Developing a descriptive system for a new preference-based measure of health-related quality of life for children

Katherine Stevens

Accepted: 28 July 2009 / Published online: 20 August 2009

Abstract

Objectives The use of preference-based measures (PBMs) of health-related quality of life (HRQoL) is increasing in health care resource allocation decisions. Whilst there are measures widely used for this purpose in adults, research in the paediatric field is more limited. This paper reports on how the descriptive system for a new paediatric generic PBM of HRQoL was developed from dimensions identified in previous research.

Methods Existing scales from the paediatric literature were reviewed for suitability, and scales were also developed empirically, based on qualitative interview data from children, by taking adverbial phrases and confirming the ordinality by a ranking exercise with children. The resulting scales were applied to the dimensions from the previous research.

Results No suitable scales were found in the paediatric literature, so the empirically derived scales were used resulting in seven different types. Children were successfully able to rank these to determine the ordinality, and these types were applied to the dimensions.

Conclusions This work has empirically developed a descriptive system for the dimensions of HRQoL identified in previous research. Further research is needed to test the descriptive system on a paediatric population and reduce the number of dimensions to be amenable to health state valuation.

K. Stevens (✉)
Health Economics and Decision Science, School of Health
and Related Research (ScHARR), The University of Sheffield,
Regent Court, 30 Regent Street, Sheffield S1 4DA, UK
e-mail: K.Stevens@Sheffield.ac.uk

Keywords Health care rationing · Paediatrics · Quality-adjusted life years · Quality of life

Abbreviations

PBM	Preference-based measure
HRQoL	Health-related quality of life
QALYs	Quality-adjusted life years

Introduction

The use of preference-based measures (PBMs) of health-related quality of life (HRQoL) is increasing in health care resource allocation decisions. In the United Kingdom in particular, the National Institute for Health and Clinical Excellence (NICE) specifies that for its reference case, a PBM be used to quantify the benefits of an intervention [1]. PBMs allow the calculation of quality-adjusted life years (QALYs) by combining length of life with quality of life, which can be used in economic evaluation as part of a decision-making process. Whilst there are PBMs widely used for this purpose in adults, research in the paediatric field is more limited [2].

Research by Stevens [3] reported on the first stage in the development of a new generic paediatric PBM for children age 7–11 years, in order to start addressing this gap. The paper reported on the process of identifying relevant dimensions of health-related quality of life (HRQoL) for inclusion in the new measure. They were identified by undertaking qualitative interviews with children aged 7–11 years with a wide range of acute and chronic health conditions, to find out how their health affected their lives. The children were divided into two age groups according to

Table 1 Dimensions of health-related quality of life [3]

	(7–9 years)	(9–11 years)
1	Worried	Worried
	Scared	
2	Sad	Sad
	Upset	Upset
		Unhappy
		Miserable
3	Annoyed	Annoyed
	Frustrated	Frustrated
		Angry
4	Hurt	Hurt
	Pain	Pain
5	School work	Learning
6	Daily routine	Daily routine
7	Tired	Tired
	Weak	Weak
	Drowsy	Energy
		Weary
8	Joining in activities that want to	Joining in activities that want to
9	Sleep	Sleep
10	Jealous	
11		Embarrassed

their school year (7–9 and 9–11 years). Each group was sampled, interviewed and analysed independently to explore whether these groups identified the same dimensions and therefore shared a common HRQoL framework. The research found that they did share a common framework as the dimensions identified were almost identical; therefore, a measure could be developed for the age group 7–11 years as a whole. Eleven dimensions were identified from the interviews, covering social, emotional and physical aspects of HRQoL. These dimensions are reproduced from Stevens [3] in Table 1.

Having identified these dimensions, the next stage in the development of a paediatric PBM was to create a descriptive system based on these dimensions that is suitable for use in economic evaluation. This paper reports on how this descriptive system was developed. The aim was to begin to develop a descriptive system suitable for health state valuation, based on the dimensions identified from the previous interview work [3].

Background

Existing non-preference-based quality of life measures have generally taken an approach to descriptive system development whereby a series of items or statements are developed using focus groups, the literature or interviews. Work is then undertaken to develop order and scales for these items, or response options could be based on Likert-scale-type responses [4]. These are then reduced or sorted into factors or dimensions using psychometric techniques. Reduction of items is common as generally long lists of items are generated, which are too long to have each item in the final questionnaire, hence testing is useful to identify redundant items (e.g. if items are not used or are very similar to another item), incomprehensible or ambiguous items and to test the internal consistency of a scale [4]. Factor analysis or Rasch techniques can be used to do this and can also be used as complements rather than alternatives [5].

Stevens [3] took a different approach to the development of the dimensions, in that the dimensions of paediatric health-related quality of life were determined directly from qualitative interviews and analysis. The qualitative work provides supporting evidence as to why the dimensions arose, and the terminology of the dimensions is based on the terminology used in the interviews. There is very little guidance in the literature about how to develop levels for dimensions directly. One way could be to consider the use of standard response scales from the literature.

Most of the existing measures use categorical response scales for their items, including those based on options relating to frequency (e.g. never, sometimes, often), the intensity/severity of a dimension (e.g. a little, moderately, a lot) or the level of agreement with something (strongly agree, disagree etc.), also known as a Likert scale. [4].

Existing generic PBMs take different approaches when using scales. The EQ-5D takes the severity approach, using three levels for each dimension; the Health Utilities Index (HUI)2/3 has a mixture of both (severity and frequency); and the SF-36 (used to obtain the SF-6D) has a mixture of both, but is mainly a frequency-based approach [6]. The levels on the EQ-5D descriptive system (a generic preference-based measure for adults) were developed to be ordinal and were developed using an expert panel. The developers also recommend using severity-based scales although they do not justify why [7].

The choice of scale can make a substantial difference to the descriptive system. For example, a frequency-based scale may not capture the range of how something can affect a person: for example, you can always be worrying, but only at a low level, which is different to being extremely worried. Equally, a scale based on severity may not adequately describe frequency. Another type of scale that is used in health status measures is the level of agreement, which asks a respondent how much they agree (or disagree) with a statement. This type of scale does not really make sense for a preference-based measure as you do not want a separate scale for each item level. There is also a scale that asks you to indicate how much something bothers you;

however, again this is not suitable for a preference-based measure as it is not useful for societal valuation, but may be useful for individual clinical decision-making.

The majority of scales used in existing paediatric measures are categorical response-type scales with a variety of response options, and the vast majority are frequency based rather than severity [8]. Most do not give any explanation as to how the levels or scales were developed. Those with a shorter recall period, the 16D/17D and HUI2/3 are statement based [9, 10].

There is not much empirical work in the paediatric field with regard to the use of response options and children's ability to understand and use them across ages. [8] Many existing measures use response options with between 3 and 7 points and there is literature that has shown that the number of categories used by raters should be in the region of between 5 and 7 as a maximum [4, 11]. Some measures use the same number of response options for each question, and some use different numbers of response options. The HUI2/3 and the 16D/17D use descriptive statements instead; however, these are still ordinal [9, 10]. There are also developmental differences in children's ability to understand and respond to items on a Likert scale. Eight-year-old children can accurately use a 5 or 7 point scale to rate their health status, whereas younger children tend to use more extreme responses. Some instruments have used visual aids to help with this, for example the Child Health and Illness Profile, which uses graduated circle sizes for the response options [12].

Another important feature of descriptive system development is the recall period. This is the time frame respondents are asked to think about when completing a questionnaire. In existing paediatric generic measures, there is a whole mixture of recall periods, from several weeks to the current day. More research is needed in this area about what is appropriate for children and different health conditions [8, 13].

Many of the existing paediatric instruments based on a frequency approach ask questions about how often something has been the case over the past few weeks. The evidence from the qualitative interviews undertaken in previous work by Stevens [3] is that children are able to recall information about their health and understand and describe it well, but often have difficulty remembering when they had a particular health problem or when an event had occurred. The advantage of asking about HRQoL today is that you are focusing on a point in time, and you also remove any potential problems with recall bias as children are thinking about the present time. The disadvantage is that this may miss important episodes in the context of a clinical trial for example, particularly in episodic conditions.

The main constraint in designing a descriptive system for a preference-based measure is that the health states defined by the system should be amenable to valuation. Ideally, each dimension needs to contain levels (response scales) that are ordered within it to fit this criteria well. There are also constraints on the number of dimensions that can be included due to limitations on people's ability to process information [11]. This paper reports on how levels were developed for the dimensions identified in previous work [3] to form a descriptive system amenable to valuation.

Methods

The first stage in developing the levels (response scales) for the dimensions was to determine whether they should be frequency based or severity based. To do this, data from the original qualitative work for developing the dimensions was used [3]. All the interview transcripts were reviewed, and adverbial phrases were extracted when the children were describing the dimensions and the way in which something was described, for example "it's a bit annoying" or "it's quite annoying". Phrases were extracted for each dimension separately and this was used to determine whether the dimension was about severity or frequency. In this way, the decision was based on the data.

Once this had been determined, the next step was to develop the scales for each dimension. Scales were developed based on the qualitative interview data from children and using guidance from the methodological literature [4] together with what is required for a PBM (i.e. ordinal levels within each dimension) [14]. The principles from the literature are as follows:

- Items should be clear, relevant and understandable.
- Scales will be developed with 5–7 levels with a view to reduction in further testing.
- Language should be kept simple.
- Double-barrelled questions will be avoided (asking two different things within one question.)
- Negatively worded items will be avoided, using positive wording styles instead.
- Vague quantifiers will be avoided, although this can be very difficult in practice.

In addition, the following approach was also used because of the qualitative data being used and the constraints of a PBM:

- The qualitative interviews were used to guide the wording of the levels, by analysing how the children described the problem, e.g. "It hurts a bit" and "It hurts a lot".
- Levels were ordinal, using an adjectival scale with discrete responses.
- Language was based on the qualitative data.

From the original qualitative work, there were alternative wording terms used to describe the dimensions, for example pain and hurt. Where more than one term existed, the alternative wordings were each developed into separate questions for future-testing work about which was the most appropriate.

Not all terms were used as alternatives, as sometimes words were used by the older age group and so were more complex, for example miserable. As the measure was being developed for the two age groups combined (as they were found to have a common HRQoL framework in the earlier research) [3], where there was a choice over wording, the wording used by the younger age group was selected.

The final questions developed were as follows. Worried and scared were developed as separate questions, and sad and upset were developed as separate questions. Miserable is just a more sophisticated wording style by the older children and was therefore not included. Unhappy was felt not to be a good term for use in a questionnaire as it is negatively worded and so was not included. Annoyed, frustrated and angry were all developed as separate questions. Hurt and pain were developed as separate questions. School work and learning were referred to as the same thing in the interviews, therefore the younger children's terminology was used (i.e. school work). Daily routine was the same for both age groups, so this was developed into a question. Tired and weak were developed into questions as drowsy and weary were not in common across age groups, and energy is the opposite meaning. Joining in activities was the same for both age groups, so this was developed into a question. Sleep was the same for both age groups, so this was developed into a question. Finally, jealous and embarrassed were both developed into questions.

This resulted in seventeen questions in total: Worrying; Sad; Weak; Angry; Pain; Frustrated; Hurting; School Work; Upset; Tired; Annoyed; Scared; Sleep; Embarrassed; Jealous; Daily Routine and Joining in activities.

As described previously, the qualitative data were used to develop levels (response scales) for each of these 17 questions. In addition, the wording used tried to incorporate the ways in which children had described the dimensions, for example for worried, sad, angry, weak and embarrassed, children were often using the term "feel". For hurt and pain, they were describing it in terms of it hurting or having pain.

Whilst the scales developed would be based on children's descriptions, the ordinality of these scales needed to be confirmed. As children have been involved at every stage of the development of this measure, and the measure is intended for children, it was important to verify the order of the scales with them.

The ordinality of the scales developed was tested by asking children to rank the levels in order of their severity. Children were sampled from the same two schools used in the original qualitative work [3].

Levels (response scales) were created for each question by applying the scales developed. These scales were applied to all seventeen questions. Piloting of the ranking work with children demonstrated that 17 ranking exercises was infeasible for them to do in one sitting, and so a subset of the scales from the questions were ranked, making sure each type of scale developed was covered. This assumes that the ordinality of the scale is independent of the item (question).

Cards were created for each question being tested, with each card displaying a level, and these were put together into a coloured envelope, one for each question/scale being tested. Children were asked to choose an envelope, one at a time and asked to rank the levels on the cards in order of severity (how bad they thought they were) from best to worst. Ties were allowed. Where children ranked levels as equal, they were asked if they had a preference for the wording. The ranking work was first piloted on 10 children aged 7–11 years (5 male and 5 female). They were able to complete the tasks successfully and advised on the size of the cards, the font used and the colours of the cards.

For the main study, 31 children were sampled from both schools involved in the research, and each child carried out the same number of ranking exercises. The aim of the sampling was to get an equal balance across gender and all year groups and to include both schools equally. The number of children included in the study was based on what was possible given resource constraints, as there was only one researcher undertaking this work, with a limited time period. Ethical approval and consent from the parents of children in both schools had already been obtained when the qualitative work was undertaken [3]. Children were sampled from those where parents had given their consent for the researcher to approach the child to ask if they would like to participate in the research. Children were approached one by one, and the study was explained to them with the aid of an information leaflet, which they could take and keep. The children had an opportunity to ask any questions they liked before being asked if they would like to take part. If children consented to take part, they were given the ranking tasks to do. All children carried the task out by themselves with the researcher sat with them in the school library or the dining room. The children's rankings for each of the sets were recorded by the researcher, along with any comments on preferences for wording where levels were ranked equally.

Analysis

The rank data was analysed by looking at the mean ranking and variation (standard deviation) and by using Kendall's coefficient of concordance test statistic. The approach of looking at the mean ranking is similar to the work undertaken by Keller et al. [15] as part of their work testing the equivalence of translations of widely used response choice labels, where they looked at the mean response choice ratings by country and language.

The Kendall statistic is between 0 and 1 and is a measure of the agreement between rankings, 0 means there is no agreement between rankings. It measures the extent to which ordering by each of two (or more) variables would arrange the observations into the same numerical order [16].

The rank data was coded using the mid rank method [17, 18] as this is more appropriate for this type of analysis and ensures that the sum of ranks is maintained. That is, a rank of 1 was coded as 1, a rank of 2 was coded as 2 and where rankings were tied, each tied ranking was given a value of the midpoint of the previous and next ranks. For example, a ranking sequence where the second and third cards were ranked equally was coded as 1, 2.5, 2.5, 4 and 5.

Where there was a very small difference between mean rankings, this was taken to mean that only one statement was needed for the descriptive system. A difference in mean ranking of less than 0.20 (chosen as a very low and conservative estimate) was taken to be a small difference. Whilst a difference of 0.20 was an arbitrary choice, this was chosen as the aim was to be conservative so that any removal of levels due to redundancy was based on a clear overlap.

In order to choose between the statements, the variation and the preferences of children for the wording was looked at, with the least amount of variation taking priority.

Results

For every dimension, severity arose as the predominant characteristic. In a couple of dimensions (worrying and angry/annoyed/frustrated), frequency arose in one case in each. For worrying, this was a mixture of the two "I always get a bit worried". For angry/annoyed/frustrated, it was frequency "it's always annoying". For sleep, one child described it in frequency terms "can't get to sleep that often". In the schoolwork, activities and daily routine dimensions, children were describing how much they could or couldn't do something, which again indicated a severity approach.

As the vast majority of dimensions and evidence within dimensions steered towards a severity-based approach, the dimension scales developed were based on this.

The adverbs and adverbial phrases used to describe the dimensions in the qualitative data are listed below.

at all	a little bit	a bit	quite	quite a lot
much	a lot	very	very much	really

The only wording not included in this list was "kinda", as this is a colloquial word and was felt to be not appropriate to include.

Applying these phrases to the dimensions resulted in seven types of scale, some of which were very similar, but had subtle differences depending on how the dimension fitted with the wording. There were therefore 7 unique scales to test in the ranking work, and it was felt appropriate that each child should rank each one. Figure 1 gives the 7 scales tested and the dimensions (questions) to which each scale applies.

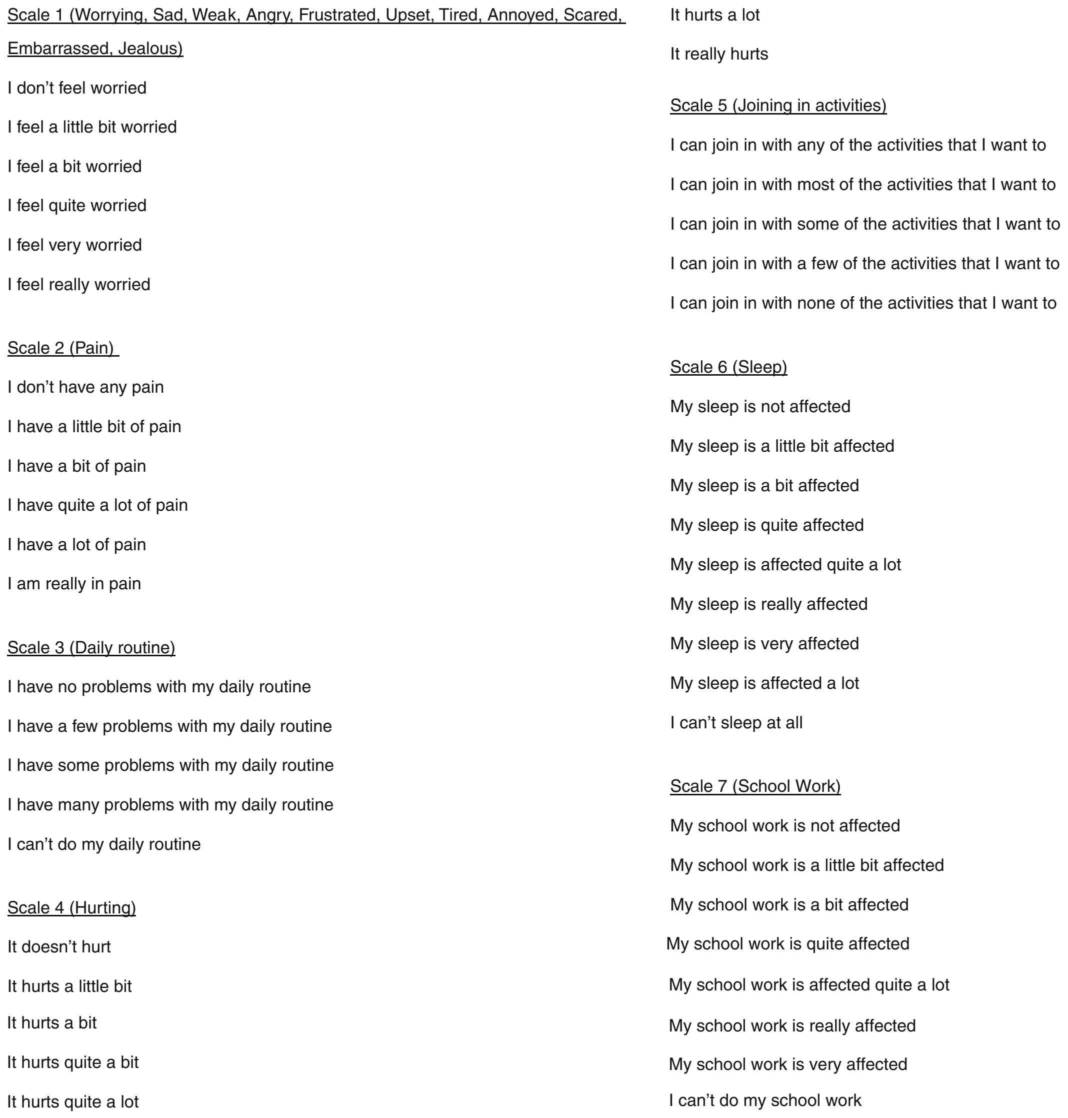
All 31 children consented to take part in the ranking, and all children completed all 7 ranking tasks. The characteristics of the sample are presented in Table 2. Table 3 presents the mean rank order, standard deviation and difference in mean rank for each of the 7 scales.

Table 4 shows the Kendall coefficient for each scale, which was very high for all scales. The lowest was for scale 3 (school work). An agreement of 0.81–1.00 is suggested to be almost perfect agreement for the Kappa statistic, which is another statistical measure of agreement [19].

The difference in the mean rank order was very low for the statements highlighted in bold in Table 3 ("My sleep is very affected" and "My sleep is really affected" had a difference of 0.05, "My school work is very affected" and "My school work is really affected" had a difference of 0.0. "I feel very worried" and "I feel really worried" had a difference of 0.16).

As there was such a small difference between these mean rankings, it indicated that only one statement was needed for the descriptive system. The preferences of children when these statements were ranked equally are presented in Table 5. The choice made over these three sets of statements where the difference in mean rank order was low was as follows:

1. Sleep: "Really" had a lower standard deviation and a smaller range (presented in Table 3). The preferences of the children were equal. Therefore, "My sleep is really affected" was chosen.
2. School work: "Really" had a lower standard deviation and a smaller range (presented in Table 3). "Very" has one more vote. Therefore, "My school work is really affected" was chosen.
3. Worried: "Really" and "very" have the same standard deviation and range (presented in Table 3). "Very" is preferred by one vote. Therefore, "I feel very worried" was chosen.



Scale 1 (Worrying, Sad, Weak, Angry, Frustrated, Upset, Tired, Annoyed, Scared, Embarrassed, Jealous)

I don't feel worried

I feel a little bit worried

I feel a bit worried

I feel quite worried

I feel very worried

I feel really worried

Scale 2 (Pain)

I don't have any pain

I have a little bit of pain

I have a bit of pain

I have quite a lot of pain

I have a lot of pain

I am really in pain

Scale 3 (Daily routine)

I have no problems with my daily routine

I have a few problems with my daily routine

I have some problems with my daily routine

I have many problems with my daily routine

I can't do my daily routine

Scale 4 (Hurting)

It doesn't hurt

It hurts a little bit

It hurts a bit

It hurts quite a bit

It hurts quite a lot

It hurts a lot

It really hurts

Scale 5 (Joining in activities)

I can join in with any of the activities that I want to

I can join in with most of the activities that I want to

I can join in with some of the activities that I want to

I can join in with a few of the activities that I want to

I can join in with none of the activities that I want to

Scale 6 (Sleep)

My sleep is not affected

My sleep is a little bit affected

My sleep is a bit affected

My sleep is quite affected

My sleep is affected quite a lot

My sleep is really affected

My sleep is very affected

My sleep is affected a lot

I can't sleep at all

Scale 7 (School Work)

My school work is not affected

My school work is a little bit affected

My school work is a bit affected

My school work is quite affected

My school work is affected quite a lot

My school work is really affected

My school work is very affected

I can't do my school work

Fig. 1 Scales tested (and applicable wording for questions)

Table 2 Characteristics of the sample

Characteristic	*N*
Hunter's Bar Junior School	16
Firs Hill Community Primary School	15
Male	15
Female	16
Y3 (age 7–8 years)	8
Y4 (age 8–9 years)	8
Y5 (age 9–10 years)	8
Y6 (age 10–11 years)	7
White	17
Mixed/dual heritage	2
Asian or Asian British	12
Black or Black British	0
Chinese	0

The results of this ranking exercise were then applied to the scales on all questions in order to form the draft descriptive system.

Discussion

A draft descriptive system has been developed from the dimensions formed from the original qualitative work [3]. This descriptive system is based on the qualitative data and is for both age groups combined. It contains 17 questions, some of which are alternative wording for the same dimensions, as further testing is required to determine the best wording. Instead of developing scales empirically, a scale could have been used from the paediatric literature; however, the only severity-based scale in the literature for

Table 3 Mean rank order, standard deviation (SD) and difference in mean rank, for each set of statements ($n = 31$)

Level	Mean rank order	SD	Difference
I can join in with any of the activities that I want to	1.10	0.30	0.92
I can join in with most of the activities that I want to	2.02	0.49	1.06
I can join in with some of the activities that I want to	3.08	0.43	0.73
I can join in with a few of the activities that I want to	3.81	0.46	1.19
I can join in with none of the activities that I want to	5.00	0.00	
My sleep is not affected	1.00	0.00	1.52
My sleep is a little bit affected	2.52	0.71	0.26
My sleep is a bit affected	2.77	0.59	1.05
My sleep is quite affected	3.82	0.75	1.26
My sleep is affected quite a lot	5.08	0.50	1.23
My sleep is affected a lot	6.31	0.69	0.92
My sleep is very affected	7.23	0.92	**0.05**
My sleep is really affected	7.27	0.76	1.73
I can't sleep at all	9.00	0.00	
My school work is not affected	1.19	1.08	1.32
My school work is a little bit affected	2.52	0.70	0.32
My school work is a bit affected	2.84	0.66	1.02
My school work is quite affected	3.85	0.83	1.16
My school work is affected quite a lot	5.02	0.70	1.27
My school work is very affected	6.29	1.08	**0.00**
My school work is really affected	6.29	0.69	1.71
I can't do my school work	8.00	0.00	
I don't feel worried	1.00	0.00	1.27
I feel a little bit worried	2.27	0.48	0.73
I feel a bit worried	3.00	0.55	0.73
I feel quite worried	3.73	0.60	1.69
I feel very worried	5.42	0.45	**0.16**
I feel really worried	5.58	0.45	
I don't have any pain	1.00	0.00	1.29
I have a little bit of pain	2.29	0.42	0.42
I have a bit of pain	2.71	0.42	1.58
I have quite a lot of pain	4.29	0.48	0.79
I have a lot of pain	5.08	0.59	0.55
I am really in pain	5.63	0.66	
I have no problems with my daily routine	1.00	0.00	1.27
I have a few problems with my daily routine	2.27	0.40	0.45
I have some problems with my daily routine	2.73	0.40	1.31
I have many problems with my daily routine	4.03	0.18	0.94
I can't do my daily routine	4.97	0.18	
It doesn't hurt	1.00	0.00	1.34
It hurts a little bit	2.34	0.57	0.55
It hurts a bit	2.89	0.59	0.89
It hurts quite a bit	3.77	0.48	1.52
It hurts quite a lot	5.29	0.51	0.66
It hurts a lot	5.95	0.57	0.81
It really hurts	6.76	0.56	

Table 4 Kendall coefficient

Set	Kendall coefficient
1	0.925
2	0.939
3	0.880
4	0.918
5	0.914
6	0.954
7	0.933

Table 5 Preference of children when statements were ranked equally

	Statement	Children's preference (*n* preferring each statement)
1	My sleep is very affected	1
	My sleep is really affected	1
2	My school work is very affected	3
	My school work is really affected	2
3	I feel very worried	3
	I feel really worried	2

paediatric generic instruments is the scale from the KIDSCREEN [20]. This scale is for children aged 8–18 years and uses the scale:

Not at all	slightly	moderately	very	extremely

The words "slightly", "moderately" and "extremely" never appeared in the qualitative interviews undertaken in the original qualitative research [3] and seem complex for young children and so this was felt not to be a suitable option.

The dimensions contain levels (response scales) that are based on severity, which was determined empirically from the qualitative data. The original interviews contained a good mix of acute and chronic conditions such as sickness, fever, flu, pneumonia, hearing problems, vision problems, asthma, weak wrists and ankles, eczema, hyperactive fits and abnormal muscle growth. Children with these problems all described the dimensions mainly in terms of severity, whether they had acute or chronic conditions.

The ranking exercise worked well with children, and they were successfully able to complete the tasks with a 100% completion rate. The ordering of the statements resulting from the analysis made sense at face value, and there was very good agreement in the rankings by children. Whilst the sample size was quite low in this study, the high agreement in rankings gives confidence in the results produced.

The advantages of this new descriptive system being developed with children are that the language is appropriate to children of this age group, which will aid self-completion and the content validity is more likely to be increased. Whilst this has been developed with children age 7–11 years, it may be the case that the descriptive system is also suitable for other paediatric age groups; however, this cannot be determined without further empirical testing.

In comparison with the only other existing paediatric generic preference-based measure, the HUI2, all the dimensions in the new measure are based on severity, whereas the HUI2 contains a mixture of severity and frequency-based items. However, both measures are statement based (rather than having an item and then a standard response scale). This makes the descriptive systems more amenable to valuation as a health state can be formed from these statements, whereas the language may be clumsy with a standard response scale as the item and response scale are separate.

The spacing of the scales is not necessarily even; however, they do not have to be equally spaced as ultimately this will be a preference-based instrument, and those levels that are too close will drop out in future-testing work. It is also possible that there are too many levels as whilst the principle was to aim for 5–7 levels, a few of the scales have more than this number (sleep and school work with 9 and 8, respectively); however, in scale development, it is usual to start with too many levels and then reduce these down. These issues will be addressed in future work.

Conclusion

This work has empirically developed a descriptive system for the dimensions of HRQoL identified in the original interview work. As the methods were based on using the data from children, the content validity should be increased. Seventeen questions are included within the descriptive system, some of which are alternative wordings for the same dimension. Further research is needed to test these alternative wordings on a paediatric population and to test the psychometric performance of this descriptive system. In addition, due to the constraints of PBMs, the number of dimensions will need to be reduced to be amenable to valuation. Further research is required to do this.

Acknowledgments This research was funded as part of an MRC Special Training Fellowship in Health Services and Health of the Public Research. Thanks go to The Children and Young People's Directorate at Sheffield City Council, the staff of Firs Hill Community Primary School and Hunters Bar Junior School, to the parents who gave their consent, to the staff at the schools and to all the children who took part in the research.

References

1. NICE (National Institute for Clinical Excellence). (April 2004). Guide to the methods of technology appraisal.
2. McCabe, C. (May 2003). Estimating preference weights for a paediatric health state classification (HUI2) and a comparison of methods. Ph.D.Thesis. The University of Sheffield.
3. Stevens, K. J. (2008). Working with children to develop dimensions for a preference based generic paediatric health related quality of life measure. Health Economics and Decision Science Discussion Paper 08/04. Available from http://www.shef.ac.uk/scharr/sections/heds/discussion.html Accessed 14/08/2008.
4. Streiner, D. L., Norman, G. R., & Health Measurement Scales. (1995). *A practical guide to their development and use* (2nd ed.). Oxford: Oxford University Press.
5. Tennant, A., McKenna, S. P., & Hagell, P. (2004). Application of Rasch analysis in the development and application of quality of life instruments. *Value in Health, 7*(Supplement 1), S22–S26.
6. Brazier, J. E., Ratcliffe, J., Salomon, J., & Tsuchiya, A. (2007). *Measuring and valuing health benefits for economic evaluation*. Oxford: Oxford University Press.
7. Kind, P., Brooks, R., & Rabin, R. (Eds.). (2005). *EQ-5D concepts and methods, a developmental history. Chapter 3*. Berlin: Springer.
8. Eiser, C., & Morse, R. (2001). Quality-of-life measures in chronic diseases of childhood. *Health Technology Assessment, 5*(4), 1–156.
9. Apajasalo, M., Sintonen, H., Holmberg, C., et al. (1996). Quality of life in early adolescence: A sixteen-dimensional health-related measure (16D). *Quality of Life Research, 5*, 205–211.
10. Health Utilities Index. http://healthutilities.biz/ Accessed 14/07/2008.
11. Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review, 63*(2), 81–97.
12. Riley, A. W., Forrest, C. B., Rebok, G. W., et al. (2004). The child report form of the CHIP-child edition: reliability and validity. *Medical Care, 42*(3), 221–231.
13. Matza, L. S., Swensen, A. R., Flood, E. M., et al. (2004). Assessment of health related quality of life in children: A review of conceptual, methodological, and regulatory issues. *Value in Health, 7*(1), 79–92.
14. Brazier, J. E., Deverill, M., Green, C., et al. (1999). A review of the use of health status measures in economic evaluation. *Health Technology Assessment*, *3*(9), 1–164.
15. Keller, S. D., Ware, J. E., Gandek, B., et al. (1998). Testing the equivalence of translations of widely used response choice labels: Results from the IQOLA project. *Journal of Clinical Epidemiology, 51*(11), 933–944.
16. Bland, M., & Peacock, J. (2001). *Methods based on rank order. In statistical questions in evidence-based medicine*. Oxford: Oxford Medical Publications, Oxford University Press.
17. Argyrous, G. (2006). *Rank-order tests for two or more samples. In statistics for research with a guide to SPSS* (2nd ed.). London: Sage.
18. Hinton, P. R. (1995). *Statistics explained. A guide for social science students*. London: Routledge.
19. Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*, 159–174.
20. MAPI research trust. Patient-Reported Outcome and Quality of Life Instruments Database. Available from: www.proqolid.org Accessed 13/04/2006.