SUPPLEMENTARY DIGITAL MATERIAL 4

Control analysis: 50 questionnaires from each country.

We report here the results of a control analysis in which each of the seven countries contributed 50 questionnaires to the dataset (350 questionnaires in total). To this aim, 50 questionnaires were randomly extracted from the complete set of 250 Italian questionnaires.

As in the analysis reported in the main text, item 5 showed disordered categories and its fit to the model was quite poor (IN-MNSQ = 1.62, IN-ZSTD = 5.30; OUT-MNSQ = 2.53, OUT-ZSTD = 5.74). In the subsequent analyses, fit of item 13 was also poor (OUT-MNSQ = 1.76, OUT-ZSTD = 2.55), as well as that of item 6 (OUT-MNSQ = 1.72, OUT-ZSTD = 2.71) and that of item 10 (OUT-MNSQ = 1.41, OUT-ZSTD = 4.84). The remaining 16 items showed good fit to the model (IN-MNSQ range: 0.89 – 1.13; OUT-MNSQ: 0.81 – 1.17).

The principal component analysis of the model's residuals confirmed some amount of multidimensionality. The eigenvalue of the first principal component (2.22) is the same as that found in the primary analysis.

Persons' reliability was 0.79 and the Cronbach's alpha 0.86, which was also comparable to the main analysis.

The following table compares the items' calibration from the main and control analyses. In addition, the standard error of the calibrations (SE) is also provided.

	50 questionnaires per country			Full sample			
item number	calibration	SE	item rank	calibration	SE	item rank	Δ
1	-0.49	0.11	6	-0.52	0.09	6	-0.03
2	-0.10	0.10	8	-0.03	0.08	8	0.07
3	0.77	0.10	12*	0.86	0.08	13*	0.09
4	0.31	0.10	9	0.15	0.08	9	-0.16
5	0.56	0.10	11	0.62	0.08	11	0.06
6	0.89	0.10	14	0.87	0.08	14	-0.02
7	-0.72	0.10	5	-0.74	0.08	5	-0.02
8	0.82	0.10	13*	0.74	0.08	12*	-0.08
9	0.43	0.09	10	0.38	0.08	10	-0.05
10	-0.73	0.11	4	-0.95	0.09	4	-0.22
11	-0.98	0.12	3	-1.17	0.09	3	-0.19
12	-0.20	0.12	7	-0.16	0.09	7	0.04
13	-1.43	0.13	2	-1.29	0.10	2	0.14
14	0.97	0.14	15	1.02	0.10	15	0.05
15	1.74	0.17	16	1.97	0.13	16	0.23
16	-1.85	0.15	1	-1.76	0.11	1	0.09

The split items procedure (see main text) was not applied here. The maximum difference in the two items' calibrations (Δ) is 0.23 logit, well below 0.5 logit (the conventional threshold flagging an appreciable difference between two measures). It is also noteworthy that the items' raking (from the item with the smallest calibration to the one with the highest) is essentially the same in the two analyses, with only items 3 and 8 with inverted ranks (*) in the two analyses. The order of the "ruler's ticks" is invariant regarding the patients' sample. Agreement between the two calibrations is high.

Differential item functioning of ISYQOL International: detailed results.

The table reports the items affected by a large (i.e. > 0.5 logit) and significant (i.e. p < 0.01) differential item functioning (DIF) for nationality, brace and gender. No DIF was found for age and disease severity.

Nationality						
Item Number	**Group 1**	**calibration**	**Group 2**	**calibration**	**DIF size**	**p-value**
2	French CAN	-0.84	Overall	-0.03	-0.81	0.003
3	English CAN	0.90	Overall	0.15	0.75	0.008
3	French CAN	0.13	Overall	0.86	-0.73	0.007
4	Greece	1.04	Overall	0.15	0.89	0.003
5	Greece	-0.64	Overall	0.62	-1.26	< 0.001
5	Turkey	-0.46	Overall	0.62	-1.08	< 0.001
7	Poland	-1.94	Overall	-0.74	-1.20	< 0.001
10	Poland	-0.21	Overall	-0.95	0.74	0.006
12	Turkey	-1.27	Overall	-0.16	-1.11	0.002
Brace						
Item Number	**Group 1**	**calibration**	**Group 2**	**calibration**	**DIF size**	**Prob.**
1	Brace No	-0.87	Brace Yes	-0.36	-0.51	0.006
2	Brace No	-0.45	Brace Yes	0.23	-0.68	< 0.001
Gender						
Item Number	**Group 1**	**calibration**	**Group 2**	**calibration**	**DIF size**	**Prob.**
7	Female	-1.71	Male	-0.62	-1.09	< 0.001
10	Female	-0.23	Male	-1.05	0.82	0.005
14	Female	2.89	Male	0.91	1.98	0.004
16	Female	-2.72	Male	-1.64	-1.08	0.004

Calibration: item's calibration in the two groups contrasted in the DIF analysis (Group 1 vs Group 2). DIF size: item's calibration in Group 1 – item's calibration in Group 2. Calibration and DIF size are given in logit. The Student's t-test tests the null hypothesis *"DIF size = 0 logits"*. The corresponding type 1 error probability is given by p-value. French CAN: French Canada; English CAN: English Canada; Brace: participants not wearing the brace (Brace No) vs participants wearing the brace (Brace Yes).