Quality of Life Research, 7, pp. 323–335

A model of equivalence in the cultural adaptation of HRQoL instruments: the universalist approach

M. Herdman,* J. Fox-Rushby and X. Badia

Catalan Institute of Public Health, Universitat de Barcelona, Barcelona, Spain (M. Herdman, X. Badia); Department of Public Health and Policy, London School of Hygiene and Tropical Medicine, London, UK (J. Fox-Rushby)

The health-related quality of life (HRQoL) literature presents a confused picture of what 'equivalence' in the cross-cultural use of HRQoL questionnaires means and how it can be assessed. Much of this confusion can be attributed to the 'absolutist' approach to the cross-cultural adaptation of HRQoL questionnaires. The purpose of this paper is to provide a model of equivalence from a universalist perspective and to link this to the translation and adaptation of HRQoL questionnaires. The model evolved from reviews of the HRQoL and other literatures, interviews and discussions with researchers working in HRQoL and related areas and practical experience in the adaptation and development of HRQoL instruments. The model incorporates six key types of equivalence. For each type of equivalence the paper provides a definition, proposes various strategies for examining whether and how types of equivalence can be achieved, illustrates the relationships between them and suggests the order in which they should be tested. The principal conclusions are: (1) that a universalist approach to the cross-cultural adaptation of HRQoL instruments requires that six types of equivalence be taken into account; (2) that these are sufficient to describe and explain the nature of the cross-cultural adaptation process; (3) that this approach requires careful qualitative research in target cultures, particularly in the assessment of conceptual equivalence; and (4) that this qualitative work will provide information which will be fundamental in deciding whether to adapt an existing instrument and which instrument to adapt. It should also result in a more sensitive adaptation of existing instruments and provide valuable information for interpreting the results obtained using HRQoL instruments in the target culture.

Qual. Life Res. 7: 323–335 © 1998 Lippincott-Raven Publishers

Key words: Equivalence; health-related quality of life; questionnaire; cross-cultural; adaptation.

*To whom correspondence should be addressed at Catalan Institute of Public Health, Universitat de Barcelona, Campus de Bellvitge, Ctra de al Feixa Llarga S/N, 08907 L'Hospitalet de Llobregat, Barcelona, Spain. Tel: 343 4024250; Fax: 343 4024258; email: mherdman@bell.ub.es

Introduction

In a previous paper[1] we presented evidence of confusion in discussions of equivalence between different language versions of generic health-related quality of life (HRQoL) questionnaires. There were, for example, references to 19 different types of equivalence and substantial variation in the way some types of equivalence were defined. We argued that much of that variation, particularly in relation to conceptual evidence, could be traced to the 'absolutist' approach[2] adopted in much cross-cultural work in the HRQoL field.

The absolutist approach makes the initial assumption that there will be a nil or negligible change in the content and organization of concepts such as HRQoL across cultures, and that careful attention to linguistic elements will make a questionnaire developed for use in one culture acceptable for use in another culture. We have argued that such an assumption should be supported by strong theoretical and empirical evidence and that this is currently not the case in the HRQoL field.

We therefore recommended that researchers adopt a universalist approach to cross-cultural research. Such an approach does not make the prior assumption that constructs will be the same across cultures and, consequently, implies a need to establish whether the concept exists and is interpreted similarly in the two cultures and, if so, the degree to which it is interpreted similarly. The universalist approach aims to elicit those aspects of a concept which are genuinely universal across cultures and to use only those in developing instruments which measure the concept in different cultures. The model of equivalence presented here aims to reflect this approach, an approach which may be particularly important when HRQoL questionnaires are increasingly 'exported' to cultures markedly dissimilar to the originating culture. One outcome of adopting a universalist perspective may

be to suggest that a questionnaire is not suitable for translation into the target language before translation takes place. The model presented here will also be useful in suggesting which existing instruments are most suitable for use in other cultures, as well as providing information which will be useful in interpreting the results obtained with any instrument.

The objectives of this paper are (1) to present a model of equivalence based on the universalist approach, (2) to provide a definition for each type of equivalence included in the model, (3) to propose various strategies for examining whether and how the different types of equivalence can be achieved (4) to illustrate the relationships between them, and to suggest the order in which they should be tested. This model will also place current translation methodologies within a broader context and provide a cogent basis for critiquing and reviewing those methodologies, as well as highlighting the implications for cross-cultural research in the HRQoL field.

Developing the model

In developing the model, we have drawn to a large extent on references to cross-cultural work in the HRQoL field[3–5] and other literature,[6] on our combined experience in the area of translation and questionnaire development[7–9], and on interviews and discussions with researchers working in HRQoL and related areas.

A number of criteria guided our choice of the types of equivalence to be included in the model. The universalist approach emphasizes the possibility of cross-cultural variance in the nature of multidimensional concepts such as HRQoL, making it important, therefore, (1) to investigate which domains are important to the concept in the target culture and the relationships between them (conceptual equivalence), (2) to examine critically the items used to tap those domains as the relevance of items may vary across cultures (item equivalence), (3) to ensure that any translation which takes place leads to semantically equivalent items (semantic equivalence), (4) to ensure that the measurement methods used are appropriate to the culture in question (operational equivalence) and (5) to examine the outcome of the process in terms of instrument behaviour (measurement equivalence). As each of these different types of equivalence is important to the process as a whole it is important to be able to summarise the process and findings with a final type of equivalence (functional equivalence).

Types of equivalence and their assessment

Conceptual equivalence

Investigating conceptual equivalence essentially involves exploring the ways in which different populations conceptualize health and quality of life (QoL) and the values they place on different domains of health and QoL. The universalist approach to cross-cultural research in particular guards against the automatic assumption that domains which are relevant to HRQoL in one culture will be (equally) relevant in all cultures and implies that such claims should be tested empirically before adapting an instrument. It has been shown, for example, that the domain of family relationships may well be more important in Spain than in North America[10] and, in Kenya, that the nature and range of familial relations is different to that represented in existing instruments.[9] The definition of conceptual equivalence offered here is at odds with many of the definitions offered in the HRQoL literature,[11,12] but is an important theoretical baseline within the universalist approach.

Conceptual equivalence between questionnaires will be achieved when the questionnaire has the same relationship to the underlying concept in both cultures, primarily in terms of the domains included and the emphasis placed on different domains. Before the degree of conceptual equivalence can be decided, however, careful research will be necessary to determine how health and QoL are conceptualized in other cultures, particularly in terms of the nature of and emphasis on particular domains. This stage of the process provides a background against which the legitimacy of adapting a questionnaire can be judged, as well as providing a context for the interpretation of results. This definition of conceptual equivalence also implies that it will not be possible to achieve or examine conceptual equivalence simply through translation and the *post hoc* analysis of results obtained using the questionnaire.

Methods for investigating conceptual equivalence. An initial assessment of the conceptual equivalence of HRQoL in the source and target cultures will involve examining the nature of the HRQoL concept in both cultures. Where there has been substantial research on the nature of HRQoL in the source culture, information on its form and content can be obtained through literature reviews concerning the theoretical[13] and empirical explorations of the concept *per se*,[14] as well as through reviews of instrument development.

In the target culture, there is a range of potential

research approaches to establishing local perceptions of HRQoL. The first resort is to local literature, particularly general ethnographies, as well as publications on perceptions of health, well-being, illness and disease in the target language and communities. In some cases, questionnaires dealing with similar or related topics (for example, the expression of emotions) may already have been developed in the target culture and may indicate initial points of convergence or divergence with the conceptual structure of existing instruments. Secondly, it is possible to consult experts in the target culture, though it should be remembered that the aim is to obtain a picture of the cultural environment in which the instrument may be employed. The range of experts consulted should therefore be broad, including, for example, anthropologists, medical sociologists, linguists and QoL experts, as well as health professionals. The third approach is to involve a wider representation of the general population in an investigation of beliefs and behaviours regarding health and QoL. For example, the WHOQOL Group[15] used focus groups in which they posed questions such as 'What are words or phrases that describe 'quality of life'? A more dynamic and interactive approach to language was presented by Amuyunzu *et al.*[16] Other authors[17] have gone further and recommended 'protracted periods of participant observation and repeated open-ended unstructured interviews with participants over a period of time'. Such an approach has the advantage that concepts can be gleaned from the target group's perspective, thus increasing the likelihood that their views are captured. Such ethnographic strategies are described in Lonner and Berry.[18] Researchers from different cultures could also undertake the research in both the target and source cultures, as a useful way of identifying biases in their own views of health and HRQoL. This approach will be particularly important when there is little published literature or when the literature is dated or reflects only one point of view. The assessment of conceptual equivalence is most likely to need all three approaches. Once domains which are important to the HRQoL concept in the target culture are selected, it may be possible to make an initial assessment of the relative importance of those domains by asking a representative group of respondents to prioritize them in order of importance.[14]

Possible outcomes of investigating conceptual equivalence. In broad terms, investigations of conceptual equivalence have four possible outcomes.

(1) The domains employed in the source instrument are equally relevant and important to the concept in the target culture and the differing emphasis placed on different domains is also equally appropriate in both cultures, indicating that the construct employed in the original questionnaire is likely to be equally valid in the target culture.
(2) Though the domains incorporated in the original instrument are also relevant to the HRQoL concept in the target culture, the importance of the domains varies between the two cultures (i.e. the emphasis which should be placed on different domains varies).
(3) One or more of the domains used in the original instrument are not relevant to the concept of HRQoL in the target culture or domains which are relevant to the concept in the target culture are not included in the source instrument.
(4) The domains of HRQoL are different in the source and target cultures.

In the case of outcome (2) it may be possible to weight the domains in order to reflect differing importance;[10] while in the case of outcome (3) it might be possible to use the relevant domains from the original instrument, though careful attention should be paid to the effects on the psychometric properties of the instrument if only some domains are used, careful retesting of the validity and reliability will be required, and what remains will only allow a partial comparison across cultures. In the case of outcome (4), the questionnaire should not be considered for adaptation and the adaptation of another instrument should be considered or a culture-specific questionnaire should be developed.

Item equivalence

Item equivalence concerns the way in which domains are sampled. In the same way that the relevance of domains to the HRQoL concept may vary across cultures, the validity of items as measures of a particular domain may also vary. Item equivalence exists when items estimate the same parameters on the latent trait being measured and when they are equally relevant and acceptable in both cultures.

The relevance of items will vary across cultures, for example items which ask about the use of sleeping pills will not be relevant in cultures where sleeping pills do not exist or items which ask about the ability to look after a garden will not be appropriate in cultures where a large part of the population do not have gardens.[8] Even items which ask about activities considered universal may in fact be asking very different things in different cultures. Actvities of daily

living,[19] such as getting dressed may not only involve a different series of movements in different cultures, but the inability to dress oneself might be perceived as more serious in some cultures than in others. The inability to dress oneself may therefore indicate very different levels of incapacity and/or emotional impact in different cultures.

As well as varying in relevance, items will often also vary in acceptability. Items may prove offensive to members of the target culture or may deal with taboo subjects.[5] When this is the case, it may be that careful rewording of the item will provide a solution, but occasionally items will have to be omitted, in which case it would be necessary to revalidate the questionnaire and to retest its psychometric properties before using it in a study population.

When items, or parts of items need to be replaced, it is very important to be clear about the purpose of the item. When the item's purpose can be clearly expressed, there is a greater likelihood that replacement items can be developed which measure the same trait to the same extent as the original item. In this context, the use of 'mapping sentences', as suggested within the facet theory approach, can be very useful in making explicit the underlying purpose of the original item.[20]

Investigating item equivalence. Investigating item equivalence involves making an initial qualitative examination of the relevance of items as well as the psychometric properties of the item in the target culture. In some cases the irrelevance of an item may be obvious, while in other cases the degree of relevance will be less obvious. For example, though doing vigorous activities it may be relevant to some extent in all cultures, it will vary in relevance across cultures.[5]

The relevance of items can be examined in different ways and the method chosen will depend on the type of information required to determine an item's relevance. A review of the available data on lifestyle patterns and habits might, for example, suggest that a certain item would be largely irrelevant, if it was found that it applied to only a small proportion of a given population. Hunt and McKenna[12] pointed out, for example, that the proportion of people engaging in vigorous activities was less than 20% in Britain and that the proportions were lower in France, Spain and Italy. In other cases or where available literature is scant, a structured process such as the Delphi technique[21] could be used to elicit 'expert' judgement, particularly that of anthropologists and sociologists familiar with the target culture, regarding the potential relevance of items. In many instances, however, the most appropriate source of information will be members of the target population itself, who could be accessed through a variety of research methods,[22] in which, for example, samples of the target population could be asked to discuss the relevance of items to themselves and to people they know or to rank items in a particular domain. Detailed research using the above three approaches may reveal possible alternatives or conclusive evidence for removal.

This initial investigation of the properties of items will obviously provide only a provisional idea of those properties. A more exact picture can be gained through the use of more sophisticated methods of psychometric testing once the pilot version of the questionnaire has been used in larger samples. In this respect Rasch item analysis[23] can be particularly useful in determining the extent to which a given item measures a given trait, whilst the assessment of internal consistency through the use of Cronbach's α[24] indicates the extent to which items may be measuring the same underlying trait.

Possible outcomes of investigating item equivalence. There are four possible outcomes of investigating item equivalence.

(1) Items can be used in the target version without modification (other than translation).
(2) Items require minor modifications, but may be used more or less in their original form.
(3) Replacement items must be used.
(4) Neither existing nor replacement items can be used because they deal with subjects which are considered offensive or taboo. In the case of outcomes (1) and (2) researchers should go on to translate and test the items further. In the case of outcome (3) the initial choice of the substitution items will be largely a question of considered judgement by researchers, based on literature reviews, expert opinion and/or input from members of the target population. In the case of outcome (4) it may be possible to omit items, but in this case it would be necessary to revalidate the questionnaire and to re-examine its psychometric properties.

Semantic equivalence

Semantic equivalence is concerned with the transfer of meaning across languages, and with achieving a similar effect on respondents in different languages. This means taking into account a number of different types of meaning. Linguists and translators in other

Table 1. A guide to different types of meaning

Type of meaning	Explanation
Referential meaning	Concerns the ideas or objects in the world that a word or words refer to, e.g. the part of the body referred to by 'abdomen'
Connotative meaning	Concerns the emotional response evoked by a word e.g. disease
Stylistic (social) meaning	Many words are used in specific social or stylistic contexts and are inappropriate in other contexts. Stylistic or social meaning can be transmitted via the following aspects of language: (1) geographical dialect, (2) technical language, (3) levels of formality, (4) poetic language, (5) time-restricted (old-fashioned or modern) and (6) age or sex restricted. A good example is the word 'blue' as it is used in some HRQoL questionnaires; it is geographically restricted, age restricted and time restricted
Affective meaning	Affective meaning refers to the way in which the words used can reflect the views and feelings of the writer (e.g. choosing between the words 'fat', 'plump' and 'chubby')
Reflected meaning	Words may sometimes be humorous or offensive because of their meaning when used in other contexts
Collocative meaning	Taking a word out of its usual context can give it extra force or meaning simply because of the strangeness of the collocation
Thematic meaning	In which a particular meaning is given to a message by the way it is organized e.g. word order can be used to highlight the importance of some aspects of the text whilst diminishing other aspects. Could be important in emphasising aspects such as the time frame of the questionnaire. Different languages emphasise words or ideas by using different parts of sentences, which will often argue against maintaining similar word order

Adapted from Barnwell K. *Introduction to Semantics and Translation*. High Wycombe Summer Institute of Linguistics, 1980.

fields have developed detailed classifications to aid the consideration of meaning, (see Table 1).

Not all of these types of meaning will be relevant in all cases and some will be more important than others, but achieving semantic equivalence will mean taking many of them into account. A clear example of the way in which this checklist of possible influences can be used is in reference to the phrase 'leisure activities'. In this case, the 'ideas or objects in the world which the word refers to' will be that broad group of activities which people do in their spare time. However, it should be noted that the referential meaning will vary between individuals, sub-groups and cultures as the word leisure activities will denote different activities for different people. The connotative meaning may vary in the sense that, for some people, leisure activities are something to be actively enjoyed, whilst others see them merely as a way of relaxing. The phrase itself may be more acceptable and more familiar to certain segments of a population than to others (social meaning) and, whereas inclusion in a questionnaire might be acceptable, its use in an interview situation might not be (collocative meaning).

One of the most important aspects of meaning is ensuring that the level of the language used is appropriate to the needs of the target population. For example, we have found during the translation of HRQoL questionnaires into Spanish that a direct translation of the word 'abdomen' was not understood by a large number of the respondents, as the register (level of language) was too high. The solution was to replace it with two words which together covered the same region as the original English word.

Methods for achieving semantic equivalence. Prior to any translation, it is important that key words or expressions within the questionnaire should be clearly understood. It may be the case that the developers of the original instrument provided descriptions of the ideas behind the language used in their questionnaire. The WHOQOL Group,[15] for example, provided descriptions of what each of the domains is intended to cover. Latterly, the EuroQol Group[25] have been

exploring the possibility of providing more detailed descriptions to show the range of ideas intended to be covered (or not) within key words/phrases. When such information is not available this should be the first task of any translator. This 'semantic rewrite' finds support in the work of Bible translators[26], and examples of a similar kind of exercise are provided in Sartorius and Kuyken.[27] When translators become aware of a variety of possible meanings, they can contact the original developers for clarification (and it is likely that this questioning will help some developers clarify the aims of their original version). Questionnaire developers should at least be aware that any instrument is likely to be translated and should be prepared to provide clear definitions to translation teams.

Establishing the meaning of items, words or phrases in the source language is one of the most common problems faced when translating HRQoL questionnaires. The word 'distress', for example, can be difficult to translate into other languages, principally because different cultures have different ways of classifying the experience of suffering. There may not be a word which is close in meaning to 'distress' or there may be a range of words referring to experiences which are similar to distress. A first step in exploring the meaning of a word is that of defining the 'semantic space' in which the word is located.[27] This step helps to define the relation of a particular word to other words which have some aspects in common with it. The lexical relationships of particular words or phrases, in both the source and target languages, can be further explored using checklists such as that shown in Table 2.

Although an awareness of the different types of meaning outlined above can be very useful in helping to clarify and think through translation issues, much of the translator's task cannot be reduced to a mere technical exercise and there is almost always an element of 'art' involved. The ability to express the original message as accurately, clearly and naturally as possible, the ability to find the right tone and register (level of language) and an awareness of the impression that his or her translation will make on the reader in the target language are all hallmarks of a good translator. Finding such translators may not be easy, however. Experience indicates that personal recommendations are often worth following up, in particular when people are located in the same geographical area where the questionnaires will be used, though it can also be useful to contact university translation/linguistics departments and local Chambers of Commerce. Translators should be made very aware of who is likely to be the target audience in order to ensure that they use a register (and dialect) which is appropriate to that audience. When there is a big difference in the level of education between translators and the target audience, it is more likely that the

Table 2. Checklist for establishing lexical relationships between words

Checklist for establishing lexical relationships between words
Determine the generic set of words which the target word belongs to involves finding those words with which the target word has shared components of meaning. Once the generic set has been established, the ways in which the target word differs from other words in that set should also be explored e.g. 'distress' in relation to 'discomfort'
Determine words to which the target word stands in a hierarchical relationship. A hierarchical relationship is the move from generality to specificity e.g. 'anxiety' in relation to 'distress'.
Determine the part-whole relationship (though not one of hierarchy), whereby an object is part of a larger or more complex whole, e.g. the stomach is part of the abdomen
Establishing synonyms and antonyms of the target word can be helpful in determining the meaning of the target word and may suggest different possible translations ,e.g. 'distress' is the antonym of 'well-being'
Derived forms are forms of words with the same lexical root, e.g. know, knowledge. Shared lexical roots can sometimes give insight into the nature and meaning of a word
Common and acceptable collocations. It is not always appropriate to use a word in conjunction with another e.g. in English it is appropriate to say 'I am in great pain', whereas it is not common or appropriate to say 'I feel great pain' and even less so to say 'I have great pain'

Adapted from Barnwell K. *Introduction to Semantics and Translation*. High Wycombe Summer Institute of Linguistics, 1980.

level of language used by the translators will be perceived as less natural by respondents to the questionnaire.

Further tests of translation quality could be achieved by asking translators who are not involved in the original process to comment on the semantic equivalence of the source and translated versions, whilst the linguistic quality of the translated items could be assessed by linguists who are experts in the target language (who need not be translators). Finally, asking a sample of the target population to paraphrase translated items will also provide an insight into the similarity of meaning between the target and source versions.[22]

Few of these ideas are currently written into translation guidelines. In addition to this, the HRQoL field has not addressed issues such as the use of protocols in translation meetings and the development of criteria for decision making in those meetings. Both are areas which would be of benefit, particularly to those with little experience in this field. On a more general level, greater discussion of the problems encountered in translation exercises and the ways in which problems are resolved would be helpful in exploring the interaction between language and culture in the field of health and QoL.

Possible outcomes of examining semantic equivalence. There are three possible outcomes for the testing of semantic equivalence (translation).

(1) Items are easy to translate.
(2) Items are difficult to translate.
(3) Items are impossible to translate.

Although all translated items should obviously be tested on a sample of the target population for comprehensibility, accuracy and naturalness before being incorporated into the final version of the questionnaire, the translation process will have highlighted potentially problematic items and special attention should be paid to these when they are tested on population samples. A balance should always be sought between accuracy and naturalness and items which fail to meet either of these criteria should either be retranslated or, as a last resort, replaced. Achieving a good translation will always involve repeated translations and reformulations. It may also be possible to replace items which prove impossible to translate, in the same way as described in the section on item equivalence.

Operational equivalence

Operational equivalence refers to the possibility of using a similar questionnaire format, instructions, mode of administration and measurement methods. Equivalence will be attained when these elements do not affect the results. For example, different levels of literacy affect the ability to use self-completed forms and the spread of ownership of telephones will determine whether telephone surveys can be used. Researchers should also be aware of the customs of addressing people. For example, in some places it may be inappropriate for a young person to ask an older person about certain issues (or vice versa). In other cases, there are examples in the literature of the format affecting response patterns in unexpected ways. For example the IQOLA Group found that levels of missing data were higher among Swedish respondents than American and British respondents in that part of the SF-36 which uses a grid format, though the levels were similar in items printed singly.[28]

Measurement methods (e.g. yes/no questions, visual analogue scales, Likert scales and paired comparisons)[29] may likewise not be equally suitable for use in all cultures. For example, when respondents are not even used to seeing maps of their own area, why should the idea of indicating sensation 'intensity' on a visual analogue scale make any sense, let alone the same sense? In other places, a direct 'yes' or 'no' is customarily avoided.[30] This highlights the need for some prior idea of the feasibility and relevance of different techniques.

Other important factors to be taken into account include the time frame of the questionnaire. Time is most often included as a frame of reference for all questions (e.g. a person is asked to think of the previous week) or used as a response category regarding the frequency of an experience (e.g. all, most, some or none of the time). There are many cultures which do not share the same 'chronological' conception of time presented in HRQoL questionnaires which view time as linear and geometric. For example, where people do not differentiate past, present and future, life is only discernable by 'today' and a question which focuses on the previous week can only be thought about in terms of events that have happened that day.[2]

Methods for investigating operational equivalence. A prior awareness of the potential suitability of different operational options will save later frustration. The methods used to investigate the possibility of operational equivalence will obviously depend on the particular aspect of questionnaire operationalization being examined and some operational effects may well appear only once the questionnaire is used. Similar methods to those used in investigating earlier aspects of equivalence will, nevertheless, be useful.

Apart from literature reviews regarding instrument use in a given culture, another important source of information would be the type of instruments used and the experience accumulated by researchers in other fields. This could also indicate the options available as regards time frames. Literacy rates would provide a good indication of the possibilities of successfully using written questionnaires. Anthropological and/or sociological data on cultural norms regarding forms of address and appropriate ways of framing questions will also provide insights into the best ways of obtaining information.

Interviews and the testing of proposed methods with samples of the target population will also indicate the extent to which these are likely to be successful, as well as potentially providing useful expressions, for example for time frames. In Spain, we have found that direct translations of expressions such as 'over the last week' and 'over the last month' were interpreted by approximately 50% of respondents in pilot testing as meaning over the last full calendar week, not the previous 7 days. This could obviously affect results when treatment effects are to be measured over time.

Outcomes of assessing operational equivalence. The possible outcomes of assessing operational equivalence are as follows.

(1) The same methods (mode of administration, measurement methods, format, time frame, etc.) can be used.
(2) Some aspects of operationalisation need to be different.
(3) It is impossible to achieve operational equivalence.

When the same operational methods can be used, it will nevertheless be important to test for possible systematic biases in the response patterns once the instrument has been used. When different operational methods need to be used, it will be important to review the available literature on the possibility of obtaining results which are unaffected by the mode of operationalization, for example, whether systematic differences between results obtained using telephone interviews and postal questionnaires, or using different formats are likely to make results non-comparable. Finally, in some cases, it may be impossible to achieve sufficient levels of operational equivalence, for example when response modes are incompatible or are likely to be affected by systematic biases or where time frames such as 'over the last week' are unusable.

Measurement equivalence

The aim of investigating measurement equivalence is to ensure that different language versions of the same instrument achieve acceptable levels in terms of their psychometric properties, primarily in terms of their reliability, responsiveness and construct validity (including an instrument's discriminant, evaluative and predictive properties). The degree of measurement equivalence is therefore defined as the extent to which the psychometric properties of different language versions of the same instrument are similar, though not all of these properties will be expected to be the same, as discussed below.

In the case of instruments which are scaled to provide scoring norms, the investigation of measurement equivalence should also include an examination of the scoring norms in the target culture, where this is feasible. Obtaining scoring norms for the target population may also provide insights into the degree to which items and, in some cases, domains are given equal importance in the source and target cultures.[10]

Methods for examining measurement equivalence. As there are numerous discussions in the literature[31,32] of the methods for examining the psychometric properties of HRQoL instruments, it would not be appropriate to examine them in depth here. We will limit ourselves to mentioning some of the possibilities in each case and to discussing some of the results which might be expected when they are used to establish the existence of measurement equivalence across cultures. Tests of reliability would include tests such as Cronbach's α[24] for internal consistency and the intraclass correlation coefficient[33] for test–retest reliability. Tests of responsiveness would include tests such as the paired *t*-statistic, the effect size statistic and the responsiveness statistic.[32] Tests of construct validity[34] are designed to determine the extent to which measures behave as they are expected to behave and include many different types of analysis, such as, for example, the extent to which instruments can discriminate between patients and healthy respondents or the extent to which instruments correlate highly with instruments purporting to measure the same trait (convergent validity) and less highly with instruments designed to measure dissimilar traits (divergent validity). Finally, in the case of multidimensional instruments, factor analysis can be used to examine the factor structure of a given instrument, to determine the extent to which items load onto the same factors in different language versions of the same instrument and to examine the extent to which a

similar amount of variance can be explained by a similar number and definition of factors.[35] The techniques for scaling instruments are well-documented in the literature,[10,36] though one of the commonest approaches is the use of Thurstone's[37] method of paired comparisons.

Outcomes of measuring measurement equivalence. It should be noted that, within the context of cross-cultural studies, although it is undoubtedly important to achieve very similar or equivalent results in some of these areas, notably in terms of reliability (internal consistency and test–retest reliability), in other cases it might be unrealistic to expect similar results, and similar results might even indicate that the response patterns were confounding results. A good example would be that of effect size:[38] effect sizes using the same instrument might very well be expected to differ in different circumstances, as they will be affected by a number of factors external to the instrument itself. In the case of effect sizes established by pre–post testing after an intervention of known efficacy, these might be expected to differ across cultures where the quality of care on offer differed or where the availability of follow-up facilities was superior in one location or where emotional responses to treatment differ due to cultural variations in attitudes to medical treatment and/or to illness itself. Problems in categorizing populations according to sociodemographic characteristics might similarly lead to difficulties in establishing whether an instrument's discriminative and evaluative properties, are similar or not, in particular where patterns of ill-health are different. These points highlight the need for a very careful interpretation of the results in the light of knowledge regarding the relevant aspects of the culture(s) in question.

Item weights may also be expected to differ in some cases across cultures, as they reflect the comparative importance given to different items in different cultures, though again it should be ascertained that the methodology used to obtain the item weights is equally applicable in both cultures. Whether or not it is preferable to obtain scoring norms through the use of population-based weighting tasks or through the use of other methods such as item response theory and whether these methods would give comparable results has not received very much attention in the HRQoL field. Finally, although factor analytic techniques can be useful in examining the similarity of the underlying instrument structure across cultures, it should be remembered that 'Factor analysis cannot substitute for preparatory qualitative and conceptual work. It is well known that the establishment of factors depends on the number of correlated items one has included, and that such a balance of items needs very thorough ethnographic conceptual work before drawing up a list of items'.[39] The model presented here aims to ensure that such ethnographic work is incorporated early in the adaptation process.

Functional equivalence

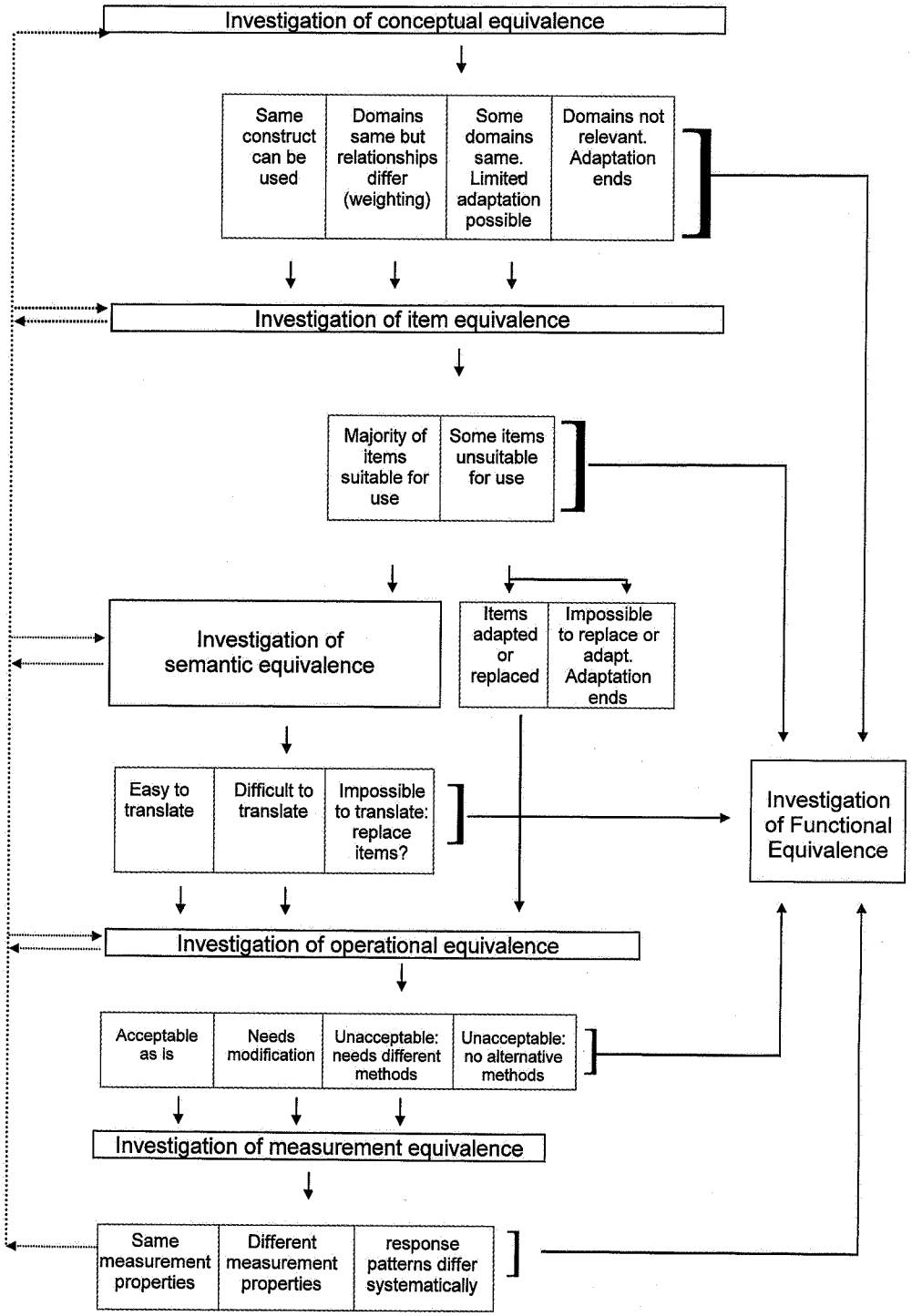
This final type of equivalence is intended to highlight the fact that all parts of the process outlined here are important in achieving cross-culturally equivalent questionnaires. Functional equivalence can therefore be defined as the extent to which an instrument does what it is supposed to do equally well in two or more cultures. If a measure is designed to measure QoL, how successful is it in measuring 'QoL' in a given culture? The answer to this question, when a universalist approach is adopted, depends on two things: firstly how 'QoL' is defined or conceptualized in a particular culture and, secondly, how successful the instrument is in measuring the trait it is supposed to measure. To be able to demonstrate functional equivalence, it is therefore necessary to be able to say, firstly, how the underlying trait is defined or conceptualized in the target culture, secondly, how well the instrument design reflects that underlying trait; and, finally, how the results obtained from a given instrument compare across cultures.

Assessing the degree of functional equivalence achieved is therefore a question of assessing the degree to which the other types of equivalence included in the model presented here have been achieved.

Outcomes of assessing functional equivalence. The possible outcomes of assessing functional equivalence can be categorized into three broad types according to the extent to which it is justifiable to compare and aggregate results across cultures.

(1) If a reasonable degree of equivalence is achieved in all areas, then the results are strictly comparable and may be aggregated.
(2) If all types of equivalence are achieved to a reasonable level except conceptual equivalence, we would argue that the results are not strictly comparable, as features of HRQoL which are important to a given culture may not be incorporated into the questionnaire, in which case the instrument cannot be said to be successfully measuring the concept in that culture. At the very least, reports on the translation of HRQoL instruments into other languages should indicate the degree to which conceptual equivalence has been

Figure 1. Model for assessing cross-cultural equivalence in HRQoL questionnaires.



examined and whether or not there is reason to believe that the most important domains are represented adequately in the translated instrument.

(3) Failure to achieve a reasonable level of equivalence in any of the other areas means that the results are not comparable and should not be aggregated, unless there is the possiblity, when the differences are systematic, of transforming the results to make them comparable.

The adaptation process

Figure 1 details the steps for deciding whether a given questionnaire is suitable for adaptation into another language/culture and for carrying out the adaptation. It also shows the order in which the different types of equivalence should be tested. It is also important to note that the adaptation process is an iterative process. The use of techniques such as factor analysis will, for example, reveal how well the final instrument performs in terms of relating to the underlying concept. Other techniques, such as item response theory,[23] which can only be employed once the questionnaire has been administered to a relatively large sample, may also suggest that changes to items are required.

We have suggested testing conceptual equivalence at the beginning of the process to reflect the importance of this stage within the universalist approach. It will be very difficult to know how conceptually appropriate a given questionnaire will be for use in the target culture without some degree of qualitative work to establish a working idea of what the concept means in the target culture. This is clearly preferable to leaving the investigation of conceptual equivalence to the end of the adaptation process, when questionnaire content has already been decided.

Discussion

The model presented here is based on the universalist assumption that what is important to the HRQoL concept may differ across cultures and, as such, is intended to explore possible cross-cultural differences in the concept. The opposing, absolutist view claims (implicitly or explicitly) that there will be no variation in the HRQoL concept across cultures and that rigorous translation will be sufficient to ensure cross-culturally comparable questionnaires. We have suggested that a model of six types of equivalence will be useful in approaching the adaptation of HRQoL questionnaires from a universalist perspective and have provided definitions for those six types of equivalence as well as methods for their investigation. This model aims to clarify some of the confusion surrounding discussions of equivalence in the HRQoL literature. We limited the types of equivalence included in the model to six, in the belief that they provide a sufficiently complete framework for the examination and achievement of equivalence and because it ensures that the model is relatively simple, whilst remaining faithful to a universalist approach.

The definition of conceptual equivalence is particularly important and differs from many of those found in the HRQoL literature.[5,11] It is important because it requires that those wishing to adapt an existing questionnaire take explicit account of the cultural factors which may make adaptation invalid, and that this should be done before beginning adaptation. We have suggested some ways in which this might be done. Although the inclusion of this stage may make the adaptation process lengthier, it also ensures that adapted versions are likely to be relevant to the target population, something which we believe many current translations methodologies overlook. This stage would also be useful in helping researchers to decide which existing questionnaire might be most appropriate for use in the target culture. In cases where the investigation of conceptual equivalence indicates that adaptation will not be appropriate, researchers should consider developing a culture-specific instrument.

Many existing translation methodologies rely heavily on techniques such as forward and back translation, lay panel testing and psychometric or statistical analyses of instrument behaviour in an attempt to achieve or demonstrate equivalence, but do not perform any initial investigation of conceptual equivalence, as it is defined here. Such methodologies also tend to be rigid with regard to features such as item substitution. The assumption underlying such approaches is that HRQoL instruments will be equally valid in any culture. Other approaches are arguably more sensitive to the dangers of imposing one culture's norms and values onto another. The simultaneous development of instruments in different cultures could, at least in theory, give equal weight to the norms and values of the different cultures involved, though the extent to which this happens will depend to a large degree on the way in which constructs are elaborated and items chosen. Even in the case of simultaneous cross-cultural instrument development, however, we would argue that initial investigations of conceptual equivalence should be as thorough as possible.

Some caveats: throughout this paper, we have often used the expressions 'source culture' and 'target culture' simply for ease of reference (and presumably because of the overriding tendency in the field to translate existing questionnaires). Where simultaneous questionnaire development occurs, however, these expressions are obviously not appropriate, as all cultures involved are 'source' and 'target' cultures. It should also be remembered that obtaining exact equivalence in any of those areas of equivalence mentioned is not a feasible objective; the aim is simply to get as close as possible, though the parameters for

deciding 'how close is close enough' need to be more clearly defined. Finally, although we have referred throughout the paper to the notion of 'culture', many of the comments and suggestions included here would be equally applicable at the level of subcultures or to social categories based on age and sex, for example.

The three principal conclusions of this study are that, firstly, from a universalist perspective, six types of equivalence are sufficient for examining not only how to achieve equivalence, but also whether adaptation should occur at all, secondly, that assumptions of conceptual equivalence should be justified not simply by examining the measurement properties and correlations obtained using translated questionnaires, but also by evaluating the questionnaires themselves in the light of knowledge regarding the nature of the HRQoL concept in other cultures and, thirdly, that there is a concomitant need for more qualitative work in investigating the nature of the HRQoL concept in all cultures, but particularly in those cultures where very little or no such work has been carried out. At the very least, such work would provide researchers working in other cultures with a more solid foundation for making decisions on which questionnaire to adapt, as well as providing a context for the interpretation of the results. Such an approach will lead the international HRQoL field into truly cross-cultural research, research that studies both the similarities and differences amongst cultures, rather than simply imposing notions of health and QoL across cultures.

Acknowledgements

This research was funded by the Catalan Institute of Public Health. Dr Fox-Rushby also received funding in the form of a fellowship from the Economic and Social Research Council in the UK. We would also like to thank Professor Donald Patrick, Kirsten Johnson, Isaac Mwanzo and an anonymous referee for valuable comments on earlier drafts of this paper.

References

1. Herdman MJ, Fox-Rushby J, Badia X. Equivalence and the translation and adaptation of health-related quality of life questionnaires. *Qual Life Res* 1997; **6**: 237–247.
2. Berry JW, Poortinga YH, Segall MH, Dasen PR. *Cross-Cultural Psychology: Research and Applications*. Cambridge University Press, 1992.
3. Bullinger M, Anderson R, Cella D, Aaronson N. Developing and evaluating cross-cultural instruments: From minimum requirements to optimal models. *Qual Life Res* 1993; **2**: 451–459.
4. Patrick DL, Wild DJ, Johnson ES, Wagner TH, Martin MA. Cross-cultural validation of quality of life measures. In: Orley J, Kuyken W, eds. *Quality of Life Assessment: International Perspectives*. Heidelberg: Springer-Verlag, 1994: 19–32.
5. Hunt SM. Cross-cultural comparability of measures and other issues related to multi-country studies. *Br J Med Econ* 1993; **6c**: 27–34.
6. Hui CH, Triandis HC. Measurement in cross-cultural psychology: a review and comparison of strategies. *J Cross-Cult Psychol* 1985; **16**: 131–152.
7. Badia X, Salamero M, Alonso J, Ollé A. *La Medida de la Salud: Guía de Escalas de Medición en Español*. Barcelona: PPU, 1996.
8. Adaptación de una medida de la disfunción relacionada con la enfermedad: La versión Española del sickness Impact Profile. *Med Clin* 1994; **102**: 90–95.
9. Fox-Rushby J, Mwenesi H, Parker M *et al.* Questioning premises: health-related quality of life in Kenya. *Qual Life Res* 1995; **4**: 428–429.
10. Badia X, Alonso J. Re-scaling the Spanish version of the Sickness Impact Profile: an opportunity for the assessment of cross-cultural equivalence. *J Clin Epidemiol* 1995; **48**(7): 949–957.
11. Leplège A, Verdier A. The adaptation of health status measures: methodological aspects of the translation procedure. In: Shumaker SA, Berzon R, eds. *The International Assessment of Health-related Quality of Life: Theory, Translation, Measurement and Analysis*. Oxford: Rapid Communications, 1995; 93–101.
12. Hunt S, McKenna S. Cross-cultural comparability of QoL measures. *Brit J Med Econ* 1992; **4**: 17–23.
13. Shumaker SA, Naughton MJ. The international assessment of health-related quality of life: a theoretical perspective. In: Shumaker S, Berzon R, eds. *The International Assessment of Health-related Quality of Life: Theory, Translation, Measurement and Analysis*. Oxford: Rapid Communications, 1995.
14. Bowling A. What things are important in people's lives? A survey of the public's judgements to inform scales of health-related quality of life. *Soc Sci Med* 1995; **41**(10): 1447–1462.
15. WHOQOL Group. The development of the WHO Quality of Life Assessment Instrument (the WHOQOL). In: Orley J, Kuyken W, eds. *Quality of Life Assessment: International Perspectives*. Heidelberg: Springer-Verlag, 1994: 41–60.
16. Amuyunzu M, Allen T, Mwenesi H *et al.* (1995) The resonance of language: health terms in Kenya. *Qual Life Res* 1995; **4**: 388.
17. Fox-Rushby J, Parker M. Culture and the measurement of health-related quality of life. *Eur Rev Appl Psychol* 1995; 45: 257–263.
18. Lonner WJ, Berry JW, eds. *Field Methods in Cross-cultural Research*. Beverly Hills: Sage Publications, 1986.
19. Katz S, Ford AB, Moskowitz RW, *et al.* Studies of illness in age: the index of ADL, a standardized measure of biological and psychosocial function. *JAMA* 1963; **185**: 914–919.
20. Borg I, Shye S. *Facet Theory: Form and Content. Advanced Quantitative Methods in the Social Sciences*, 1995: Vol. 5. Newbury Park, CA: Sage, 1995.
21. Sackman H. *Delphi Critique: Expert Opinion, Forecasting and Group Processes*. Lexington MA: Lexington Books, 1975.

22. Acquadro C, Jambon B, Ellis D, Marquis P. Language and translation issues. In: Spilker B, ed. *Quality of Life and Pharmacoeconomics in Clinical Trials*, 2nd edn. Philadelphia: Lippincott-Raven, 1996: 575–585.
23. Wright BD, Store MH. *Best Test Design*. Chicago: Mesa Press, 1979.
24. Cronbach LJ. Coefficient alpha and the internal structure of tests. *Psychometrika* 1951; **16**: 297–334.
25. Fox-Rushby J. *First Steps Towards Assessing Semantic Equivalence in the EQ5D: Results of a questionnaire Survey to members of the EuroQol Group*. In: Nord E (editor). Oslo EuroQol Plenary Meeting, *Discussion papers*, 1997.
26. Barnwell K. *Bible Translation: An Introductory Course in Translation Principles*, 3rd edn. Dallas: Summer Institute of Linguistics, 1992.
27. Sartorius N, Kuyken W. Translation of health status instruments. In: Orley J, Kuyken W, eds. *Quality of Life Assessment: International Perspectives* Heidleberg: Springer-Verlag, 1994.
28. Ware J, Keller S, Gandek B, Brazier JE, Sullivan M. Evaluating translations of health Status questionnaires. *Int J Tech Assoc Health Care* 1995; **11**(3): 525–551.
29. Torgerson WS. *Theory and Methods of Scaling*. New York: Wiley, 1958.
30. Sartorius N. Cross-cultural psychiatry. In: Kisker KP, Meyer JE, Muller C, Stromgren E eds. *Psychiatrie der Gegenwart*. Berlin: Springer-Verlag, 1979.
31. Hays RD, Anderson R, Revicki D. Psychometric considerations in evaluating health-related quality of life measures. *Qual Life Res* 1993; **2**: 441–449.
32. Deyo RA, Diehr P, Patrick DL. Reproducibility and responsiveness of health status measures. *Control Clin Trials* 1991; **12**: 142S–158S.
33. Bartko JJ. The intraclass correlation coefficient as a measure of reliability. *Psychol Rep* 1966; **19**: 3–11.
34. Nunnally JC. Psychometric Theory, 2nd edn. New York: McGraw-Hill, 1978
35. Child D. *The Essentials of Factor Analysis*. London, New York: Holt, Rinehart and Winston, 1970.
36. Bucquet D, Condor S, Ritchie K. The French version of the Nottingham Health Profile: a comparison of item weights with those of the source version. *Soc Sci Med* 1990; **30**: 829–835.
37. Thurstone LL. A law of comparative judgement. *Psychol Rev* 1927; **34**: 273–286.
38. Kazis LE, Anderson JJ, Meenan RF. Effect sizes for interpreting changes in health status. *Med Care* 1989; **27**: S178–189.
39. Sell H. The Subjective Well-being Inventory (SUBI). *Int J Mental Health* 1994; **23**(3): 89–102.

(Received 11 August 1997; accepted 19 January 1998)