# Cross-cultural validation of the Turkish Four-Dimensional Symptom Questionnaire (4DSQ) using differential item and test functioning (DIF and DTF) analysis

Article (Accepted Version)

# Assessing measurement equivalence of the Turkish Four-Dimensional Symptom Questionnaire (4DSQ) to the original Dutch instrument. A cross-cultural validation study using differential item and test functioning (DIF and DTF) analysis

Berend Terluin[1], Pemra C Unalan[2], Nurver Turfaner Sipahioglu[3], Seda Arslan Özkul[2], Harm WJ van Marwijk[1,4]

**1** Department of General Practice and Elderly Care Medicine, and the EMGO Institute for Health and Care Research, VU University Medical Center, Van der Boechorststraat 7, 1081 BT Amsterdam, The Netherlands

**2** Department of Family Medicine, Marmara University Medical Faculty, Tıbbiye Cad. 34688 sküdar, Istanbul, Turkey

**3 .** Department of Family Medicine, Cerrahpasa Medical Faculty, Istanbul University, 34303, Cerrahpasa, Istanbul, Turkey

**4.** Primary Care Research Centre, Institute of Population Health, Williamson Building, Oxford Road, Manchester M13 9PL, United Kingdom

Email addresses:

BT: b.terluin@vumc.nl

PCU: punalan@marmara.edu.tr

NTS: nurver@doctor.com, nurver@istanbul.edu.tr

HWJvM: harm.vanmarwijk@manchester.ac.uk

Abstract

Background

The Four-Dimensional Symptom Questionnaire (4DSQ) is originally a Dutch 50 item questionnaire developed in primary care to assess distress, depression, anxiety and somatization. We aimed to develop and validate a Turkish translation of the 4DSQ.

Methods

The questionnaire was translated using forward and backward translation, and pilot testing. Turkish 4DSQ-data were collected in 352 consecutive adult primary care patients. For comparison, gender and age matched Dutch reference data were drawn from a larger existing dataset. We used differential item and test functioning (DIF and DTF) analysis to validate the Turkish translation to the original Dutch questionnaire. Through additional inquiry we tried to obtain more insight in the background of DIF in some items.

Results

Twenty-one items displayed DIF but this impacted only the distress and depression scores. Inquiry among Turkish people revealed that the reason for DTF in the distress scale was probably related to unfavourable socio-economic circumstances. On the other hand, the likely explanation for DTF in the depression scale appeared to be grounded in culturally and religiously determined optimistic beliefs. Raising the distress cut-offs by 2 points and lowering the depression cut-offs by 1 point ensures that individual Turkish 4DSQ scores be correctly interpreted.

Conclusions

The Turkish translation of the 4DSQ (named: “Dört-Boyutlu Yakınma Listesi”, 4BYL) measures the same constructs as the original Dutch questionnaire. Turkish anxiety and somatization scores can be interpreted in the same way as Dutch scores. However, when interpreting Turkish distress and depression scores, DTF should be taken into account.

**Key words**: Distress, depression, anxiety, somatization, cross-cultural validation, differential and test functioning, cultural beliefs, religious beliefs

The Four-Dimensional Symptom Questionnaire (4DSQ) is used as a tool to aid the evaluation of patients with (suspected) mental health problems in primary care settings. Filling in the scale and discussing the results with a health care provider enhances the acknowledgement of emotional issues in patients, and helps to detect moderate and severe depressive and anxiety disorders that need specific attention (1). The 4DSQ is an originally Dutch questionnaire, developed in primary care (2) to measure distress, depression, anxiety and somatization symptoms (3). Its first focus on a generic distress dimension is unique for primary care and covers the emotional consequences of stress and coping. Distress reflects the amount of difficulty a person is experiencing having to deal with psychosocial problems, adversities and negative life events. A high score in this dimension, combined with low scores on depression, anxiety and somatization, reflects pure 'stress'. The depression dimension taps on symptoms of moderate and severe depressive disorder, and reflects the probability of having a depressive disorder severe enough to warrant specific treatment (4). The anxiety dimension encompasses the kind of symptoms that are characteristic of anxiety disorders, and the anxiety score reflects the probability of having one or more anxiety disorders severe enough to warrant specific treatment (5). The somatization dimension covers the kind of physical symptoms that usually are manifestations of bodily distress, i.e., the way the body reacts to distress. The somatization score reflects the severity of this reaction. High scores on depression, anxiety or somatization are usually combined with high distress scores.

We aimed to develop and validate the Turkish 4DSQ against the original Dutch 4DSQ. This paper describes the procedure of translating the 4DSQ into the Turkish language (i.e., linguistic validation), and the subsequent assessment of measurement equivalence of that translation compared with the original Dutch 4DSQ (i.e., psychometric validation). Measurement equivalence (or invariance) implies that a measurement instrument measures the same construct in the same way across different groups (6). When a translation can be shown to be equivalent to the original instrument, the validity of the original instrument can be assumed to apply to the translated instrument. The scores can be assumed to possess the same meaning, mean scores across different groups can be compared in a meaningful way, and cut-off points can be applied with the same consequences.

## METHODS

### Translation – linguistic validation

We created a Turkish version of the 4DSQ largely in accordance with the recommendations of the MAPI Research Institute (7). The purpose of the procedure was to obtain a Turkish version as similar as possible to the original questionnaire, that would be acceptable to native Turkish speakers in and outside Turkey. The process started with conceptual analysis of the source instrument. A Turkish family physician, living and working in Istanbul, acted as a consultant reviewer and coordinator between the translators and clinicians involved in the process. The developer of the questionnaire (BT) was contacted so that his involvement from the early beginning would ensure greater coherence of the final version. The original Dutch 4DSQ was forward translated by a psychiatrist, born in Turkey and living and working in the Netherlands for about ten years, who was a native speaker of the target Turkish language and fluent both in Dutch and English. A second forward translation, created by an unknown translator in the Netherlands, was already available. We reviewed both forward translations, discussed differences among the translator, developer, and three Turkish reviewers. A consensus translation was then presented to an independent translator for back-translation. The back translator was a Turkish medical secretary, born in the Netherlands, who had lived and attended school in this country till the age of 13. The back-translation was then reviewed and compared with the original questionnaire by the consultant and developer. This lead to the establishment of a revised preliminary Turkish version that was subsequently presented to the reviewers for clinical review. They independently scrutinized each question of the Turkish version and identified several items for further discussion. As a result some stylistic changes and alternative wordings were realized. The resulting version was then pre-tested by each reviewer/physician in at least ten patients visiting the family medicine outpatient clinic. Then the reviewers analysed the responses to identify necessary modifications. They decided to adjust two items for cultural reasons to make these clearer. After this adjustment, the Turkish 4DSQ was finalized and baptized the “Dört Boyutlu Yakınma Listesi” (4BYL).

Differential item functioning

We used differential item functioning (DIF) analysis to establish measurement equivalence, i.e., whether the Turkish translation actually measured the same constructs as the original Dutch 4DSQ. DIF analysis assumes that the responses to the items of a scale (e.g., a depression scale) reflect an underlying latent trait (e.g., depression). The method examines whether these item responses, in relation to the underlying latent depression trait, are the same in different groups (8-10). When the responses to the depression items, in relation to the underlying depression trait, can be demonstrated to be the same in Turkish and Dutch primary

care patients (i.e., when the depression items ‘function’ the same way in both groups), it can be assumed that the depression scale measures the same construct in both groups and that Turkish depression scores and cut-off points can be interpreted in the same way as Dutch depression scores.

The relation between an item and the underlying latent trait, of which the responses to that item are assumed to be the expression, is characterized by the item parameters ‘severity’ and ‘discrimination’ and can be visualized by the item characteristics curve (ICC). Figure 1 shows an ICC of a dichotomous item (i.e., an item with two response options: yes/no). The curve displays the probability of a positive response as a function of the underlying trait. If this were the latent trait of depression, the probability of a positive response to this depression item increases with an increasing level of depression severity. The ‘severity’ parameter of the item is the level of the latent trait associated with a 50% probability of a positive response. It reflects the severity of the item in terms of the level of depression required to endorse the item. The ‘discrimination’ parameter of the item is represented by the slope of the curve. The steeper the curve, the better able is the item to distinguish people who are high on the latent trait (i.e., who are more severely depressed) from people who are low on the trait (i.e., who are less severely depressed). Because the ‘discrimination’ parameter also reflects the correlation between the item and the latent trait, this parameter tells us how well an item measures the latent trait. Different items have different ‘severity’ and ‘discrimination’ parameters in a given group of respondents.

Differential item functioning (DIF, i.e., when an item does not ‘function’ the same way in different groups) means that an item has different ‘severity’ or ‘discrimination’ parameters (or both) across different groups. An item can be more ‘severe’ for one group than for another group. This is called uniform DIF because the item is uniformly (i.e., over the whole range of the latent trait) more severe for one group than the other (Figure 2, left panel). In contrast, an item can be more ‘discriminative’ in one group than another group. This is called non-uniform DIF because the item is relatively more severe for one group in one part of the latent trait scale, but more severe for the other group in the other part of the scale (Figure 2, right penal). Non-uniform DIF suggests that the item does not measure the latent trait equally well in both groups. In the group in which the item demonstrates the lowest discrimination (i.e., in which the slope of the ICC is more gentle), the item either measures (partly) something different than the latent trait or the item score contains more measurement error (i.e., the item is less reliable). DIF-analysis aims to examine whether or not the severity and discrimination of an item is the same in different groups. If the items of a scale have the same severity and

discrimination parameters in different groups, then it can be assumed that the scale measures the same construct in these groups. However, when a depression item is less severe in group A than in group B (Figure 2, left panel), people in group A tend to get higher depression scores than people in group B while having the same true level of the latent depression trait. The presence of DIF prevents the meaningful comparison of depression scores across different groups. When a depression item is less discriminative in group A than in group B (Figure 2, right panel), because in group A the item is measuring something else or has more measurement error, people tend to get higher depression scores when their true level of depression is relatively low. But, when the true level of depression is relatively high, people in group A tend to get lower depression scores because the item does not measure depression as well as it does in group B.

In DIF analysis the item parameters severity and discrimination are compared between two (or more) groups. All DIF-analysis methods need to match the groups according to the latent trait. Therefore, they need a (measurable) variable that approximates the (immeasurable) latent trait. This so-called 'matching variable' is usually somehow constructed based on the information of the item scores. However, when some of the items contain DIF, the matching variable will contain DIF and will not provide an unbiased approximation of the latent trait. Therefore, the matching variable needs to be 'purified', i.e., DIF needs to be removed from the matching variable. This is accomplished in different ways in different DIF-analysis methods.

## Participants

Turkish 4DSQ-data were collected in consecutive adult patients at their first visit to Marmara University Family Medicine outpatient clinics in Istanbul, Turkey. Patients were personally approached in the waiting room, and informed about the study. If they accepted to join the study written informed consent was obtained. Patients were explicitly instructed not to skip any questions. The study protocol was approved by the Marmara University Medical Faculty Ethical Committee (Ref. 70737436-050.06.04). The Dutch reference 4DSQ data were drawn from a larger database of primary care patients with suspected mental health problems, who had completed a 4DSQ within the framework of routine care in Health Centre De Spil in Almere, the Netherlands. An age and gender matched sample of patients was randomly selected from this database.

## Measurement

The 4DSQ contains 50 items, measuring distress (16 items), depression (6 items), anxiety (12 items) and somatization (16 items). The 4DSQ asks how often in the past week respondents have experienced certain symptoms. Responses can be provided on a 5-point scale from "no" to "very often or constantly". These responses are coded on a 3-point scale for the purpose of calculating scale scores: "no" = 0, "sometimes" = 1, "regularly"/"often"/"very often or constantly" = 2. The 4DSQ was completed as a pen-and-paper version in both groups.

Analysis

*Initial analyses*. Missing item scores were imputed using the response function method, a method that accounts for both differences between items and persons (11). Differences in gender and age across the groups were compared using Chi-square test and t-test respectively. Differences in mean 4DSQ scale scores were tested using t-tests. In addition, we calculated Cronbach's alpha as a measure of reliability and obtained 2000 bootstrap estimates of the difference between the groups using the 'psych' package (12) in R, a statistical program that is freely available (13).

*Unidimensionality*. As DIF analysis assumes the scales are unidimensional, we assessed unidimensionality by multi-group confirmatory factor analysis (CFA) using the 'lavaan' package in R (14). We fitted one-factor models for each scale, allowing for correlations between residual variances of items sharing specific content. To account for the ordinal character of the item scores, the items were treated as ordered variables. As indicators of unidimensionality we inspected the mean and variance adjusted model chi-square statistic (and its degrees of freedom), the comparative fit index (CFI), the Tucker-Lewis index (TLI) and the root mean square error of approximation (RMSEA). Criteria for unidimensionality were >0.95 for the CFI and TLI, and <0.06 for the RMSEA (15). We assessed 'configural invariance' without constraints on item loadings and intercepts as the DIF-analysis would address the equivalents of loadings (i.e., discrimination parameters) and intercepts (i.e., severity parameters). Configural invariance thus means that the same items load onto the same factors across the groups.

*Differential item functioning.* For DIF-analysis several methods are available without there being any consensus in the literature about which method is superior (9). For that reason, some experts recommend to use more than one method (16). We chose two methodologically

different methods, the non-parametric Mantel-Haenszel (M-H) method (17) and the parametric hybrid ordinal logistic regression (HOLR) method (18).

The M-H-method uses the ordinary sum score of the items of a scale as 'matching variable' to start with. Then mean item scores of the groups are compared for every level of the matching variable. After standardization of the sum scores over de groups, a 'standardized mean difference' (SMD) is calculated. Conventionally, an SMD of 5% of the item score range (in this case 0.1 points) indicates a clinically important degree of DIF, provided that this effect is also statistically significant (19). We chose $p < 0.001$ to account for multiple testing. When one or more items are found to have DIF, the item with the most severe DIF is removed from the matching variable as a way of purification of the matching variable. Then the analysis is repeated. Purification and reanalysis is repeated until no further items with DIF are discovered. The M-H-method detects mainly uniform DIF. We used the program jMetrik 2.1, which is freely available on the Internet (www.itemanalysis.com) (20).

For the HOLR-analysis we used the package 'lordif' in R (18). The HOLR-method combines item response theory (IRT) with ordinal logistic regression (OLR) (18). OLR models the odds of endorsing each of the categories of an ordinal scale (in this case the item's response categories) as a function of one or more 'determinants', in this case the latent trait and group membership. Two-parameter logistic (2PL) IRT-analysis is used to calculate theta-scores, which are subsequently used as matching variable. In the absence of DIF, the item responses are solely determined by the matching variable, and group membership does not have an additional predictive value in the regression model. However, when group membership as a determinant does result in a substantial improvement of the prediction of the item responses, uniform DIF is present. Inclusion of the interaction term between matching variable and group membership allows for the testing of non-uniform DIF. We used a significant ($p < 0.001$) increase in the model's explained variance (McFadden's $R^2$) by 2% or more as criterion for total DIF (i.e., uniform DIF and non-uniform DIF together) (18). When DIF is detected, the HOLR-method 'purifies' the matching variable by estimating group-specific parameters for the DIF-laden item, and re-estimation of the theta-scores. An advantage of this method is that no items need to be removed from the matching variable and all available item information can be utilized. After recalculating the matching variable, the analysis is repeated until the same items are flagged for DIF in two consecutive iterations.

*Differential test functioning.* To evaluate the effect of item level DIF on the 4DSQ scale scores we compared the raw scale scores (i.e., the ordinary sum of the item scores) with

estimates of the DIF-free scores across both groups. We used Rasch analysis, a one parameter (1PL) IRT model, to obtain theta-scores (21). We used the DIF-free items as anchor-items to estimate theta-scores in both groups on the same scale. The item parameters of the DIF-laden items were estimated separately for Turkish and Dutch patients. The raw (i.e., DIF-laden) scale scores by group were then plotted against the DIF-free theta-scores. The effect of item level DIF on the scale score (i.e., DTF) was evidenced by the distance between the group-specific curves. We used jMetrik 2.1 to perform the Rasch analyses.

### Additional inquiry

To obtain insight in the background of the DIF discovered, we presented the results to a convenience sample of Turkish speaking people in our personal networks, both living in Turkey and in the Netherlands. They were asked to reflect on the meaning of the DIF-items by cognitive interview method in order to discern why these items were either more or less severe for Turkish people or seemed not to measure exactly the same constructs as the Dutch counterpart items did. Results were discussed in the research team.

## RESULTS

### Initial analysis

A total of 352 Turkish patients (73% female) agreed to participate. Their mean age was 37.4 (SD=14.5). The matched Dutch sample consisted of 352 patients (73% female) with a mean age of 38.3 (SD=14.5). There were no significant differences between the groups regarding gender (Chi-square=0.000, p=0.000) and mean age (t(702)=0.813, p=0.417). In the Dutch sample 145 item scores (0.82%) were missing. In the Turkish sample only 2 item scores (0.0001%) were missing. The missing item scores were successfully imputed. The 4DSQ scales demonstrated good reliability as evidenced by Cronbach's alpha values well above 0.80 (Table 1). However, the Turkish values were significantly lower than the Dutch values. Probably this is connected to the lower mean 4DSQ scores and consequently the lower variability of the scores in Turkish patients (also shown in Table 1). Lower mean scores in Turkish patients had to be expected because the Turkish patients were, unlike the Dutch patients, not specifically selected because of a (suspected) mental health problem.

Unidimensionality

Multi-group CFA confirmed one-factor models for the 4DSQ-scales in both groups (Table 2). The residual covariance of four item pairs and one item triplet needed to be freely estimated in order to obtain the required model fit. The model chi-square statistic suggested rejection of perfect fit, however chi-square is sensitive to the sample size (22). The other indices, CFI, TLI and RMSEA, suggested adequate fit of the data to a one-factor model.

DIF-analysis

The M-H-analysis identified DIF in 20 out of the 50 items, whereas the HOLR-method identified 12 items with DIF (both methods identified 21 items with DIF; Table 3). Ten items were more severe for Turkish patients while eleven items were less severe. Three items (#47, #48, #49) exhibited mixed uniform and non-uniform DIF. The Turkish versions of item #47 and #48 possessed greater discrimination than the corresponding Dutch items, indication that the Turkish items fitted the distress scale better than the Dutch items. The Turkish item #49 was less discriminative than the Dutch item, but still acceptable as an indicator of anxiety. The other DIF-items demonstrated predominantly uniform DIF.

The distress item with the largest amount of total DIF ($\Delta R^2 = 6.47\%$), item #47 ("fleeting images of any upsetting event(s)"), was responsible for a mean increase in raw distress score of 0.38 scale points, holding the true level of distress constant. The worst depression item (total DIF: ($\Delta R^2 = 8.49\%$), item #35 ("feeling there is no escape from your situation"), was responsible for a mean decrease in raw distress score of 0.29 scale points, holding the true level of depression constant.

DTF-analysis

Figure 3 shows the raw 4DSQ scale scores as functions of the theta scores, by group. Note that the theta scores produced by the Rasch analysis reflected the unbiased position of the patients on the latent traits underlying the 4DSQ-scales. The impact of item level DIF on the total scale score was apparent by the vertical distance between the curves for Turkish and Dutch patients.

With respect to the distress scale, we can see that the Turkish patients obtained a higher distress score (about 2 scale points on a scale range of 32 points) than the Dutch patients when they had the same true level of distress (represented by the theta score). This difference was about the same across the whole range of the distress scale except for the extremes. If left uncorrected, this DIF will result in some overrating of distress in Turkish patients.

Regarding the depression scale, we can see a relatively large difference between the raw scale scores of Turkish and Dutch patients (about 1.5 points on a scale range of 12 points). Given a true level of depression represented by a theta score of 0 (this is the mean severity level of the symptoms), which corresponds to a total score of about 6 in Dutch patients, Turkish patients scored on average 1.5 points lower than Dutch patients (4.5 versus 6). If left uncorrected, this DIF will lead to underrating of depression and underdetection of depressive disorder in Turkish patients, compared to Dutch patients. Six or more points on the depression scale is an important cut-off point for detecting moderate-severe depressive disorder (4).
With respect to the anxiety and somatization scales, the item level DIF did not have any substantial impact on the raw scale score.

Background of DIF
Results were discussed with 9 persons living in Turkey and 20 native Turkish speaking immigrants living in the Netherlands for more than 20 years. The discussions focused on the items that were responsible for differential functioning of the distress and depression scales, i.e., the distress items that were less severe for Turkish patients and the depression items that were more severe for Turkish patients.

*Distress*. Low mood (item #17, "feeling down or depressed") appeared to be difficult to translate into Turkish. Our translation of "keyifsizlik/isteksizlik", literally meaning "malaise/reluctance", turned out to be less severe for Turkish patients than the Dutch word "neerslachtigheid" is for Dutch patients. The other DIF-laden distress items (#22, #37, #41, #47, and #48) appeared to be correctly translated. Our informants suggested that the difficult socio-economic situation in Turkey, given the economic crisis, had made Turkish people more sensitive to specific features of distress as described by these items.

*Depression*. The most problematic depression item was item #35 ("feeling that there is no escape from your situation"). The translation appeared to be linguistically correct. However, our informants suggested that in the Turkish culture it is considered to be a shame to be that hopeless. One woman informant said that "every bad thing has its worse". This basic optimism appeared also to have a religious Islamic dimension too, as observed by the difference in responses between religious and not much religious informants. A common expression is "When one door closes, God opens another door". Religious beliefs also appeared to be at play in the response of Turkish people to item #28 ("feeling that everything

is meaningless"). Our informants expressed a deeply felt conviction that "every creature in the world has an aim and its life has a meaning". The feeling that everything is meaningless implies criticism towards the world's creator, as the whole world is God's work. The informants suggested that in item #34 ("can't enjoy anything anymore") probably the use of the word "zevk" ("pleasure") was responsible for DIF because "zevk" processes stronger physical connotations than "enjoy" (in Dutch: "genieten"). This would make it more difficult for Turkish people to admit any loss of pleasure.

## DISCUSSION

### Summary of findings

We found that the Turkish translation of the 4DSQ contained 21 items that functioned differently from their original Dutch counterpart items. These Turkish items differed mainly in 'severity' (10 items were more severe and 11 items were less severe). Only three items also differed in 'discrimination'. Because the most prevalent type of differential item functioning (DIF) was 'uniform' (i.e., concerning the item's severity), we can conclude that the Turkish 4DSQ items and hence the Turkish 4DSQ scales measure the same constructs as the Dutch items and scales. However, this does not imply that the Turkish and Dutch 4DSQ scores are completely equivalent. Our differential test functioning (DTF) analysis showed that the raw anxiety and somatization scores were related to the DIF-adjusted 'true' anxiety and somatization scores in identical ways in Turkish and Dutch patients. Therefore, we can conclude that, on the scale level, the Turkish 4DSQ anxiety and somatization scales are equivalent to the corresponding Dutch scales. This is the case despite the existence of non-equivalence at the item level. The likely reason why item level DIF does not impact the scale score, is that the effect of items that are more severe for Turkish patients are counteracted by the effect of items that are less severe, causing DIF to cancel out on the scale level. Unfortunately, such cancelling out of DIF effects did not occur with the distress and depression scales. In case of the distress scale, most DIF-laden items (6 out of 9) were less severe for Turkish patients, causing them to score higher on the distress scale compared to their true level of distress (as estimated by the theta-score), and compared to Dutch patients. The depression scale also suffered from imbalanced DIF. Three out of four DIF-laden items were more difficult for Turkish patients than for Dutch patients, causing lower raw depression scores in comparison to their true level of depression as estimated by the theta-score. The

DTF analysis revealed that the effect of DIF on the scale level occurred only in the mild and moderate range of the depression trait. Turkish patients with a moderate degree of depression (corresponding to a theta-score of 0 or a raw depression score of 6 in Dutch patients) scored on average about 1.5 scale points lower than Dutch patients with an equivalent true depression level. The reason was that for Turkish patients the threshold to score on items #28, #34 and #35 were much higher than for Dutch patients. However, when the depression was really severe (corresponding to a theta-score of 1 or a raw depression score of 8-9 in Dutch patients) Turkish patients scored just like Dutch patients. Thus, because some depression items are more severe for Turkish patients than for Dutch patients, mild and moderate raw depression scores in Turkish patients, do not have the exact meaning as in Dutch patients and tend to underestimate the true level of depression.

Explanations for DIF

Whenever an item functions differently in two groups (i.e., when DIF occurs), there must be a reason for this that must be found in differences between the groups involved. Our primary interest concerned differences in language (translation) and culture. However, other differences between the groups, such as religion, marital status, educational level, or occupational status, might also be responsible. It should be noted that differences in gender and age have been controlled for by matching the Dutch sample on these characteristics. Other variables were not assessed. Differences in severity spectrum could not be responsible for the occurrence of DIF as the DIF detection methods controlled for those differences (through the matching variable).

The linguistic validation procedure, with forward and back translation and pilot reviews, provides some protection against plain wrong translations. Nevertheless, a translation might not catch the exact meaning and nuance of the original item. That is, the translated item acquires a slightly different meaning than the original item. The translated item may still be a good indicator of the trait that is intended to be measured, but the translated item represents a more or less severe level of the trait than the original item. In many instances the exact meaning and nuance of a given word or expression in one language is difficult or sometimes even impossible to grasp in another language. This is true for depression-related words and expressions in the Turkish language (23). This kind of DIF is not always a big problem on the level of the scale score provided that DIF-items that are more severe balance DIF-items that are less severe. This is the case with the Turkish 4DSQ anxiety and somatization scales.

A special situation occurs when a translated item, which may be perfectly translated from a linguistic point of view, acquires a different cultural loading. This was the case in 3 of the 6 depression items. Two items expressed severe pessimistic/desperate thoughts and feelings, unacceptable in Turkish people. Therefore the threshold for experiencing and reporting such thoughts and feelings are much higher for Turkish patients than for Dutch patients. In other words, whereas Dutch patients fall prey to pessimism and despair at relatively low levels of depression, Turkish patients need to be really severely depressed before they start experiencing and reporting these pessimistic thoughts and feelings.
There is some supportive evidence in the literature. In a comparison of depressive symptom profiles between native Dutch people and Turkish-Dutch immigrants the item "feeling trapped" appeared to be much more severe for Turkish-Dutch people than for native Dutch (24). "Feeling trapped" and "feeling there is no escape" probably refer to the same pessimistic, desperate mind set. In a study comparing depressive symptoms across British and Turkish psychiatric outpatient samples the same phenomenon was suggested as Turkish patients scored lower on pessimism (25).

## Implications for practice

Now that we have established that the Turkish translation of the 4DSQ measures the same constructs in primary care patients as the original Dutch questionnaire, the Turkish 4DSQ can be used to measure distress, depression, anxiety and somatization. In addition, the Turkish anxiety and somatization scales were found to be equivalent to the corresponding Dutch scales and, thus, the scores can be interpreted in the same way as the Dutch scores. Hence, the Dutch cut-off points of the anxiety and somatization scales have the same meaning in the Turkish scales. However, when interpreting Turkish distress and depression scores, it should be realized that Turkish patients tend to score higher on the distress scale and lower on the depression scale compared to Dutch patients. In order to retain the same meaning of the cut-off points, those of the Turkish distress scale should be raised by 2 points. In addition, the cut-off points of the depression scale should be lowered by 1 point.
In future research, other items could be tested to replace the worst DIF-items of the distress and depression scales. For the time being, adjustment of cut-off points for distress and depression seems to be a practical solution.

## Conclusions

The Turkish 4DSQ is a valid questionnaire to assess distress, depression, anxiety and somatization, provided that the cut-off points for distress and depression be adjusted.

Acknowledgement
Mehmet Akman, MD. MPh

Table 1. Cronbach's alpha values, mean 4DSQ-scores and standard deviations of the study groups

| | | Cronbach's alpha | | | | Mean scores (SD) | | |
|---|---|---|---|---|---|---|---|---|
| 4DSQ-scales | scale range | Turkish | Dutch | p | | Turkish | Dutch | p |
| Distress | 0-32 | 0.90 | 0.93 | 0.002 | | 12.6 (8.1) | 19.9 (9.0) | .000 |
| Depression | 0-12 | 0.86 | 0.92 | 0.001 | | 3.0 (3.4) | 4.4 (4.2) | .000 |
| Anxiety | 0-24 | 0.84 | 0.92 | 0.000 | | 5.2 (5.1) | 7.2 (6.9) | .000 |
| Somatization | 0-32 | 0.86 | 0.89 | 0.035 | | 11.1 (7.0) | 15.3 (8.2) | .000 |

Tabel 2. Results of the confirmative factor analysis (CFA)

| 4DKL-scales | Chi-square | df | p | CFI | TLI | RMSEA | 90% CI |
|---|---|---|---|---|---|---|---|
| Distress[a] | 442.41 | 204 | 0.000 | 0.994 | 0.993 | 0.058 | 0.050-0.065 |
| Depression[b] | 30.99 | 16 | 0.014 | 0.999 | 0.999 | 0.052 | 0.023-0.079 |
| Anxiety | 236.93 | 108 | 0.000 | 0.993 | 0.991 | 0.058 | 0.048-0.068 |
| Somatization[c] | 356.90 | 200 | 0.000 | 0.989 | 0.987 | 0.047 | 0.039-0.055 |

Chi-square = mean and variance adjusted model chi-square statistic

CFI = comparative fit index

TLI = Tucker-Lewis index

RMSEA = root mean square error of approximation

90% CI = 90% confidence interval RMSEA

[a] residual correlation between item pairs #20-#39 (sleep problems), and #47-#48 (items referring to (an) upsetting event(s))

[b] residual correlation between item pairs #33-#46 (suicidal ideation)

[c] residual correlation between item pair #15-#16 (thoracic symptoms), and item triplet #9-#12-#13 (gastro-intestinal symptoms)

Table 3. Items identified as having differential item functioning (DIF)

| Scale / item # | English description | Turkish description | M-H-method | | | HOLR-method | |
|---|---|---|---|---|---|---|---|
| Distress | | | direction | SMD | | direction | $\Delta R^2$ |
| | During the past week, did you suffer from: | Geçtiğimiz hafta aşağıdaki belirtilerden şikayetiniz oldu mu? | | | | | |
| # 17 | feeling down or depressed? | keyifsizlik / isteksizlik | + | 0.40 | | + | 3.87 |
| # 19 | worry? | birşeyleri kafaya takıp durmak | - | -0.16 | | | |
| # 20 | disturbed sleep? | huzursuz uyuma | - | -0.18 | | | |
| # 22 | lack of energy? | bitkinlik | + | 0.23 | | | |
| | During the past week, did you feel: | Geçtiğimiz hafta aşağıdaki duyguları yaşadınız mı? | | | | | |
| # 26 | easily irritated? | çarçabuk asabileşmek | - | -0.32 | | - | 5.36 |
| | During the past week, did you: | Geçtiğimiz hafta aşağıdaki durumları hissettiniz mi? | | | | | |
| # 37 | no longer feel like doing anything? | artık içinizden hiç bir şey yapmak gelmediğini / hiç bir şeyden zevk almadığınızı | + | 0.19 | | | |
| | During the past week: | Geçen hafta: | | | | | |
| # 41 | did you easily become emotional? | Çabuk duygusallaştığınız oldu mu? | + | 0.31 | | + | 2.60 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| # 47 | did you ever have fleeting images of any upsetting event(s) that you have experienced? | Birden, daha önce başınızdan geçmiş ağır bir olayın görüntüleri veya izleri zihninize (aklınıza) doluştu mu? | + | 0.38 | | + | 6.47 |
| # 48 | did you ever have to do your best to put aside thoughts about any upsetting event(s)? | Daha önce başınızdan geçmiş ağır bir olayı zihninizden uzaklaştırmak (aklınızdan çıkarmak) için olağanüstü bir çaba harcamak zorunda kaldınız mı? | + | 0.26 | | + | 3.45 |
| Depression | | | | | | | |
| | During the past week, did you feel: | Geçtiğimiz hafta aşağıdaki duyguları yaşadınız mı? | | | | | |
| # 28 | that everything is meaningless? | herşeyin manasız olduğunu | - | -0.15 | | | |
| # 34 | that you can't enjoy anything anymore? | hiç bir şeyden zevk almadığınızı | - | -0.18 | | | |
| # 35 | that there is no escape from your situation? | hiç bir çıkış yolunuzun kalmadığını | - | -0.29 | | - | 8.49 |
| | During the past week: | Geçen hafta: | | | | | |
| # 46 | did you ever think "I wish I was dead"? | Keşke ölsem dediğiniz oldu mu? | | | | + | 2.55 |
| Anxiety | | | | | | | |
| | During the past week, did you suffer from: | Geçtiğimiz hafta aşağıdaki belirtilerden şikayetiniz oldu mu? | | | | | |
| # 21 | a vague feeling of fear? | sebepsiz / yersiz korkular | - | -0.25 | | | |
| # 23 | trembling when with other people? | başkalarının yanında sıkılma / bunalma | + | 0.44 | | + | 10.46 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | During the past week, did you feel: | Geçtiğimiz hafta aşağıdaki duyguları yaşadınız mı? | | | | | |
| # 27 | frightened? | korku içinde olma | - | -0.28 | | - | 3.58 |
| | During the past week: | Geçen hafta: | | | | | |
| # 49 | did you have to avoid certain places because they frightened you? | Korktuğunuz için belirli yerlerden geçmemek / oralarda bulunmamak için çaba harcadınız mı? | + | 0.19 | | + | 3.84 |
| Somatization | | | | | | | |
| | During the past week, did you suffer from: | Geçtiğimiz hafta aşağıdaki belirtilerden şikayetiniz oldu mu? | | | | | |
| # 1 | dizziness or feeling light-headed? | baş dönmesi veya kafanızda bir hafiflik hissi | - | -0.35 | | - | 2.49 |
| # 6 | excessive sweating? | aşırı terleme | - | -0.17 | | | |
| # 11 | shortness of breath? | bunaltı | + | 0.19 | | | |
| # 14 | tingling in the fingers? | parmaklarda karıncalanma | + | 0.16 | | | |

direction: - = more severe for Turkish patients, + = less severe for Turkish patients; SMD = standardized mean difference; $\Delta R^2$ = difference in $R^2$ value (x100)

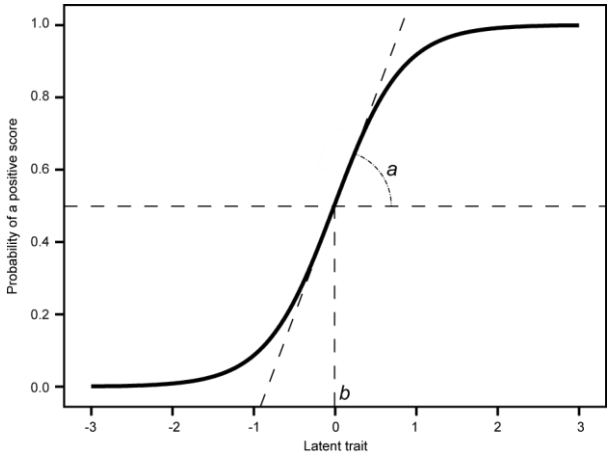


Figure 1. Item characteristic curve (ICC) of a dichotomous item. The curve displays the probability of a positive response to the item as a function of the underlying trait. The 'severity' of the item (*b*) is the level of the latent trait where the probability of a positive response is 50%. The 'discrimination' of the item is the slope of the curve (*a*). The latent trait scale is an arbitrary scale.

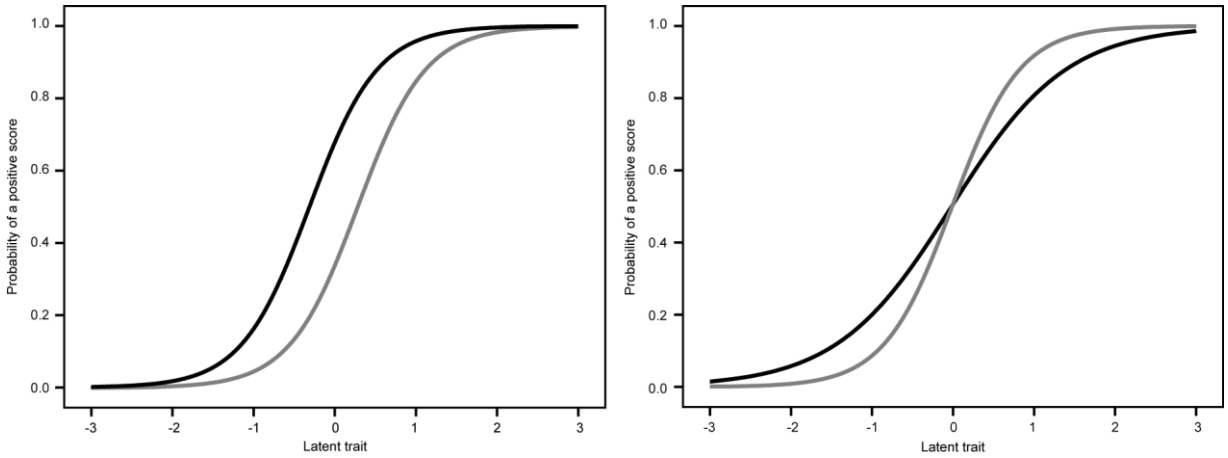


Figure 2. Examples of uniform DIF (left panel) and non-uniform DIF (right panel). Both figures display the ICC's of one item in two groups in which the item differs in 'severity' (left panel) or in 'discrimination' (right panel). In uniform DIF (left panel) group A (black ICC) has a uniformly higher (or equal) probability of a positive score than group B (grey ICC). In non-uniform DIF (right panel) group A (black ICC) has a higher probability of a positive score than group B (grey ICC) in the lower part of the latent trait scale while, reversely, group B has a higher probability of a positive score than group A in the higher part of the scale.

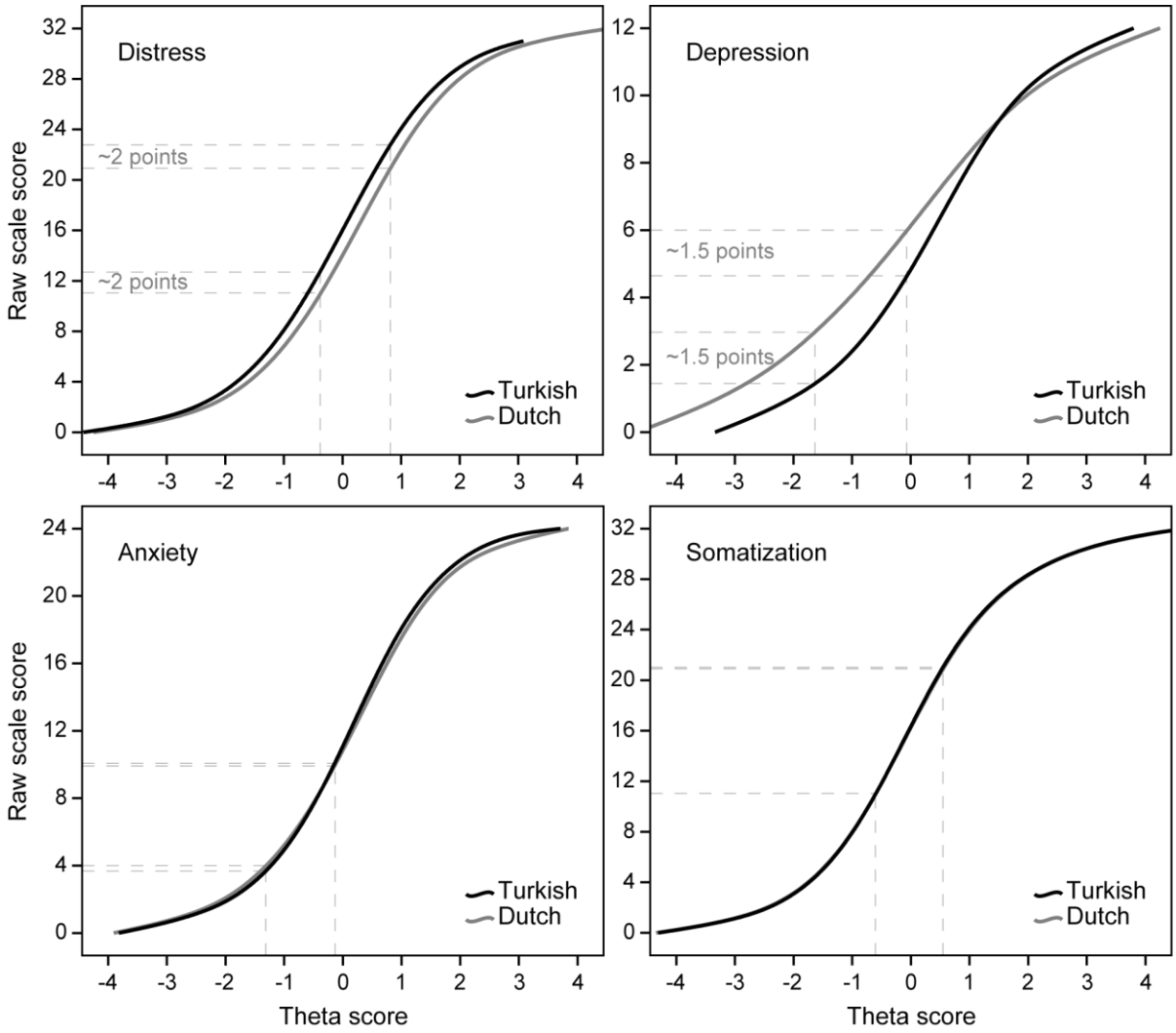


Figure 3. Raw 4DSQ scale scores as a function of the DIF-adjusted theta scores for distress, depression, anxiety and somatization, by language group (solid black curves: Turkish, solid grey curves: Dutch). Conventional Dutch cut-off points and corresponding Turkish cut-off points are indicated by dashed grey lines. The vertical distance between the Dutch and corresponding Turkish cut-off points indicate differential test functioning.

Reference List

(1) Terluin B, Terluin M, Prince K, Van Marwijk HWJ. De Vierdimensionale Klachtenlijst (4DKL) spoort psychische problemen op [The Four-Dimensional Symptom Questionnaire (4DSQ) detects psychological problems (English translation available on: http://www.emgo.nl/researchtools/4DSQ-cme-article.pdf)]. Huisarts Wet 2008;51:251-5.

(2) Terluin B. De Vierdimensionale Klachtenlijst (4DKL). Een vragenlijst voor het meten van distress, depressie, angst en somatisatie [The Four-Dimensional Symptom Questionnaire (4DSQ). A questionnaire to measure distress, depression, anxiety, and somatization]. Huisarts Wet 1996;39:538-47.

(3) 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.

(4) Terluin B, Brouwers EPM, Van Marwijk HWJ, Verhaak PFM, van der Horst HE. Detecting depressive and anxiety disorders in distressed patients in primary care; comparative diagnostic accuracy of the Four-Dimensional Symptom Questionnaire (4DSQ) and the Hospital Anxiety and Depression Scale (HADS). BMC Fam Pract 2009;10:58.

(5) Terluin B, Oosterbaan DB, Brouwers EPM, van Straten A, van de Ven PM, Langerak W, et al. To what extent does the anxiety scale of the Four-Dimensional Symptom Questionnaire (4DSQ) detect specific types of anxiety disorder in primary care? A psychometric study. BMC Psychiatry 2014;14:121.

(6) Borsboom D. When does measurement invariance matter? Med Care 2006;44(Suppl 3):S176-S181.

(7) Acquadro C, Conway K, Giroudet C, Mear I. Linguistic validation manual for patient-reported outcomes (PRO) instruments. Lyon: MAPI Research Institute; 2004.

(8) Zumbo BD. A Handbook on the theory and methods of differential item functioning (DIF): logistic regression modeling as a unitary framework for binary and likert-type (ordinal) item scores. Ottawa ON: Directorate of Human Resources Research and Evaluation, Department of National Defense; 1999.

(9) Teresi JA. Different approaches to differential item functioning in health applications. Advantages, disadvantages and some neglected topics. Med Care 2006;44(11 Suppl 3):S152-S170.

(10) 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;12:373-85.

(11) van Ginkel JR, van der Ark LA. SPSS syntax for missing value imputation in test and questionnaire data. Applied Psychological Measurement 2005;29:152-3.

(12) Revelle W. Package 'psych' (available on: http://personality-project.org/r/psych.manual.pdf). 2013.

(13) R Development Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2012.

(14) Rosseel Y. lavaan: an R package for structural equation modeling. Journal of Statistical Software 2012;48:2.

(15) Hu L, Bentler PM. Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. Structural Equation Modeling 1999;6:1-55.

(16) Hambleton RK. Good practices for identifying differential item functioning. Med Care 2006;44(11 Suppl 3):S182-S188.

(17) Michaelides MP. An illustration of a Mantel-Haenszel procedure to flag misbehaving common items in test equating. Practical Assessment, Research & Evaluation 2008;13:7.

(18) 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.

(19) Dorans NJ, Schmitt AP, Bleistein CA. The standardization approach to assessing comprehensive differential item functioning. Journal of Educational Measurement 1992;29:309-19.

(20) Gotzmann A, Bahry LM. Review of 'jMetrik'. Research & Practice in Assessment 2012;7:56-8.

(21) Bond TG, Fox CM. Applying the Rasch model: fundamental measurement in the human sciences. Second edition. New York: Routledge; 2007.

(22) Bentler PM. Comparative fit indices in structural equation models. Psychological Bulletin 1990;107:238-46.

(23) Borra R. Depressive disorder among Turkish women in the Netherlands: a qualitative study of idioms of distress. Transcult Psychiatry 2011;48:660-74.

(24) Schrier AC, de Wit MAS, Rijmen F, Tuinebreijer WC, Verhoeff AP, Kupka RW, et al. Similarity in depressive symptom profile in a population-based study of migrants in the Netherlands. Soc Psychiatry Psychiatr Epidemiol 2010;45:941-51.

(25) Ulusahin A, Basoglu M, Paykel ES. A cross-cultural comparative study of depressive symptoms in British and Turkish clinical samples. Soc Psychiatry Psychiatr Epidemiol 1994;29:31-9.