would you rate your overall quality of life during the past week?) present an adequate fit to

TABLE I  
Summary of person and item statistics

Measure	Error	INFIT		OUTFIT		PTBIS CORR.
		MNSQ	ZSTD	MNSQ	ZSTD	
Summary of persons						
Mean	0.54	1.04	0.0	1.00	-0.1	
S.D.	1.01	0.48	1.0	0.51	1.0	
Max	3.84	2.40	2.3	3.24	3.1	
Min	-0.99	0.32	-2.3	0.23	-1.8	
Summary of persons						
Mean	0.00	1.03	-0.1	1.00	-0.1	
S.D.	0.99	0.44	1.0	0.56	1.0	
Max	1.94	2.84	3.5	3.35	3.8	
Min	-2.16	0.57	-1.6	0.55	-1.3	
Items statistics	Measure	INFIT		OUTFIT		
	Entry number	MNSQ	ZSTD	MNSQ	ZSTD	
1	1.86	0.92	-0.2	0.92	-0.2	0.46
2	1.94	0.92	-0.2	0.81	-0.4	0.48
3	0.51	0.85	-0.9	0.78	-0.7	0.53
4	0.88	0.80	-0.9	0.74	-0.8	0.59

TABLE I  
Continued

Items statistics	Entry number	Measure	Error	INFIT		OUTFIT		
				MNSQ	ZSTD	MNSQ	ZSTD	
	5	-0.10	0.23	0.87	-0.7	0.79	-0.5	0.48
	6	0.98	0.13	0.81	-0.6	0.74	-0.7	0.75
	7	0.87	0.12	0.93	-0.2	0.91	-0.2	0.67
	8	-0.87	0.15	1.18	0.4	1.17	0.2	0.40
	9	0.68	0.12	0.90	-0.3	0.85	-0.4	0.58
	10	0.66	0.13	0.90	-0.3	0.92	-0.2	0.63
	11	-0.27	0.13	1.13	0.4	1.12	0.2	0.43
	12	0.21	0.12	0.60	-1.5	0.56	-1.2	0.65
	13	-0.63	0.14	0.93	-0.2	0.86	-0.2	0.47
	14	-1.67	0.20	1.06	0.1	0.97	0.0	0.34
	15	-2.16	0.25	1.12	0.2	0.92	-0.1	0.26
	16	-0.43	0.13	1.30	0.7	10.19	0.3	0.39
	17	-2.16	0.25	1.43	0.5	1.09	0.1	0.26
	18	0.45	0.12	0.71	-1.0	0.78	-0.6	0.57
	19	0.58	0.12	0.57	-1.6	0.55	-1.3	0.77
	20	-0.61	0.14	0.85	-0.4	0.73	-0.5	0.55
	21	-0.04	0.12	0.68	-1.1	0.69	-0.8	0.55
	22	-0.04	0.12	0.77	-0.7	0.80	-0.5	0.52
	23	-0.45	0.13	0.78	-0.6	0.84	-0.3	0.44
	24	-0.05	0.12	0.79	-0.7	0.74	-0.6	0.53
	25	-0.27	0.13	1.26	0.7	1.13	0.2	0.45

---

26	0.41	0.12	0.87	-0.4	0.83	-0.4	0.63
27	-0.46	0.13	0.89	-0.3	0.81	-0.4	0.61
28	0.77	0.12	0.78	-0.7	0.72	-0.7	0.70
29	-0.16	0.12	0.79	-0.7	0.70	-0.7	0.65
30	-0.96	0.15	1.36	0.7	1.25	0.3	0.27
31	-1.63	0.20	1.13	0.2	0.97	0.0	0.22
32	1.24	0.12	2.23	2.5	2.87	3.1	-0.08
33	0.91	0.11	2.84	3.5	3.35	3.8	-0.18

---

the model. One can see ('ptbis corr.' column) that in these two items that the item-total correlation is low and negative.

#### *Item and Response Category Calibration*

The item calibration for EORTC QLQ-C30 provides us with information about the place that each item and their response alternatives occupy in the HRQOL continuum. Due to the fact that the EORTC QLQ-C30 has two, four and five response category items, parameters are estimated for one, two three and four item steps, respectively. As discussed previously, item steps are defined as the transition between two adjacent response categories, and their value is expressed by the threshold parameter between categories ( $\tau_m$ ) or by the location of step  $m$  related to the scale value of the item.

Table I shows the location values of each item in the HRQOL continuum ('measure' column). What is desirable for the items to work optimally is that they should cover the continuum adequately at the point where we attempt to measure the persons. The EORTC QLQ-C30 items work like this, since according to their location values their distribution is sufficiently wide to collect the persons' abilities variability (this will be seen more clearly later on in the persons and items distribution map, Figure 4). We can observe that the items with the lowest ability are numbers 15 and 17, with values of  $-2.16$  logits. The item with the highest ability is number 2 with a value of  $1.94$  logits. Item 1 (Do you have any trouble doing strenuous activities, like carrying a heavy shopping bag or a suitcase?) and 2 (Do you have any trouble taking a long walk?) are situated on the side of the continuum which is closest to a high HRQOL. That is to say, the persons with the highest scores in this item will be those who have a better HRQOL. On the other hand, those items which are closest to a low HRQOL are number 15 (Have you vomited?) and 17 (Have you had diarrhoea?), that is to say, those persons with a high HRQOL are expected to score highly in those items, whilst those with a low HRQOL will have lower scores. Let us remember that items and persons are measured on an interval scale with a common unit, the logit, derived from a function of the probability of a positive answer, given person and item parameters.

In Table II is shown the value of each item step ( $\delta_{i1}$ ,  $\delta_{i2}$ ,  $\delta_{i3}$  y  $\delta_{i4}$ ). In the case of those items with two response categories, the threshold

TABLE II  
Calibration of item and steps

Items	$\delta_i$	1° step $\delta_{i1} = \delta_i + \tau_1$	2° step $\delta_{i2} = \delta_i + \tau_2$	3° step $\delta_{i3} = \delta_i + \tau_3$	4° step $\delta_{i4} = \delta_i + \tau_4$
1	1.86	*	*	*	*
2	1.94	*	*	*	*
3	0.51	*	*	*	*
4	0.88	*	*	*	*
5	-0.10	*	*	*	*
6	0.98	0.50	0.73	1.71	*
7	0.87	0.39	0.62	1.60	*
8	-0.87	-1.35	-1.12	-0.14	*
9	0.68	0.20	0.43	1.41	*
10	0.66	0.18	0.41	1.39	*
11	-0.27	-0.75	-0.52	0.46	*
12	0.21	-0.27	-0.04	0.94	*
13	-0.63	-1.11	-0.88	0.10	*
14	-1.67	-2.15	-1.92	-0.94	*
15	-2.16	-2.64	-2.41	-1.43	*
16	-0.43	-0.91	-0.68	0.30	*
17	-2.16	-2.64	-2.41	-1.43	*
18	0.45	-0.03	0.20	1.18	*
19	0.58	0.10	0.33	1.31	*
20	-0.61	-1.09	-0.86	0.12	*
21	-0.04	-0.52	-0.29	0.69	*
22	-0.04	-0.52	-0.29	0.69	*
23	-0.45	-0.93	-0.70	0.28	*
24	-0.05	-0.53	-0.30	0.68	*
25	-0.27	-0.75	-0.52	0.46	*
26	0.41	-0.07	0.16	1.14	*
27	-0.46	-0.94	-0.71	0.27	*
28	0.77	0.29	0.52	1.50	*
29	-0.16	-0.64	-0.41	0.57	*
30	-0.96	-1.44	-1.21	-0.23	*
31	-1.63	-2.11	-1.88	-0.90	*
32	1.24	0.10	0.89	1.36	2.62
33	0.91	-0.23	0.56	1.03	2.29

Two categories of response items:

$$\tau_1 = 0$$

Four categories of response items:

$$\tau_1 = -0.48, \tau_2 = -0.25, \tau_3 = 0.73$$

Five categories of response items:

$$\tau_1 = -1.14, \tau_2 = -0.35, \tau_3 = 0.12, \tau_4 = 1.38$$

parameter of the first step is defined by the intersection between the two categories and its value is always equal to 0 ( $\delta_{i1} = 0$ ). For the items with four response alternatives there are three threshold parameters. In every item one can appreciate how the values for each item step appear in a particular order ( $\delta_{i1} < \delta_{i2} < \delta_{i3}$ ) which means that the response alternatives work correctly. Finally, those items with five response categories (items 32 and 33) have four steps, which as we can see in Table II, also appear in order ( $\delta_{i1} < \delta_{i2} < \delta_{i3} < \delta_{i4}$ ).

#### *Category Probability Curves*

The Category Probability Curves (CPC) are represented starting from the ability threshold parameters of the item steps. In the RSM these parameters all have the same value, and the difference between each item lies in the place they occupy within the ability continuum. Therefore, the CPCs of each item have the same shape. However, in the EORTC QLQ-C30 there are items with one, three or four threshold parameters, which make it necessary to represent a CPC for each of these item types.

The interest in studying these curves arises from two main aspects: (a) the curves provide information about how the response alternatives work. To be specific, the fact that all curves are at some point the most probable ones indicates that the response alternatives work correctly; (b) the intersections between the curves (thresholds) will define the 'more probable response areas' within the continuum. That is to say, they define within the continuum, the different regions in which the persons will respond with a higher probability to the response category represented in that particular region.

In the RSM a hypothesis is established about the data, according to the order of the steps without order in which the unidimensional continuum is divided into categories. Thus, the estimation of the value of the steps without order is sufficient evidence to conclude that the empirical order is not consistent with the theoretical order, the categories are not working as planned (Andrich et al., 1997). As one can see in Figure 1, in the items with five response categories, all of these constitute the most likely response alternatives at some point within the HRQOL continuum. Therefore, with these response

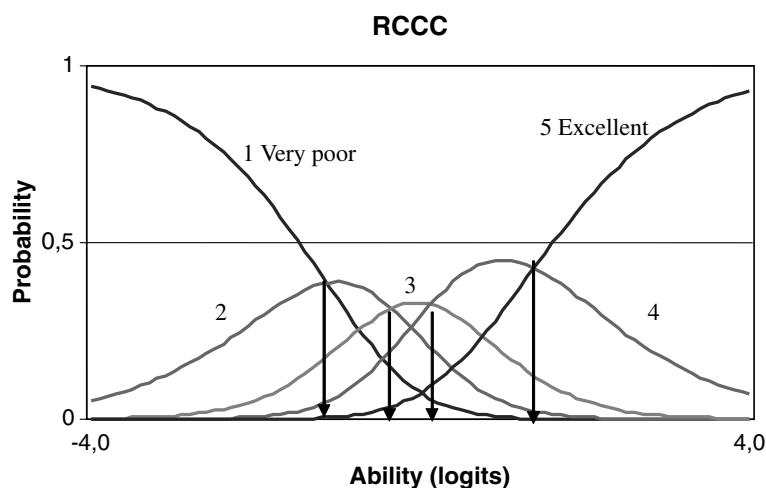


Figure 1. CPC of items with five categories of responses.

categories, not leaving any of them out, one can collect information adequately about the content of these items.

With regard to the continuum regions which define the alternatives 'Very poor' and 'Excellent' are the ones which occupy a larger area within the continuum. Conversely, the intermediate alternative is the most probable one in a very narrow region of the continuum.

In the case of those items with four response categories (Figure 2), they are all at some point in the continuum the most likely alternatives. However, one can appreciate that the alternative 'quite a bit' is the most probable response in a very narrow area of the continuum. This shows that only those persons situated within the region of the continuum defined by this alternative will respond using the 'quite a bit' alternative. Conversely, the other response alternatives are indeed the most likely ones for a greater number of persons, since, as we can see in this figure, the continuum areas which define their CPCs are wider.

For the test items with two response categories (the first five items), the CPCs are shown in Figure 3. We can observe that these alternatives are the most likely in different parts of the continuum. To be specific, they define areas with the same extension on both sides of the continuum. This result indicates that these response categories work correctly.

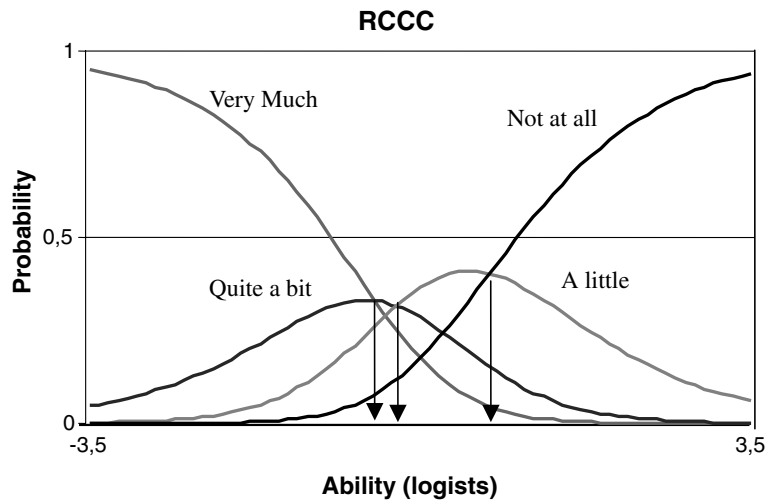


Figure 2. CPC of items with two categories of responses.

*Distribution Map for Persons and Items*

In Figure 4 we can see the persons and items which have been calibrated and measured with the logit scale. The 'persons' column shows

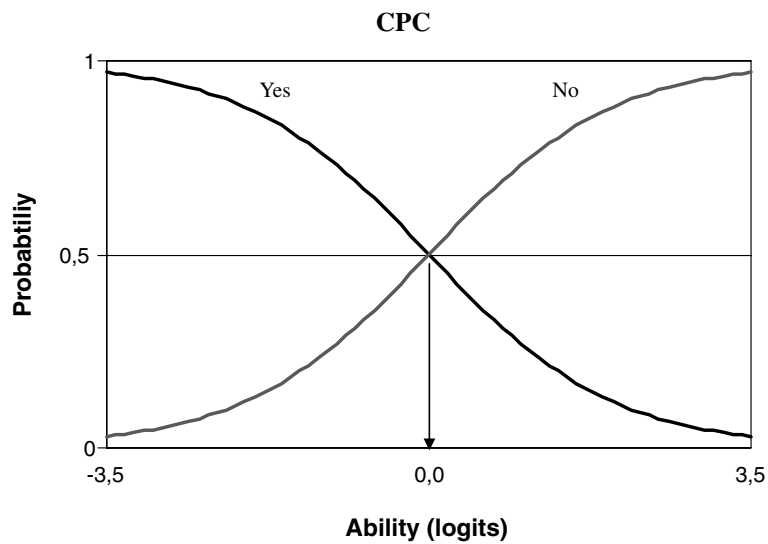


Figure 3. CPC of items with two categories of responses.

the location of the persons within the HRQOL continuum. The 'items center' column indicates the location of the item within the continuum, its ability value (D points out the location of dichotomic items and X of polytomic ones, moreover, the items numbers appear in this central column). The items bottom column shows the minimum

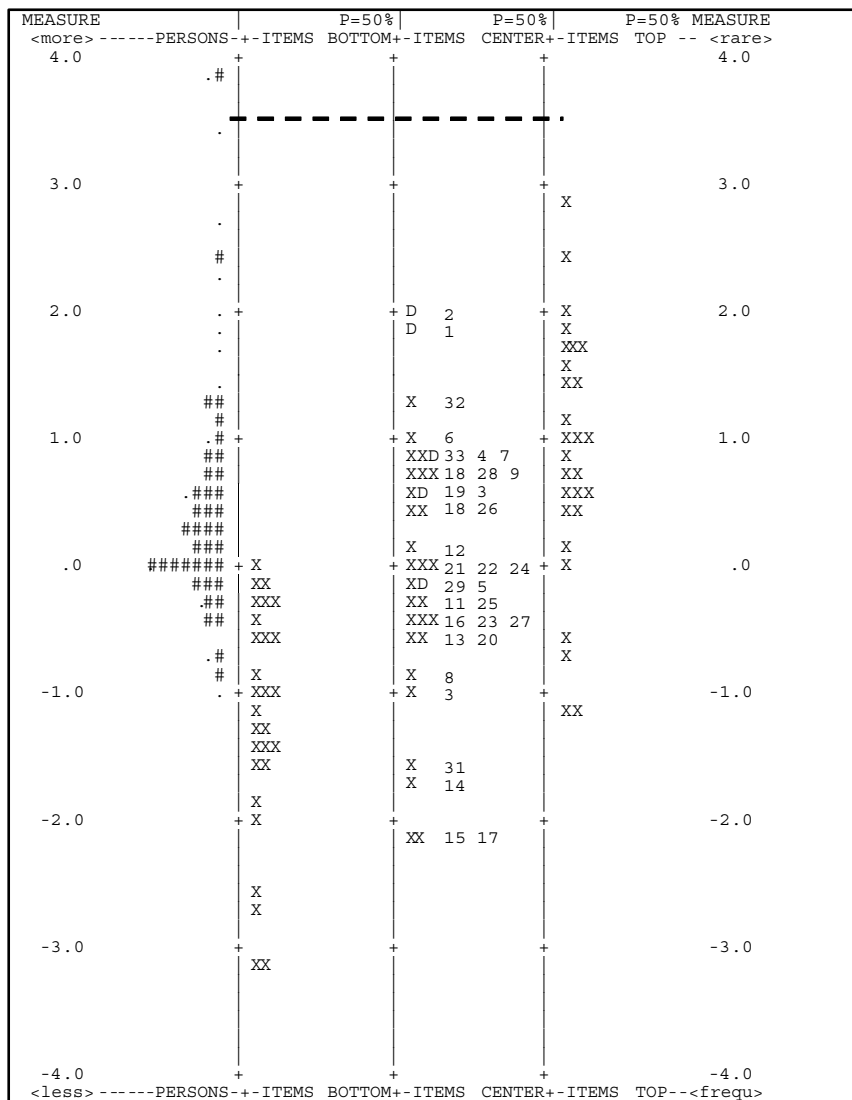


Figure 4. Distribution map for persons and items.

functioning level of the items in the continuum (item score 1), whilst the 'items top' column shows the items maximum functioning level within the continuum (item score four or five depending on the type of item).

When analysing the EORTC QLQ-C30 persons and items map, we can see that the location of nearly all the patients within the continuum is in the area covered by the items. Therefore, the items are used to measure those patients presenting high variability in their HRQOL.

Starting from the position of the item category with the lowest score (item bottom column) and from the location of the persons (persons column), we can deduce that the patients do not have a very negative score in HRQOL. This can be appreciated in the item bottom column of the continuum, since very few patients are located there. However, many patients are located (above the dotted line in Figure 4) in the part of the continuum defined by the highest item score (item top column).

#### *Measure of the Persons*

The RSM gives us the relationship between the scores obtained with the EORTC QLQC-30 test and those measures assigned to the persons. This relationship can be seen in Table III. In this table we can see the measures corresponding to every possible score that the persons can get in a test and which could well be used as the test scoring table.

The measure column shows the ability of each person in the variable that has been measured: HRQOL. In the S.D. (standard deviation) column we record the estimation error for each ability level. This error (which indicates the accuracy of the measure, that is to say, the equivalent of the reliability in CTTs) varies with regard to location within the continuum. The score column corresponds to the possible scores that can be obtained in the EORTC QLQC-30. For the minimum score (33), that is, responding to the category with a value of 1 point in the 33 items, the RSM forecasts a measure value of  $-5.22$  logits and a standard error of 1.40; and for the highest score (124), that is, responding to the category with the highest score in the 33 items, the measure would be 5.26 logits and the standard error 1.42. For a score of 81 points in the EORTC QLQC-30, the ability

TABLE III  
Measures on complete test

Score	Measure	S.E.	Score	Measure	S.E.	Score	Measure	S.E.
33	-5.22E	1.40	64	-78	0.23	95	0.69	23
34	-4.54	0.99	65	-0.73	0.22	96	0.74	23
35	-3.87	0.69	66	-0.68	0.22	97	0.80	23
36	-3.48	0.57	67	-0.63	0.22	98	0.85	23
37	-3.20	0.49	68	-0.58	0.22	99	0.91	24
38	-2.98	0.45	69	-0.53	0.22	100	0.96	24
39	-2.80	0.41	70	-0.48	0.22	101	1.02	24
40	-2.64	0.39	71	-0.44	0.22	102	1.08	25
41	-2.50	0.36	72	-0.39	0.22	103	1.14	25
42	-2.38	0.35	73	-0.34	0.22	104	1.21	26
43	-2.26	0.33	74	-0.30	0.21	105	1.28	26
44	-2.15	0.32	75	-0.25	0.21	106	1.35	27
45	-2.05	0.31	76	-0.20	0.21	107	1.42	27
46	-1.96	0.30	77	-0.16	0.21	108	1.49	28
47	-1.87	0.29	78	-0.11	0.21	109	1.57	29
48	-1.78	0.29	79	-0.07	0.21	110	1.66	30
49	-1.70	0.28	80	-0.02	0.21	111	1.75	31
50	-1.63	0.27	81	0.02	0.21	112	1.85	32
51	-1.55	0.27	82	0.07	0.21	113	1.95	33
52	-1.48	0.26	83	0.11	0.21	114	2.06	34
53	-1.41	0.26	84	0.16	0.21	115	2.19	36
54	-1.35	0.25	85	0.21	0.22	116	2.32	38
55	-1.28	0.25	86	0.25	0.22	117	2.47	40

TABLE III  
Continued

Score	Measure	S.E.	Score	Measure	S.E.	Score	Measure	S.E.
56	-1.22	0.25	87	0.30	0.22	118	2.65	43
57	-1.16	0.24	88	0.35	0.22	119	2.85	47
58	-1.10	0.24	89	0.39	0.22	120	3.10	52
59	-1.05	0.24	90	0.44	0.22	121	3.41	60
60	-0.99	0.23	91	0.49	0.22	122	3.84	72
61	-0.94	0.23	92	0.54	0.22	123	4.55	1.01
62	-0.88	0.23	93	0.59	0.22	124	5.26E	1.42
63	-0.83	0.23	94	0.64	0.23			

measure forecast is 0.02 logits and the standard error 0.21 (the highest precision in the test is achieved at this point in the continuum).

The ability values given by this model for each person are preferable to the test scores (obtained by adding all the scores reached in each item), since for each level of the ability continuum we have the degree of precision of the measure. With the scores obtained in the test there is no specific information about the precision throughout the continuum, but of the reliability of the test as a whole. As we have guessed, this last utilisation has been carried out by the CTT.

## DISCUSSION

In this study we have presented the result of the application of a measure model based on the IRT, namely the RSM, to measure the HRQOL by means of the EORTC QLQC-30. As Revicki and Cella (Revicki and Cella, 1997) point out, the application of IRT models in the HRQOL measure has a series of advantages and disadvantages. Amongst the most outstanding metric advantages highlighted in this research are the following: (1) joint measure of persons and items in the same scale (logits). Starting from these models we can obtain the scale or ability values not only for persons but for items as well, which is very useful for their analysis. (2) IRT polytomous models inform us of the way in which the items and the different response categories work by the use of CPCs, as we have shown here. Moreover, we can add other advantages that have not been explored in this study, such as the possibility of identifying misfitting response patterns which could affect the validity of the measure.

In this research paper we have shown that the RSM is a viable and feasible way of approaching the HRQOL measure by studying the metric properties of the EORTC QLQ-C30 items. This should make us reflect on the possibility of using RTI-based models as opposed to classic approaches to measure (based on CTT), above all, considering the advantages of the application of these models (measures invariance, joint measure of persons and items, study of the way in which response categories operate, etc.)

There are also some disadvantages in IRT models when compared with the CTT (Hambleton, 1985). The latter presents great limitations, but is a simple model (in terms of easy calculation in the

estimation of metric properties), as well as flexible and very well known. The IRT solves many of the CTT deficiencies, but its models are more complex (requiring more complicated calculations and depending on software for its application), more rigid (if there is no fit between data and model it is not possible to establish a measure) and little known (at least in applied fields). These issues have hindered the extended use of IRT. In the specific case of the HRQOL measure this is highlighted in the few articles published on this subject. None the less, there are more and more periodicals which thanks to their publication of articles on this subject, are making the application of these models known to many and generating a greater interest in them in health researchers.

On the other hand, despite the IRT metric demands already mentioned with regard to the CTT, the results of this study show the viability of the EORTC QLQ-C30 test to measure HRQOL in this type of patients, supporting other studies carried out under the CTT which have yielded satisfactory results (Aaronson et al., 1993; Ringdal and Ringdal, 1993; Porzsolt et al., 1996; Sprangers et al., 1996; Godoy et al., 1999) The items analysis has shown that most of them are adequate to measure HRQOL. However, we must point out the lack of fit found in the health perception and quality of life perception items (items 32 and 33). Therefore, the lack of fit may be due to a violation of the RMS supposition about the homogenous discrimination of these two items with regard to the rest.

An analysis of the content of these two items shows that both ask in a global way about the health and the quality of life. This can originate a more negative HRQOL global perception (the item measures are: item32 = 1.24; item33 = 0.91) than their perception of symptoms and alterations of daily life, as consequence of the illness. That is to say, the answers to these items include a component of 'stigmatization' that doesn't appear in the rest of test items (previous items ask for symptoms and specific behaviors). This can be affecting the unidimensionality. This would not only explain the misfit, but also the low correlations and the negative sign of item32 and item33.

Finally, and along the same lines as other authors (Andrich et al., 1997; Bjorner et al., 2003; Lai et al., 2003; Hagquist and Andrich, 2003) would like to highlight in this article the importance of measuring HRQOL using IRT models, since such applications can be carried out as item banks and computerised adaptation tests, to

generate more precise tests which are easier to interpret in clinical practice and user- friendly for patients.

## APPENDIX

### **Items of the adapted Spanish version of EORTC-QLQ-C30 (vers. 2.0).**

A complete version of EORTC QLQ-C30 (version 2.0) can be found at <http://www.eortc.be/home/qol/ExplQLQ-C30.htm>.

### **Items with two categories of responses.**

1. Do you have any trouble doing strenuous activities, like carrying a heavy shopping bag or a suitcase?
2. Do you have any trouble taking a **long** walk?
3. Do you have any trouble taking a **short** walk outside of the house?
4. Do you have to stay in a bed or a chair for most of the day?
5. Do you need help with eating, dressing, washing yourself or using the toilet?

### **Items with four categories of responses.**

6. Were you limited in doing either your work or other daily activities?
7. Were you limited in pursuing your hobbies or other leisure time activities?
8. Were you short of breath?
9. Have you had pain?
10. Did you need to rest?
11. Have you had trouble sleeping?
12. Have you felt weak?
13. Have you lacked appetite?
14. Have you felt nauseated?
15. Have you vomited?
16. Have you been constipated?
17. Have you had diarrhoea?
18. Were you tired?
19. Did pain interfere with your daily activities?
20. Have you had difficulty in concentrating on things, like reading a newspaper or watching television?
21. Did you feel tense?

22. Did you worry?
23. Did you feel irritable?
24. Did you feel depressed?
25. Have you had difficulty remembering things?
26. Has your physical condition interfered with your **family** life?
27. Has your medical treatment interfered with your **family** life?
28. Has your physical condition treatment interfered with your **social** activities?
29. Has your medical treatment interfered with your **social** activities?
30. Has your physical condition caused you financial difficulties?
31. Has your medical treatment caused you financial difficulties?

**Items with five categories of responses.**

32. How would you rate your overall **health** during the past week?
33. How would you rate your overall **quality of life** during the past week?

**Note:** The EORTC QLQ-C30 is a copyrighted questionnaire. Requests for permission to use the questionnaire and for scoring instructions should be addressed to the Quality of Life Unit, EORTC Data Center, Avenue E. Mounier 83, Bte 11, 1200 Brussels, Belgium.

## NOTES

\*This study has been carried out thanks to a project financed by Andalusian Health Service, Andalusian Regional Government, Spain (Servicio Andaluz de Salud. Junta de Andalucía. España. Exp. N° 114//00).

## REFERENCES

- Aaronson, N. K., S. Ahmezdai, B. Bergman, 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 National Cancer Institute* 85, pp. 365–376.
- Andrich, D. JHAL de Jong and B.E. Sheridan: 1997, 'Diagnostic opportunities with the Rasch model for ordered response categories', in J. Rost and R. Langeheine (eds.), *Applications of Latent Trait and Latent Class Models in the Social Sciences* (New York, Vaxmann)

- Andrich, D.: 1978a, 'A rating formulation for ordered response categories', *Psychometrika* 43, pp. 561–573.
- Andrich, D.: 1978b, 'Scaling attitude items constructed and scored in the Likert Tradition', *Educational and Psychological Measurement* 38, pp. 665–680.
- Arrarás, J. I., J. J. Illarramendi, and J. J. Valerdi.: 1995, 'El cuestionario de calidad de vida para cáncer de la EORTC, QLQ-C30. Estudio estadístico de validación con una muestra española', *Revista of Psicol. Salud* 7(1), pp. 13–34.
- Bjorner, J., M. Kosinski and J. Ware: 2003, 'Calibration of an item pool for assessing the burden of headaches: An application of item response theory to the Headache Impact Test (HIT<sup>TM</sup>)', *Qualifying Life Research*, 12(8), pp. 913–933.
- Bond, T. G. and C. M. Fox: 2001, *Applying the Rasch Model* (Mahwah, NJ, LEA).
- Godoy, M. J., A. J. Rojas, J. L. Garcia Puche and J. Cabrera: 1999, 'Fiabilidad y validez de la versión española del EORTC QLQ-C30: medida de la calidad de vida en pacientes oncológicos avanzados', *Revista–Psic Salud* 11(1–2), pp. 125–139.
- Hagquist, C. and D. Andrich: 2003, 'Measuring subjective health among adolescents in Sweden', *Social Indicator Research* 68(2), pp. 201–220.
- Hambleton, R. K.: 1985, *Item Response Theory: Principles and Applications* (Kluwer Academic Publishers Boston).
- Higginson, I. and M. McCarthy: 1989, 'Evaluation of palliative care: Step to quality assurance?', *Palliative Medicine* 3, pp. 267–274.
- Lai, J. D. Cella, C. H. Chang, R. K. Bode and A. W. Heinemann: 2003, 'Item Banking to improve, sorter and computerize self-reported fatigue: an illustration of steps to create a core item bank from the FACIT-Fatigue Scale', *Qualifying Life Research*, 12(5), pp. 485–501.
- Linacre, J. and B. Wright: 1988, *BIGSTEPS ver 2.82. Computer Program*. www.winstep.com.
- Linacre, J. M. and B. D. Wright: 1998, *A User's Guide to BIGSTEPS* (Chicago, Mesa Press)
- Lunz, M. E., B. Wright and M. J. Linacre: 1990, 'Measuring the impact of judge severity on examination scores', *Applied Measur Education* 3(4), pp. 331–345.
- Masters, G. N.: 1982, 'A Rasch model for partial credit scoring', *Psychometrika* 47, pp. 149–174.
- Porzolt, F., C. Wöpl, C. E. Rist, R. Koza, G. Büchele and W. Gaus: 1996, 'Comparison of three instruemnts (QLQ-C30, SF-36 y QWB-/) measuring health-related Quality of Life/Quality of Well-being', *Psychooncology* 5, pp. 103–117.
- Revicki, D. A. and F. Cella: 1997, 'Health status assessment for the twenty-first century: Item response theory, item banking and computer adaptive testing', *Quality Life Research* 6, pp. 595–600.
- Ringdal, G. I. and K. Ringdal.: 1993, 'Testing the EORTC Quality of life Questionnaire on cancer patients with heterogeneous diagnoses', *Qualifying Life Research* 2, pp. 129–3140.
- Rojas, A. J., A. González, J. L. Padilla and C. Pérez-Meléndez: 2002, 'Two strategies for fitting real data to Rasch polytomous models', *Journal of Applied Measurement* 3(2), pp. 129–145.
- Siegrist, J. and y Junge, A.: 1989, 'Conceptual and methodological problems in research on the quality of life in clinical medicine', *Social Medicine* 29, pp. 463–468.
- Spitzer, W. O., A. Dobson, J. Hall, E. Chesterman and J. Levi: 1981, 'Measuring the quality of life of cancer patients', *Journal of Chron Disease.*, 34, pp. 585–597.

- Sprangers, M. A., M. Goenvold J. I. Arraras, J. Franklin, A. Te Velde, M. Muller, L. Franzini, A. Williams, H. De Haes, P. Hopwood, A. Cull, N. K. Aaronson.: 1996, 'The EORTC Breast cancer-specific Quality of Life questionnaire module: First results from a three country field study', *Journal of Clinical Oncology* 14, pp. 2756–2768.
- Toscani, F.: 1996, 'Classification and staging of terminal cancer patients: Rationale and objectives of a multicentre cohort prospective study and methods used', *Cooperative Research Group on Palliative Medicine. Support Care Cancer* 4(1), pp. 56–60.
- Wright, B. D. and G. N. Masters: 1982, *Rating Scales Analysis*. Chicago (Mesa Press).

*Methodology of Social Sciences*  
*University of Almeria*  
*Spain*  
*E-mail: arojas@ual.es*

Antonio J. Rojas Tejada

*Technical Scientist on Information*  
*Systems and Research*  
*FADA (Andalusian Foundation for the*  
*Attention to the Drugs Addiction) and*  
*Methodology of Social Sciences*  
*University of Huelva*  
*Spain*

Oscar M. Lozano Rojas