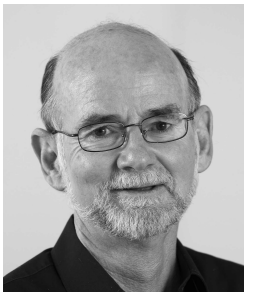
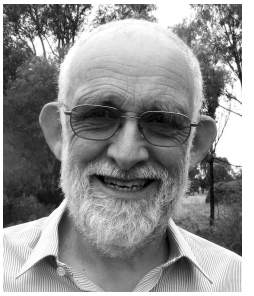
Editorial

Jeopardizing Validity by Mismeasurement of Quality of Life





Practitioners and policy-makers have a common interest in an instrument that will measure the value people place on different health states. For practitioners, a patient's valuation can guide individual decisions. For senior managers and policy-makers the information can help them decide which services to provide and which technologies to install. It is therefore disappointing to encounter results, such as those from the Leung study in this issue of *The Journal*[1], that show 2 reputable and widely used instruments that purport to provide such information giving different results. It raises a number of questions: How can such an outcome arise? What should practitioners and policy-makers do in view of such a finding?

Measuring patient quality of life (QOL) is a relatively new venture — in particular, measurement using multi-attribute utility (MAU) instruments, like those used in the Leung study. MAU are questionnaires in which response categories are scored using utility weights that purport to measure the strength of preference for a health state. In principle, the application of these weights permits comparison of dissimilar health states. A higher score for health state A versus health state B does not have clinical meaning. Rather, a higher score indicates that there is a subjective preference for health state A by the person or persons whose judgment was used to create the utility weights. MAU instruments are now widely employed in the calculation of quality-adjusted life-years (QALY) or, more correctly, preference-adjusted life-years. Used in economic evaluation studies, QALY serve to rank dissimilar services (cost utility analysis). A variant of the QALY — the disability-adjusted life-year — is used to estimate the burden of disease (for example, the recent World Health Organization Burden of Disease Study[2]).

The field is currently dominated by a limited number of MAU instruments. Two of these, the EuroQol-5D (EQ-5D) and Short Form-6D (SF-6D), were developed in the UK. Three health utility index (HUI) instruments were developed in Canada; a 15D instrument in Finland; 4 assessment of quality of life (AQoL) instruments in Australia, and a quality of wellbeing (QWB) instrument in the USA. (For a review of these, see Brazier, *et al*[3] and Richardson, *et al*[4].)

The instruments have been influential. Between 2005 and 2010, 1682 studies listed on the *Web of Science* used one or more of these instruments[5]. However, usage has been highly concentrated, with 63.2% of studies using the EQ-5D and only 8.8% using the SF-6D. Muscular skeletal disease and arthritis account for 16.4% of the total, followed by cancer (6.4%), degenerative diseases and the elderly (6.4%), and psychiatric health (6%). By 2010 fourteen countries had recommended or mandated the use of MAU instruments in their official guidelines for the evaluation of pharmaceuticals. In every case the EQ-5D was one of the recommended instruments. The HUI was recommend or noted in 7 countries and the SF-6D in 4 countries.

The article by Leung, *et al*[1] points to a general problem with these instruments. In their comparison of the EQ-5D and SF-6D the authors find dissimilar distributions of utilities, different ceiling effects, different mean scores, and a low correlation between the 2 sets of results. This is a disturbing finding, since the 2 instruments purport to measure the same thing — utility (the strength of preference for a health state). If comparable differences were found in the measurement of weight using, say, an electronic versus a spring scale, then it would be reasonably concluded that one or both scales were defective. Nevertheless, both the EQ-5D and the SF-6D have coexisted with relatively little criticism for 15 years, and both have been widely used for the evaluation of health services and products.

The results from Leung, *et al* are broadly consistent with those found in the major multi-instrument comparison (MIC) study by Richardson, *et al*[4], which compared utility scores from the major instruments in 6 countries and 7 disease categories. While Leung, *et al* found a ceiling effect of 20% among patients with psoriatic arthritis using the

See EQ-5D and SF-6D in PsA, *page 859*

The Journal of Rheumatology 2013; 40:6; doi:10.3899/jrheum.130226

EQ-5D, and no ceiling effect for the SF-6D, the MIC study found a ceiling effect of 39% for non-patient members of the population in Australia using the EQ-5D, but only 2% using the SF-6D. The intraclass correlation (ICC) reported by Leung, *et al* was 0.43; in the Australian branch of the MIC the ICC for the EQ-5D and SF-6D was higher at 0.66, but this is partly attributable to the wider range of observations. Leung, *et al* find a stronger correlation with general health and external measures using the SF-6D. In the MIC study the SF-6D is more highly correlated with the SF-36, with 3 indices of subjective well-being and with the ICECAP measure of individual capabilities. The EQ-5D also produced significantly lower scores for the least healthy, reflecting a set of utility weights that allows scores to fall to a (probably meaningless) value of –0.59 (using the Dolan algorithm) on a scale where 1.0 is best health and 0.0 represents death.

These discrepancies are not limited to the comparison of the EQ-5D and SF-6D. Before the MIC study cited above, only 2 multi-instrument comparisons had been conducted, which included 5 instruments. In an early Australian comparison, 956 hospital and general respondents were administered the EQ-5D, SF-6D, 15D, HUI 3, and AQoL-4D[6]. The proportion of instrument variation explained by other instruments varied from 41% to 59%, leaving an average of 44% unexplained. The highest explanatory power was achieved by 15D, followed by AQoL. In a more recent US study, 3844 adults were surveyed to compare the EQ-5D, QWBSA, HUI 2, HUI 3, and SF-6D. A weaker association was found compared to results in Australia (reflecting the use of only general population respondents). Overall, 53% of instrument variance was not explained[7].

Generally, researchers conducting multi-instrument comparisons have concluded that the utilities derived from them are "not equivalent," that translation between them will result in "low precision," and that comparisons between them "warrant caution."

Results from the MIC study support these conclusions. The regression of different MAU scores upon the dimension scores of the SF-36 produce significantly different results. The two SF-36 dimensions Pain and Physical Function statistically "explain" 44%, 31%, and 14% of the variation in the scores of the EQ-5D, SF-6D, and AQoL-8D, respectively. The SF-36 dimensions General Health, Vitality, Social Function, and Mental Health together explain 27%, 54%, and 66% of variation of the 3 instruments, respectively. The differences indicate that the instruments measure different "constructs": different concepts of what constitutes "health."

These results highlight an unsatisfactory state of measurement theory and practice with respect to QOL. Economists who have created the instruments have focused almost exclusively upon the measurement of utility, a key and absorbing topic in orthodox economics. They have largely ignored the reliability and validity of the questionnaires used to obtain the health state description, despite the existence of well established psychometric methods for achieving this. Not surprisingly, the ad hoc descriptive systems of the major instruments differ significantly, calling into question the accuracy of the measured benefits of medical and pharmaceutical services. As a consequence, the precision of clinical measurement in evaluation studies may well be offset by the unreliability of the MAU instrument, and acceptance or rejection of a therapy by a national health service may be contingent upon the QOL instrument chosen.

The first step in rectifying this problem is a recognition that the problem exists. In this context studies such as Leung, *et al* are important. The next step is for the academic community and regulatory authorities to jointly determine, in operational terms, the concept (or concepts) of QOL that should be incorporated in evaluation studies and to determine which, if any, of the existing instruments measure this concept with acceptable precision.

It is possible that no one instrument will ultimately be satisfactory for all health states. However, the use of multiple instruments will leave the problem of achieving comparability of measurement unresolved — the very problem MAU instruments sought to overcome.

In the interim, researchers concerned with the QOL and its use in evaluation studies have little choice but to exercise their discretion in the selection of a measurement instrument. The key judgment is whether or not the instrument's descriptive system appears capable of describing the health states of interest. That is, the users of these instruments should undertake formal or informal content analyses to determine the a priori likelihood of the instrument measuring the health states of interest. A sensible precaution is to employ at least 2 instruments as a form of reliability test. When results concur, confidence is increased. When they conflict, the approach adopted by Leung, *et al* is justified: base conclusions upon the instrument with greatest face validity, albeit with reduced confidence. Impatience with the need for an additional instrument and additional questions is analogous to impatience with the need for a control group in an RCT.

JEFF RICHARDSON, BA (Hons), PhD,
Centre for Health Economics,
Monash University,
Clayton, Victoria;
NEIL A. DAY, BA (Hons), DipEd, DipSocScDataAnalysis, MA,
Centre for Program Evaluation,
University of Melbourne,
Parkville, Victoria
Australia.

Address correspondence to N.A. Day;
E-mail: neilathertonday@bigpond.com

REFERENCES

1. Leung Y-Y, Png M-E, Wee H-L, Thumboo J. Comparison of EQ-5D and SF-6D utility scores in multiethnic Asian patients with psoriatic arthriis: A cross-sectional study. J Rheumatol 2013;40:859-65.
2. Horton R. Global burden of disease 2010: Understanding disease, injury, and risk. Lancet 2012;380:2053-4.
3. Brazier J, Ratcliffe J, Salomon JA, Tsuchiya A. Measuring and valuing health benefits for economic evaluation. Oxford: Oxford University Press; 2007.
4. Richardson J, Iezzi A, Khan MA, Maxwell A. Cross-national comparison of twelve quality of life instruments. MIC Paper 2. Melbourne: Centre for Health Economics, Monash University; 2012.
5. Richardson J, McKie J, Bariola E. In: Culyer A, editor. Encyclopedia of health economics. San Diego: Elsevier Science; in press. Available from: http://www.buseco.monash.edu.au/centres/che/pubs/researchpaper64.pdf
6. Hawthorne G, Richardson J, Day NA. A comparison of the Assessment of Quality of Life (AQoL) with four other generic utility instruments. Ann Med 2001;33:358-370.
7. Fryback DG, Palta M, Cherepanov D, Bolt D, Kim JS. Comparison of 5 health related quality of life indexes using item response theory analysis. Med Decis Making 2010;30:5-15.

J Rheumatol 2013;40:758–60; doi:10.3899/jrheum.130226