



Why the factorial structure of the SCL-90-R is unstable: Comparing patient groups with different levels of psychological distress using Mokken Scale Analysis

Muirne C.S. Paap^{a,b,*}, Rob R. Meijer^c, Peggy T. Cohen-Kettenis^d, Hertha Richter-Appelt^e, Griet de Cuyper^f, Baudewijntje P.C. Kreukels^d, Geir Pedersen^g, Sigmund Karterud^b, Ulrik F. Malt^a, Ira R. Haraldsen^{a,*}

^a Department of Neuropsychiatry and Psychosomatic Medicine, Oslo University Hospital, Norway

^b Institute of Clinical Medicine, University of Oslo, Norway

^c Department of Psychometrics and Statistical Techniques, Faculty of Behavioural and Social Sciences, University of Groningen, The Netherlands

^d Department of Medical Psychology, VU University Medical Center, Amsterdam, The Netherlands

^e Institute for Sex Research and Forensic Psychiatry, University Hospital Hamburg-Eppendorf, Hamburg, Germany

^f Department of Sexology and Gender Problems, University Hospital Gent, Belgium

^g Department for Personality Psychiatry, Clinic for Mental Health and Addiction, Oslo University Hospital, Norway

ARTICLE INFO

Article history:

Received 3 August 2011

Received in revised form
7 March 2012

Accepted 9 March 2012

Keywords:

Item response theory

Validity

Personality disorder

Questionnaire evaluation

Gender identity disorder

Depression

ABSTRACT

Since its introduction, there has been a debate about the validity of the factorial structure of the SCL-90-R. In this study we investigate whether the lack of agreement with respect to the dimensionality can be partly explained by important variables that might differ between samples such as level of psychological distress, the variance of the SCL-90-R scores and sex. Three samples were included: a sample of severely psychiatrically disturbed patients ($n=3078$), a sample of persons with Gender Incongruence (GI; $n=410$) and a sample of depressed patients ($n=223$). A unidimensional pattern of findings were found for the GI sample. For the severely disturbed and depressed sample, a multidimensional pattern was found. In the depressed sample sex differences were found in dimensionality: we found a unidimensional pattern for the females, and a multidimensional one for the males. Our analyses suggest that previously reported conflicting findings with regard to the dimensional structure of the SCL-90-R may be due to at least two factors: (a) level of self-reported distress, and (b) sex. Subscale scores should be used with care in patient groups with low self-reported level of distress.

© 2012 Elsevier Ireland Ltd All rights reserved.

1. Introduction

The Symptom Checklist-90-Revised (SCL-90-R) (Derogatis, 1994) was designed to cover nine different dimensions of psychological distress; the mean item score across all 90 items with theoretical values ranging from 0 through 4 is referred to as the Global Severity Index (GSI), which is widely used as a global index for psychological distress. Since the introduction of the SCL-90 (-R), there has been a debate about the validity of the factorial structure, which was aptly expressed in the title of the paper 'Factor structure of the SCL-90-R: is there one?' (Cyr et al., 1985). More than two decades have passed since the publication of that

paper; however, the debate has still not abated, as recent publications have demonstrated (Olsen et al., 2004; Arrindell et al., 2006; Elliott et al., 2006; Hafkenscheid et al., 2007). On one hand, there is a group of researchers that firmly believe in the multidimensionality of the instrument (Arrindell et al., 2004b, 2004a, 2006), whereas another group has pointed out that alternative models with only one or at most a few factors show an equally good or better fit (Hafkenscheid, 2004; Hafkenscheid et al., 2007). In a recent paper, Paap et al. (2011b) proposed a new scale solution of 7 scales based on a study involving patients referred for a personality disorder (PD); scales were built on two start items that reflected the content of the disorder that corresponded with the specific scale. The new solution included 60 of the 90 items clustered in seven scales: Depression, Agoraphobia, Physical Complaints, Obsessive-Compulsive, Hostility (unchanged), Distrust and Psychoticism. The authors found that most of the new scales discriminated reliably between patients with moderately low scores to moderately high scores.

* Correspondence to: Clinic for Gender Identity Disorder, Department of Neuropsychiatry and Psychosomatic Medicine, Rikshospitalet, NO-0027 Oslo, Norway. Tel.: +47 23074160; fax: +47 23074170.

E-mail addresses: muirne@nxdomain.nl (M.C.S. Paap), i.h.haraldsen@medisin.uio.no (I.R. Haraldsen).

The items forming the GSI showed low scalability, and the authors concluded that their research findings lent support for a multidimensional model of the SCL-90-R. The authors speculated that the lack of agreement between studies might be due to several factors, such as difference in variance, the existence of structure generating factors, differences in the interpretation of the fit indices, and, finally, the chosen analytic strategy (Paap et al., 2011b).

In the current study, we investigate whether the findings in the study by Paap et al. can be generalized to other patient groups by comparing the dimensionality of the PD sample to that of a sample of persons with Gender Incongruence (GI) and a sample of depressed outpatients. The term 'GI' signifies the incongruence between one's gender identity on one hand, and one's assigned gender and/or one's congenital primary and secondary sex characteristics on the other hand (Kreukels et al., 2010; Meyer-Bahlburg, 2010).¹ Following Kreukels et al., we use GI when referring to patients who have not yet been diagnosed with GID (APA, 1994) or transsexualism (WHO, 1992). We expect the reported level of psychological distress (estimated by the GSI) to be lower in the GI sample than in the depressed sample and PD sample. Haraldsen and Dahl (2000) showed that patients diagnosed with GID had slightly elevated GSI scores when compared to healthy adults, but did not reach the value of 1.0 which is the cut-off for clinically significant symptoms ($GSI_{GID}=0.6$, $GSI_{controls}=0.4$). In contrast, depressed outpatients have been found to exceed the cut-off ($GSI_{DEP}=1.4$) (Leinonen and Niemi, 2007), and so have the patients in the PD sample used in the study by Paap et al. ($GSI_{PD}=1.5$). Our main research questions are:

- (1) Is the dimensionality of the SCL-90-R similar for patient groups that differ in level of reported psychological distress?
- (2) Are the different factorial solutions found in the literature due to a difference in variance in reported psychological distress?

Following Paap et al. (2011b) and Meijer et al. (2011), Mokken Scale Analysis (MSA; Mokken, 1971) was used to analyze the data. MSA is a nonparametric Item Response Theory (IRT) approach that can be used to explore and test hypotheses about the dimensionality of a data-set, while at the same time resulting in scales adhering to a measurement model.

2. Methods

2.1. Participants

2.1.1. Personality disorder sample: PD_{low} and PD_{high}

This sample consisted of 3078 patients admitted to 14 different day hospitals participating in the Norwegian Network of Personality-Focused Treatment Programs (Karterud et al., 1998), treated in the period from January 1993 through July 2007. This sample was also used in the study by Paap et al. (2011b). Sex ratio and age are depicted in Table 1. Seventy-nine percent were diagnosed with at least one personality disorder (PD). Of the PDs, Avoidant PD was most common (39%), followed by Borderline PD (24%). Extensive comorbidity was common in this group. All patients had at least one axis I disorder. The majority of the patients fulfilled criteria for either Major Depressive Disorder or Dysthymic Disorder (69%), and almost half of the patients were phobic (45% fulfilled criteria for at least one of the following: Agoraphobia, Social Phobia or Specific Phobia). We refer to Paap et al. (2011b) and Karterud et al. (2003) for sociodemographic and diagnostic details. Patients admitted before 1996 were diagnosed according to the DSM-III-R

(APA, 1987) and patients admitted from 1996 onwards according to the DSM-IV (APA, 1994).

In the study by Paap et al. (2011b), two subgroups were created based on clinical criteria: the first group existed of patients with a clinical disorder (CD) only ($GSI=1.3$), and the second group of patients diagnosed with a PD in addition to a CD ($GSI=1.6$). Since the focus of the current study is on the impact of psychological distress on dimensionality, we chose to use a different criterion to create two subgroups in the current study. To maximize the difference in GSI scores in the resulting subgroups while at the same time create subgroups that showed similar variance of GSI scores as the GI and depression samples, the total group of 3078 patients was divided along the median GSI-score (1.53). The group consisting of patients with a GSI-score through 1.53 are referred to as the PD_{low} group ($n=1528$, mean age= 35 ± 9 years) and the group of patients with a GSI-score of 1.53 or higher as the PD_{high} group ($n=1550$, mean age= 35 ± 9 years).

All participating hospitals complied with the diagnostic and data collection procedures required for membership in the Norwegian Network. All data registered by the different hospitals were collected regularly in a central, anonymised database, administrated by the Department of Personality Psychiatry, Oslo University Hospital. All patients gave written consent and the procedures were approved by the State Data Inspectorate and the Regional Committee for Medical Research and Ethics.

2.1.2. Gender incongruence sample

This sample consisted of 410 persons referred to four Gender Identity Disorder (GID) clinics: Ghent (Belgium), Hamburg (Germany), Amsterdam (the Netherlands) and Oslo (Norway). The data collection took place within the framework of the 'European Network for the Investigation of Gender Incongruence' (ENIG) initiative (Kreukels et al., 2010). This network was created in order to improve comparability of data pertaining to gender incongruence (GI) and GID across clinics, as well as diagnostic transparency (Paap et al., 2011a). The ENIG study includes applicants that were seen at GID clinics in Ghent, Hamburg, Amsterdam, and Oslo from the start of January 2007. In the current study all new applicants that were seen between January 2007 and December 2009 and whose data had been entered in the database, were at least 16 years of age at their first visit, and who had filled out the SCL-90-R were included. Sex ratio (reported sex corresponds to natal sex) and age are depicted in Table 1. At the time of data analysis, 56% of the total sample had been diagnosed with GID, 10% with another disorder pertaining to gender incongruent feelings (such as transvestic fetishism or GID NOS) and the remaining 34% had not yet received a diagnosis. The four participating clinics complied with the diagnostic and data collection procedures required for membership in the ENIG initiative. All data registered by the different clinics were collected regularly in a central, anonymised database, administrated at the Oslo University Hospital. All patients gave written consent and the procedures were approved by the regional ethical committees.

2.1.3. Depression sample

This sample consisted of 223 patients who had been referred to the Department of Neuropsychiatry and Psychosomatic Medicine at Oslo University Hospital and fulfilled the DSM-IV (APA, 1994) criteria for Major Depressive Disorder or Dysthymic Disorder. The patients were at least 18 years old at the first visit, and were seen between January 2005 and December 2008. Sex ratio and age are depicted in Table 1. Seventy-four percent of the patients fulfilled criteria for at least one other axis I disorder, of which a phobic disorder was most common (46% fulfilled criteria for either Agoraphobia, Social Phobia or Specific Phobia), followed by Generalised Anxiety Disorder (37%). The M.I.N.I. (Sheehan and Lecrubier, 1994) was used to screen for axis I disorders. All patients gave written consent and the procedures were approved by the State Data Inspectorate and the Regional Committee for Medical Research and Ethics.

2.2. Measures

All patients completed a number of self-report measures prior to or directly after one of the first consultations, including the Symptom Checklist 90-Revised (SCL-90-R; Derogatis, 1994). The instrument was designed to measure nine symptom dimensions (comprising a total of 83 items): somatization (Som), interpersonal sensitivity (Int), depression (Dep), anxiety (Anx), phobic anxiety (Pho), obsession-compulsion (Obs), hostility (Hos), paranoid ideation (Par), and psychoticism (Psy), and includes 7 additional items. Each item is scored on a scale ranging from 0 ('not at all') through 4 ('extremely'). The mean score on all 90 items (including the 7 additional items) is referred to as the Global Severity Index (GSI; range 0–4) and is widely used as a global index for psychological distress.

2.3. Investigating dimensionality: Mokken Scale Analysis (MSA)

To investigate the dimensionality of the SCL-90-R, MSA was used (Mokken, 1971, 1997). MSA is a nonparametric Item Response Theory (IRT) approach that can be used to explore and test hypotheses about the dimensionality of a data-set. MSA can be used in a confirmatory or exploratory way. In either case, it assesses whether the

¹ Note that the DSM5 Work Group for Sexual and Gender Identity Disorders first considered proposing the replacement of "GID" with the term "Gender Incongruence", but that the current recommendation is replacing "GID" with "Gender Dysphoria". However, gender incongruent feelings are still considered core to the phenomenon of GID/Gender Dysphoria. In this paper, we use the term "GI" to signify people who experience gender incongruent feelings, but who have not yet been diagnosed.

Table 1
Descriptive statistics for the four samples.

	PD _{high}		PD _{low}		GI		Depression	
	Males	Females	Males	Females	Males	Females	Males	Females
N	386	1164	459	1069	264	146	94	129
Mean age ± S.D.	37 ± 9	34 ± 9	37 ± 9	35 ± 9	35 ± 12	27 ± 10	47 ± 14	44 ± 13
Mean GSI ± S.D.	2.0 ± 0.38	2.1 ± 0.40	1.0 ± 0.34	1.1 ± 0.34	0.5 ± 0.46	0.6 ± 0.54	1.2 ± 0.50	1.3 ± 0.62
Skewness GSI	1.00	0.88	−0.52	−0.69	1.24	1.72	0.07	0.74
Kurtosis GSI	0.89	0.49	−0.42	−0.27	1.45	3.50	−0.80	0.52

clusters of items (dimensions) that are found or are tested adhere to a measurement model called the Monotone Homogeneity Model (MHM). This model allows for the ordering of respondents on an underlying dimension (cluster of items) using the unweighted sum score (Sijtsma and Molenaar, 2002; Meijer and Baneke, 2004; Sijtsma et al., 2008; Wismeijer et al., 2008). A scale fulfilling the criteria of the Mokken's Monotone Homogeneity Model (MHM) measures one latent trait only (unidimensionality), is made up of items which the participant approaches in a way that is independent of the previous items (local independence), and results in a scale where the participants tend to score higher on items when they have a high latent trait score (monotonicity). In summary, MSA is a method that can be used for dimensionality testing while at the same time resulting in scales adhering to a measurement model. In this study we did not investigate whether items retained in the Mokken scales showed invariant item ordering (IIO). Although some studies showed that sets of items in clinical scales are in agreement with this model, in general IIO is a restrictive assumption that may impact the construct validity of the scale (for a discussion see Sijtsma et al. (2011), Meijer and Egberink (in press)). MSA was applied using the software package Mokken Scale Analysis for Polytomous items (MSP5.0) (Molenaar and Sijtsma, 2000).

In order to determine whether the scale or scales are unidimensional, scalability coefficients are calculated. These coefficients are calculated between item-pairs (H_{ij}), on the item-level (H_i) and on the scale-level (H). There are some parallels between H_i , which is based on the H_{ij} s, and other popular coefficients such as the item-rest correlation used in Classical Test Theory (CTT) and the item discrimination parameter used in parametric Item Response Theory (IRT). Similar to the item-rest correlation, H_i expresses the degree to which an item is related to other items in the scale. However, unlike the item-rest correlation, the H_i coefficient is a 'corrected' correlation: the correlation between items is divided by the maximum expected correlation given the items' univariate score-frequency distributions (Molenaar, 1997). An important advantage of this statistic is that it avoids problems with respect to the distorting effect of difference in item-score distributions on inter-item correlations; more traditional methods that are based on inter-item correlations, such as Principal Components Analysis (PCA), produce artifactual 'difficulty factors' as soon as the items have different distributions of items scores, in particular when items have only a few answer categories (Wismeijer et al., 2008). Similar to the item discrimination parameter, a high value of H_i indicates that the item distinguishes well between people with relatively low latent trait values and people with relatively high latent trait values. H is based on the H_{ij} s and expresses the degree to which the total score accurately orders persons on the latent trait scale (Sijtsma and Molenaar, 2002). A scale is considered acceptable if $0.3 \leq H < 0.4$, good if $0.4 \leq H < 0.5$, and strong if $H \geq 0.5$ (Mokken, 1971; Sijtsma and Molenaar, 2002).

The algorithm that MSP5.0 uses to build one or more scales is called Algorithm for Item Selection (AISP); it aims to find unidimensional clusters of items. These clusters were identified by running the AISP several times in a row, each time increasing the lower bound scalability coefficient (also known as the user-specified constant, c). The higher the value of c , the more confidence we have in the ordering of persons by means of their total scale score (Mokken, 1971; Sijtsma and Molenaar, 2002). Following Sijtsma and Molenaar (2002), we ran the AISP repeatedly for increasing values of c (0, 0.1, 0.2, 0.25, 0.3, 0.35, 0.4, 0.5) and set the maximum number of scales to 10. The resulting sequence of outcomes indicates whether the data-set is unidimensional or multidimensional. Sijtsma and Molenaar (2002); pp. 81–82 provide the following guidelines. In case of a unidimensional scale, the typical sequence is: (1) most or all items are in one scale, (2) one smaller scale is found, and (3) one or a few small scales are found and several items are excluded. In multidimensional datasets the typical sequence is: (1) most or all items are in one scale, (2) two or more scales are formed, and (3) two or more smaller scales are formed and several items are excluded. See Wismeijer (2011) for a recent empirical application.

3. Results

3.1. Missing data: two-way imputation

Less than 1% of the data were missing in each of the data-sets. Following Paap et al. (2011b), we used Two-Way imputation

(Bernaards and Sijtsma, 2000), which allows the user to transform an incomplete data-file into a complete one by using all available information about the proficiency of the respondent and the 'difficulty' of the item (Sijtsma and van der Ark, 2003). This method is easy to implement using SPSS (SPSS, 2007), using the syntax provided by van Ginkel and van der Ark (2005).

3.2. Description of the data

Table 1 shows sample size, mean ages and mean GSI score for males and females separately within each group. The mean GSI score was highest for the PD_{high}, followed by the PD_{low} group and depression sample, and finally the GI sample. Sex differences in mean GSI scores were small (0.1 for each group). Tables 2–5 show the correlations between the subscales of the SCL-90-R, mean subscale scores with SD, and Cronbach's alpha, for the PD_{high}, PD_{low}, GI and depression group respectively.

On the whole, the mean correlations between the subscales were of weak to medium strength, ranging between 0.16 for the phobic anxiety (Pho) scale in the PD_{low} group to 0.35 for the anxiety (Anx) and psychoticism (Psy) scales in the PD_{high} group. The hostility (Hos) and Pho scales had the lowest mean correlations. When comparing Tables 2 and 3, it can be seen that the correlations and S.D.'s for the Pho and Psy scales show the largest difference (correlations: 0.07; S.D.'s: 0.31). However, the difference in correlations and S.D.'s for the obsessive-compulsive (Obs) and interpersonal sensitivity (Int) was very small (correlations: 0.01; S.D.'s: 0.01 and 0.04, respectively).

Table 4 shows that the mean correlations were a lot higher for the GI sample than for the PD groups (in spite of similar S.D.'s for most subscales), ranging between 0.59 for the Pho scale to 0.73 for the Int scale. The differences in the mean scores on the Hos, Par and Psy scales between the GI sample on one hand, and the PD_{low} group on the other hand, were small (0.1, 0.2 and 0.1, respectively).

Inspection of Table 5 reveals that the mean correlations for the depression sample are not as low as those for the PD groups, and not as high as those for the GI sample, ranging from 0.40 (Hos) to 0.59 (Int). Furthermore, the mean subscale scores are highly similar to those of the PD_{high} group for most subscales. The difference is largest for the Hos scale: 1.5 for the depression sample and 1.0 for the PD_{high} group.

3.3. Dimensionality of the SCL-90-R

3.3.1. Mokken Scale Analysis

Four (groups) × eight (different values of c) = 32 exploratory analyses were performed using the SEARCH-procedure. A summary of the findings can be seen in Table 6. Several important findings can be noted. Firstly, at $c=0$, five scales were found for both PD groups. This result is a strong indication for multidimensionality. Additionally, less than half of the items ended up in the first scale. As the value of c increased from 0.10 to 0.20, the number of scales increased sharply for both PD groups, and

Table 2
Correlations on the SCL-90-R subscales, mean subscale scores with S.D., Cronbach's alpha (α), PD_{high} group.

	Som	Obs	Int	Anx	Pho	Dep	Hos	Par	Psy
Somatization (Som)	1	0.31	0.08	0.51	0.31	0.25	0.11	0.11	0.18
Obsessive-compulsive (Obs)		1	0.25	0.38	0.22	0.38	0.20	0.28	0.37
Interpersonal sensitivity (Int)			1	0.23	0.34	0.38	0.19	0.54	0.43
Anxiety (Anx)				1	0.46	0.36	0.20	0.29	0.38
Phobic Anxiety (Pho)					1	0.12	0.05	0.16	0.17
Depression (Dep)						1	0.15	0.28	0.36
Hostility (Hos)							1	0.38	0.33
Paranoid ideation (Par)								1	0.55
Psychoticism (Psy)									1
Mean correlation	0.23	0.30	0.31	0.35	0.23	0.29	0.20	0.32	0.35
Mean subscale score	2.2	2.4	2.4	2.3	1.8	2.7	1.0	1.8	1.2
S.D.	0.75	0.60	0.68	0.64	0.96	0.50	0.81	0.80	0.60
α	0.81	0.68	0.72	0.72	0.81	0.67	0.79	0.67	0.69

Table 3
Correlations on the SCL-90-R subscales, mean subscale scores with S.D., Cronbach's alpha (α), PD_{low} group.

	Som	Obs	Int	Anx	Pho	Dep	Hos	Par	Psy
Somatization (Som)	1	0.23	0.03	0.43	0.22	0.23	0.09	0.02	0.09
Obsessive-compulsive (Obs)		1	0.40	0.31	0.11	0.57	0.20	0.30	0.33
Interpersonal sensitivity (Int)			1	0.26	0.24	0.48	0.23	0.51	0.44
Anxiety (Anx)				1	0.44	0.37	0.12	0.16	0.28
Phobic Anxiety (Pho)					1	0.08	0.02	0.08	0.08
Depression (Dep)						1	0.15	0.26	0.41
Hostility (Hos)							1	0.33	0.18
Paranoid ideation (Par)								1	0.42
Psychoticism (Psy)									1
Mean correlation	0.17	0.31	0.32	0.30	0.16	0.32	0.17	0.26	0.28
Mean subscale score	1.1	1.4	1.3	1.1	0.7	1.6	0.5	0.7	0.5
S.D.	0.62	0.61	0.64	0.56	0.65	0.63	0.45	0.57	0.33
α	0.78	0.73	0.73	0.73	0.76	0.79	0.65	0.61	0.47

Table 4
Correlations on the SCL-90-R subscales, mean subscale scores with S.D., Cronbach's alpha (α), GI group.

	Som	Obs	Int	Anx	Pho	Dep	Hos	Par	Psy
Somatization (Som)	1	0.69	0.60	0.75	0.60	0.64	0.56	0.55	0.56
Obsessive-compulsive (Obs)		1	0.76	0.78	0.61	0.80	0.68	0.68	0.71
Interpersonal sensitivity (Int)			1	0.74	0.69	0.81	0.67	0.80	0.76
Anxiety (Anx)				1	0.68	0.77	0.61	0.64	0.68
Phobic Anxiety (Pho)					1	0.56	0.49	0.55	0.52
Depression (Dep)						1	0.64	0.68	0.73
Hostility (Hos)							1	0.64	0.61
Paranoid ideation (Par)								1	0.75
Psychoticism (Psy)									1
Mean correlation	0.62	0.71	0.73	0.71	0.59	0.70	0.61	0.66	0.67
Mean subscale score	0.4	0.7	0.7	0.5	0.3	0.9	0.4	0.5	0.4
S.D.	0.49	0.61	0.69	0.54	0.56	0.74	0.51	0.61	0.44
α	0.85	0.85	0.87	0.87	0.84	0.90	0.79	0.78	0.72

Table 5
Correlations on the SCL-90-R subscales, mean subscale scores with S.D., Cronbach's alpha (α), depression group.

	Som	Obs	Int	Anx	Pho	Dep	Hos	Par	Psy
Somatization (Som)	1	0.45	0.32	0.53	0.43	0.38	0.24	0.22	0.33
Obsessive-compulsive (Obs)		1	0.63	0.65	0.51	0.72	0.44	0.48	0.58
Interpersonal sensitivity (Int)			1	0.57	0.67	0.69	0.48	0.73	0.60
Anxiety (Anx)				1	0.64	0.71	0.40	0.42	0.57
Phobic Anxiety (Pho)					1	0.60	0.25	0.47	0.45
Depression (Dep)						1	0.40	0.48	0.57
Hostility (Hos)							1	0.53	0.48
Paranoid ideation (Par)								1	0.69
Psychoticism (Psy)									1
Mean correlation	0.36	0.56	0.59	0.56	0.50	0.57	0.40	0.50	0.53
Mean subscale score	1.5	1.6	1.1	1.3	0.8	1.8	0.5	0.6	0.6
S.D.	0.85	0.83	0.83	0.82	0.86	0.83	0.63	0.68	0.48
α	0.86	0.85	0.85	0.86	0.84	0.87	0.82	0.77	0.69

Table 6

Results of the Mokken Scale Analyses using the Algorithm for Item Selection: number (No.) of scales, number of items in the first scale and number of excluded items for 8 levels of c , reported separately for the four samples.

	$c=0$	$c=0.10$	$c=0.20$	$c=0.25$	$c=0.30$	$c=0.35$	$c=0.40$	$c=0.50$
<i>PD_{high}</i>								
No. scales	5	4	10	10	10	10	10	10
No. items 1st scale	39	38	21	7	6	6	4	3
No. excluded items ^a	0	7	7	21	39	54	59	67
<i>PD_{low}</i>								
No. scales	5	8	10	10	10	10	10	10
No. items 1st scale	38	38	22	16	12	6	4	2
No. excluded items ^a	1	1	16	32	39	54	59	69
<i>GI</i>								
No. scales	2	2	2	3	5	5	9	10
No. items 1st scale	86	86	85	82	74	62	46	17
No. excluded items ^a	1	1	2	2	7	13	16	42
<i>Depression</i>								
No. scales	4	4	4	6	8	10	10	10
No. items 1st scale	71	71	71	60	53	39	30	9
No. excluded items ^a	0	0	4	5	8	10	24	43

Note: c is the lowerbound scalability coefficient H specified by the user.

^a Either rejected due to negative H with one of the scale items or excluded due to lowerbound and/or significance criteria.

the number of items in the first scale dropped by a third. As the value of c increased further, the number of items in the first scale continued dropping. This was accompanied by an increasing number of items being excluded.

In contrast, at $c=0$, only two scales were found for the GI group, one large scale including 86 items and one smaller scale including three items. This scale structure (one dominant scale with one or several very small scales) persisted throughout all analyses. The scale solution remained largely unchanged until $c=0.30$ was reached; as c increased from 0.30 to 0.50, the number of items in the first scale decreased slightly, and the number of scales increased. Overall, this pattern indicates unidimensionality.

The pattern for the depression sample was less clear-cut than for the other samples. At $c=0$, four scales were found, and at $c=0.30$, as many as eight scales were found. However, the first scale remained the dominating one throughout all analyses. At this stage of the analyses, the pattern of scale solutions for the DEP sample did indicate multidimensionality.

3.3.2. Sex differences

Since there were considerable differences in sex ratio between the four groups, we repeated the above mentioned analyses for each sex separately. For the depression group, the patterns of outcomes for increasing levels of c were very different for both sexes. The pattern of the male depressed patients was highly similar to that of the PD groups (many smaller scales, first scale relatively small), whereas the pattern for the females was similar to that of the GI sample (one large dominant scale emerged accompanied by one smaller one). This is illustrated in Fig. 1. To explore potential explanations for these differences, we compared the comorbidity rates for females and males, and inspected the correlations between the original subscales. The percentage of females diagnosed with agoraphobia were similar to that of the percentage of males (30%). The depressed females were, however, diagnosed more frequently with specific phobia (19% versus 11% of the males) and social phobia (32% versus 27% of the males). As could be expected given the outcomes of the MSP analyses, the correlations between the subscales were higher for the females than for the males. The largest differences could be found for the Som, Hos, Par and Psy scales: the mean subscale correlation for these scales was 0.17–0.20 higher for the females than for the

males. For the PD groups and the GI sample, only small differences in scale solutions were found, which did not impact the pattern of outcomes and as a consequence will not be reported here.

4. Discussion

Studies reporting on the dimensionality of the SCL-90-R have had very diverse outcomes. To this day, the original 9-scale solution (Derogatis, 1994) remains controversial (Schwarzwalder et al., 1991; Holi et al., 1998; Vassend and Skrandal, 1999; Schmitz et al., 2000; Olsen et al., 2004; Arrindell et al., 2006; Elliott et al., 2006; Hafkenscheid et al., 2007; Paap et al., 2011b). Here, we wanted to identify factors that could help explain the inconsistent findings in the literature. The main purpose of this study was to compare the dimensionality of the SCL-90-R in three different patient groups, using Mokken Scale Analysis (MSA). We wanted to ascertain whether the dimensional structure depends on (a) the level of psychological distress (GSI score), (b) the variance in SCL-90-R scores, and (c) the primary diagnosis in a particular patient group.

Our results indicated that the dimensional structure in fact depends on the level of psychological distress as measured by the Global Severity Index (GSI). We found support for the unidimensionality of the SCL-90-R when analyzing the data from the Gender Incongruence sample, which was characterized by a low level of psychological distress. In contrast, Paap et al. (2011b) found support for the multidimensionality of the SCL-90-R based on a sample of patients that reported a high level of psychological distress. These findings are directly comparable, since the same analytic strategy was used. Note that we deliberately chose not to simply test the 7-dimensional structure reported by Paap et al. (2011b) using a confirmatory analysis.

Merely investigating the H -values for scales produced by a confirmatory approach does not suffice when one wants to investigate the underlying dimensionality. The H -values found for the subscales in a confirmatory analysis can be of a satisfactory size, as we found in this study, but this does not rule out that a unidimensional solution would show superior “fit”; H -values tend to increase when the construct becomes narrower but this does not imply the solution with the highest H -values is to be preferred.

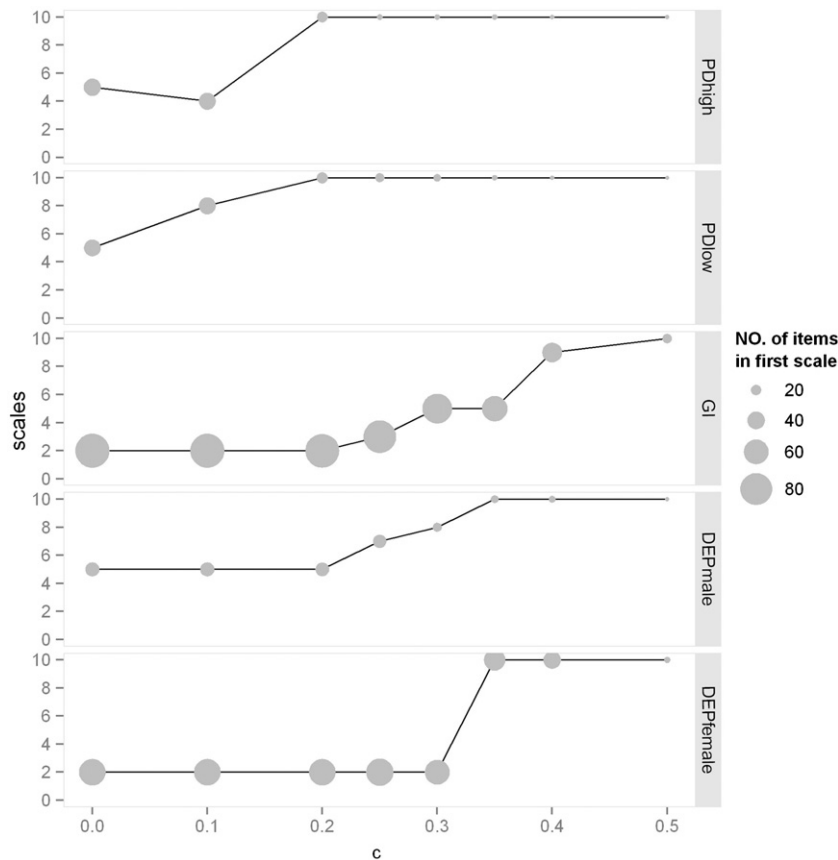


Fig. 1. Number of scales (y-axis) and number of items in the first scale (size of dots) for different levels of c (x-axis), with separate panels for the different groups. The GI sample and the female DEP group show a typical unidimensional pattern for increasing c : (1) most or all items are in one scale (2) one smaller scale is found, and (3) one or a few small scales are found and several items are excluded.

Recent studies that examined the dimensionality of the SCL-90 (-R) in community samples, found the instrument to be either unidimensional or found one very strong and dominant factor with one or two very small residual ones (Holi et al., 1998; Vassend and Skrandal, 1999; Olsen et al., 2004). One possible cause for such largely 'unidimensional findings' could be a lack of variance in reported psychological distress in these samples. To rule out this explanation, we divided the personality disorder (PD) sample used in the study by Paap et al. (2011b) in two subgroups by means of a median split based on the GSI score. This way we obtained two subgroups that had a smaller variance than the original sample; a variance that was now comparable to that in the GI group. At the same time, both subgroups still had a much higher mean GSI score than the GI group. Our results clearly showed support for a multidimensional solution in both PD data sets,² in spite of the diminished variance. Therefore it is unlikely that the largely 'unidimensional findings' reported by others using samples characterized by low levels of psychological distress can be merely explained by a lack of variance in SCL-90-R scores.

To test the generalizability of our findings, we investigated the dimensionality in a third sample, consisting of depressed outpatients. This sample was characterized by an intermediate level of reported psychological distress. In this sample, we found an effect of sex on dimensionality; the depressed males demonstrated a dimensional structure that was highly similar to that of

the PD groups, whereas the depressed females resembled the GI patients, interpreting the SCL-90-R largely as a unidimensional construct. This is an important finding for several reasons. First of all, these sex differences could underlie 'intermediate' scale solutions (neither convincingly unidimensional nor multidimensional) such as was the case in our depression sample. Second, our finding demonstrates that finding factorial invariance for sex in one patient group is not necessarily generalizable to another patient group. Finally, it illustrates the importance of taking sex into account when investigating the dimensionality of self-report instruments such as the SCL-90-R. Most of the studies that have reported on the factorial structure/dimensionality of the SCL-90-R, have only reported sex ratio in the sample(s) used and/or sex differences in subscale and GSI scores. Only very few studies investigated the actual sex effect on the dimensionality or final scale solution. Exceptions are Vassend and Skrandal (1999), who demonstrated factorial invariance for sex, and Olsen et al. (2004), who showed that there were two items in the SCL-90-R that were sex biased ('having to do things slowly' and 'crying easily').

Unidimensionality as indicated by MSA indicates that most items in the scale correlate relatively highly with each other. In other words: our results suggest that depressed females do not differentiate as much between different types of psychological complaints as depressed men do. Interestingly, Armour et al. (2011) found that the subscales of a PTSD screening measure they investigated correlated more strongly in the female group than in the male group. It may be that the gender difference in dimensionality is typical for affective disorders; this needs further investigation.

² To make sure this finding was not simply caused by a large sample size (e.g. higher power to detect multidimensionality), we repeated the analyses for random subsets of the two PD groups that were similar in size to the GI/DEP groups. This procedure did not affect the results.

Our analyses suggest that differences in variance of SCL-90-R scores are unlikely to have a big impact on the dimensionality. We found that sex and level of psychological distress (measured by the GSI) were related to dimensional structure. In what way the main diagnosis and degree of comorbidity impacts the dimensional structure remains unresolved. Future studies are needed to investigate whether the sex effect on dimensionality is generalizable to other patient groups or whether it is typical for depressed patients with moderate levels of psychological distress. Our results suggest that total scores (GSI) can be reliably used in patient groups with low self-reported level of distress, such as GI patients, but subscale scores may be unreliable.

Acknowledgments

We thank Jan van Bebber for imputating the missing data, Xi X. Zhao for preparing Fig. 1, and Mitzi Paap, Frøydis Hellem and Thomas Mengshoel for helpful discussions. We thank the patients and staff from the GID clinics in Amsterdam, Oslo, Hamburg and Ghent, as well as from the Department of Neuropsychiatry and Psychosomatic Medicine at Oslo University Hospital for their contribution to this study. Finally, we thank the patients and staff from the following treatment units in the Norwegian Network of Personality-Focused Treatment Programs: Department for Personality Psychiatry, Oslo University Hospital; the Group Therapy Unit, Lillestrøm District Psychiatric Center, Akershus University Hospital; the Unit for Group Therapy, District Psychiatric Center, Lovisenlund, Sørlandet Hospital HF, Kristiansand; the Outpatient Clinic, Department of Mental Health, Sanderud, Innlandet Hospital Health Authority; the Group Therapy Unit, Outpatient Clinic, Drammen Psychiatric Center; the Unit for Group Therapy, Vestfold Mental Health Care Trust, Tønsberg; the Group Therapy Unit, Alna District Psychiatric Center, Department of Psychiatry, Aker University Hospital, Oslo; the Årstad Day Unit, Fjell & Årstad District Psychiatric Center, Bergen; the Bergenhus Day Unit, District Psychiatric Center, Bergen; the Unit for Group Therapy, Skien District Psychiatric Center, Telemark Hospital Health Authority; Day Treatment Unit, Furuset District Psychiatric Center, Aker University Hospital, Oslo; the Group Therapy Unit, Ringerike Psychiatric Center, Hønefoss; the Outpatient Clinic in Farsund, District Psychiatric Center, Farsund, and the Unit for Group Therapy, Jessheim District Psychiatric Center, Akershus University Hospital HF. The study was supported by the South-Eastern Norway Regional Health Authority, the Norwegian Research Council, and the University of Oslo. The founding sources did not participate in the collection of data, the interpretation of the results or the writing of the manuscript. They have not taken part in the decision of submitting the manuscript for publication.

References

- APA. 1987. *Diagnostic and Statistical Manual of Mental Disorders* (3rd ed., revised) (DSM-III-R). American Psychiatric Association, Washington, DC.
- APA. 1994. *Diagnostic and Statistical Manual of Mental Disorders* (4th ed.) (DSM-IV). American Psychiatric Association, Washington, DC.
- Armour, C., Elhai, J.D., Layne, C.M., Shevlin, M., Duraković-Belko, E., Djapo, N., Pynoos, R.S., 2011. Gender differences in the factor structure of posttraumatic stress disorder symptoms in war-exposed adolescents. *Journal of Anxiety Disorders* 25, 604–611.
- Arrindell, W.A., Barelids, D.P., Janssen, I.C., Buwalda, F.M., van der Ende, J., 2006. Invariance of SCL-90-R dimensions of symptom distress in patients with peripartum pelvic pain (PPPP) syndrome. *British Journal of Clinical Psychology* 45, 377–391.
- Arrindell, W.A., Boomsma, A., Ettema, H., Stewart, R., 2004a. Nog meer steun voor het multidimensionale karakter van de SCL-90-R [Even more support for the multidimensional nature of the SCL-90-R]. *De Psycholoog* 39, 368–371.
- Arrindell, W.A., Boomsma, A., Ettema, H., Stewart, R., 2004b. Verdere steun voor het multidimensionale karakter van de SCL-90-R [Further support for the multidimensional nature of the SCL-90-R]. *De Psycholoog* 39, 195–201.
- Bernaards, C.A., Sijtsma, K., 2000. Influence of simple imputation and EM methods on factor analysis when item nonresponse in questionnaire data is nonignorable. *Multivariate Behavioral Research* 35, 321–364.
- Cyr, J.J., McKenna-Foley, J.M., Peacock, E., 1985. Factor structure of the SCL-90-R: is there one? *Journal of Personality Assessment* 49, 571–578.
- Derogatis, L.R., 1994. SCL-90-R: Administration, Scoring and Procedures Manual. National Computer Systems, Minneapolis, MN.
- Elliott, R., Fox, C.M., Belyukova, S.A., Stone, G.E., Gunderson, J., Zhang, X., 2006. Deconstructing therapy outcome measurement with rasch analysis of a measure of general clinical distress: the Symptom Checklist-90-revised. *Psychological Assessment* 18, 359–372.
- Hafkenscheid, A., 2004. Hoe multidimensionaal is de Symptom Checklist (SCL-90) nu eigenlijk? [How multidimensional is the Symptom Checklist (SCL-90) really?]. *De Psycholoog* 39, 191–194.
- Hafkenscheid, A., Maassen, G., Veeninga, A., 2007. The dimensions of the Dutch SCL-90: more than one, but how many? *Netherlands Journal of Psychology* 63, 25–30.
- Haraldsen, I.R., Dahl, A.A., 2000. Symptom profiles of gender dysphoric patients of transsexual type compared to patients with personality disorders and healthy adults. *Acta Psychiatrica Scandinavica* 102, 276–281.
- Holi, M.M., Sammallahti, P.R., Aalberg, V.A., 1998. A Finnish validation study of the SCL-90. *Acta Psychiatrica Scandinavica* 97, 42–46.
- Karterud, S., Pedersen, G., Bjordal, E., Brabrand, J., Friis, S., Haaseth, O., Haavaldsen, G., Irion, T., Leirvag, H., Torum, E., Urnes, O., 2003. Day treatment of patients with personality disorders: experiences from a Norwegian treatment research network. *Journal of Personality Disorders* 17, 243–262.
- Karterud, S., Pedersen, G., Friis, S., Urnes, Ø., Irion, T., Brabrand, J., Falkum, L.R., Leirvåg, H., 1998. The Norwegian network of psychotherapeutic day hospitals. *Therapeutic Communities* 19, 15–28.
- Kreukels, B.P.C., Haraldsen, I.R., De Cuyper, G., Richter-Appelt, H., Gijs, L., Cohen Kettens, P.T., 2010. A European network for the investigation of gender incongruence: the ENIG initiative. *European Psychiatry*. <http://dx.doi.org/10.1016/j.eurpsy.2010.04.009>.
- Leinonen, E., Niemi, H., 2007. The influence of educational information on depressed outpatients treated with escitalopram: a semi-naturalistic study. *Nordic Journal of Psychiatry* 61, 109–114.
- Meijer, R.R., Baneke, J.J., 2004. Analyzing psychopathology items: a case for nonparametric item response theory modeling. *Psychological Methods* 9, 354–368.
- Meijer, R.R., de Vries, R.M., van Bruggen, V., 2011. An evaluation of the brief symptom inventory-18 using item response theory: which items are most strongly related to psychological distress? *Psychological Assessment* 23, 193–202.
- Meijer, R.R., Egberink, I.J.L. Investigating Invariant Item Ordering in Personality and Clinical Scales: Some empirical findings and a Discussion. *Educational and Psychological Measurement*, <http://dx.doi.org/10.1177/0013164411429344>, in press.
- Meyer-Bahlburg, H., 2010. From mental disorder to iatrogenic hypogonadism: dilemmas in conceptualizing gender identity variants as psychiatric conditions. *Archives of Sexual Behavior* 39, 461–476.
- Mokken, R.J., 1971. *A Theory and Procedure of Scale Analysis*. Mouton, The Hague.
- Mokken, R.J., 1997. Nonparametric models for dichotomous responses. In: van der Linden, W.J., Hambleton, R.K. (Eds.), *Handbook of Modern Item Response Theory*. Springer, New York, pp. 351–367.
- Molenaar, I.W., 1997. Nonparametric models for polytomous responses. In: van der Linden, W.J., Hambleton, R.K. (Eds.), *Handbook of Modern Item Response Theory*. Springer, New York, pp. 369–380.
- Molenaar, I.W., Sijtsma, K., 2000. MSP5 for Windows. iecProGAMMA. Groningen, The Netherlands.
- Olsen, L.R., Mortensen, E.L., Bech, P., 2004. The SCL-90 and SCL-90R versions validated by item response models in a Danish community sample. *Acta Psychiatrica Scandinavica* 110, 225–229.
- Paap, M.C.S., Kreukels, B.P.C., Cohen-Kettens, P.T., Richter-Appelt, H., de Cuyper, G., Haraldsen, I.R., 2011a. Assessing the utility of diagnostic criteria: a multisite study on gender identity disorder. *Journal of Sexual Medicine* 8, 180–190.
- Paap, M.C.S., Meijer, R.R., van Bebber, J., Pedersen, G., Karterud, S., Hellem, F.M., Haraldsen, I.R., 2011b. A study of the dimensionality and measurement precision of the SCL-90-R using item response theory. *International Journal of Methods in Psychiatric Research* 20, e39–e55.
- Schmitz, N., Hartkamp, N., Kiuse, J., Franke, G.H., Reister, G., Tress, W., 2000. The Symptom Check-List-90-R (SCL-90-R): a German validation study. *Quality of Life Research* 9, 185–193.
- Schwarzwald, J., Weisenberg, M., Solomon, Z., 1991. Factor invariance of SCL-90-R: the case of combat stress reaction. *Psychological Assessment* 3, 385–390.
- Sheehan, D.V., Lecrubier, Y., 1994. *Mini International Neuropsychiatric Interview (M.I.N.I.)*. University of South Florida Institute for Research in Psychiatry/INSERM-Hôpital de la Salpêtrière, Tampa, FL/Paris.
- Sijtsma, K., Emons, W.H., Bouwmeester, S., Nyklicek, I., Roorda, L.D., 2008. Nonparametric IRT analysis of Quality-of-Life Scales and its application to the World Health Organization Quality-of-Life Scale (WHOQOL-Bref). *Quality of Life Research* 17, 275–290.

- Sijtsma, K., Meijer, R.R., van der Ark, L.A., 2011. Mokken scale analysis as time goes by: an update for scaling practitioners. *Personality and Individual Differences* 50, 31–37.
- Sijtsma, K., Molenaar, I.W., 2002. *Introduction to Nonparametric Item Response Theory*. Sage Publications, Thousand Oaks.
- Sijtsma, K., van der Ark, L.A., 2003. Investigation and treatment of missing item scores in test and questionnaire data. *Multivariate Behavioral Research* 38, 505–528.
- SPSS, 2007. *SPSS for Windows, Rel. 16.0.1*. SPSS Inc., Chicago.
- van Ginkel, J.R., van der Ark, L.A., 2005. SPSS syntax for missing value imputation in test and questionnaire data. *Applied Psychological Measurement* 29, 152–153.
- Vassend, O., Skrondal, A., 1999. The problem of structural indeterminacy in multidimensional symptom report instruments. The case of SCL-90-R. *Behaviour Research and Therapy* 37, 685–701.
- WHO, 1992. *The ICD–10 Classification of Mental and Behavioral Disorders: Clinical Descriptions and Diagnostic Guidelines*. World Health Organization, Geneva.
- Wismeijer, A.A.J., 2011. Dimensionality analysis of the thought suppression inventory: combining EFA, MSA, and CFA. *Journal of Psychopathology and Behavioral Assessment*.
- Wismeijer, A.A.J., Sijtsma, K., van Assen, M.A.L.M., Vingerhoets, A.J.J.M., 2008. A comparative study of the dimensionality of the self-concealment scale using principal components analysis and Mokken scale analysis. *Journal of Personality Assessment* 90, 323–334.