SUPPLEMENTARY DIGITAL MATERIAL 3

Details on the Rasch analysis

This appendix details the Rasch analysis run in the current study. As mentioned in the main text, the Rasch analysis assesses the following questionnaire's characteristics:

1. categories' order,
2. items' fit to the model,
3. dimensionality,
4. differential item functioning,
5. persons' reliability,
6. items' map.

Ordered categories. Ordered categories, an assumption of the Rasch model, mean that categories have been numbered so that higher numerals (e.g. the score of ISYQOL items) imply more of the variable (e.g. health-related quality of life, HRQOL). This assumption can be easily verified by showing that the higher the participants' scores, the higher their measures.

Items' fit to the model. Infit (IN) and outfit (OUT) means square (MNSQ) and z-standardised (ZSTD) statistics were calculated for each item to evaluate if each of them fits well the model of Rasch. The MNSQ returns the amplitude of data departure from the model's expectations, while the ZSTD returns the statistical significance (i.e. the type I probability) of this departure. MNSQ within the 0.6 - 1.4 (1) range indicates that data departure from the model is reasonable (e.g. not too large), and ZSTD within -1.96 and 1.96 indicates that the departure is not significant.

Dimensionality. Another assumption of the analysis is that the questionnaire is unidimensional, which means that the only variable affecting the items' scores is the one grabbed by the Rasch model. Here, it is assumed that the measures returned by the Rasch analysis of ISYQOL data are measures of HRQOL. Unidimensionality thus means that the scores of the ISYQOL items *only depend* on HRQOL.

Dimensionality is usually tested by running a principal component analysis (PCA) on the models' residuals. Unidimensionality is inferred if the variance taken into account by the first principal component is small enough. In practical terms, this is indicated by an eigenvalue of the first principal component < 2. In the case multidimensionality is found, the procedure detailed by Smith (2) can be adopted to test if this causes artefacts in the persons' measures. If this does not happen, multidimensionality can be safely ignored.

Following this procedure, patients' measures returned by the items with positive loadings on the first principal component are contrasted to those returned by the items with negative loadings. In plain words, patients' measures from items with a positive correlation with the additional variable pointed out by the PCA are compared with those from the items negatively correlating with it. Given that the hidden variable has opposite effects on the score of items with positive and negative loadings (i.e. increases the score of the former and decreases that of the latter), a significant difference between the two sets of measures points out that the additional variable found by the PCA affects the patients' estimation. For practical purposes, if measures obtained with the two sets of items are significantly different in < 5% of patients, multidimensionality is not considered an issue.

Differential item functioning. The main aim of the current work is to evaluate if ISYQOL international provides a measure of HRQOL that is equivalent across cultures. As reported above, Rasch analysis assumes that the only variable affecting the questionnaire's score is that modelled by the model of Rasch (HRQOL, in the ISYQOL case). This assumption means that nationality should not affect *by itself* (i.e. without affecting HRQOL) the score of the ISYQOL items.

Consider an Italian and a Polish girl of the same age and both wearing the brace, and let us assume that their HRQOL is known and that it is precisely the same. Since their HRQOL level the same, the girls' score to the ISYQOL items is expected to be the same. Imagine that the Italian girl scores 2 and the Polish one scores 0 on the same item. There is another variable in addition to HRQOL (which, as we said, is precisely the same in the two girls) that affects the item's score independently from HRQOL. Gender, age and treatment are the same in the two girls. Therefore, nationality, which is different between the two participants, could bias the girls' answers to this item. In this condition, DIF for the item is concluded.

DIF was tested for each ISYQOL item as usual in Rasch analysis. Briefly, an item is affected by DIF for a variable if its calibration is significantly different between two groups of participants and when this difference is > 0.5 logit. As done in the case of multiple comparisons, DIF for nations was tested for each nation against all nations combined.

DIF for culture and nationality is quite common, and thus it was expected for ISYQOL. In alignment with the main aim of the current work, we decided to correct any DIF for nations by applying the "item splitting" procedure (3).

According to this method, the different translations of the items with DIF are handled as different items (4). For simplicity, consider that two countries only took part in the study (e.g. Italy and Poland) and assume that item 10 showed DIF for nationality, with the calibration of the Polish translation being different from that of the Italian one. Item 10 is thus split into two separate items: one (the Polish translation of item 10) administered to Polish patients only and the other (the Italian version of the item) administered to Italian patients. A subsequent Rasch analysis is run on the new dataset containing two versions of item 10 ("10 – Poland" and "10 – Italy"), with Italians with missing values on "10 – Poland" and, conversely, Polish participants with missing values on "10 – Italy". A different calibration is obtained for item "10 – Italy" and item "10 – Poland", thus taking into account that the same score on item 10 does not reflect the same amount of HRQOL in Polish patients and Italians. Alternate forms of the score-to-measure table are eventually available (see below), with the score-to-measure conversion for Italians using the calibration of item "10 – Italy" and that for the Polish patients using that of item "10 – Poland".

In addition to nationality, DIF was also tested for age (≤ 12 vs > 12 years), brace (not wearing vs wearing the brace), disease severity (Cobb's angle ≤ 30 ° vs > 30 °) and gender (males vs females).

Preparing alternate forms of a questionnaire that consider all DIF would be unpractical. For example, suppose one item was affected by DIF for nationality (with Spanish patients and Turkish respondents different from the whole group), brace and gender. In that case, 12 other score-to-measure tables should be arranged (e.g. one for male patients from Spain without the brace, a second for female patients from Spain without the brace…). However, similarly to multidimensionality, DIF could be of no harm for measures from a practical point of view. DIF impact on measures can be tested following the procedure described by Lange and colleagues (5,6) and taken up by Tennant and Pallant (7). According to these Authors, DIF can be ignored if no more than 5% of the patient's measures returned by the items affected by DIF are significantly different from those obtained with a set of pure items (i.e. items free of DIF for any of the variables reported above).

This second solution has been adopted here to consider the consequences of any DIF for age, brace, severity and gender.

Persons' reliability. ISYQOL reliability was estimated with the persons' reliability of the Rasch analysis (extreme and non-extreme observations) and Cronbach's alpha.

The items' map. The items' map shows the position of the items' calibration and the frequency distribution of the patients' measures along a line representing the variable continuum (here HRQOL).

This graph is full of information. For example, the floor/ceiling effect of the questionnaire is immediately apparent from the persons' distribution along the line of the variable. The items' distribution along this line is also essential. For instance, a wide gap between two consecutive items flags a range of the variable poorly measured by the questionnaire.

Score-to-measure conversion. For questionnaires consistent with the Rasch model, it is good practice to provide a table reporting the questionnaire's total score conversion into the corresponding interval measure.

These measures are provided in logits (i.e. the accepted measurement unit in the Rasch framework), but they are often expressed on a 0-100% scale with arbitrary units. It is worth stressing that they are interval measures in either case. The score-to-measure table also reports the corresponding standard error for each measure, which reflects the measurement's precision.

This table is addressed to scholars, who, for example, could benefit from these interval measures to run parametric statistics and clinicians, who could use these measures and their errors to assess if a single patient is significantly different between two consecutive measures.

REFERENCES

1. Wright B. Reasonable mean-square fit values. Rasch Meas Trans. 1994;8:370.

2. Smith Jr EV. Detecting and evaluating the impact of multidimensionality using item fit statistics and principal component analysis of residuals. J Appl Meas. 2002;3(2):205–31.

3. Kyngdon A. Is combining samples productive. Quick Check Tests DIF Rasch Meas Trans. 2011;25(2):1324–5.

4. Roorda LD, Jones CA, Waltz M, Lankhorst GJ, Bouter LM, van der Eijken JW, et al. Satisfactory cross cultural equivalence of the Dutch WOMAC in patients with hip osteoarthritis waiting for arthroplasty. Ann Rheum Dis. gennaio 2004;63(1):36–42.

5. Lange R, Irwin HJ, Houran J. Top-down purification of Tobacyk's Revised Paranormal Belief Scale. Personal Individ Differ. 2000;29(1):131–56.

6. Lange R, Thalbourne MA, Houran J, Lester D. Depressive response sets due to gender and culture-based differential item functioning. Personal Individ Differ. 2002;33(6):937–54.

7. Tennant A, Pallant J. DIF matters: A practical approach to test if differential item functioning makes a difference. Rasch Meas Trans. 2007;20(4):1082–4.