Volume 8 • Number 2 • 2005
VALUE IN HEALTH

The Translation and Cultural Adaptation of Patient-Reported Outcome Measures

Stephen P. McKenna, PhD, Lynda C. Doward, MRes

Galen Research, Manchester, UK

Promotion of good practice guidelines is clearly a laudable pursuit for any aspect of scientific endeavor; not least in the area of patient-reported outcome (PRO) instrument adaptation, where extemporary approaches to translation are commonplace. Nevertheless, what Wild and colleagues have produced are not so much "principles of good practice" as a synthesis of the opinions of a number of organizations on how to translate PRO measures [1]. Indeed, in several instances more than one opinion is expressed. No scientific evidence is provided to support these opinions and the suggested methodology falls short of including a meaningful test of the quality of the translations produced.

Rather than reviewing all published translation recommendations as reported, the authors have neglected a considerable body of literature that utilizes a rather different methodology to the forward–backward translation approach [2–4]. This is the two-panel approach that employs professional and lay panel meetings rather than back translation. This approach has been used in the development of 20 disease-specific quality of life measures in up to 30 languages [5].

Indeed it is not clear how the concept of back translation came into existence. With the need for standardized PROs for use in multinational studies it was probably recognized that there was a requirement for determining the quality of the target version by some sort of "scientific" method. Despite this need, back translation has no clear scientific basis and its use casts doubts on the ability of translators. One could question what right the test-developers or adaptors have to question the competence of members of another profession. If the translation is good, the back translation may look nothing like the source questionnaire. Consequently, little information of any value will be obtained from a back translation, whereas many misleading impressions may result. A classic example of the problem with relying on back translation was provided by Fukuhara and colleagues [6] whose attempt to translate the SF-36 into Japanese failed and they had to resort to employing the equivalent of a lay translation panel (as used by the European Group for Quality of Life Assessment and Health Measurement [2,3]) to overcome the problems. The researchers concluded that adherence to the International Quality of Life Assessment project guidelines did not ensure an adequate translation, reflecting the linguistic and cultural differences between Japan and the United States. The answer must be to produce quality *in* the translation, in addition to checking it a posteriori [4]. Producing quality requires checking and rechecking throughout the process, as well as after it, to see if the instrument functions as required with "real" people.

It should also be noted that translation is just one step in the production of new language versions of an instrument. Full adaptation requires that the scaling and psychometric properties of the new language version are also assessed [4]. It is indeed bad practice to simply assume that a translated version shares the same psychometric properties as the original source version. It may be that this is an area still to be addressed by the working group, although this seems unlikely as the guidelines they reviewed do not include such quality controls and such validation is rarely reported.

It is here that Wild et al.'s proposed methodology is particularly weak. Virtually no attempt is made to validate the process—a fundamental requirement of any research. Cognitive debriefing interviews are generally conducted with as few as five patients and no other form of assessment is proposed. Given that PRO measures are sometimes written by test-developers rather than being derived from patients, this is hardly an adequate test of even face and content validity. No recommendation is made for retesting scalability, reproducibility, construct validity or equivalence to the source measure. It is essential (at a minimum) to show that language adaptations are

Address correspondence to: Dr. Stephen P. McKenna, Galen Research, Manchester Science Park, Lloyd Street North, Manchester, M15 6SE, UK. E-mail: smckenna@galen-research.com

 1098-3015/05/89 89–91

reproducible and valid in themselves. Such retesting is routinely conducted in the adaptation of some instruments. For example, all adaptations of needs-based quality of life measures are formally tested for unidimensionality, reproducibility and construct validity [see for example; 7–9].

It is now also possible to test whether adapted measures are equivalent to the original source version by application of item response theory [10]. The probability of being able to affirm a PRO item for patients at the same level of ability (or, e.g., with the same level of quality of life) should remain the same across language versions. Thus, comparability of language versions is dependent on both the conceptual equivalence of items and also on the construct value equivalence of those items. An item may cover the same concept in both source and target language but it may well be valued differently in each culture. Assessment of such "differential item functioning" yields crucial information about the equivalence of language versions and provides valuable information on the validity of pooling data across countries; an issue which is of particular importance for multicenter clinical trials [11].

But why should it be expected that well translated questionnaires will be equivalent across languages? A language is a specific way of putting life into words; and words change with their context. They generate representations that may be universal or culture-specific. Therefore, it is more appropriate to consider the process as one of adaptation rather than translation [4]. Extending this idea, it is likely that in the relatively near future it will be recognized that translating source measures into new languages is both inefficient and scientifically unhelpful. Where carefully constructed unidimensional scales are produced based on a coherent measurement model it is possible to develop language versions that consist of items that are specific to, and work well in, each country. These need not be the same items—although it is likely that there will be items in common across languages. The aim would be to have a set of items that had the same value in each language (derived from the application of item response theory). Only then would construct equivalence be guaranteed.

Although such work is currently being piloted it is likely to be some time before such a radical (but necessary) development would be acceptable to health authorities. In the meantime, how is it possible to determine the most appropriate method of translating PRO measures? A study is currently underway to compare two Swedish versions of the Rheumatoid Arthritis Quality of Life (RAQoL) questionnaire [12,13]. The opportunity arose to make the comparison due to the unauthorized translation and publication of a Swedish version of the RAQoL [14,15] that occurred simultaneously with the production of an authorized adaptation. The former applied back translation and the latter the two-panel methodology to produce the translations. Blind assessments of the final item translations are being made by Swedish bilinguals and patients with rheumatoid arthritis (RA). A test-retest survey with RA patients is planned to compare psychometric properties of the two versions and Rasch analysis will be used to compare the scaling properties of each version to the source questionnaire. Initial results from the study indicate that items translated using the two-panel method are statistically significantly preferred to those based on back translation. This is initial evidence that the two-panel method has better face validity than the use of back translation.

We strongly feel that it is essential to collect evidence before asserting that back translation, an untested method—however, widely implemented—represents "principles of good practice."

References

1 Wild D, Grove A, Martin M, et al. Principles of good practice for the translation and cultural adaptation process for patient reported outcomes (PRO) measures: report of the ISPOR task force for translation and cultural adaptation. Value Health 2005;8:94–104.
2 Swaine-Verdier A, Doward LC, Hagell P, et al. Adapting quality of life instruments. Value Health 2004;7(Suppl. 1):S27–30.
3 Hunt SM, Alonso J, Bucquet D, et al. Cross-cultural adaptation of health measures. Health Policy 1991;19:33–44.
4 European Group for Quality of Life Assessment and Health Measurement. European Guide to the Nottingham Health Profile. Brookwood, Surrey: Brookwood Medical Publications, 1993.
5 McKenna SP, Doward LC, Meads DM, et al. Summary of needs-based quality of life instruments. Value Health 2004;7(Suppl. 1):S39–40.
6 Fukuhara S, Bito S, Green J, et al. Translation, adaption, and validation of the SF-36 health survey for use in Japan. J Clin Epidemiol 1998;51:1037–44.
7 McKenna SP, Doward LC, Kohlmann T, et al. International development of the quality of life in depression scale (QLDS). J Affect Disord 2001;63:189–99.
8 McKenna SP, Doward LC, Alonso J, et al. The QoL-AGHDA: an instrument for the assessment of

quality of life in adults with growth hormone deficiency. Qual Life Res 1999;8:373–83.
9 Whalley D, McKenna SP, Dewar AL, et al. A new instrument for assessing quality of life in atopic dermatitis: international development of the quality of life index for atopic dermatitis (QoLIAD). Br J Dermatol 2004;150:274–83.
10 Tennant A, McKenna SP, Hagell P. Application of Rasch analysis in the development and application of quality of life instruments. Value Health 2004; 7(Suppl. 1):S22–6.
11 Smith RM. Applications of Rasch Measurement. Sacramento: JAM Press, 1992.
12 de Jong Z, van der Heijde D, McKenna SP, Whalley D. The reliability and construct validity of the RAQoL: a rheumatoid arthritis-specific quality of life instrument. Br J Rheumatol 1997;36:878–83.
13 Whalley D, McKenna SP, de Jong Z, van der Heijde D. Quality of life in rheumatoid arthritis. Br J Rheumatol 1997;36:884–8.
14 Eberhardt K, Duckberg S, Larsson BM, et al. Measuring health related quality of life in patients with rheumatoid arthritis—reliability, validity, and responsiveness of a Swedish version of RAQoL. Scand J Rheumatol 2002;31:6–12.
15 McKenna SP, Hedin PJ. Adapting the rheumatoid arthritis quality of life instrument (RAQoL) for use in Sweden. Scand J Rheumatol 2003;32:1–3.