284/ METHODOLOGY FOR CHOICE OF ITEMS: DESIGNING A QUALITY OF LIFE INSTRUMENT FOR HIP AND KNEE OSTEARTHITIS
Francis Guerlin, Epidemiologie et Evaluation Cliniques, Hopital Marin, Nancy, France; Joel Coste, Departement Biostatistique et Informatique Medicale, Hopital Cochin, Paris, France; Jacques Pochet, Service de Medecine Interne, Hopital Louis Mourier, Paris, France; Nathalie Releu-Roule, Laboratoire de Biostatistique, Faculte de Medecine, Besancon, France; Elisabeth Spitz, Departement de Psychologie, Universite de Metz, Metz, France; Michelle Baumann, Ecole de Sante Publique Uper, Faculte de Medecine, Vandoeuvre les Nancy, France; Cyrille Tarquinio, Departement de Psychologie, Universite de Metz, Metz, France

In the construction of composite health status measurement scales, item generation targets the content of the scale. It proceeds usually in three stages: item elicitation to obtain a coverage of content in all potential affected areas, testing instrument scaling properties, and item selection to ensure feasibility and acceptability of the scale. The objective of this study was to investigate the relative contribution of several methods of item development and choice, with an application to designing a lower limbs osteoarthritis (OA) QoL scale. This study was conducted with the working hypothesis that various item generation methods may each contribute the addition of relevant statements and words. The methods tested included items with patient individual or focus group interview, health professional individual or focus group interview, and content analysis of an instrument. A frame for content analysis was developed by a sociologist team. Two other sociologist teams operated independently a semantic theme content analysis of tape recorded and transcribed interviews, generated a list of words, sorted and verified by hierarchical cluster analysis. This allowed to identify 140 items pertaining to QoL dimensions. The separate contribution of each generation method confirmed the working hypothesis. The primary stages of scale construction include concept specification and content description. The comparison of methods helps contribute to future methodological guidelines to generate quality of life questionnaires.

285/ HOW MUCH IS "VERY MUCH?" DEVELOPING A RATING SCALE FOR PORTUGUESE SPEAKING COUNTRIES
Benjamin J. Arnold, S. Eremencko, C. H. Chang, D. Cell, CORE, ENH, Evanston, IL; J. L. Pires, Faculdade de Psicologia, Universidade do Porto, Porto, Portugal; M. Doro, Bone Marrow Transplant Unit, Univerisidade Hospital, Curitiba, Parana, Brazil; R. Harris, CORE, ENH, Evanston, IL

Developing linguistically appropriate and psychometrically sound response scales for use in questionnaires is critical to ensure scores obtained from the summation of responses are meaningful. Translating these answer categories while retaining their linguistic and measurement equivalence poses great challenges. In this study, we examined the translated 5-point Likert-type rating scale used in the Portuguese Functional Assessment of Cancer Therapy-General (FACT-G) 27-item questionnaire. We applied Andrich's rating scale model (RSM) to explore the answer category structure, identify possible problematic answer categories, investigate alternative categories, and develop one set of equivalent translations to be used in Brazil (BR) and Portugal (POR). A Portuguese version was translated and tested in both BR (n=81) and POR (n=81). Through RSM analyses, we found the translated answer categories (nem um pouco, um pouco, mais ou menos, bastante, muito) did not represent equal intervals on the two highest categories. We then carried out cognitive debriefing to pinpoint the problem and find better answer categories. A survey with 15 possible alternatives to cover the answer category continuum was compiled. Using a Q-sort method, 62 subjects (32 BR; 30 POR) arranged the alternatives in order of increasing intensity and were encouraged to comment. Data were tabulated using median ranking for each answer category. Results showed that high-end (bastante and muito) answer categories were perceived as having the same intensity for the subjects as a whole. Independent of literacy, health status, and country. The answer category set (nem um pouco, um pouco, mais ou menos, muito, muito) was then finalized based on the median ratings and feedback of patients. This method produced a Portuguese language rating scale with categories judged to be of equal interval across both countries.

286/ SHOULD SYMPTOMS BE SCALED FOR INTENSITY, FREQUENCY, OR BOTH?
Chih-Hung Chang, Center on Outcomes, Research and Education, David Cell, Susan Clarke, Evanston Northwestern Healthcare, Evanston, IL; Jamie H. von Roenn, Northwestern University Medical School, Chicago, IL; Alan W. Heinemann, Rehabilitation Institute of Chicago, Chicago, IL

The purpose of the study was to evaluate the comparability between two 5-point Likert-type rating scales commonly used in symptom measurement. One measured intensity (category ranging from "not at all" to "very much") and one measured frequency ("none of the time" to "all of the time"). Data were collected as part of a larger study to develop a fatigue computerized adaptive testing platform. Thirteen questions comprising the Functional Assessment of Chronic Illness Therapy (FACT)-Fatigue were asked using both response formats and separated by "interference" questions. Patients (N=161) were from three disease groups: Cancer (n=66), stroke (n=51) and HIV (n=52). These two rating scales were calibrated separately using Andrich's rating scale model to distance and compare the order of steps from one response category to its higher (adjacent) category. Each item was evaluated for its position on the fatigue measurement continuum relative to the other items. The comparability of measured fatigue from each rating scale was also evaluated. The order of the two sets of steps measures was in the expected direction and was comparable across both rating scales. There was a high correlation (r=.86, p<.001) between two sets of person fatigue measures measured by intensity and frequency, also supporting their measure equivalence regardless of rating scale being used. There was not a major distinction between frequency versus intensity measures in assessing fatigue in chronically ill patients, suggesting results from one or the other approach are roughly comparable.

287/ TEST-RETEST RELIABILITY AND INTERPRETATION OF UTILITY SCORES: EVIDENCE FROM ELECTIVE TOTAL HIP ARTHROPLASTY
David H. Feeny, Chris Blanchard, Institute of Health Economics, Edmonton, AB, Canada; Jeffrey Mahon, Department of Epidemiology and Biostatistics, Robert Bourque, Cecil Rorabeck, Department of Surgery, Larry Stitt, Susan Webber-Bogniet, Department of Epidemiology and Biostatistics, University of Western Ontario, London, ON, Canada

The first purpose is to assess test-retest reliability of standard gamble(SG)utility scores for repeated assessments in a prospective study. The second purpose is to provide evidence on the marker-scale approach in interpreting utility scores. SG scores for three hypothetical marker states (mild, moderate, and severe osteoarthritis(OA)) and the patient's current state were obtained at multiple points in time in a longitudinal study of elective total hip arthroplasty (THA). Patients were assessed repeatedly throughout the process of waiting to see a surgeon, waiting for surgery, and recovery after THA. Test-retest reliability was assessed using the intraclass correlation coefficient (ICC). Analysis of covariance (ANCOVA) was used to assess the effects of time on SG scores. The mean (SD) time from first assessment (A1) to A2 was 5.2 months (2.0); the mean from A2 to A3 was 5.1 (1.8), from A3 to A4 5.6 (2.3). ICCs varied from 0.55 to 0.83. Time was insignificant (<.05) in 8 of 6 ANCOVAs. Mean (SD) baseline SG scores for mild, moderate, and severe marker states were 0.71 (0.29), 0.61 (0.32), and 0.39 (0.33). Mean SG scores for current health improved from 0.92 (0.13) at baseline to 0.78 (0.23) at follow-up after THA. Test-retest reliability of SG scores over an extended period of time is acceptable at the group level. The marker states assist in interpreting results. On average, THA converted moderate OA to better than mild.