Psychometric validation of the Swedish Four-Dimensional Symptom Questionnaire (4DSQ) using differential item and test functioning (DIF and DTF) analysis

Berend Terluin, MD, PhD
Almere (NL), August 23, 2017

Abstract

This study aimed to validate the Swedish translation of the Dutch Four-Dimensional Symptom Questionnaire (4DSQ) using differential item and test functioning (DIF and DTF) analysis. Two methods were used to detect DIF: the Mantel-Haenszel (MH) method and the hybrid ordinal logistic regression (HOLR) method. DTF was assessed by plotting raw scale scores against DIF-adjusted Rasch theta scores. Seven items (in 3 scales) demonstrated DIF, but the impact on the scale score (DTF) was negligible. The Swedish 4DSQ scores can be interpreted in the same way as the Dutch scores.

Introduction

The Four-Dimensional Symptom Questionnaire (4DSQ) is a self-report questionnaire measuring distress, depression, anxiety and somatization [1]. The questionnaire has been developed in Dutch primary care and was translated into the Swedish language. This report concerns the psychometric validation of the Swedish version of the 4DSQ against the original Dutch questionnaire.

Methods

Translation

The Swedish translation of the 4DSQ was conducted by a team of primary care professionals led by Liisa Lönnberg, in collaboration with the author of the 4DSQ. A procedure of forward and backward translation was used. First, two independent Swedish translations were made, one from the original Dutch version, and another from the English version. Both translations were compared and a single Swedish translation was decided upon. Subsequently, this Swedish translation was independently back-translated into the original Dutch language. Differences between the back-translation and the original 4DSQ were then scrutinized and discussed among all members of the translation team to obtain maximum linguistic validity of the Swedish version.

Data for psychometric analyses

The focal (Swedish) group consisted of 252 primary care patients. The percentage females was 73%. The mean age was 45.6 years (SD = 14.8, range 17-83). The reference (Dutch) group consisted of a gender and age matched sample (n = 252) drawn from a data pool of primary care patients with (suspected) mental health problems (73% females, mean age 45.1 years (SD = 15.4, range 15-83)). In both groups 4DSQ data had been collected as part of routine care.

In the Swedish data only 56 item scores were missing (0.4% of all item scores); in the Dutch data 91 item scores were missing (0.7% of all item scores). Missing item scores were imputed using the response function method [2].

Analysis

Whether the Swedish 4DSQ measures the same constructs in the same way as the Dutch 4DSQ was examined using differential item functioning (DIF) analysis and differential test functioning (DTF) analysis.

The idea behind DIF-analysis as a way to validate questionnaire translations, is that the translated scale measures the same as the original scale when the translated items can be shown to "function" the same as the original items [3]. Two methods were used for the detection of DIF: the Mantel-Haenszel (MH) method as implemented in the freeware statistical program jMetrik 2.1.0 (www.itemanalysis.com) and the hybrid ordinal logistic regression (HOLR) method as implemented in the R-package "lordif" [4]. Criteria for DIF were a standardized mean difference (SMD) >0.1 (and p<0.001) for the MH-method and a delta R-square >0.02 (and p<0.001) for the HOLR-method.

To what extent does the item level DIF impact on the scale scores? This is of practical importance when one would like to apply the 4DSQ in Swedish patient samples and interpret the scale scores. To examine the impact of DIF, the DIF-items were split into two items, one for each group. For instance, if item 1 was found to have DIF, the item was split into item 1F for the Swedish (focal) group and item 1R for the Dutch (reference) group. Dutch patients would have missing values for item 1F and Swedish patients for item 1R. Next, for each scale, concurrent calibration was performed using Rasch-analysis (in jMetrik) with the DIF-free items of the scale as anchor items. Simultaneously, Rasch IRT-scores were estimated for all patients on the same theta-scale. Note that the theta-scores represent estimates of the true levels of attributes, adjusted for DIF. Finally, raw scale scores were plotted as a function of the DIF-free theta-scores by group.

Table 1 Items detected with DIF

Scale	Item#	Short Swedish description	MH-method[a]	HOLR-method[b]	Direction[c]
Distress	17	nedstämdhet	0.24	0.021	less
	22	orkeslöshet	0.23	–	less
	25	spänd	– 0.18	0.025	more
Anxiety	21	en oförklarlig känsla av rädsla	–	0.034	more
	24	ångest- eller panikattacker	0.22	–	less
	27	ängslig	0.19	–	less
Somatization	12	illamående, eller en orolig mage	0.24	–	less

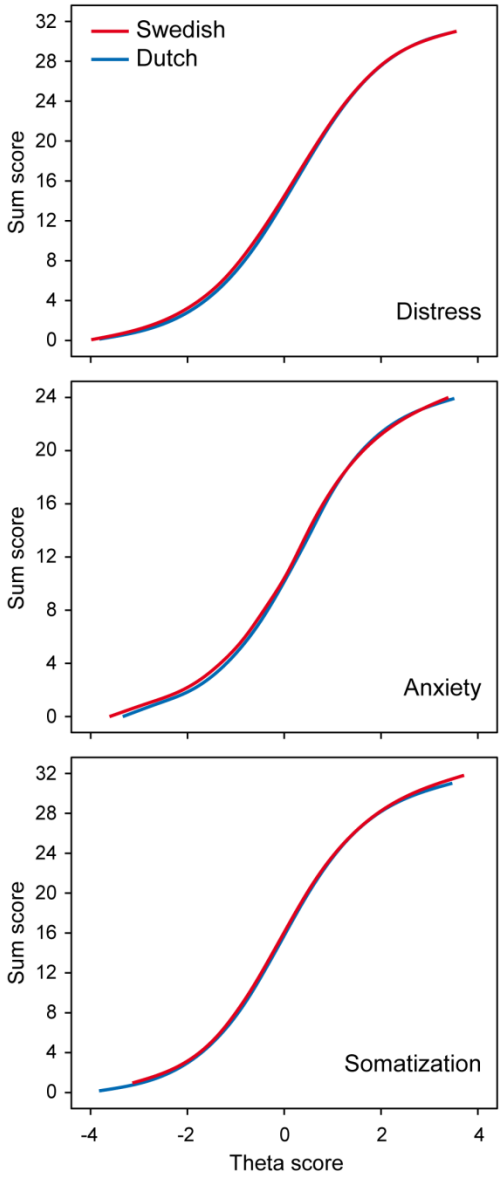
[a] standardized mean difference

[b] delta R-squared

[c] Direction of DIF: item was less/more severe for Swedish patients

Results

DIF was detected in three 4DSQ scales (Table 1). The depression scale was DIF-free. A total of 7 items were flagged for DIF by either method. The MH-method detected 6 DIF-items and the HOLR-method detected 3 items. Only 2 DIF-items were detected by both methods. Five items were found to be less severe for Swedish patients, meaning that Swedish patients had a lower threshold than Dutch patients for endorsing these items. On the other hand, 2 items appeared to be more severe for Swedish patients than for Dutch patients. More severe means that the item represents a more severe symptom for Swedish patients. An item that is less severe might lead to higher scores than can be explained by the true level of distress, anxiety or somatization, whereas an item that is more severe can lead to a lower score. For instance, item 25 was more severe for Swedish patients, causing them to score on average 0.18 points lower on the distress scale, adjusted for differences in the distribution of true distress across the Swedish and Dutch groups. However, items 17 and 22 caused Swedish people to score a little higher than Dutch patients.



The results of the DTF analysis are shown in Figure 1. It is apparent that the item-level DIF had a negligible effect on the scale scores.

Conclusion

The Swedish 4DSQ measures the same constructs (distress, depression, anxiety and somatization) as the original Dutch 4DSQ, practically in the same way. The Swedish scores can be interpreted in the same way as the Dutch scores. The Swedish 4DSQ can use the same cut-off points as the original 4DSQ.

Acknowledgement

The Swedish team consisted of L. Lönnberg, B. Hultberg, G.N. Krums-Vabins, L. Persson, and J. Eggenkamp.

Figure 1 DTF, scale impact of DIF: scales' sum scores as a function of the true (theta) score by group (Swedish: red curves, Dutch: blue curves)

Reference List

1. Terluin B, Van Marwijk HWJ, Adèr HJ, De Vet HCW, Penninx BWJH, Hermens MLM *et al.*: **The Four-Dimensional Symptom Questionnaire (4DSQ): a validation study of a multidimensional self-report questionnaire to assess distress, depression, anxiety and somatization.** *BMC Psychiatry* 2006, **6:** 34.
2. van Ginkel JR, van der Ark LA: **SPSS syntax for missing value imputation in test and questionnaire data.** *Applied Psychological Measurement* 2005, **29:** 152-153.
3. Petersen MA, Groenvold M, Bjorner JB, Aaronson N, Conroy T, Cull A *et al.*: **Use of differential item functioning analysis to assess the equivalence of translations of a questionnaire.** *Qual Life Res* 2003, **12:** 373-385.
4. Choi SW, Gibbons LE, Crane PK: **lordif: An R package for detecting differential item functioning using iterative hybrid ordinal logistic regression/item response theory and Monte Carlo simulations.** *J Stat Softw* 2011, **39:** 1-30.