Translation and cross-cultural validation of the Turkish, Moroccan Arabic and Moroccan Berber versions of the 48-item Symptom Questionnaire (SQ-48)

Kovács, V.; Carlier, I.V.E.; Zitman, F.G.; Hemert, A.M. van; Giltay, E.J.

RESEARCH ARTICLE

WILEY

Translation and cross-cultural validation of the Turkish, Moroccan Arabic and Moroccan Berber versions of the 48-item Symptom Questionnaire (SQ-48)

Viktória Kovács [iD] | **Ingrid V. E. Carlier** [iD] | **Frans G. Zitman**[†] |
Albert M. van Hemert | **Erik J. Giltay**

Department of Psychiatry, Leiden University Medical Centre, Leiden, The Netherlands

Correspondence
Viktória Kovács, Department of Psychiatry, Leiden University Medical Centre, Albinusdreef 2, 2300 RC Leiden, The Netherlands.
Email: vituskovacs@gmail.com

Funding information
Rivierduinen

Abstract

We aimed to validate cross-culturally the Turkish, Moroccan Arabic and Moroccan Berber versions of the 48-item Symptom Questionnaire (SQ-48). Its psychometric properties were assessed in four samples: patients ($n = 150$) and controls ($n = 103$) with Turkish or Moroccan origins ($n = 103$) and patients ($n = 189$) and controls ($n = 463$) with native Dutch origins. Internal consistency and discriminatory power of SQ-48 subscales across groups were adequate to high. However, immigrant groups scored on average higher than Dutch native groups, but there was full configural, metric and partial scalar invariance in the immigrant groups. Although the SQ-48 is a valid measure of psychopathology in immigrant groups of Turkish and Moroccan origins, their cut-off values should likely be higher compared to natives.

KEYWORDS
cross-cultural validity, Moroccan, psychopathology, Routine Outcome Monitoring (ROM), Symptom Questionnaire-48 (SQ-48), Turkish

1 | INTRODUCTION

Moroccan and Turkish people are currently the two largest groups of immigrants in the Netherlands (Harmsen et al., 2018). Several studies have shown that especially anxiety and depression were more prevalent among Turkish and Moroccan immigrants compared to native Dutch samples (Bengi-Arslan et al., 2002; de Wit et al., 2008; Ikram et al., 2014; Schrier et al., 2010; Schrier et al., 2017; van der Wurff et al., 2004). The question remains whether these higher prevalence rates show a real difference in the presence of symptoms, or whether this difference reflects different cultural interpretation of symptoms in self-report questionnaires (Galenkamp et al., 2017).

The first generation of Moroccan and Turkish immigrants generally have great difficulty understanding the Dutch language and there is substantial illiteracy, in particular among Berber-speaking Moroccan subgroups (Dagevos et al., 2003; Hoopman et al., 2009). Because of this, immigrant people who are not fluent in the native language of that country are often excluded from studies, especially in mental health care research (Streiner & Norman, 2003). This results in a shortage of research data on mental health in ethnic minorities. Because of the ethnic variations in psychopathology, this disparity should be urgently addressed.

Consequently, to promote culturally sensitive mental health care, it is essential to translate, culturally adapt and validate mental health instruments for people with diverse cultural and linguistic backgrounds (Sidani et al., 2010). For instance, in order to reach the

[†]Deceased.

Moroccan population in the Netherlands as widely as possible, it is necessary that questionnaires are available in Moroccan Arabic and Moroccan Berber. Developing these questionnaires is complicated by the fact that both languages are mainly oral and do not exist in written form. The latter is by no means uncommon, as about half of the world's approximately 7000 languages have no written form (Lewis & Fenning, 2015). The questionnaires must also be made available in Turkish, because a large part of the first-generation Turkish migrants are illiterate or low literate. It is therefore essential that we create a way that allows the illiterate and people without written language to complete questionnaires without being able to read. One way to do this is to develop a website on which questionnaires are made available with audiovisual support.

Currently, a well-translated and culturally adapted instrument for measuring psychopathology is lacking for the Turkish, Moroccan Arabic and Moroccan Berber populations in the Netherlands. Such an instrument is important, because it is often used by default in mental health care for the purpose of screening for psychopathology and/or Routine Outcome Monitoring (ROM). This study aims to address this by focusing on the cross-validation of three translations of the Dutch 48-item Symptom Questionnaire (SQ-48), a generic self-report questionnaire on psychopathology (Carlier et al., 2012, 2017, 2019). It can be used in clinical settings as well as for research or benchmarking purposes. It is available in the public domain, free for everyone to use. SQ-48 was developed in 2012 and has become one of the most widely used tools in Dutch mental health care and related research.

The aims of this study were (1) to translate and cross-culturally adapt the SQ-48 into Turkish, Moroccan Arabic and Moroccan Berber; (2) to develop a special website on which questionnaires like SQ-48 are made available with audiovisual support (see Section 2); (3) to examine psychometric properties; and (4) to examine measurement invariance across immigrant and Dutch native groups, in order to demonstrate that SQ-48 can be interpreted in the same way for different ethnic groups. We hypothesized that the seven-factor psychopathology structure of SQ-48 would be confirmed across the immigrant groups and that at least configural invariance would be found.

2 | METHODS

2.1 | Participants and procedure

This Dutch multicentre research project was an extension of the NormQuest project, which is part of the Leiden Routine Outcome Monitoring Study from the Department of Psychiatry of Leiden University Medical Centre (LUMC). The NormQuest project aimed to validate a set of questionnaires for ROM in the Netherlands (Schulte-van Maaren et al., 2012, 2013). Initially, NormQuest was limited to Dutch native speakers. The current study is an extension for the most frequent non-native migrant languages in the Netherlands, that is, Turkish, Moroccan Arabic and Moroccan Berber. Our study

> **Key Practitioner Message**
> - The Turkish, Moroccan Arabic and Moroccan Berber versions of the SQ-48 have satisfactory psychometric properties.
> - The AUC values showed that the subscales and total score of the SQ-48 performed acceptable to excellent in discriminating between psychiatric and non-psychiatric participants within each immigrant and Dutch native group.
> - The seven-factor psychopathology structure of SQ-48 was supported across the immigrant and Dutch native groups, implying that the same concepts were measured.
> - Full configural, metric and partial scalar invariance was obtained across the immigrant and Dutch native groups, which means that the SQ-48 is suitable for cross-cultural comparisons across groups, although some noninvariant items should be interpreted with caution.
> - Immigrant groups scored on average higher on all SQ-48 subscales compared to their native counterparts.

sample had participants from four subsamples: (1) psychiatric outpatients with Turkish or Moroccan background; (2) non-psychiatric subjects with Turkish or Moroccan background; (3) native Dutch psychiatric outpatients; and (4) native Dutch non-psychiatric subjects.

2.1.1 | Turkish and Moroccan groups

The Turkish-, Moroccan Arabic- and Moroccan Berber-speaking participants were recruited from psychiatric outpatient clinics and non-psychiatric settings. Subjects received a 20-euro gift voucher for participating.

Psychiatric outpatients group
This sample consisted of 90 Turkish-, 31 Moroccan Arabic- and 29 Moroccan Berber-speaking outpatients. Participants were recruited between October 2014 and December 2015 from eight Dutch mental health care providers: i-psy The Hague, i-psy Amsterdam, Riagg Rijnmond Schiedam & Rotterdam, FortaGroep, GGZ Rivierduinen, Esens GGZ, GGZ Divers and the Cirya Foundation. The inclusion criteria were (1) presence of at least moderate mood, anxiety or somatoform symptoms; (2) being born in Turkey or Morocco or having at least one parent or grandparent born in Turkey or Morocco; (3) being between 18 and 65 years old; and (4) being proficient in the Turkish, Moroccan Arabic or Moroccan Berber language. Patients with severe psychiatric problems or with suicidal ideations were excluded. Clinical diagnoses were determined by psychiatrists or clinical psychologists.

Non-psychiatric reference group

This sample consisted of 42 Turkish-, 42 Moroccan Arabic- and 19 Moroccan Berber-speaking healthy participants who were recruited between September 2015 and December 2015 via contact persons from Turkish and Moroccan associations, mosques and community centres. The same inclusion criteria were used as for the outpatient sample with an additional criterion that participants had no mood, anxiety or somatoform symptoms.

2.1.2 | Native Dutch groups

This sample was derived from the NormQuest research project of LUMC. This project has been described in detail elsewhere (Schulte-van Maaren et al., 2013). Shortly, the NormQuest project collected ROM data from Dutch native psychiatric outpatients and non-psychiatric participants.

Psychiatric outpatient group

This sample consisted of 189 native Dutch psychiatric outpatients (age range 18–65 years) from the Dutch Regional Mental Health Provider GGZ Rivierduinen (RD) and the Department of Psychiatry of LUMC. They were referred for mood, anxiety or somatoform disorders between October 2010 and February 2011. Excluded patients were (1) those who suffered from hearing impairment or limited cognitive abilities and (2) those who had insufficient knowledge of the Dutch language or were illiterate. For a description of characteristics of this sample, see Table 1.

All patients referred for were enrolled in ROM. The procedure for the web-based ROM program is described elsewhere in detail (see Carlier et al., 2012; de Beurs et al., 2011; van Noorden et al., 2010). ROM is part of the standard diagnostic and treatment procedure and includes an optional battery of observer-rated measures, self-report questionnaires and a standardized diagnostic interview (MINI International Neuropsychiatric Interview Plus, MINI Plus 5.0.0) (Sheehan et al., 1998; van Vliet & de Beurs, 2007). Measurements were scheduled before, during and after treatment. For this study, we only used baseline data of the SQ-48.

Non-psychiatric reference group

This sample consisted of 463 native Dutch participants who fully completed the SQ-48 and whose parents were born in the Netherlands. Participants (18–65 years) were recruited between November 2009 and January 2011 from eight Dutch university-affiliated general practices in the vicinity of Leiden. Randomly selected subjects who were registered in the general practices were asked to participate. ROM measures were done as described above. Participants received travel allowance and a 30-euro gift voucher (Schulte-van Maaren et al., 2013). The same eligibility criteria were used as for the outpatient group sample with an additional criterion that participants have not been treated in a secondary psychiatric care centre in the last 6 months. For a description of the characteristics of the sample, see Table 1.

2.2 | Measurement and translations

SQ-48 is a 48-item self-report questionnaire that takes an average of 6 min to complete (Carlier et al., 2012). It has nine subscales, of which seven measure different psychopathology constructs (i.e., aggression/hostility, agoraphobia, fear, cognitive complaints, depression, somatic complaints and social phobia) and two additional subscales relating to vitality/optimism and work/study. The latter two scales were not included in the present study because these subscales do not measure psychopathology and tend to cluster rather separately from the negative affect subscales.

Items are rated on a 5-point Likert scale ranging from 0 (*never*) to 4 (*very often*). The subscales are scored by summing the ratings (range from 0 to 148), with a higher score indicating higher levels of psychological stress. The SQ-48 has good internal consistency as well as good convergent and divergent validity among native clinical and non-clinical samples (Carlier et al., 2012). The SQ-48 also has excellent test–retest reliability and good responsivity to therapeutic change. The SQ-48 is a useful tool for screening and monitoring in clinical setting and a useful tool for research purposes (Carlier et al., 2017). The Dutch SQ-48 was translated into English according to the guidelines for translation and cultural adaptation of questionnaires (Carlier et al., 2012; Wild et al., 2005).

2.2.1 | Translation and cross-cultural adaption of the SQ-48

The three translated SQ-48 versions (Turkish, Moroccan Arabic and Moroccan Berber; see Supporting Information S1–S3) were made in a five-step process according to established guidelines (e.g., Beaton et al., 2000; Guillemin et al., 1993; Kortmann, 2005; van Widenfelt et al., 2005; Wild et al., 2005). The five steps were discussed as follows (see also Figure 1).

Step 1: Forward translation

Two independent translators, whose native language is the target language, translated the Dutch questionnaire into the target language. One of these two translators was a medical student who had relevant knowledge about psychiatry. The other was a professional translator, without clinical knowledge. This resulted in two translated versions per language, encompassing both the nuances of clinical and spoken language. This step went without major problems for the different languages, with one exception: the Moroccan Berber version. The Moroccan Berber language has no official script. Therefore, some items were formulated very differently according to which city the translator came from (Al Hoceima or Nador). One translator used the Tifinaghe-Latin-Ircam transcription, and the other translator used letters and numbers from Arabic to indicate how to pronounce certain items. Moroccan Arabic is also an oral language, but for the Moroccan Arabic version, it was possible to translate the questionnaire into standard Arabic.

TABLE 1 Sociodemographic characteristics of psychiatric outpatient and non-psychiatric reference groups.

|	Turkish (n = 132)		Moroccan Arabic (n = 73)		Moroccan Berber (n = 48)		Dutch (n = 652)	
	Patient (n = 90)	Reference (n = 42)	Patient (n = 31)	Reference (n = 42)	Patient (n = 29)	Reference (n = 19)	Patient (n = 189)	Reference (n = 463)
Age (in years, SD)	44.4 (9.5)	27.2 (8.4)	45.0 (8.6)	32.9 (12.9)	42.8 (9.6)	28.6 (12.6)	38.3 (13.1)	39.62 (12.6)
Age range	18–65	18–48	23–64	18–57	25–63	19–53	18–65	18–64
Gender (female) (n, %)	62 (68.9)	28 (66.7)	15 (48.4)	25 (59.5)	14 (48.3)	10 (52.6)	117 (61.9)	304 (65.7)
Marital status (n, %)								
Married/living with partner	39 (43.3)	12 (28.6)	17 (54.8)	17 (40.5)	17 (58.6)	7 (36.8)	92 (48.7)	319 (68.9)
Divorced/widowed	31 (34.4)	2 (4.8)	13 (41.9)	8 (19.0)	8 (27.6)	1 (5.3)	21 (11.1)	23 (5.0)
Unmarried	20 (22.3)	28 (66.6)	1 (3.3)	17 (40.5)	4 (13.8)	11 (57.9)	76 (40.2)	121 (26.1)
Housing situation (n, %)								
Living alone	22 (24.4)	6 (14.3)	7 (22.6)	3 (7.1)	6 (20.7)	1 (5.2)	61 (32.3)	85 (18.4)
Living with partner	35 (38.9)	12 (28.6)	17 (54.8)	17 (40.5)	18 (62.1)	6 (31.6)	94 (49.7)	322 (69.5)
Living with family	33 (36.7)	24 (57.1)	7 (22.6)	22 (52.4)	5 (17.2)	12 (63.2)	34 (18.0)	56 (12.1)
Educational status (n, %)								
None or low	40 (44.4)	0 (0.0)	9 (29.0)	3 (7.1)	19 (65.5)	3 (15.8)	11 (5.8)	18 (3.9)
Intermediate	42 (46.7)	23 (54.8)	14 (45.2)	27 (64.3)	4 (13.8)	12 (63.2)	121 (64.0)	266 (57.5)
High	8 (8.9)	19 (45.2)	8 (25.8)	12 (28.6)	6 (20.7)	4 (21.1)	57 (30.2)	179 (38.7)
Employment status (n, %)								
Working	13 (14.4)	26 (61.9)	7 (22.6)	24 (57.1)	6 (20.7)	9 (47.4)	84 (44.4)	376 (81.2)
Non-working	77 (85.6)	16 (38.1)	24 (77.4)	18 (42.9)	23 (79.3)	10 (52.6)	105 (55.6)	87 (18.8)
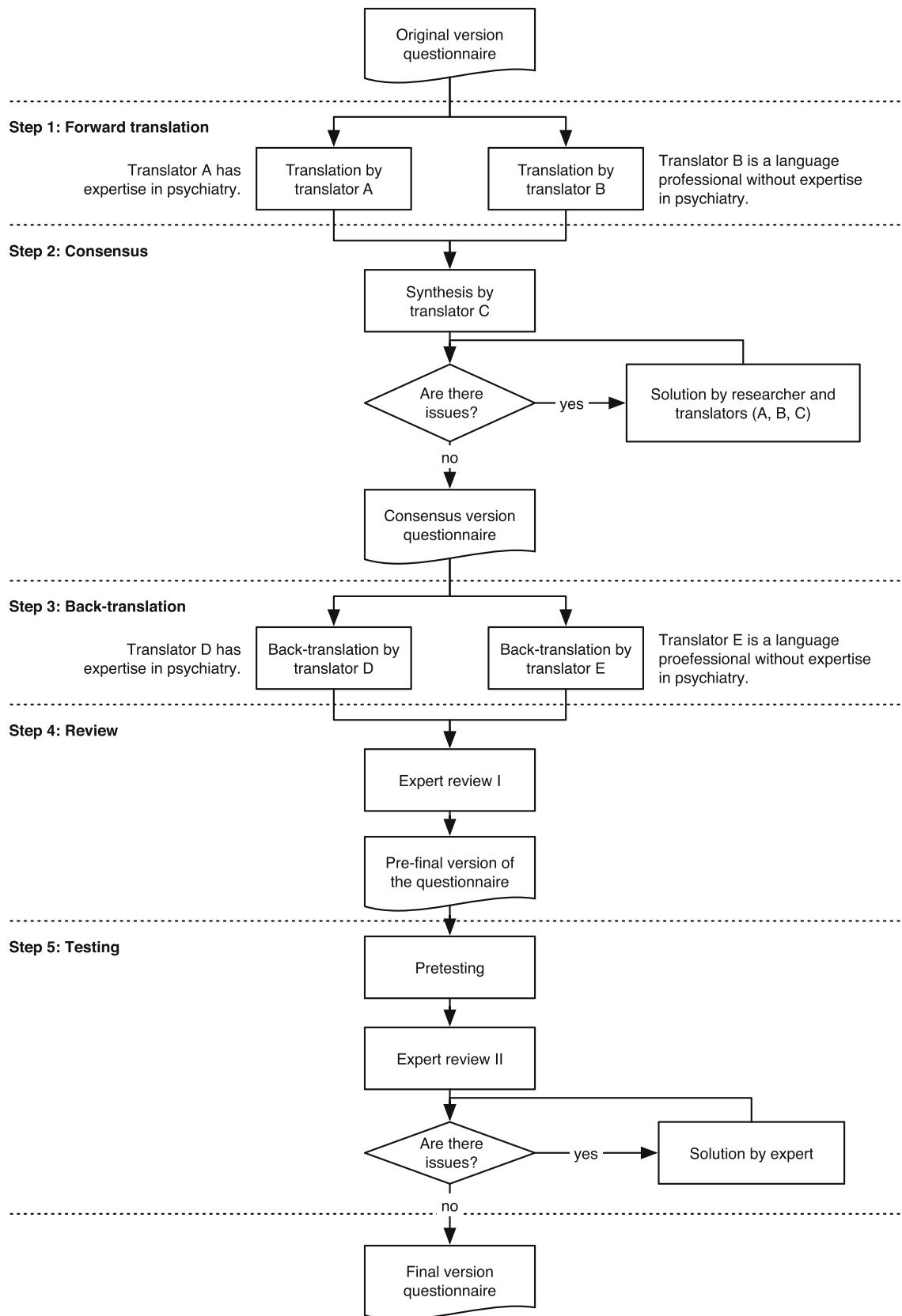
FIGURE 1 Flowchart translation and cross-cultural adaption process. A visual representation of the guidelines from Guillemin et al. (1993), Beaton et al. (2000), Kortmann (2005), van Widenfelt et al. (2005) and Wild et al. (2005).

Step 2: Consensus version

A third translator was responsible for comparing the two translated versions and the original Dutch version. A meeting was organized with the three translators to discuss any uncertainties and differences and to produce a consensus translation of the questionnaire. Although making a consensus version led to considerable discussions for the Moroccan Berber version, it was possible to create a consensus version that took into account the different dialects. For the Moroccan Arabic version, notes were added to each item about how to pronounce them in Moroccan Arabic.

Step 3: Back translation

The consensus translation was subsequently translated back into Dutch by two new translators. This resulted in two independently back-translated versions per language. Similar to the translators in Step 1, one of the translators had relevant clinical knowledge, and the other did not. Both translators were not allowed to see the original version of the questionnaire before or during their translation.

Step 4: Expert review

A meeting was held with several stakeholders: a researcher (VK), health care professionals (bilingual psychiatrist or psychologist), translators and language experts. During the expert meeting, there were a number of items that required a lot of discussion. During the Turkish expert meeting, Items 3, 12, 26, 28 and 48 created problems. They were not clear enough or the experts had problems finding the correct translation, to make sure the items would be understandable to those without education. The Moroccan Arabic expert meeting went more smoothly. The experts only had problems with Item 12. Most problems occurred during the expert meeting of the Berber version. Some items were difficult to translate, because, for those items, there are no words in the Berber language. This was the case for Items 40 and 47. In these cases, we created a description: Item 40 'I felt down or depressed' was eventually translated as 'It feels like I have had a black life'; and Item 47 'I could not concentrate well' was translated as 'Whatever I do, my mind is absent'.

During this expert meeting, a preliminary version of the questionnaire was compiled for testing within a small sample (Step 5). In addition, a list of dubious items was created: These items caused discussion, were unclear or could be culturally irrelevant. The questionnaire was read aloud and recorded for the Moroccan Arabic and Moroccan Berber versions.

Step 5: Pilot study/testing

The preliminary version was presented to healthy first-generation Turkish and Moroccan immigrants (five participants per language). After completing the questionnaire, participants were asked how they experienced it and which items they found difficult or unclear. When the participants assessed an item as unclear, they were asked to formulate the item in such a way that the meaning of it was clear for them. Also, the list of dubious items was discussed to determine if the items expressed the meaning as intended. Eventually, the bilingual professional who participated in the expert team processed all the input from the pilot study and compiled the final version. After this, the translated versions in three languages were ready to be tested, as described in the current paper.

2.2.2 | Website with audiovisual support

Native Dutch participants were selected for literacy and completed the questionnaire in a written format, either digital on a computer screen or with pencil and paper.

As mentioned, the Moroccan Arabic and Moroccan Berber languages are spoken languages without written form. In addition, illiteracy is quite common among the first-generation Turkish and Moroccan immigrants in the Netherlands. Therefore, we developed a website on which the questionnaire is presented as audio, with a visual aid to facilitate scoring, referred to as the '*Glass method*'. Items are read aloud, and the participants can click on the glass that is filled to a certain level to indicate to what extent they experience this particular item. The Glass method has been developed by the Dutch Leiden University and was further developed by the Psychiatry Department of LUMC. In addition to reading aloud, items were written out and highlighted when pronounced. Participants could click on the written and highlighted answer categories as well, instead of the glass. Even for the low-literate people, this was the preferred interaction. Figure 2 is a representation of an item from the Moroccan Arabic SQ-48 in this website.

2.3 | Statistical analysis

We have categorized the previously described samples into eight groups: (1) Turkish psychiatric; (2) Turkish non-psychiatric; (3) Moroccan Arabic psychiatric; (4) Moroccan Arabic non-psychiatric; (5) Moroccan Berber psychiatric; (6) Moroccan Berber non-psychiatric; (7) Dutch psychiatric; and (8) Dutch non-psychiatric. Descriptive statistics on outcome variables are given for each of these eight groups, including internal consistency and receiver operating characteristic (ROC) analyses.

Baseline categorical variables are presented as number (percentage), and continuous variables are presented as mean (±SD, with range). Differences in sociodemographic and clinical characteristics between groups were analysed using analysis of variance (ANOVA) for the continuous variables and chi-square test for the categorical variables.

The internal consistency for the SQ-48 subscales and total score was calculated with Cronbach's alpha (Cronbach, 1951), of which values of 0.7 or higher indicate acceptable internal consistency within each group (Nunnally & Bernstein, 1994).

The areas under the ROC curves (AUCs) were calculated to indicate the discriminative capacities of the SQ-48 subscales. It is a measure of accuracy and has a range from 0.5 to 1.0. AUC was calculated with ROC analysis of the combined data of the psychiatric and non-psychiatric groups. Values between 0.5 and 0.7 represent poor

FIGURE 2 Representation of the Glass method.

accuracy, between 0.7 and 0.8 indicate acceptable accuracy, between 0.8 and 0.9 indicate good accuracy and over 0.9 indicate excellent accuracy (Agresti, 2003). These statistical analyses were performed with IBM SPSS 23.0 (SPSS Inc., Chicago, Illinois).

We used multi-group confirmatory factor analysis (CFA) to evaluate the measurement invariance of SQ-48 across immigrant and Dutch native groups in RStudio (Version 3.2.4; Foundation for Statistical Computing, Vienna, Austria, 2016; https://www.R-project.org/), with the packages 'lavaan' (Version 0.5-23.1097) and 'semTools' (Version 3.2.4).

To analyse measurement invariance, we reduced the eight groups to three groups, to avoid small sample sizes: (1) combined Turkish psychiatric and non-psychiatric; (2) combined Moroccan Arabic and Moroccan Berber psychiatric and non-psychiatric; and (3) combined Dutch psychiatric and non-psychiatric.

All variables were treated as continuous. Rhemtulla et al. (2012) suggested that treating Likert-type scale variables as continuous is acceptable when the number of response categories is five or more. Because the distribution of our continuous data was not normal, we chose the robust maximum likelihood (ML) fitting function. The robust ML estimation was provided using the Satorra–Bentler scaled chi-square statistics ($S-B\chi^2$; Satorra & Bentler, 1988). This estimator can better handle non-normality and smaller sample sizes (Chou & Bentler, 1995; Satorra & Bentler, 1994).

We used several indices for evaluating the model fit in addition to $S-B\chi^2$: the comparative fit index (CFI), the Tucker–Lewis index (TLI), the standardized root mean square residual (SRMR) and the root mean square error of approximation (RMSEA) and its 90% confidence interval (CI; Hu & Bentler, 1999). Due to unequal and small sample sizes among the immigrant groups, the fit indices RMSEA and CFI had the most added value when evaluating the model fit. To evaluate the fit indices, the following cut-off criteria have been chosen: (1) for CFI and TLI, values of >0.9 indicate good model fit, 0.8–0.9 indicate acceptable fit and <0.8 indicate poor fit; (2) for RMSEA, values of <0.05 indicate close fit, 0.05–0.08 indicate

adequate fit and >0.10 indicate poor fit; and (3) for SRMR, values of <0.08 indicate acceptable fit (Browne & Cudeck, 1992; Hu & Bentler, 1999).

The multi-group CFA procedure for testing measurement invariance was developed by Meredith (1993). During this procedure, a series of analyses were performed to evaluate the increasingly restrictive levels of measurement invariance. The first step is to find a baseline model for each group. The second step is to see if the factor structure is the same for each group, which represents *configural invariance*. The third step is *metric/weak invariance*, to test whether factor loadings of the same items across groups are the same, which shows that the construct has the same meaning across groups. The fourth step is to test whether the intercepts of each item are the same across groups, which is *scalar/strong invariance*. The fifth/last step is to test whether the error variances for the variables are the same for all groups, which is *strict invariance* (Little, 2013).

After each step, a chi-square difference test was performed, to see whether the model has been improved or worsened over the previous model. If the $\Delta S-B\chi^2$ value is significant, it suggests that the constraints in the more restrictive model do not hold and, therefore, the two models are not equivalent across groups. However, the use of chi-square test is somewhat problematic because chi square is sensitive to sample sizes. Therefore, in addition to the Satorra–Bentler chi-square test for assessing the fit of the models, we used the change in CFI (ΔCFI). Cheung and Rensvold (2002) suggested a ΔCFI greater than 0.01 to be indicative of non-equivalence. The other fit indices (RMSEA and SRMR overlap in 90% of RMSEA's CIs) are also evaluated (Cheung & Rensvold, 2002).

Finally, *partial invariance* was examined if no full measurement invariance could be achieved at a particular step described above. In order to test partial measurement invariance, the sources of misspecified items are first identified by modification indices, and then they are allowed to differ between groups (Byrne et al., 1989; Steenkamp & Baumgartner, 1998; Vandenberg & Lance, 2000).

3 | RESULTS

3.1 | Sociodemographic and clinical characteristics of the samples

The characteristics of the eight subsamples are presented in Table 1. Significant age differences were detected in the immigrant groups as compared to the native Dutch group. Turkish and Moroccan Arabic psychiatric outpatients tended to be older than native Dutch psychiatric outpatients, whereas Turkish and Moroccan Berber non-psychiatric participants tended to be younger than native Dutch non-psychiatric participants. There were no significant gender differences among groups.

3.2 | Internal consistency

The internal consistency results and the optimal cut-off values of the native Dutch group are presented in Table 2. Cronbach's alpha coefficients of SQ-48 subscales ranged from 0.72 to 0.98 across groups, indicating overall adequate to high internal consistency. All subscales met or exceeded the 0.70 level recommended for group comparison. The highest values in each group were obtained in the Anxiety subscale (Turkish $\alpha = 0.92$, Moroccan Arabic $\alpha = 0.92$, Moroccan Berber $\alpha = 0.93$ and native Dutch $\alpha = 0.92$), and the lowest values were obtained in the Aggression subscale (Turkish $\alpha = 0.83$, Moroccan Arabic $\alpha = 0.72$, Moroccan Berber $\alpha = 0.89$ and native Dutch $\alpha = 0.79$).

3.3 | Percentiles, mean values and standard deviations

The mean scores (M); standard deviations (SDs); the 25th, 50th and 75th percentile scores; and the AUC values for the different subsamples are presented in Table 3 and Figure 3. The results show that immigrant groups, both psychiatric and non-psychiatric, scored on average higher on almost every subscale, compared to their Dutch counterpart group.

The AUC values showed that the subscales and total score of SQ-48 performed acceptable to excellent in discriminating between psychiatric and non-psychiatric participants (AUC values ranging from 0.71 to 0.97) in each group. For the whole sample, the best performing subscales were 'Cognitive complaints' and 'Depression', and the least performing subscale was 'Aggression' with still acceptable accuracy.

3.4 | Testing measurement invariance

The results for the multi-group models tested across groups are presented in Table 4. In each case, we compared one immigrant group with the native Dutch group: Turkish versus native Dutch; and combined Moroccan Arabic and Moroccan Berber versus native Dutch.

Before comparing the groups, we ran CFAs in order to assess the SQ-48 structure on each of the samples separately. All the indices in the three samples met the criteria. The results showed a moderate to adequate fit for all samples (Turkish: S–Bχ^2(608) = 909.8; RMSEA = 0.061; SRMR = 0.055; CFI = 0.931; TLI = 0.924; Moroccan Arabic and Moroccan Berber: S–Bχ^2(608) = 850.2; RMSEA = 0.057; SRMR = 0.055; CFI = 0.936; TLI = 0.930; Dutch: S–Bχ^2(608) = 1472.7; RMSEA = 0.047; SRMR = 0.052; CFI = 0.891; TLI = 0.880).

We then proceeded with the testing of measurement invariance.

3.4.1 | Turkish versus Dutch

First, for the two groups together, we found the following fit data for the overall model: S–Bχ^2(608) = 1660.1; RMSEA = 0.058; SRMR = 0.046; CFI = 0.935; TLI = 0.929. The results showed adequate fit with the data.

Second, a configural invariance model was evaluated. When the same factor structure was specified for both groups simultaneously, the results indicated a good overall fit, supporting an equivalent factor structure among the Turkish and Dutch groups: S–Bχ^2(1216) = 2456.6; RMSEA = 0.060; SRMR = 0.052; CFI = 0.921; TLI = 0.914.

TABLE 2 Internal consistency in combined psychiatric outpatient and non-psychiatric reference groups.

		Cronbach's alpha			
	Number of items	Turkish $n = 132$	Moroccan Arabic $n = 73$	Moroccan Berber $n = 48$	Dutch $n = 652$
Aggression	4	0.83	0.72	0.89	0.79
Agoraphobia	4	0.90	0.89	0.85	0.83
Anxiety	6	0.92	0.92	0.93	0.92
Cognitive complaints	5	0.83	0.91	0.86	0.88
Depression	6	0.90	0.92	0.89	0.93
Somatic complaints	7	0.91	0.91	0.92	0.89
Social phobia	5	0.91	0.92	0.90	0.91
Total	37	0.96	0.98	0.98	0.97

TABLE 3 Percentiles, mean values and standard deviations of the SQ-48 in four samples.

	Non-psychiatric reference group					Psychiatric outpatient group				
	n	P25	P50 (median)	P75	Mean ± SD	AUC	*n*	P25	P50 (median)	P75	Mean ± SD
Aggression											
Turkish	42	2	4	6	4.1 ± 3.1	0.71	90	3	7	11	7.3 ± 4.4
Moroccan Arabic	42	0.8	2	5	2.8 ± 2.8	0.87	31	5	8	9	7.4 ± 2.7
Moroccan Berber	19	0	1	4	2.6 ± 3.4	0.84	29	4.5	9	12	8.3 ± 4.5
Dutch	463	0	1	2	1.2 ± 1.7	0.73	189	1	3	6	3.6 ± 3.4
Agoraphobia											
Turkish	42	0	1	2	1.0 ± 1.3	0.91	90	4	7.7	10.3	7.5 ± 4.7
Moroccan Arabic	42	0	1	2	1.3 ± 1.6	0.94	31	6	8	11	8.0 ± 4.0
Moroccan Berber	19	0	1	2	1.8 ± 3.1	0.77	29	1	6	9	6.0 ± 4.7
Dutch	463	0	0	0	0.4 ± 1.1	0.73	189	0	1	4	2.5 ± 3.4
Anxiety											
Turkish	42	3	4.5	11	6.6 ± 5.0	0.88	90	12	17	20	15.8 ± 5.6
Moroccan Arabic	42	2	6	9	5.9 ± 4.2	0.96	31	15	17	19	16.5 ± 3.8
Moroccan Berber	19	2	4	7	5.1 ± 4.9	0.84	29	9.5	16	19	14.0 ± 6.6
Dutch	463	1	3	6	3.7 ± 3.9	0.87	189	7.5	13	17	12.1 ± 5.9
Cognitive complaints											
Turkish	42	5	7	8.3	7.1 ± 3.4	0.90	90	10.8	13	17	13.6 ± 3.9
Moroccan Arabic	42	3	4.5	7.3	5.0 ± 2.8	0.97	31	11	14	17	14.0 ± 3.9
Moroccan Berber	19	2	5	8	5.4 ± 3.7	0.86	29	9	12	15	11.6 ± 4.4
Dutch	463	1	3	6	4.1 ± 3.4	0.89	189	8	12	15	11.2 ± 4.6
Depression											
Turkish	42	1	4.5	7	4.5 ± 4.1	0.91	90	10	14	18	13.9 ± 5.4
Moroccan Arabic	42	1	2.5	6	3.6 ± 3.5	0.93	31	8	15	19	13.6 ± 5.5
Moroccan Berber	19	2	3	5	4.4 ± 5.1	0.85	29	7	11	14.5	11.2 ± 5.6
Dutch	463	0	1	3	2.1 ± 2.5	0.90	189	6	12	16	11.2 ± 6.4
Somatic complaints											
Turkish	42	0.8	2	7	3.6 ± 3.9	0.93	90	10	15	19	14.5 ± 6.2
Moroccan Arabic	42	1	3	6.3	4.0 ± 3.6	0.93	31	11	14	18	14.3 ± 5.9
Moroccan Berber	19	0	2	6	4.4 ± 5.5	0.80	29	5	12	16	11.7 ± 7.0
Dutch	463	0	0	2	1.7 ± 3.1	0.74	189	1	3	10	6.3 ± 6.3
Social phobia											
Turkish	42	1	3	5.3	3.7 ± 3.3	0.85	90	6	11	16.3	10.9 ± 5.8
Moroccan Arabic	42	0	2	6	3.0 ± 3.3	0.90	31	6	14	16	11.4 ± 5.3
Moroccan Berber	19	1	3	5	3.6 ± 3.8	0.82	29	4	10	14.5	10.1 ± 5.7
Dutch	463	0	1	4	2.4 ± 3.0	0.83	189	4	7	11.5	7.8 ± 4.9
SQ-48 total											
Turkish	42	16.5	25	44.5	30.7 ± 20.1	0.91	90	59.8	83	106.5	83.6 ± 29.8
Moroccan Arabic	42	10.5	25.5	37.3	25.7 ± 17.5	0.96	31	63	88	105	85.3 ± 25.5
Moroccan Berber	19	11	19	40	27.4 ± 26.8	0.86	29	53	78	93	72.9 ± 33.2
Dutch	463	5	11	22	15.6 ± 14.4	0.89	189	36.5	54	70.5	54.8 ± 26.0

Abbreviations: AUC, area under the receiver operating characteristic curve; SQ-48, 48-item Symptom Questionnaire.

Third, we tested the equivalence of the factor loadings for the Turkish and Dutch groups. The global fit indices suggested adequate fit of the metric invariance model: S−Bχ²(1246) = 2581.9; RMSEA = 0.061; SRMR = 0.065; CFI = 0.916; TLI = 0.910. The model had good fit indices, but the Satorra–Bentler chi-square test was significant. The CFI and the RMSEA are better fit indices to
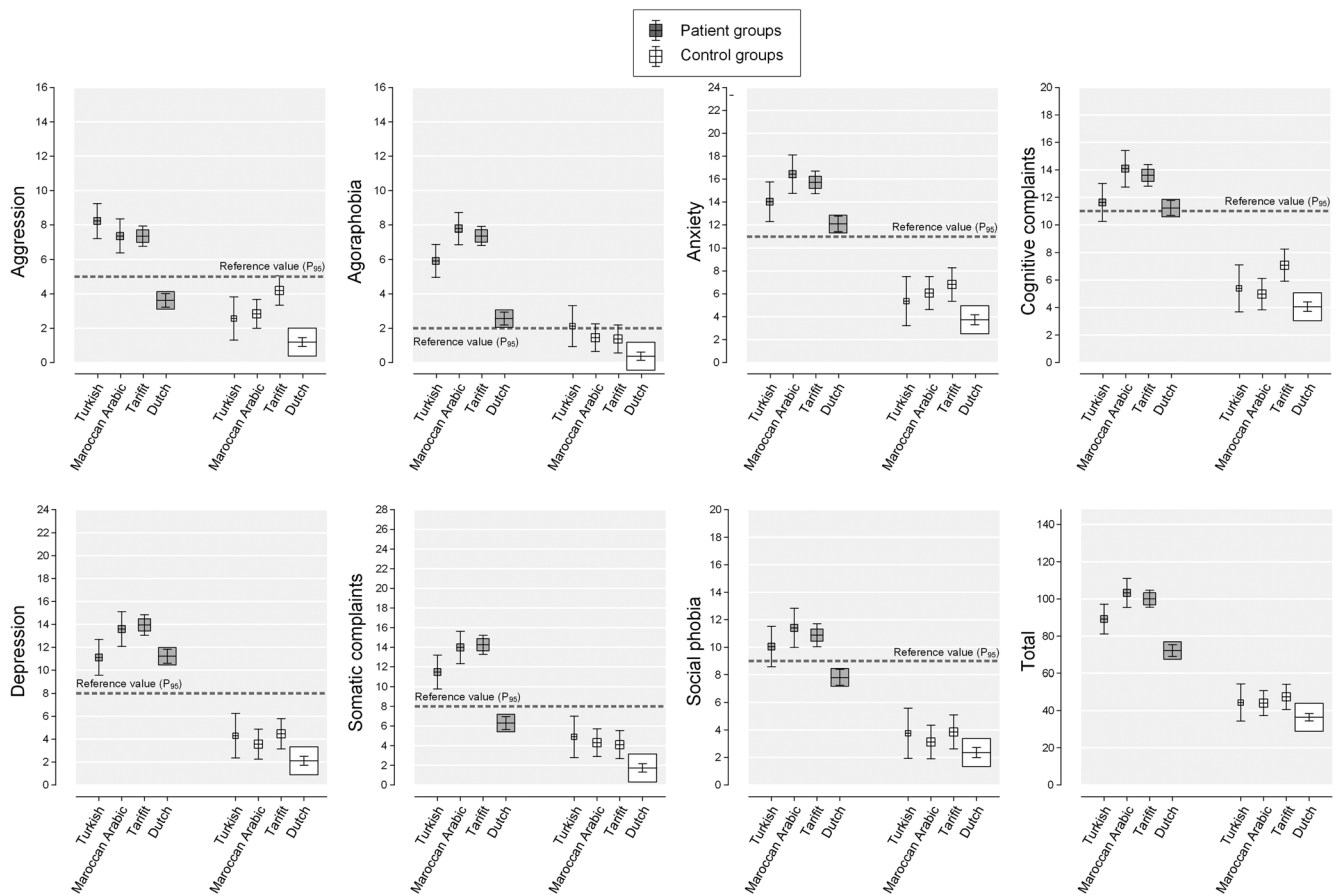
FIGURE 3 Distribution of the scores between the four samples.

evaluate invariance between the models, because the chi-square test is more sensitive for small sample sizes. The ΔCFI between metric and configural models supported metric invariance (ΔCFI = 0.005). This is less than the cut-off criteria of ΔCFI. These results support a weak measurement equivalence across ethnic groups.

Fourth, we examined the equivalence of the intercepts for the Turkish and Dutch groups. Our analyses failed to support the intercept equivalence, suggesting that one or more parameters were not equivalent across groups (ΔCFI = 0.024). The examination of the modification indices indicated that Item 24 ('I was afraid or anxious'), Item 3 ('I was irritable and dissatisfied'), Item 26 ('I was shaking or trembling'), Item 28 ('I was scared') and Item 44 ('I struggled to get the day started') may be noninvariant across groups. When the intercepts of Items 24, 3, 26, 28 and 44 were freely estimated, partial scalar invariance could be established: $S–B\chi^2(1271) = 2817.3$; RMSEA = 0.064; SRMR = 0.066; CFI = 0.906; TLI = 0.901. It resulted in a ΔCFI of 0.01. The observed difference in item means between ethnic groups is due to factor mean differences only, except for Items 24, 3, 26, 28 and 44. Participants in the Turkish group scored systematically higher for these items, compared to participants in the native Dutch group.

Fifth, we tested whether the error variances are the same for the Turkish and native Dutch groups. The decrease in model fit from the partial scalar invariance model to the partial strict model was not significant: $S–B\chi^2(1308) = 2336.6$; RMSEA = 0.065; SRMR = 0.079; CFI = 0.900; TLI = 0.898. The ΔCFI between partial scalar and partial strict models supported partial strict invariance (ΔCFI = 0.006).

3.4.2 | Moroccan Arabic and Moroccan Berber versus Dutch

For the two ethnic groups together, we found the following fit data for the overall model: $S–B\chi^2(608) = 1651.6$; RMSEA = 0.059; SRMR = 0.046; CFI = 0.928; TLI = 0.921. The results showed adequate fit with the data.

The multi-group model testing configural invariance [$S–B\chi^2(1216) = 2363.7$; RMSEA = 0.060; SRMR = 0.052; CFI = 0.922; TLI = 0.915] fits the data well, as did the metric invariance model [$S–B\chi^2(1246) = 2454.5$; RMSEA = 0.060; SRMR = 0.063; CFI = 0.919; TLI = 0.913]. Although the Satorra–Bentler chi-square difference test between the configural invariance and metric invariance models was statistically significant, $\Delta S–B\chi^2(30) = 90.8$, $p = 0.000$, the other fit indices were good and the metric invariance model did not result in a significantly worse fit (ΔCFI = 0.003). The metric invariance between the ethnic groups was attained.

TABLE 4 Tests of measurement invariance of the SQ-48 across ethnic groups.

Model and constraints	df	S–Bχ²	RMSEA [90% CI]	SRMR	CFI	TLI	Δdf	ΔS–Bχ²	ΔCFI	ΔRMSEA
Single-group solutions										
Turkish (n = 132)	608	909.8	0.061 [0.054, 0.069]	0.055	0.931	0.924	NA	NA	NA	NA
Moroccan Arabic and Berber (n = 121)	608	850.2	0.057 [0.049, 0.065]	0.055	0.936	0.930	NA	NA	NA	NA
Dutch (n = 652)	608	1472.7	0.047 [0.044, 0.049]	0.052	0.891	0.880	NA	NA	NA	NA
Measurement invariance										
Dutch (n = 652) versus Turkish (n = 132)										
Overall	608	1660.1	0.058 [0.055, 0.062]	0.046	0.935	0.929	NA	NA	NA	NA
Configural invariance	1216	2456.6	0.060 [0.056, 0.063]	0.052	0.921	0.914	NA	NA	NA	NA
Metric invariance	1246	2581.9	0.061 [0.058, 0.064]	0.065	0.916	0.910	30	166.4*	0.005	0.001
Scalar invariance	1276	3097.3	0.068 [0.065, 0.071]	0.069	0.892	0.888	30	193.1	0.024	0.007
Partial scalar invariance with Items 24, 3, 26, 28 and 44	1271	2817.3	0.064 [0.061, 0.067]	0.066	0.906	0.901	25	305.6	0.010	0.003
Partial strict invariance	1308	2336.6	0.065 [0.061, 0.069]	0.079	0.900	0.898	37	41.1	0.006	0.001
Dutch (n = 652) versus Moroccan Arabic and Berber (n = 121)										
Overall	608	1651.6	0.059 [0.056, 0.063]	0.046	0.928	0.921	NA	NA	NA	NA
Configural invariance	1216	2363.7	0.060 [0.056, 0.063]	0.052	0.922	0.915	NA	NA	NA	NA
Metric invariance	1246	2454.5	0.060 [0.057, 0.064]	0.063	0.919	0.913	30	90.8*	0.003	0.000
Scalar invariance	1276	2746.9	0.064 [0.061, 0.067]	0.066	0.906	0.902	30	162.8	0.013	0.004
Partial scalar invariance with Items 28 and 8	1274	2668.4	0.063 [0.059, 0.066]	0.065	0.910	0.905	28	185.0	0.009	0.003
Partial strict invariance	1311	2145.9	0.060 [0.055, 0.064]	0.073	0.915	0.914	37	31.1	0.005	0.003

Abbreviations: ΔCFI, change in comparative fit index; Δdf, degree of freedom for scaled (Satorra–Bentler) chi-square difference test; ΔRMSEA, change in RMSEA; ΔS–Bχ², scaled (Satorra–Bentler) chi-square difference test statistic; CFI, comparative fit index; CI, confidence interval; Configural invariance, equal form; df, degrees of freedom; Metric invariance, equal factor loadings; NA, not applicable; RMSEA, root mean square error of approximation; S–Bχ², Satorra–Bentler chi square; Scalar invariance, equal indicator thresholds; SQ-48, 48-item Symptom Questionnaire; SRMR, standardized root mean square residual; Strict invariance, equal indicator error variances; TLI, Tucker–Lewis index.
*$p \leq 0.05$.

Next, the model for scalar invariance was run. This model had acceptable fit [S–Bχ²(1276) = 2746.9; RMSEA = 0.064; SRMR = 0.066; CFI = 0.906; TLI = 0.902], but the ΔCFI was larger than 0.01, indicating that full scalar invariance was not met (ΔCFI = 0.013). Allowing the intercepts for Item 28 ('I was scared') and Item 8 ('I did not dare to cross open spaces, such as a public square') to vary across the groups, as indicated by the modification indices, produced a model with adequate fit [S–Bχ²(1276) = 2668.4; RMSEA = 0.063; SRMR = 0.065; CFI = 0.910; TLI = 0.905] and resulted in a ΔCFI smaller than 0.01, in order to support partial scalar invariance. Participants from the Moroccan Arabic and Moroccan Berber group scored systematically higher compared to participants in the native Dutch group for these items.

Lastly, we tested whether the error variances are the same for the Moroccan Arabic, Moroccan Berber and native Dutch groups. The decrease in model fit from the partial scalar invariance model to the partial strict model was not significant: S–Bχ²(1311) = 2145.9; RMSEA = 0.060; SRMR = 0.073; CFI = 0.915; TLI = 0.914. The ΔCFI between partial scalar and partial strict models supported partial strict invariance (ΔCFI = 0.005).

4 | DISCUSSION

4.1 | Main findings

We have translated and validated the SQ-48 into three languages according to established guidelines: Turkish, Moroccan Arabic and Moroccan Berber. Internal consistency of SQ-48 subscales across groups was adequate to high. The seven-factor psychopathology structure of SQ-48 fitted the data adequately in the total sample and within each subgroup. The AUC values showed acceptable to excellent discrimination between psychiatric and non-psychiatric participants within each group. Furthermore, we established full configural and metric invariance and partial scalar invariance across the groups. Finally, the Turkish, Moroccan Arabic and Moroccan Berber psychiatric and non-psychiatric groups scored higher on average on almost every subscale compared to their Dutch counterparts. Thus, when Dutch cut-off values (Carlier et al., 2012) are used in these immigrant groups, this will lead to extra false positives.

Our results supported only partial scalar invariance for the Turkish and Moroccan groups. According to Chen (2008), intercept

noninvariance could occur due to (a) social desirability reasons or social norms; (b) particular groups displaying a propensity to respond more strongly to an item despite having the same latent trait or factor mean; and (c) certain groups having different reference points when making statements about themselves. The question remains whether the different scores between immigrant groups and native Dutch group are a measurement bias or whether immigrant groups actually score higher on the specific items, because they truly experience more complaints/problems, for reasons we discuss underneath. An acceptable degree of partial invariance is less than 20% freed parameters (Byrne et al., 1989; Crockett et al., 2005; Dimitrov, 2010). We can conclude that there is acceptable partial invariance, with invariance of all but two intercepts by the combined Moroccan Arabic and Moroccan Berber group and all but five intercepts by the Turkish group. Overall, the SQ-48 is suitable for cross-cultural comparisons across these immigrant groups, but the noninvariant items should be cautiously interpreted.

Compared to native Dutch, severity scores were on average high in immigrant groups. This suggests that the Dutch cut-off values are not directly applicable to the immigrant groups. Future research with larger sample sizes of the immigrant groups could provide more clarity on this subject.

Our results are in line with previous findings that immigrants tend to score higher on mental health questionnaires compared to the native group. Deisenhammer et al. (2012) showed, for example, that Turkish patients (living in Turkey) and migrated Turkish patients (living in Austria) scored higher on the Beck Depression Inventory (BDI), the Montgomery–Asberg Depression Rating Scale (MADRS) and the Bradford Somatic Inventory (BSI) compared to Austrian patients. Terluin et al. (2016) showed that Turkish patients tend to score higher on the distress scale and lower on the depression scale, compared to Dutch patients on the Turkish Four-Dimensional Symptom Questionnaire (4DSQ). As a result, the researchers adjusted the cut-off points of the scales for the Turkish version of 4DSQ. In the study by Cicero (2016), ethnic groups tended to have higher scores compared to White participants on the Schizotypal Personality Questionnaire. Wind et al. (2017) showed that Dutch patients systematically reported milder symptom severity on most items of the Harvard Trauma Questionnaire (HTQ) compared to refugee patients. Fassaert et al. (2009) showed that Turkish and Moroccan participants had significantly higher scores on the Kessler Psychological Distress Scale (K10) compared to ethnic Dutch participants. They determined higher cut-off scores for the ethnic groups than for ethnic Dutch participants.

The lack of full scalar invariance, found in our study, is consistent with the results of recent cross-cultural studies evaluating measurement invariance of mental health questionnaires. Cicero (2016) showed, for example, that the Schizotypal Personality Questionnaire had configural, metric and partial scalar invariance. Wind et al. (2017) showed that the HTQ and the Hopkins Symptom Checklist-25 (HSCL-25) have been interpreted in the same way by refugee patients as by native Dutch patients. They have found partial scalar invariance. Rasmussen et al. (2015) could not confirm metric and scalar measurement

invariance of HTQ in a multinational clinical sample. They do not recommend standard cut-off scores for posttraumatic stress disorder measures and concluded that there is possibly a difference in response style between ethnic groups.

The tendency that immigrant groups score higher on questionnaires than the native groups is clearly reflected in the various studies. A possible explanation is that immigrant groups actually experience more intensive complaints/problems as a result of migration compared to the native group. Migration to a new country entails a number of stressors, such as separation from family, loss of social contacts, acculturation stress, coping with the new norms and learning a new language (Bhugra, 2003; Bhugra & Minas, 2007). In our study, we included both first-generation and second-generation immigrants, which may reduce the negative effects of migration. The method of administering the questionnaire may also explain why immigrants score higher on the questionnaire. In our study, the immigrant groups were read the questionnaire with audiovisual support, while the native Dutch group read the questionnaire themselves.

4.2 | Strengths and limitations

Our study had the following strengths. First, we performed the translation and cross-cultural adaptation process according to high standards (e.g., Beaton et al., 2000; Guillemin et al., 1993; Kortmann, 2005; van Widenfelt et al., 2005; Wild et al., 2005). Second, we included both psychiatric and non-psychiatric participants. Third, our immigrant patients were collected from eight mental health providers across the Netherlands, which increased external validity. Fourth, our study is one of the few to provide evidence for partial strong measurement invariance of a mental health tool among ethnic minority groups.

Some limitations need to be taken into account when interpreting our findings. First, our study had a small sample size for Turkish, Moroccan Arabic and Moroccan Berber groups. It is recommended to replicate the multi-group CFA with a larger sample size and to test the Moroccan Arabic and Moroccan Berber versions separately. Second, our CFI and TLI values tended to fall in the acceptable rather than good range. Small sample sizes of our immigrant groups may have likely decreased the fit and statistical power of analyses, which adversely affected measurement invariance (Brown, 2006). Third, our study used the newly developed 'Glass method' in combination with text to speech, which seemed adequate for low-literate/illiterate people and languages without written form. Nevertheless, we recommend further research and development in this context regarding online methods for these target groups with special needs.

5 | CONCLUSION

We developed and validated the Turkish, Moroccan Arabic and Moroccan Berber versions of the SQ-48. These are now available in the public domain and can be freely downloaded for everyone to be

used. Psychopathology measured by the SQ-48 can largely be interpreted in the same way for patients from different ethnic backgrounds. After all, the translated SQ-48 versions possess the same seven-factor psychopathology structure as the original version, with equivalent pattern of factor loadings and acceptable partial invariant factor variances. So, we have created a well-translated and culturally adapted questionnaire for the Turkish, Moroccan Arabic and Moroccan Berber immigrant groups, which can be used as a screening tool or for monitoring treatment progress. This is a welcome addition to clinical practice as measuring instruments adapted for immigrant groups are scarce. Future research with larger sample sizes is recommended to confirm our results and develop norm values for these ethnic minority groups.

ACKNOWLEDGEMENTS

CONFLICT OF INTEREST STATEMENT

The authors declare no conflict of interest.

ORCID

Viktória Kovács https://orcid.org/0000-0002-4327-7961

Ingrid V. E. Carlier https://orcid.org/0000-0001-6237-1438

REFERENCES

Agresti, A. (2003). *Categorical data analysis*. John Wiley & Sons.

Beaton, D. E., Bombardier, C., Guillemin, F., & Ferraz, M. B. (2000). Guidelines for the process of cross-cultural adaptation of self-report measures. *Spine (Phila Pa 1976)*, *25*(24), 3186–3191. https://doi.org/10.1097/00007632-200012150-00014

Bengi-Arslan, L., Verhulst, F. C., & Crijnen, A. A. (2002). Prevalence and determinants of minor psychiatric disorder in Turkish immigrants living in the Netherlands. *Social Psychiatry and Psychiatric Epidemiology*, *37*(3), 118–124. https://doi.org/10.1007/s001270200003

Bhugra, D. (2003). Migration and depression. *Acta Psychiatrica Scandinavica*, *108*(Suppl. 418), 67–72. https://doi.org/10.1034/j.1600-0447.108.s418.14.x

Bhugra, D., & Minas, I. H. (2007). Mental health and global movement of people. *Lancet*, *370*, 1109–1111. https://doi.org/10.1016/S0140-6736(07)61249-5

Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. Guilford Press.

Browne, M. W., & Cudeck, R. (1992). Alternative ways of assessing model fit. *Sociological Methods & Research*, *21*(2), 230–258. https://doi.org/10.1177/0049124192021002005

Byrne, B. M., Shavelson, R. J., & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement in variance. *Psychological Bulletin*, *105*, 456–466. https://doi.org/10.1037/0033-2909.105.3.456

Carlier, I., Schulte-Van Maaren, Y., Wardenaar, K., Giltay, E., Van Noorden, M., Vergeer, P., & Zitman, F. (2012). Development and validation of the 48-item Symptom Questionnaire (SQ-48) in patients with depressive, anxiety and somatoform disorders. *Psychiatry Research*, *200*, 904–910. https://doi.org/10.1016/j.psychres.2012.07.035

Carlier, I. V. E., Kovács, V., van Noorden, M. S., van der Feltz-Cornelis, C., Mooij, N., Schulte- van Maaren, Y. W. M., van Hemert, A. M., Zitman, F. G., & Giltay, E. J. (2017). Evaluating the responsiveness to therapeutic change with routine outcome monitoring: A comparison of the Symptom Questionnaire-48 (SQ-48) with the Brief Symptom Inventory (BSI) and the Outcome Questionnaire-45 (OQ-45). *Clinical Psychology and Psychotherapy*, *24*, 61–71. https://doi.org/10.1037/0033-2909.105.3.456

Carlier, I. V. E., van Eeden, W. A., de Jong, K., Giltay, E. J., van Noorden, M. S., van der Feltz-Cornelis, C., Zitman, F. G., Kelderman, H., & van Hemert, A. M. (2019). Testing for response shift in treatment evaluation of change in self-reported psychopathology amongst secondary psychiatric care outpatients. *International Journal of Methods in Psychiatric Research*, *28*(3), 1–14, e1785. https://doi.org/10.1002/mpr.1785

Chen, F. F. (2008). What happens if we compare chopsticks with forks? The impact of making inappropriate comparions in cross-cultural research. *Journal of Personality and Social Psychology*, *95*(5), 1005–1018. https://doi.org.ezproxy.leidenuniv.nl:2048/10.1037/a0013193

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling*, *9*, 233–255. https://doi-org.ezproxy.leidenuniv.nl/10.1207/S15328007SEM0902_5

Chou, C.-P., & Bentler, P. M. (1995). Estimation and tests in structural equation modeling. In R. H. Hoyle (Ed.), *Structural equation modeling: Concepts, issues, and applications* (pp. 37–55). Sage.

Cicero, D. C. (2016). Measurement invariance of the Schizotypal Personality Questionnaire in Asian, Pacific Islander, White, and multiethnic populations. *Psychological Assessment*, *28*(4), 351–361. https://doi.org.ezproxy.leidenuniv.nl:2048/10.1037/pas0000180

Crockett, L. J., Randall, B. A., Shen, Y. L., Russell, S. T., & Driscoll, A. K. (2005). Measurement equivalence of the Center for Epidemiological Studies Depression Scale for Latino and Anglo adolescents: A national study. *Journal of Consulting and Clinical Psychology*, *73*(1), 47–58. https://doi.org.ezproxy.leidenuniv.nl:2048/10.1037/0022-006X.73.1.47

Cronbach, L. J. (1951). Coeffcient alpha and the internal structure of tests. *Psychometrika*, *16*, 297–334. https://doi.org/10.1007/BF02310555

Dagevos, J., Gijsberts, M., & van Praag, C. (2003). *Rapportage minderheden 2003. Onderwijs, arbeid en soicaal-culturele integratie*. Sociaal en Cultureel Planbureau.

De Beurs, E., Den Hollander-Gijsman, M. E., Van Rood, Y. R., Van der Wee, N. J., Giltay, E. J., Van Noorden, M. S., Van der Lem, R., Van Fenema, E., & Zitman, F. G. (2011). Routine outcome monitoring in the Netherlands: Practical experiences with a web-based strategy for the assessment of treatment outcome in clinical practice. *Clinical Psychology and Psychotherapy*, *18*, 1–12. https://doi-org.ezproxy.leidenuniv.nl/10.1002/cpp.696

De Wit, M. A. S., Tuinebreijer, W. C., Dekker, J., Beekman, A. T. F., Gorissen, W. H. M., Schrier, A. C., Penninx, B. W. J. H., Komproe, I. H., & Verhoeff, A. P. (2008). Depressive and anxiety disorders in different ethnic groups: A population based study among native Dutch, and Turkish, Moroccan and Surinamese migrants in Amsterdam. *Social Psychiatry and Psychiatric Epidemiology*, *43*, 905–912. https://doi-org.ezproxy.leidenuniv.nl/10.1007/s00127-008-0382-5

Deisenhammer, E. A., Coban-Basaran, M., Mantar, A., Prunnlechner, R., Kemmler, G., Alkin, T., & Hinterhuber, H. (2012). Ethnic and

migrational impact on the clinical manifestation of depression. *Social Psychiatry and Psychiatric Epidemiology*, *47*, 1121–1129. https://doi-org.ezproxy.leidenuniv.nl/10.1007/s00127-011-0417-1

Dimitrov, D. M. (2010). Testing for factorial invariance in the context of construct validation. *Measurement and Evaluation in Counseling and Development*, *43*, 121–149. https://doi-org.ezproxy.leidenuniv.nl/10.1007/s00127-009-0135-0

Fassaert, T., de Wit, M. A. S., Tuinebreijer, W. C., Wouters, H., Verhoeff, A. P., Beekman, A. T. F., & Dekker, J. (2009). Psychometric properties of an interviewer-administered version of the Kessler Psychological Distress scale (K10) among Dutch, Moroccan and Turkish respondents. *International Journal of Methods in Psychiatric Research*, *18*(3), 159–168. https://doi-org.ezproxy.leidenuniv.nl/10.1002/mpr.288

Galenkamp, H., Stronks, K., Snijder, M. B., & Derks, E. M. (2017). Measurement invariance testing of the PHQ-9 in a multi-ethnic population in Europe: The HELIUS study. *BMC Psychiatry*, *17*, 349. https://doi.org/10.1186/s12888-017-1506-9

Guillemin, F., Bombardier, C., & Beaton, D. (1993). Cross-cultural adaptation of health-related quality of life measures: Literature review and proposed guidelines. *Journal of Clinical Epidemiology*, *46*(12), 1417–1432. https://doi.org/10.1016/0895-4356(93)90142-N

Harmsen, C., de Heij, V., Kooiman, N., & van Roon, D. (2018). Bevolking. In *Jaarrapport Integratie 2018*. Centraal Bureau voor de Statistiek. Dutch article.

Hoopman, R., Terwee, C. B., Muller, M. J., Ory, F. G., & Aaronson, N. K. (2009). Methodological challenges in quality of life research among Turkish and Moroccan ethnic minority cancer patients: Translation, recruitment and ethical issues. *Ethnicity & Health*, *14*(3), 237–253. https://doi-org.ezproxy.leidenuniv.nl/10.1080/13557850802398832

Hu, L. T., & Bentler, P. M. (1999). Evaluating model fit. In R. H. Hoyle (Ed.), *Structural equation modeling: Concepts, issues, and applications* (pp. 76–99). Sage.

Ikram, U. Z., Snijder, M. B., Fassaert, T. J. L., Schene, A. H., Kunst, A. E., & Stronks, K. (2014). The contribution of perceived ethnic discrimination to the prevalence of depression. *European Journal of Public Health*, *25*(2), 243–248. https://doi-org.ezproxy.leidenuniv.nl/10.1093/eurpub/cku180

Kortmann, F. A. M. (2005). Meetinstrumenten in de transculturele psychiatrie. *Tijdschrift voor Psychiatrie*, *47*(11), 787–793.

Lewis, M. P., & Fenning, C. D. (Eds.). (2015). *Ethnologue: Languages of the world* (Eighteenth ed.). SIL International.

Little, T. D. (2013). *Longitudinal structural equation modelling*. Guilford Press.

Meredith, W. (1993). Measurement invariance, factor analysis, and factorial invariance. *Psychometrika*, *58*, 525–543. https://doi-org.ezproxy.leidenuniv.nl/10.1007/BF02294825

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory*. McGraw-Hill.

Rasmussen, A., Verkuilen, J., Ho, E., & Fan, Y. (2015). Posttraumatic stress disorder among refugees: Measurement invariance of Harvard Trauma Questionnaire scores across global regions and response patterns. *Psychological Assessment*, *27*(4), 1160–1170. https://doi.apa.org/doi/10.1037/pas0000115

Rhemtulla, M., Brosseau-Liard, P. É., & Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychological Methods*, *17*(3), 354–373. https://doi.org.ezproxy.leidenuniv.nl:2048/10.1037/a0029315

Satorra, A., & Bentler, P. M. (1988). Scaling corrections for chi-square statistics in covariance structure analysis. In *ASA 1988 Proceedings of the Business and Economic Statistics Section (308–313)*. American Statistical Association.

Satorra, A., & Bentler, P. M. (1994). Corrections to test statistics and standard errors in covariance structure analysis. In A. von Eye &

C. C. Clogg (Eds.), *Latent variables analysis: Applications for developmental research* (pp. 399–419). Sage.

Schrier, A. C., de Wit, M. A. S., Rijmen, F., Tunebreijer, W. C., Verhoeff, A. P., Kupka, R. W., Dekker, J., & Beekman, A. T. F. (2010). Similarity in depressive symptom profile in a population-based study of migrants in the Netherlands. *Social Psychiatry and Psychiatric Epidemiology*, *45*, 941–951. https://doi-org.ezproxy.leidenuniv.nl/10.1007/s00127-009-0135-0

Schrier, A. C., Hogerzeil, S. J., de Wit, M. A. S., & Beekman, A. T. F. (2017). Depression and anxiety in Turkish and Moroccan minorities in the Netherlands: Prevalence, symptoms, risk factors and protective factors. A systematic review. *Tijdschrift voor Psychiatrie*, *59*(1), 30–39. Dutch article.

Schulte-van Maaren, Y. W., Carlier, I. V., Giltay, E. J., van Noorden, M. S., de Waal, M. W., van der Wee, N. J., & Zitman, F. G. (2013). Reference values for mental health assessment instruments: Objectives and methods of the Leiden Routine Outcome Monitoring Study. *Journal of Evaluation in Clinical Practice*, *19*(2), 342–350. https://doi-org.ezproxy.leidenuniv.nl/10.1111/j.1365-2753.2012.01830.x

Schulte-van Maaren, Y. W., Carlier, I. V., Zitman, F. G., van Hemert, A. M., de Waal, M. W., van Noorden, M. S., & Giltay, E. J. (2012). Reference values for generic instruments used in routine outcome monitoring: The Leiden Routine Outcome Monitoring study. *BMC Psychiatry*, *12*, 203. https://doi.org/10.1186/1471-244X-12-203

Sheehan, D. V., Lecrubier, Y., Sheehan, K. H., Amorim, P., Janavs, J., Weiller, E., Hergueta, T., Baker, R., & Dunbar, G. C. (1998). The Mini-International Neuropsychiatric Interview (M.I.N.I.): The development and validation of a structured diagnostic psychiatric interview for DSM-IV and ICD-10. *The Journal of Clinical Psychiatry*, *59*(Supplement 20), 22–33.

Sidani, S., Guruge, S., Miranda, J., Ford-Gilboe, M., & Varcoe, C. (2010). Cultural adaptation and translation of measures: An integrated method. *Research in Nursing & Health*, *33*, 133–143. https://doi-org.ezproxy.leidenuniv.nl/10.1002/nur.20364

Steenkamp, J. B. E. M., & Baumgartner, H. (1998). Assessing measurement invariance in cross-national consumer research. *Journal of Consumer Research*, *25*, 78–107. https://doi-org.ezproxy.leidenuniv.nl/10.1086/209528

Streiner, D. L., & Norman, G. R. (2003). *Health measurement scales. A practical guide to their development and use*. Oxford University Press.

Terluin, B., Unalan, P. C., Sipahioglu, N. T., Ozkul, S. A., & van Marwijk, H. W. J. (2016). Cross-cultural validation of the Turkish Four-Dimensional Symptom Questionnaire (4DSQ) using differential item and test functioning (DIF and DTF) analysis. *BMC Family Practice*, *17*(53), 53. https://doi.org/10.1186/s12875-016-0449-4

Van der Wurff, F. B., Beekman, A. T. F., Dijkshoorn, H., Spijker, J. A., Smits, C. H. M., Stek, M. L., & Verhoeff, A. (2004). Prevalence and risk-factors for depression in elderly Turkish and Moroccan migrants in the Netherlands. *Journal of Affective Disorders*, *83*, 33–41. https://doi-org.ezproxy.leidenuniv.nl/10.1016/j.jad.2004.04.009

van Noorden, M. S., Giltay, E. S., Den Hollander-Gijsman, M. E., van der Wee, N. J., van Veen, T., & Zitman, F. G. (2010). Gender differences in clinical characteristics in a naturalistic sample of depressive outpatients: The Leiden Routine Outcome Monitoring Study. *Journal of Affective Disorders*, *125*(1–3), 116–123. https://doi-org.ezproxy.leidenuniv.nl/10.1016/j.jad.2009.12.007

van Vliet, I. M., & de Beurs, E. (2007). Het Mini Internationaal Neuropsychiatrisch Interview (MINI). Een kort gestructureerd diagnostisch psychiatrisch interview voor DSM-IV en ICD-10 stoornissen. *Tijdschrift voor Psychiatrie*, *49*(6), 393–397. Dutch article.

van Widenfelt, B. M., Treffers, P. D. A., de Beurs, E., Siebelink, B. M., & Koudijs, E. (2005). Translation and cross-cultural adaptation of assessment instruments used in psychological research with children and families. *Clinical Child and Family Psychology Review*, *8*(2), 135–147. https://doi-org.ezproxy.leidenuniv.nl/10.1007/s10567-005-4752-1

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, *3*, 4–70. https://journals-sagepub-com.ezproxy.leidenuniv.nl/doi/pdf/10.1177/109442810031002

Wild, D., Grove, A., Martin, M., Eremenco, S., McElroy, S., Verjee-Lorenz, A., & Erik-son, P. (2005). Principles of good practice for the translation and cultural adaptation process for Patient-Reported Outcomes (PRO) measures: Report of the ISPOR task force for translation and cultural adaptation. *Value in Health*, *8*(2), 94–104. https://doi-org.ezproxy.leidenuniv.nl/10.1111/j.1524-4733.2005.04054.x

Wind, T. R., van der Aa, N., de la Rie, S., & Knipscheer, J. (2017). The assessment of psychopathology among traumatized refugees: Measurement invariance of the Harvard Trauma Questionnaire and the Hopkins Symptom Checklist-25 across five linguistic groups. *European Journal of Psychotraumatology*, *8*(sup2), 1321357. https://doi.org/10.1080/20008198.2017.1321357

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.