Appendix S1

Item response theory – the Rasch model

Under the heading "item response theory" (IRT) various new models of measurement are subsumed (11) and are now considered as "modern test theory". One of those models is the Rasch model (RM) which is considered as a new standard in the development of HRQOL instruments (6). In modern test theory probabilistic models are used to calculate estimates which are then given on a logistic scale to assess test properties, while in classical test theory sum scores are used. Modern test theory makes stronger assumptions regarding the properties a test-instrument has to fulfil. Also, estimates are independent of the average disease severity of the investigated sample and the estimates are more precise compared to classical test theory (12). The main difference between modern and classical test theory is not about how a test-instrument has to be developed (i.e. item generation) but how the gathered information is handled statistically and which precision is assumed to be acceptable. A detailed comparison of classical and modern test theory is described by Hambelton & Jones (13).

It is possible that a test-instrument which was developed using classical test theory also meets the assumptions of modern test theory. If not, modern test theory can be used to refine the test-instrument (i.e. detecting and revising problematic items by adjusting scoring procedures). There are numerous studies applying the RM to test the psychometric properties of new or existing patient-reported outcomes (PROs) (14). The strategies which are used to improve test-instruments vary and are not always successful (15).

Statistical analysis

Data was analysed using RUMM2030 (©RummLab Pty Ltd) for the Rasch analysis and SPSS 20 (©IBM) for descriptive statistics and calculating the regression models. A significant Likelihood Ratio test ($p < 0.001$) indicated to use the partial credit scoring (16) in both samples for the Rasch analysis where the thresholds can differ across items. To assess the overall fit to the RM we calculated the item-trait interaction with a χ^2-test over 8 class intervals in sample 1 and 9 class intervals in sample 2 resulting in 46–83 observations per class. We assessed the mean fit residuals for items and persons and the corresponding standard deviation (SD); the fit residuals should be ≈0 and the SD ≈1 for both indices. The person separation index (PSI) was used as a measure for the internal reliability of the scale – it can be interpreted in analogy to Cronbach's alpha (17). A value of > 0.7 was therefore considered as evidence for good reliability.

On items basis we inspected category frequencies, thresholds, differential item functioning (DIF) and fit residuals. Fit residuals were interpreted as representing an overfit to the scale if the values were >2.5 (those items add only little information to the scale) or as representing overdiscrimination if the items were ≤2.5 (those items are poorly associated with the scale). Thresholds are given on the same logit scale as item and person estimates. They represent the border between 2 categories of an item at which the likelihood of falling in each category is 50%. In other words, if the person estimate of HRQOL impairment exactly matches a threshold, the chance for this person to fall in one of those categories is equal. Thresholds are disordered if they are not strictly increasing from one category to another (18). This is indicating that the response categories of the corresponding item are not ordinally scaled.

There are 2 types of DIF (19): uniform DIF occurs when at the same level of HRQOL impairment one group (e.g. males) is more likely to be impaired according to an item than another group (e.g. females). Researchers can adjust for this kind of DIF by splitting such an item by group, resulting in 2 items with 2 different item locations (20). Non-uniform DIF occurs when also the slope of an item differs between groups (e.g. males with low degree of HRQOL impairment are more likely to be impaired according to an item compared to females, while males with a high degree of impairment are less likely to be impaired compared to females). This kind of DIF cannot be corrected. Therefore items showing non-uniform DIF should be deleted. DIF was considered significant if the analysis of variance showed a p-value < 0.01.

Recalibration of the DLQI

There are 2 options to calibrate the DLQI according to the RM: deleting misfitting items or rescoring items by collapsing categories (21). Since deleting items leads to more information being lost we recommend to try collapsing some categories first. Therefore the following rules were applied: (a) When thresholds were disordered or when the distance between thresholds was < 0.5 logits (indicating low discrimination) the category between those thresholds was collapsed with a neighbour category. (b) A category should be collapsed with the smallest neighbour category in order to receive a regular distribution of observations (22). (c) The category "not at all" should never be collapsed because this would violate theoretical assumptions (e.g. people may have problems differentiating between "a lot" and "very much", but if they cannot tell whether they have a problem "not at all" an item should be deleted) (21).

After applying this procedure, fit indices need to be investigated again. It is likely that fit indices of the items have changed (even of those not altered during the procedure). This procedure has to be repeated until there are no thresholds left meeting criterion (a). Items which still have fit residuals outside the recommended range or which show DIF should be deleted in a next step. Before deleting items because of DIF a Bonferroni correction should be applied because of multiple testing to conservatively control for type 1 error (23).