

# The philosophy of health-related quality of life translation

G. H. Guyatt

Departments of Clinical Epidemiology and Biostatistics, and Medicine, McMaster University, Hamilton, Ontario, Canada.

**Conventional translation strategies are limited because they enshrine deficiencies of the original questionnaire and do not permit modifications that reflect differences in culture and values. An alternative philosophy would require a conceptual definition of what one wanted to measure and would allow flexibility in the methods to achieve this goal. If an investigator had considerable time and resources, he or she could replicate the process used to construct the original English language questionnaire. With very limited resources, he or she could still omit irrelevant items, include new items, and modify the wording of questions and response options. This new philosophy allows improvement in content and presentation, and moves away from the hegemony of the American middle-class outlook in quality of life questionnaires.**

*Key words:* Health-related quality of life questionnaires, resource intensive strategy, resource saving strategy, in-between strategy.

## Current philosophy

At present, investigators design translations of existing health-related quality of life (HRQL) questionnaires to replicate the originals as closely as possible. Ideally, investigators would like to create questionnaires that capture the content of the original (which, as a recent compendium of instruments demonstrates, is almost invariably in English<sup>1</sup>), with all its nuances of meaning. Investigators assume that by using this approach they will produce results that are comparable to the original questionnaire; will be most likely to maintain the measurement properties of the original (such as validity and responsiveness); and they will therefore be able to generalize across nations and legitimately aggregate data from multi-national studies.

There are a number of strengths to this

approach: (i) it is intuitively reasonable; (ii) it is flattering to the creators of the original questionnaire; (iii) it is the simplest possible approach to translation.

There are, however, a number of important limitations to the approach.

(i) Questionnaires are likely to have weaknesses even in their original English form. These may include issues of content, duplication, ambiguity, poor wording, or suboptimal response options. In attempting to maintain faithfulness to the original questionnaire, those in charge of translation efforts find themselves telling translators: 'Please do not try and improve on the original-language questionnaire'. Thus, deficiencies in the original questionnaire are enshrined in the translation.

(ii) There are items that do not translate well. Here, faithfulness to the original questionnaire is liable to force a Procrustean approach in which translators find the closest possible (but not really well-suited) meaning.

(iii) There are items that either do not translate at all, or do not make sense within the new cultural context. Possible approaches to this problem include leaving the items out of the new questionnaire, or leaving the items out of both the original and the translated questionnaire. Unfortunately, those items that investigators omit may be among the most important, or most responsive, items in the original.

(iv) Items that are important for the population for whom the questionnaire was originally created can be of trivial importance in the culture of the new language. Inclusion of such items in the new questionnaire is inefficient. Similarly, crucial items for the new culture may be absent in the original questionnaire, comprising the content validity of the translation.

(v) The current approach is consistent with the cultural hegemony of the United States, where most quality of life questionnaires have been developed. We behave as if the concerns of

---

The work reported was supported, in part, by the Ontario Ministry of Health.

non-English speaking people are only relevant to the extent that they match the concerns of the American middle-class (who are the dominant population for questionnaire development and testing).

These limitations suggest that it would be worth considering an alternative approach.

## A new philosophy

When investigators field health-related quality of life questionnaires they should focus on what they are trying to measure, whether they are studying English or non-English speaking populations. Having decided their goal, they should then look for an instrument or instruments that more or less accomplishes the task. In many cases they will be able to find a suitable instrument for English-speaking populations, but not for non-English speakers. They then face the challenge of meeting their goals in the new population.

The new approach would suggest using the English-language questionnaire as a template or guide, and nothing more. Let us say, for example, that an investigator wanted to discriminate between people in a general population with respect to aspects of their emotional function that they consider important. Of the many English-language emotional function questionnaires, she would choose the one that best meets her goal. This choice would be based on the level of emotional dysfunction in our population; the extent to which the population includes those with identifiable psychiatric illnesses; her interest in different aspects of emotional dysfunction; and the relative priorities of efficiency and comprehensiveness. She would try to accomplish the same goal in the new population.

At the start, she would acknowledge that the original questionnaire almost certainly has deficiencies, even in English, and she would hope her 'translation' would remedy, rather than replicate, these limitations. She would understand that the new population is likely to have different aspects of emotional function that they feel are important. Certain North American ways of thinking about emotional function will probably be alien to them. In areas where content does overlap, ways of presenting response options might not make sense to the new population.

Starting with these assumptions, she would not set about on a translation process that puts a high value in replicating the initial questionnaire to the

greatest extent possible. The strategy she would adopt would depend on the time and resources available. In the following section, I discuss alternative strategies that investigators can take.

### A resource intensive (Rolls-Royce) strategy

If time and resources were not an issue, investigators would replicate the process of questionnaire construction and testing that was carried out in the original language. They would begin by generating items (which could include reference to existing literature, unstructured interviews with individual patients and health care providers, and patient and health-care-provider focus groups). They would then choose the most suitable of the items they generated, often choosing on the basis of the frequency and importance of the items in the target population. At the end of the process they might find they have created a questionnaire that closely mirrors the English-language original, but they might also have a questionnaire that is very different.

### A resource-saving (Volkswagen) strategy

If resources are limited, investigators would be wise to start out assuming that the structure and content of the original questionnaire will be appropriate to measuring emotional function in the new population. Other authors in this supplement do an excellent job of describing the process of translation upon which they would then embark. The new philosophy would, however, dictate the following unconventional strategies. Translators might find what they believe are deficiencies in wording or presentation of items in the original questionnaire. Investigators would seriously consider their criticisms, and if persuaded that suggestions represented genuine improvements, would take the advice.

The translators are likely to find questions that do not translate well. Rather than trying to find the translation that best approximates the original, the investigators would consider other options. They could omit the question, or they could find a substitute that taps a related (or not-so-closely related) content area. In the context of very limited resources these substitutes would come from existing questionnaires in the new language, or

from clinical intuition of the investigators. Investigators would be equally ready to formulate substitutes for questions that do not translate at all.

When they begin pre-testing their translation, investigators are liable to find additional problems. Patients may find questions that were acceptable and relevant in the original questionnaire embarrassing, awkward, peculiar or irrelevant. Investigators would not hesitate to delete such questions. In addition, investigators would be sure to question patients involved in pre-testing regarding the comprehensiveness of the questionnaire. Should respondents identify important areas that are not explored adequately, the investigators would formulate questions that fill these gaps.

#### An in-between (Volvo) strategy

The Rolls-Royce and Volkswagen strategies represent extremes. An intermediate approach would use the strategies of the Volkswagen model for initial translation and for ultimate pre-testing, but would include an intermediate step. Like the Volkswagen model, the Volvo strategy would start with certain assumptions, but would test these assumptions more rigorously. The assumptions would have to do with the adequacy of the original questionnaire in achieving the goals in the new language. A reasonable set of assumptions would be: (i) the domains of the original instrument also represent the important domains in the new language or culture; (ii) the strategy for formulating questions will work well in the new setting; (iii) the strategy for formulating response options will also work well.

As an example, let us presume an investigative group is trying to create a comprehensive generic health-status instrument for a new language group, and their model is the Sickness Impact Profile (SIP). The investigators could start with 12 domains (as does the SIP), and assume that most of these domains can be aggregated into physical and psychosocial areas. They would assume that focusing on behaviour (rather than feelings or experiences), and wording questions accordingly, will work in the new language. Finally, they would assume that presenting patients with dichotomous response options would effectively capture the information they are seeking.

The investigators would use an initial translation based on these assumptions (and on strategies described elsewhere in the symposium on issues in translation of quality of life instruments) as the

basis for focus group discussions. They would, however, encourage the focus groups to challenge the assumptions. Do the domains represent sensible ways of aggregating items? Does framing the questions as behaviours make sense in the new language? Are dichotomous response options adequate? Whatever the answers to these questions, are items within the domains optimal, or should some be deleted, should others be substituted? Depending on the answers to these questions, the ultimate instrument could be very similar in structure and content to the original questionnaire, or substantially different.

#### **Reservations about the new philosophy—aggregating across countries**

Critics could raise at least two fundamental questions about the new philosophy: how can one aggregate results across countries, and what deductions can one make on the basis of questionnaires that, on the surface, look quite different? Both questions arise from the desirability of including all data from multinational studies in a single analysis. Currently, to secure an adequate sample size, a typical randomized trial in Europe will enrol patients from five or more countries, representing five or more linguistic and cultural groups. Investigators must be confident that they can include data from all these sites in their primary analysis.

Traditional assumptions required for aggregating across different translations of health-related quality-of-life questionnaires include the need for the same number of questions (hence the necessity for excluding items on the original questionnaire that do not translate in the new language), using the same number of response options, and the same weighting scheme. These assumptions are unnecessarily limiting. To illustrate this, I will draw an analogy with individualized questions in single-language questionnaires.

Individualized questions have been adduced as a strategy for dealing with the fact that people with the same medical problem will have different activities they undertake, and attach different values to those activities. My colleagues and I faced this problem in constructing a health-related quality-of-life instrument for patients with chronic lung disease. We found that shortness of breath on daily activities was extremely important to patients. The activities that people undertook, and the values they placed on each activity, varied

greatly with severity of illness, gender, and patient interests. We could have dealt with this problem by including a limited number of items and accepting that many items of great importance to some patients would be omitted, by including a very broad range of items (and pay the price of reducing the efficiency of the questionnaire), or individualizing the items. We implemented the third approach, and the questionnaire we constructed asked patients to identify up to five activities that were important to them and during which they experienced dyspnoea. Some patients were able to identify only three activities, and activities varied enormously across individuals. The questionnaire then asked patients to rate their shortness of breath on performing their chosen activities serially over time using a seven-point rating scale from 'No shortness of breath' to 'Extreme shortness of breath'.<sup>2</sup> When we analysed results from administration of this instrument, we dealt with the different number and nature of the activities by pro-rating scores and aggregating across subjects to determine if interventions had improved dyspnoea in daily living.

Critics have questioned the legitimacy of aggregating across patients when the individual questions or items differ. Measuring health-related quality of life, however, inevitably involves asking people different questions. Let us take a simple question such as 'How much shortness of breath do you experience going up a flight of stairs?' When answering this question, every respondent is thinking about a different flight of stairs, with a different length and grade, and a different rate of mounting the steps. Even assuming this was not the case (one could specify the number of steps, the grade, and the rate, though this is seldom, if ever, done), the importance of climbing a flight of steps would be very different for different respondents. For a respondent living in a three storey home in which her bedroom was on the upper storey and the kitchen on the main floor, difficulty climbing stairs would be a major problem. On the other hand, for a respondent who seldom leaves her single-storey apartment, and therefore mounts a flight of stairs no more than once a month, difficulty climbing stairs would not have a significant impact on her life.

This line of reasoning implies that the difference between having respondents choose different items and the traditional approach is a difference in degree, rather than in kind. Ultimately, one could determine whether individualized questions work in evaluative measures by examining their

validity and responsiveness. A number of individualized questionnaires, applied in a variety of situations, have proved to be at least as valid and responsive as conventionally structured questionnaires.<sup>3,4</sup>

The same arguments that apply to patients choosing different activities apply to countries in which, for instance, emotional function questionnaires are likely to differ in the number and content of the items. If investigators are confident that for each country the aspects of emotional function that are important to that population are included, they can accept that the results represent the content in which they are interested. The results can be pro-rated so that the maximal and minimum scores are the same, and then included in the same statistical analysis. My colleagues and I have used this approach repeatedly in our own work, and demonstrated the satisfactory measurement properties of the resulting instruments.<sup>3,5-8</sup>

There is a solution to the problem of differing format and content of questionnaires even for those who do not accept this line of reasoning. Statisticians have developed techniques for analysing data from different studies that use alternative measures to tap the same domain or content area.<sup>9</sup> The methods rely on determining differences between treatment and control groups and standardizing these differences according to the variance of the outcome measures. Investigators have used these well-described methods to conduct meta-analyses of studies using different measures of exercise capacity, well-being, and physical and emotional function. Thus, even if investigators reject the philosophy that different questionnaires from various countries can be included in the same analysis as if they were the same instrument, they can apply meta-analytic methods with a very strong theoretical foundation.

### **Reservations about the new philosophy—validity**

The new philosophy assumes that whatever the differences in content and presentation, original questionnaires and their other-language counterparts are measuring the same underlying domain. Is this assumption valid? To be sure, investigators must demand rigorous testing of the construct validity of the new-language questionnaire. The reader can find standards for validation both in the introductory papers of this symposium and from other sources.<sup>10</sup>

In brief, the validation process would begin with a careful definition of the underlying construct the instrument is trying to measure (a requirement to use the new philosophy in the first place). On the basis of this definition the investigator would make predictions about the relation between the new instrument and other measures. If she intended to use the new measure as a discriminative instrument (to measure cross-sectional differences between subjects at a point in time), she would consider correlations between instruments at a single point in time. If she were primarily interested in using the instrument for evaluative purposes (to measure change within subjects over time, such as in the context of randomized trials) she would consider correlations between changes in instrument score over time.<sup>11</sup> To the extent that observed correlations corresponded to the theoretically derived predictions, a critic could conclude that the new-language instrument had passed the tests of validity. A critic would also be wise to look for evidence that the new-language questionnaire met other tests of usefulness. An evaluative instrument's responsiveness (the ability to detect important differences, even if those differences are small) is of particular importance.<sup>12</sup>

## Conclusions

Choosing a conventional translation strategy does not free the investigator from the requirement for rigorous validation. Rigid adherence to the original-language questionnaire does not guarantee that the new-language questionnaire is as valid as the parent instrument. The validity of a conventional translation can be threatened by including items from the original which are less important in the new language/culture, from omission of items that are crucial in the new but not the original setting, and by questions that are misinterpreted or misunderstood by respondents in the new population. If there are a number of questions that do not translate, and the investigator drops these from the original instrument, its content and potentially construct validity are likely to be compromised.

There are many advantages to investigators adopting the new philosophy of health-related quality of life measurement in non-English lan-

guage cultures. They will face an additional incentive to define, in theoretical terms, what they are trying to measure. They will not be bound by the original questionnaire, and will have an opportunity to improve its content and format. The cultural hegemony of the American middle-class in measures of health-related quality of life will be broken. The new approach is likely to produce more, not less, responsive and valid instruments.

## References

1. Spilker B, Molinek FR, Johnston KA, *et al.* Quality of life bibliography and indexes. *Med Care* 1990; **28**: DS1-DS77.
2. Guyatt GH, Berman LB, Townsend M, *et al.* A measure of quality of life for clinical trials in chronic lung disease. *Thorax* 1987; **42**: 773-778.
3. Guyatt GH, Nogradi S, Halcrow S, *et al.* Development and testing of a new measure of health status for clinical trials in heart failure. *J Gen Intern Med* 1989; **4**: 101-107.
4. Tugwell P, Bombardier C, Buchanan WW, *et al.* The MACTAR Patient Preference Disability Questionnaire—an individualized functional priority approach for assessing improvement in physical disability in clinical trials in rheumatoid arthritis. *J Rheumatol* 1987; **14**: 446-451.
5. Guyatt GH, Nogradi S, Halcrow S, *et al.* Development and testing of a new measure of health status for clinical trials in heart failure. *J Gen Intern Med* 1989; **4**: 101-107.
6. Guyatt GH, Townsend M, Keller JL, *et al.* Measuring functional status in chronic lung disease: conclusions from a randomized control trial. *Respir Med* 1989; **42**: 913-920.
7. Juniper EF, Guyatt GH, Epstein RS, *et al.* Evaluation of quality of life in asthma: development of a questionnaire for use in clinical trials. *Thorax* 1992; **47**: 76-83.
8. Juniper EF, Guyatt GH, Ferrie PJ, *et al.* Measuring quality of life in asthma. *Am Rev Respir Dis* 1993; **147**: 468-479.
9. Hedges LV, Olkin I. *Statistical Methods for Meta-analysis*. New York: Academic Press, 1985, pp. 108-138.
10. Guyatt GH, Feeny DH, Patrick DL. Measuring disease-specific quality of life. *Ann Intern Med* 1993; **118**: 622-629.
11. Kirshner B, Guyatt GH. A methodological framework for assessing health indices. *J Chron Dis* 1985; **38**: 27-36.
12. Guyatt GH, Kirshner B, Jaeschke R. Measuring health status: what are the necessary measurement properties. *J Clin Epidemiol* 1992; **45**: 1341-1345.