International Journal of Methods in Psychiatric Research
Int. J. Methods Psychiatr. Res. 20(3): e39–e55 (2011)
Published online 3 August 2011 in Wiley Online Library
(wileyonlinelibrary.com) **DOI:** 10.1002/mpr.347

A study of the dimensionality and measurement precision of the SCL-90-R using item response theory

MUIRNE C. S. PAAP,[1,2] ROB R. MEIJER,[3] JAN VAN BEBBER,[4] GEIR PEDERSEN,[5] SIGMUND KARTERUD,[2,5] FRØYDIS M. HELLEM[1] & IRA R. HARALDSEN[1]

1 Department of Neuropsychiatry and Psychosomatic Medicine, Oslo University Hospital, Oslo, Norway
2 Institute of Clinical Medicine, University of Oslo, Oslo, Norway
3 Department of Psychometrics and Statistics, Faculty of Behavioural and Social Sciences, University of Groningen, Groningen, The Netherlands
4 Meurs HRM, Woerden, The Netherlands
5 Department for Personality Psychiatry, Clinic for Mental Health and Addiction, Oslo University Hospital, Oslo, Norway

Key words
item response theory, validity, personality disorder, questionnaire evaluation

Correspondence
Muirne Paap, Department of Neuropsychiatry and Psychosomatic Medicine, Oslo University Hospital, Rikshospitalet, 0027 Oslo, Norway.
Telephone (+47) 23-074-160
Fax (+47) 23-074-170
Email: muirne@nxdomain.nl

Received 19 June 2010;
revised 10 January 2011;
accepted 4 April 2011

Abstract

We used item response theory (IRT) to (a) investigate the dimensionality of the Symptom Checklist-90-Revised (SCL-90-R) in a severely disturbed patient group, (b) improve the subscales in a meaningful way and (c) investigate the measurement precision of the improved scales. The total sample comprised 3078 patients (72% women, mean age=35±9) admitted to 14 different day hospitals participating in the Norwegian Network of Personality-focused Treatment Programmes. Mokken Scale Analysis was used to investigate the dimensionality of the SCL-90-R and improve the subscales. This analysis was theory-driven: the scales were built on two start items that reflected the content of the disorder that corresponds with the specific scale. The Graded Response Model was employed to determine measurement precision. Our theory-driven IRT approach resulted in a new seven-factor solution including 60 of the 90 items clustered in seven scales: depression, agoraphobia, physical complaints, obsessive-compulsive, hostility (unchanged), distrust and psychoticism. Most of the new scales discriminated reliably between patients with moderately low scores to moderately high scores. In conclusion, we found support for the multidimensionality of the SCL-90-R in a large sample of severely disturbed patients.

Introduction

The Symptom Checklist-90-Revised (SCL-90-R) (Derogatis, 1994) is a popular self-report inventory which is widely used as an assessment instrument for psychological distress; it is both used to obtain an estimation of the general symptom level (Global Severity Index, GSI) as well as to assess a variety of dimensions of psychopathology. The 90 items were designed to cover nine different subscales (factors) of psychological distress: somatization

(Som), obsession-compulsion (Obs), interpersonal sensitivity (Int), depression (Dep), anxiety (Anx), hostility (Hos), phobic anxiety (Pho), paranoid ideation (Par), and psychoticism (Psy). Each item is scored on a scale ranging from zero ("not at all") through four ("extremely").

Even though studies have consistently shown high correlations between the SCL-90-R subscales, they have not been consistent with respect to the factorial structure (Dinning and Evans, 1977; Cyr *et al.*, 1985; Brophy *et al.*, 1988; Hafkenscheid, 1993; Holi *et al.*, 1998; Schmitz *et al.*, 2000; Olsen *et al.*, 2004; Arrindell *et al.*, 2006). The way researchers have interpreted the correlations differs as well. Some authors concluded that several of the subscales cannot be distinguished very well from each other due to the high correlations (Cyr *et al.*, 1985; Hafkenscheid, 1993, 2004). In contrast, others claim that the high correlations are a direct and valid result of the high comorbidity between certain disorders, as well as the overlap in symptomatology between specific disorders (Arrindell *et al.*, 2004a, 2004b; Arrindell *et al.*, 2006). Vassend and Skrondal (1999) pointed out that the high correlations among the subscales could be caused by an underlying structure generating factor (dimension) such as negative affectivity (NA). To test this, they used exploratory factor analysis (EFA) to compare the dimensionality for two groups: one group with a low level and one group with a high level of NA. They found eight factors in the low-NA group and only four in the high-NA group. These results demonstrate that the dimensionality of the SCL-90-R is dependent on external variables (such as level of NA).

Although most studies that report on the validity of the SCL-90-R or SCL-90 make use of a form of factor analysis, there are some exceptions. Pedersen and Karterud (2004) investigated the predictive validity of six of the nine subscales: scores on Som should be related to somatoform disorder and panic disorder, Obs to obsessive-compulsive disorder, Int to social phobia, Dep to major depression and dysthymic disorder, Anx to generalized anxiety disorder and Pho to agoraphobia. They found that Derogatis' measure of "caseness" (either a GSI score or two or more subscale scores at or above a *T*-score of 63) functioned well as a screening device for having an unspecified Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition (DSM-IV) axis I disorder. However, although they found some support for the predictive validity of the six investigated subscales (indicated by significant relationships with the associated disorder), the authors concluded that the relationships they found were not strong enough for screening purposes. Additionally, the cut-off scores had only slightly better screening properties than expected by chance for most diagnostic groups.

Only a few studies have been published on the validity of the SCL-90-R that made use of the item response theory (IRT) (Olsen *et al.*, 2004; Elliott *et al.*, 2006). IRT is a collection of mathematical models and statistical methods that has become an increasingly popular approach to the development, evaluation and administration of psychological measures (Meijer and Baneke, 2004; Reise *et al.*, 2005) and offers advantages over classical test theory (CTT) in assessing self-reported screening measures. Using IRT to investigate the internal validity of the Danish version of the SCL-90-R in a community sample, Olsen *et al.* (2004) found that the items belonging to subscales Som, Obs, Int, Dep, Anx and Pho formed a strong unidimensional scale. As is to be expected for a community sample, the mean scores on the subscales were relatively low in this study, ranging from 0.13 (Pho) to 0.63 (Obs). Elliot *et al.* (2006) used the Rasch rating scale model (an extension of the original Rasch model that requires dichotomous data) to enhance the understanding of the strengths and limitations of the SCL-90-R, using two clinical samples. In spite of their results indicating that the SCL-90-R categories advance monotonically from zero ("not at all") through four ("extremely"), the patients did not effectively discriminate between two ("moderately") and three ("quite a bit") in this study. Additionally, the authors concluded that the subscales resulted in quite poor person separation and thus might not be very useful for distinguishing between patient populations. They found one big factor measuring overall clinical distress, with two small residual subscales, measuring depressive motivational deficit and social distress.

In summary, the validity of the SCL-90-R remains unclear. The factorial structure does not seem to be invariant, the relationship between the subscales and their associated diagnoses has not been found sufficient for screening purposes and its ability to distinguish between patient populations is questionable. In this study, we propose an analytic strategy that uncovers the dimensionality of the SCL-90-R while at the same time ensuring that the content of the resulting scales reflects the content of their associated diagnoses. To evaluate the dimensionality (factorial structure), we first perform a confirmatory analysis, followed by an exploratory analysis. The starting-point of the exploratory analysis is based on DSM-IV criteria: two items are chosen per subscale that best reflect the corresponding axis I disorder (if applicable). This is the starting pair, around which the exploratory analysis builds the scale. The items are chosen by the last two authors of this paper, who have extensive experience in

the treatment of clinical patients. The chosen items reflect two distinct aspects of the disorder, if such items are available for the given subscale, thus preventing the resulting scale from becoming too narrow-band (Cronbach, 1954; Egberink and Meijer, 2011). A non-parametric IRT model (Sijtsma and Molenaar, 2002; Meijer and Baneke, 2004) is used to assess the dimensionality and a parametric IRT model (Embretson and Reise, 2000) to assess the measurement precision of the SCL-90-R. We favour IRT over more traditional methods, since it facilitates the following three aims of our study:

(i) creating clinically meaningful scales by entering two items as a starting pair around which the exploratory analysis builds the scale (non-parametric IRT);
(ii) investigating item-functioning given the estimated score on the latent trait (for example depression; both non-parametric and parametric IRT);
(iii) assessing measurement precision: can the scales reliably distinguish patients from each other across different values of the latent trait scale? (parametric IRT).

Because a scale may have different psychometric properties when applied to different populations, we split our sample in two clinically distinct subgroups and investigate whether the dimensionality is different for these two groups. The first group exists of patients with a clinical disorder (CD) only, and the second group of patients diagnosed with personality disorder (PD) in addition to a CD. Typically, behavioural patterns associated with PDs tend to be pervasive across a broad range of personal and social situations (Malt *et al.*, 2003; Pedersen and Karterud, 2010). Theoretically, this could lead to higher correlated answers on the SCL-90-R and a more unidimensional picture in the PD group. If the differences prove to be small, we will propose a scale solution that can be reliably used for both groups of patients.

Materials and methods

Participants

This study used data from patients admitted to 14 different day hospitals participating in the Norwegian Network of Personality-focused Treatment Programmes (Karterud *et al.*, 1998), treated in the period from January 1993 through July 2007. The total group of 3078 patients consisted of two subgroups: one with one diagnosis or several diagnosis on axis I only ($n_1 = 641$), which will be referred to as the CD group, and one with one diagnosis or several diagnoses on axis I as well as on axis II ($n_2 = 2437$), which will be referred to as the PD group. 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).[1]

The majority of the patients were women (72% in both groups) and the mean age was 35 years in both groups [standard deviation (SD) = 9]. In the CD group, 277 (43%) of the patients were diagnosed with one, 226 (35%) with two, and 138 (22%) with three or more axis I disorders. In the PD group, 777 (32%) of the patients were diagnosed with one, 803 (33%) with two, and 857 (35%) with three or more axis I disorders; 1661 (68%) were diagnosed with one, and 776 (32%) with two or more axis II disorders. Further details regarding socio-demographic and diagnostic characteristics are reported by Karterud *et al.* (2003).

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, anonymous database, administrated by the Department of Personality Psychiatry, Oslo (former Ullevål) 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.

Assessment

Prior to the beginning of treatment, patients completed a number of self-report measures, including the SCL-90-R (Derogatis, 1994). The instrument encompasses nine symptom subscales (comprising a total of 83 items) as well as seven additional items. The mean score on all 90 items (including the sevenadditional items) is referred to as the GSI and is widely used as a global index for psychological distress. All patients were diagnosed by means of the Mini International Neuropsychiatric Interview (MINI) (Sheehan and Lecrubier, 1994) for axis I disorders and the Structured Clinical Interview for DSM-III-R/DSM-IV Axis II Personality Disorders (SCID-II) (First *et al.*, 1995; First *et al.*, 1997) for axis II disorders. We refer to Pedersen and Karterud (2004) for more information regarding the diagnostic procedure.

Investigating dimensionality: non-parametric item response theory (NIRT)

To investigate the dimensionality of the SCL-90-R, Mokken's Monotone Homogeneity Model (MHM) was used (Mokken, 1971, 1997). This is a non-parametric item response theory (NIRT) model (Sijtsma and Molenaar, 2002). This model was tested using the software package Mokken Scale Analysis for Polytomous items (MSP5.0) (Molenaar and Sijtsma, 2000).

The basic unit in any IRT model is the Item Response Function (IRF; also known as the Item Characteristic Curve, ICC). In case of dichotomous items, the IRF depicts the relationship between the latent trait θ (x-axis) and the probability of the item being endorsed (y-axis).[2] The term "latent" is used because the trait cannot be observed directly, but can only be inferred from other variables (items in the test). Under the MHM the only demand regarding the shape of the IRFs is that the IRFs be monotone non-decreasing (monotonicity). This means that an increase in θ-level never corresponds with a decreased probability of endorsing the item.

In addition to the assumption of monotonicity, the MHM is based on the assumptions of unidimensionality and local stochastic independence. The second assumption is that the items measure one latent trait only (unidimensionality). The third assumption is that the scale consists of items which the participant approaches in a way that is independent of the previous items (local independence). Together, the assumptions result in a measurement model which can be used to order *respondents* on an underlying unidimensional scale using the unweighted sum of item scores (Sijtsma and Molenaar, 2002; Meijer and Baneke, 2004; Sijtsma *et al.*, 2008; Wismeijer *et al.*, 2008).

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). The value of H_{ij} equals the items' covariance divided by their maximum covariance 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). The H_i values are based on the H_{ij} values, and express the degree to which an item is related to other items in the scale: a high H_i value means that the item distinguishes well between people with relatively low latent trait values and people with relatively high latent trait values. Thus, H is based on the H_i values 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).

First, we performed a confirmatory analysis (option "TEST" in MSP5.0). The nine subscales as defined by Derogatis (1994) were analysed separately. In addition, the GSI was analysed to investigate the unidimensionality of the SCL-90-R. Then, exploratory analyses (option "SEARCH normal" in MSP5.0) were performed. In each analysis, all 90 items were entered. Thus, it was possible that items stemming from one subscale (e.g. Anx) could be clustered with a different subscale (e.g. Dep) in our analyses.

The algorithm that MSP5.0 uses to build one or more scales is called Automated Item Selection Procedure (AISP). If provided with a starting pair, which was the case in our study, the AISP subsequently selects one item from the remaining items that correlates positively with the starting pair, has H_{ij} values (one with each of the two items of the "starting pair") that are larger than the user-specified constant c and maximizes the H value based on all three items together. This procedure is repeated until there are no items remaining that satisfy these conditions. The higher the value of c, the more confidence we have in the ordering of persons by means of their total scale score (Egberink and Meijer, 2011). Following Sijtsma and Molenaar (2002), we ran the AISP repeatedly, starting with a low c value and increasing it with each run. The resulting sequence of outcomes indicates whether the data-set is unidimensional or multidimensional. We refer to Sijtsma and Molenaar (2002, pp. 80–82) for more detailed information about this procedure. The analyses were carried out separately for the CD and PD groups.

Choice of start items

In this study, the start items were chosen by the last two authors of this paper, who have extensive experience in the treatment of clinical patients. The procedure was as follows. First, each rater selected the two items for each scale they personally felt best reflected the content the scale was intended to measure. Subsequently, they compared their choices. For three scales (Int, Anx and Pho), the raters had chosen different start items. Finally, they discussed the reasons behind their choices and jointly made the final decision which two items should be chosen.

For the subscales corresponding clearly with a DSM-diagnosis (axis I), two items were chosen that best reflected the *diagnosis*. The following relationships between subscales and DSM-disorders were assumed in this study: Obs – obsessive-compulsive disorder, Int – social phobia, Dep – major depression and dysthymic disorder, Anx – generalized anxiety disorder, Pho – agoraphobia, Psy – any psychotic disorder. For the remaining scales (Som, Hos, Par), two items were chosen that best reflected the content of the

subscale. The two chosen items showed as little overlap in content as possible, so as to increase the chances of a multi-faceted subscale being formed.

When carrying out exploratory analyses in MSP5.0, one can opt for supplying the program with two starting items, or for letting the program choose two starting items based on the highest H_{ij} values. We performed nine exploratory analyses, each time supplying the program with two starting items stemming from one of the nine subscales as described in the Introduction section. Since the items that AISP selects for a given scale depends on the two starting items, partly different solutions may be expected when different starting items would be provided. This is an important advantage of MSP; it gives the researcher the possibility to make sure that the analysis will result in clinically meaningful scales.

Investigating measurement precision: parametric item response theory (IRT)

We applied the Graded Response Model (GRM) (Samejima, 1996) to assess the measurement precision of the individual items as well as the subscales. The GRM is a parametric IRT model which is suitable for analysing items that have ordered response categories (Hays *et al.*, 2000; Emons *et al.*, 2007). The model was implemented using the software package MULTILOG 7 (Thissen *et al.*, 2003), using program default options.

An important difference between the parametric GRM and the non-parametric MHM described in the previous section concerns the assumptions underlying the shape of the item step response functions (ISRFs). Under a non-parametric model such as the MHM, the only demand is that the ISRFs be monotonely non-decreasing. This means that an increase in θ-level never corresponds with a decreased probability of answering in category m or higher. Under a parametric model such as the GRM, the form of the ISRFs is specified beforehand. In this study a logistic function has been chosen, but other functions, such as the normal-ogive one, can be used as well (Sijtsma and Hemker, 2000). Under the GRM, each ISRF is defined by a slope parameter a (also known as the discrimination parameter) and a location parameter b (also known as "between threshold parameter", in case of polytomous items). The a parameter is related to the H_i coefficient: both reflect the degree to which the item is related to the latent trait (Egberink and Meijer, 2011). Whereas the slope parameter is held constant for all ISRFs belonging to one item, the location parameter is specific for the ISRF (and thus the number of location parameters for one item is equal to $m - 1$, the number of ISRFs for one item). In general, items with a high a contribute most information. The value of the b parameter can be interpreted as the point on the θ-scale at which the probability equals 50% of responding in category m or higher. If the b values for one item are close together, this indicates that the patient is not able to distinguish well between the response categories.

Several other types of curves can be derived from the ISRFs (Sijtsma and Hemker, 2000; Emons *et al.*, 2007). Among these are the option response curves (ORCs; also known as category characteristic curves or category response functions) and information curves. The ORCs depict the probability of responding in a specific response category conditional on θ. There is an ORC for each item category m, and at each value of θ the sum of the m probabilities is equal to 1 (Partchev, 2004). Figure 1 shows an example of the ORCs for two items from the SCL-90-R, item 89 from the Psy scale with a low a value and item 30 from Dep scale with a high a value. Moving from the left (lower values) to the right (higher values) on the θ-scale, it can be seen that for very low θ-values the "not at all" option is most likely to be chosen, for

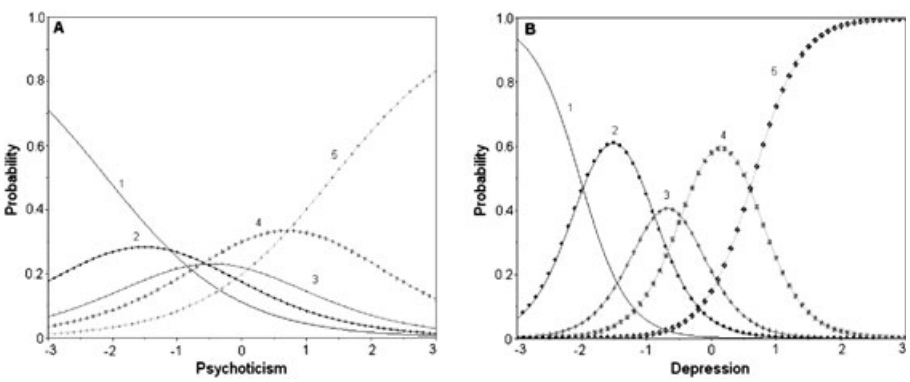

Figure 1. (A) Option response curve (ORC) for item 89 from the Psy scale with a = 1.00. (B) ORC for item 30 from the Dep scale with a = 2.80.

slightly higher θ-values the option "a little bit" and so on. A higher value of *a* implies less overlap between the curves, and thus higher measurement precision (more reliable measurement). The (parametric) IRT equivalent of reliability is item or test *information*. The item information is the inverse of the standard error of measurement, and the measurement error depends on θ (Embretson and Reise, 2000; Meijer *et al.*, 2011). This means that the reliability is not a single estimate such as in Mokken scaling or CTT, but depends on the value of θ (Egberink and Meijer, 2011). The information curve depicts the measurement precision conditionally on θ. Information curves can be generated for each item separately (item information function), as well as for the whole scale (test information function).

The information functions were used to evaluate the subscales found in the exploratory data analyses. Additionally, the *b* parameters were inspected to assess the functioning of the rating scale points.

Results

Missing data: two-way imputation

The total number of cells in the design equals the product of respondents and measured variables, that is $3078 \times 90 = 277{,}020$. Missing data occurred for 1064 (0.004%) cells. We favoured using an imputation method over list wise deletion, since the latter would have implied dropping 20% of the respondents prior to our analyses. We used Two-Way imputation (Bernaards and Sijtsma, 2000), which is a mathematically quite simple method that 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). The advantages of this method are that it is easy to implement using SPSS (van Ginkel and van der Ark, 2005), and the algorithm used is relatively simple. Since the proportion of missing data is small, we argue that the standard errors of the variables will not be substantially affected by the method of imputation. The imputation was done on the whole data-set, not for each scale separately, because we wanted to have complete data for all items, including those that do not belong to a specific subscale. The imputation was implemented using SPSS version 16 for Windows (SPSS, 2007).

Description of the data

Table 1 shows the mean item scores and the mean subscale scores for the two patient groups. Most item means and all subscale means are higher for the PD group. The GSI is also higher for the PD group. The difference in means between the two groups is largest for the interpersonal sensitivity (Int) (difference equal to 0.7) and paranoid ideation (Par) (difference equal to 0.6) scales.

Tables 2 and 3 show the correlations between the subscales of the SCL-90-R as well as some other psychometric properties, for the CD and PD group, respectively. On the whole, the correlations between the subscales were high: five of the nine mean correlations in the CD group and six in the PD group were larger than 0.50. The hostility (Hos) scale had a low mean correlation in both groups (0.33 and 0.37, respectively). When comparing Tables 2 and 3, it can be seen that the correlations for the somatization (Som) and depression (Dep) scales were quite similar for the two patient groups. To the contrary, the correlations for the phobic anxiety (Pho) scale were higher in the PD group. The other scales showed a less clear pattern of differences in correlations between the CD and PD groups.

Results of the non-parametric IRT analyses: dimensionality of the SCL-90-R

Confirmatory analysis

The *H*-value for the GSI, which comprises all 90 items, was lower than 0.3 for both patient groups, which is a first indication for multidimensionality. As can be seen from Table 2, most subscales produced an *H*-value that was at least acceptable ($H > 0.3$), with exception of the psychoticism (Psy) scale ($H = 0.26$) for the CD group. For the PD group, all scales produced acceptable *H*-values (Table 3). For the CD group there were 16 items with $H_i < 0.3$, for the PD group seven items. Note that a low H_i value does not necessarily imply the item is of bad quality. It does imply, however, that the item does not fit in well with the rest of the items in the scale. It thus seems that the existing scales show a better fit for the PD group than for the CD group.

Exploratory analyses

Nine exploratory analyses were carried out, each based on two start items stemming from one of the nine subscales (Som: 1, 42; Obs: 3, 65; Int: 37, 73; Anx: 2, 86; Pho: 50, 70; Dep: 32, 54; Hos: 24, 74; Par: 18, 83; Psy: 7, 90).

The sequence of outcomes generated by AISP at different values of *c* confirmed the multidimensionality of the data. However, the resulting scales were not completely identical to the original ones, with the exception of the Hos scale. Because only minor differences

Table 1. Mean scores for the 90 items, the nine subscales and the Global Severity Index (GSI) for the clinical disorder (CD) and personality disorder (PD) groups separately

Scale/item (number)	CD	PD	Scale/item	CD	PD
Somatization	1.5	1.7	*Depression*	1.8	2.2
Headaches (1)	1.6	1.7	Loss of sexual interest (5)	2.0	2.0
Faintness (4)	1.8	2.0	Low energy/slow (14)	2.3	2.4
Pains in heart/chest (12)	1.0	1.2	Thoughts of ending life (15)	0.5	0.9
Pains lower back (27)	1.5	1.7	Crying easily (20)	1.5	1.7
Nausea (40)	1.7	2.0	Feeling trapped (22)	0.7	1.0
Soreness of muscles (42)	2.2	2.3	Blaming yourself (26)	2.0	2.5
Trouble getting breath (48)	1.0	1.2	Feeling lonely (29)	2.0	2.6
Hot/cold spells (49)	1.4	1.6	Feeling blue (30)	2.5	2.8
Numbness (52)	1.1	1.2	Worrying too much (31)	2.6	3.0
Lump in throat (53)	1.2	1.5	No interest in things (32)	1.8	2.1
Weakness body (56)	1.6	1.9	Hopeless about future (54)	2.3	2.7
Heavy arms/legs (58)	1.6	1.8	Everything is an effort (71)	2.0	2.4
			Feeling worthless (79)	1.7	2.4
Obsessive-compulsive	1.6	2.0			
Unpleasant thoughts (3)	1.9	2.3	*Phobic anxiety*	0.9	1.3
Trouble remembering (9)	1.9	2.2	Afraid on the street (13)	0.6	1.1
Worried about sloppiness (10)	1.3	1.7	Afraid to go out alone (25)	0.7	1.1
Feeling blocked (28)	2.5	2.8	Afraid public transport (47)	1.0	1.4
Doing things slowly (38)	0.8	1.2	Having to avoid things/places/activities (50)	1.4	1.9
Having to double-check (45)	1.1	1.6	Uneasy in crowds (70)	1.2	1.8
Difficulty deciding (46)	1.8	2.4	Nervous when alone (75)	0.9	1.2
Mind going blank (51)	1.8	2.1	Afraid to faint in public (82)	0.6	0.7
Trouble concentrating (55)	2.4	2.8			
Repeating same actions (65)	0.4	0.8	*Anxiety*	1.4	1.8
			Nervousness (2)	2.5	2.9
Interpersonal sensitivity	1.3	2.0	Trembling (17)	0.9	1.2
Feeling critical of others (6)	1.3	1.8	Suddenly scared (23)	1.2	1.6
Feeling shy opposite sex (21)	1.1	1.7	Feeling fearful (33)	1.9	2.3
Feeling easily hurt (34)	2.1	2.6	Heart pounding/racing (39)	1.2	1.6
Others are unsympathetic (36)	1.3	1.9	Feeling tense (57)	2.3	2.6
People dislike you (37)	0.7	1.5	Spells of terror/panic (72)	1.2	1.5
Feeling inferior to others (41)	1.8	2.6	Can't sit still/restless (78)	1.1	1.5
Uneasy when people are watching you (61)	1.4	2.3	Something bad is going to happen to you (80)	1.2	1.8
Self-conscious with others (69)	1.2	1.9	Frightening thoughts (86)	0.6	1.0
Uncomfortable eating/drinking in public (73)	0.9	1.5			

Int. J. Methods Psychiatr. Res. 20(3): e39–e55 (2011). DOI: 10.1002/mpr

Table 1. (Continued)

Scale/item (number)	CD	PD
Hostility	0.5	0.8
Easily annoyed (11)	1.6	2.0
Temper outbursts (24)	0.3	0.7
Urges to harm someone (63)	0.1	0.5
Urges to break things (67)	0.4	0.7
Arguing frequently (74)	0.3	0.7
Shouting/throwing (81)	0.2	0.4
Additional items		
Poor appetite (19)	1.0	1.3
Overeating (60)	1.1	1.4
Trouble falling asleep (44)	2.0	2.3
Awakening early (64)	1.5	1.4
Restless sleep (66)	2.2	2.4
Thoughts of death (59)	1.2	1.6
Feelings of guilt (89)	1.9	2.4
Total scale (GSI)	1.3	1.6

Scale/item	CD	PD
Paranoid ideation	*0.8*	*1.4*
Others are to blame (8)	0.8	1.2
Most people can't be trusted (18)	0.8	1.6
Feeling watched (43)	0.8	1.6
Having beliefs that others do not share (68)	0.6	1.0
Not getting enough credit (76)	1.1	1.6
People will take advantage (83)	0.7	1.4
Psychoticism	*0.6*	*0.9*
Someone can control your thoughts (7)	0.2	0.4
Hearing voices (16)	0.1	0.1
Others knowing your private thoughts (35)	0.3	0.5
Thoughts not your own (62)	0.2	0.5
Feeling lonely with others (77)	1.6	2.2
Thoughts about sex that bother you a lot (84)	0.2	0.4
You should be punished for your sins (85)	0.3	0.6
Something is wrong with your body (87)	1.0	1.3
Never feeling close to another person (88)	1.1	1.6
Something is wrong with your mind (90)	0.8	1.4

Int. J. Methods Psychiatr. Res. 20(3): e39–e55 (2011). DOI: 10.1002/mpr

Table 2. Correlations among the SCL-90-R subscales, Cronbach's alpha (α) and *H*-values based on the confirmatory non-parametric item response theory (NIRT) analysis (clinical disorder group)

	Som	Obs	Int	Dep	Anx	Hos	Pho	Par	Psy
Somatization (Som)	1	0.55	0.43	0.56	0.69	0.31	0.44	0.35	0.40
Obsessive-compulsive (Obs)		1	0.57	0.74	0.57	0.36	0.32	0.48	0.52
Interpersonal sensitivity (Int)			1	0.66	0.53	0.38	0.45	0.67	0.61
Depression (Dep)				1	0.65	0.35	0.30	0.49	0.61
Anxiety (Anx)					1	0.39	0.56	0.43	0.54
Hostility (Hos)						1	0.13	0.46	0.40
Phobic anxiety (Pho)							1	0.22	0.25
Paranoid ideation (Par)								1	0.64
Psychoticism (Psy)									1
Mean correlation	0.46	0.51	0.54	0.54	0.54	0.35	0.33	0.47	0.50
α	0.86	0.83	0.81	0.87	0.85	0.72	0.85	0.72	0.69
H	0.36	0.38	0.35	0.39	0.41	0.42	0.49	0.32	0.26

Global Severity Index (GSI): Cronbach's alpha (α)=0.96, *H*=0.24.

Table 3. Correlations among the SCL-90-R subscales, Cronbach's alpha (α) and H-values based on the confirmatory non-parametric item response theory (NIRT) analysis (personality disorder group)

	Som	Obs	Int	Dep	Anx	Hos	Pho	Par	Psy
Somatization (Som)	1	0.57	0.43	0.56	0.70	0.31	0.54	0.42	0.47
Obsessive-compulsive (Obs)		1	0.59	0.71	0.66	0.38	0.49	0.55	0.60
Interpersonal sensitivity (Int)			1	0.68	0.59	0.36	0.55	0.68	0.63
Depression (Dep)				1	0.68	0.35	0.47	0.56	0.61
Anxiety (Anx)					1	0.37	0.67	0.56	0.62
Hostility (Hos)						1	0.25	0.48	0.44
Phobic Anxiety (Pho)							1	0.43	0.44
Paranoid ideation (Par)								1	0.67
Psychoticism (Psy)									1
Mean correlation	0.50	0.57	0.56	0.58	0.61	0.37	0.48	0.54	0.56
α	0.88	0.83	0.83	0.86	0.86	0.80	0.85	0.75	0.76
H	0.39	0.36	0.38	0.36	0.43	0.46	0.49	0.36	0.32

Global Severity Index (GSI): Cronbach's alpha (α)=0.96, *H*=0.27.

were found between the two sets of scales resulting from the separate analyses for the two clinical groups, we aimed for a final scale solution that could be used for both groups. Note that 60 of the 90 items were kept. The items that were dropped typically had low H_i values. A few items were dropped because they could not be univocally allocated to one specific subscale. Based on the results of the exploratory analyses, we recommend the following:

- Enhancing the Dep (new name Dep+) and Phob (new name: agoraphobia, Ag) scales, by adding several items from other scales.
- Not using the Anx scale as a separate scale, instead placing some of its items in other scales, such as Dep+ and Ag.
- Shortening several scales: Som, Obs and Psy (new names physical complaints; Phy, Obs–, Psy–). To Obs we would like to add one Anx item, to Psy one item of the "additional items" (Add).
- Introducing a new scale: distrust (Dis). This scale exists of several of the items of the Int and Par scales.

The psychometric properties of the seven proposed scales can be found in Table 4.

Table 4. Properties of the seven proposed subscales based on the non-parametric item response theory (NIRT) analyses

		CD[a]	PD[b]	Total group		
Subscale	Items	Mean (SD)	Mean (SD)	Mean (SD)	*H* (range H_i values)	α[c]
Dep+	Dep: 14, 15, 26, 29, 30, 31, 32, 54, 71, 79; Anx: 2, 33; Int: 34, 41; Obs: 28, 55; Psy: 77	2.0 (0.83)	2.5 (0.81)	2.4 (0.84)	0.45 (0.39–0.54)	0.93
Ag	Phob: 13, 25, 47, 50, 70, 82; Int: 73; Anx: 23, 39, 57, 72; Som: 48	1.1 (0.83)	1.5 (0.94)	1.4 (0.93)	0.47 (0.40–0.52)	0.90
Phy	Som: 4, 27, 42, 52, 56, 58	1.6 (0.95)	1.8 (0.99)	1.8 (0.99)	0.45 (0.39–0.51)	0.81
Obs-	Obs: 3, 38, 45, 46, 65; Anx: 86	1.1 (0.73)	1.5 (0.84)	1.5 (0.84)	0.38 (0.33–0.45)	0.74
Hos	Hos: 11, 24, 63, 67, 74, 81	0.5 (0.50)	0.8 (0.77)	1.3 (0.86)	0.47 (0.40–0.52)	0.80
Dis	Para: 18, 43, 83; Int: 36, 37, 61, 69	1.0 (0.74)	1.7 (0.94)	1.6 (0.95)	0.48 (0.40–0.52)	0.85
Psy–[d]	Psy: 7, 35, 62, 85, 90; Extra: 89	0.6 (0.53)	1.0 (0.72)	0.9 (0.70)	0.40 (0.36–0.42)	0.72

[a]Clinical disorder group.
[b]Personality disorder group.
[c]Cronbach's alpha.
[d]For the CD group, two smaller Psy– clusters were found, the first consisting of items 7, 35 and 62 (H=0.43) and the second of 85, 89 and 90 (H=0.45).

Results of the parametric IRT analyses

Seven analyses were carried out, one for each proposed subscale. Since the exploratory analyses resulted in scales that can be used in both groups, the parametric IRT analysis was carried out using a combined data-set, containing both the CD and the PD data. Table 5 shows the estimated discrimination (a) and location (b) parameters for each of the 60 analysed items, and Figure 2 shows the test information function for the seven subscales.

The discrimination parameter typically ranges from approximately 0.5 to 2 (Hays *et al.*, 2000), but numerous clinical studies have reported a values greater than 2.5 and often even values higher than 4.0 (Reise and Waller, 2009). Extremely high a values are undesirable, because they indicate that the construct being measured is conceptually narrow (Reise and Waller, 2009). Looking at the second column of Table 5, one can see that the estimated a parameters are of a reasonable to high magnitude (between 1.00 and 2.83). When inspecting and interpreting the b parameters and test information functions, it is important to keep in mind that it is assumed that (i) θ is normally distributed, with the mean equal to zero and a standard deviation of one and (ii) θ=0 corresponds to the mean for the total group on the subscale being analysed. Inspection of the b parameters for the Dep+ scale showed that most of the items are located left of the mean θ, indicating that most of the items are uninformative about individual differences at the range of the θ scale where a distinction is made between moderately high levels of depression and very high levels. This is reflected in the test information function, which drops sharply between θ values +1 and +2. From a similar inspection of the parameter estimates and test information functions of the remaining six subscales, it can be concluded that most scales discriminate best between patients with moderately low scores to moderately high scores. More specifically, it can be observed that the Obs, Hos, Dis and Psy scales cannot distinguish reliably between patients with no symptoms associated with the specific subscale and those with low scores, nor between those with moderately high scores and very high scores. Like Dep+, the Ag scale functions somewhat better in terms of measurement precision across the range of the latent trait, but cannot distinguish reliably between moderately high scores and very high scores. The Phy scale can only be used to reliably differentiate between persons that suffer "a little bit" and those who suffer "moderately" from physical complaints.

Discussion

When planning this study two questions emerged. First, can we reproduce the original nine-scale solution of the SCL-90-R proposed by Derogatis (1994) in a large sample consisting of severely disturbed patients? Second, if we find that there is room for improvement, what procedure do we follow in order to provide the readers with meaningful recommendations? To answer these questions we used a theory-driven IRT approach.

Table 5. Item parameters for the Graded Response Model

	Slope parameter	Location parameters			
Item	*a*	b_1	b_2	b_3	b_4
Depression+					
2	1.55 (0.06)	−2.98 (0.14)	−1.64 (0.07)	−0.67 (0.04)	0.94 (0.05)
14	1.40 (0.06)	−1.87 (0.08)	−0.85 (0.05)	−0.08 (0.04)	1.22 (0.06)
15	1.15 (0.06)	0.37 (0.05)	1.37 (0.07)	2.22 (0.11)	3.44 (0.18)
26	1.49 (0.06)	−2.03 (0.09)	−0.85 (0.05)	−0.10 (0.04)	1.15 (0.05)
28	1.66 (0.06)	−2.40 (0.10)	−1.24 (0.05)	−0.46 (0.04)	0.78 (0.04)
29	1.66 (0.06)	−1.75 (0.07)	−0.82 (0.04)	−0.13 (0.04)	0.90 (0.07)
30	2.80 (0.09)	−2.01 (0.06)	−0.99 (0.03)	−0.36 (0.03)	0.63 (0.03)
31	1.97 (0.07)	−2.35 (0.09)	−1.26 (0.05)	−0.60 (0.04)	0.51 (0.03)
32	1.60 (0.06)	−1.45 (0.06)	−0.46 (0.04)	0.35 (0.04)	1.55 (0.06)
33	1.82 (0.06)	−1.52 (0.06)	−0.57 (0.04)	0.16 (0.03)	1.28 (0.05)
34	1.62 (0.06)	−2.02 (0.08)	−0.97 (0.05)	−0.22 (0.04)	1.00 (0.05)
41	1.78 (0.06)	−1.70 (0.07)	−0.72 (0.04)	−0.07 (0.03)	0.95 (0.04)
54	1.94 (0.07)	−2.01 (0.07)	−0.92 (0.04)	−0.28 (0.03)	0.74 (0.04)
55	1.49 (0.06)	−2.42 (0.10)	−1.28 (0.06)	−0.38 (0.04)	0.92 (0.05)
71	1.82 (0.06)	−1.58 (0.06)	−0.57 (0.04)	0.11 (0.03)	1.17 (0.05)
77	1.64 (0.06)	−1.47 (0.06)	−0.40 (0.04)	0.29 (0.04)	1.45 (0.06)
79	2.04 (0.07)	−1.31 (0.05)	−0.49 (0.03)	0.06 (0.03)	1.00 (0.04)
Agoraphobia					
13	2.57 (0.10)	−0.03 (0.03)	0.61 (0.03)	1.10 (0.04)	1.79 (0.06)
23	1.69 (0.07)	−0.70 (0.04)	0.05 (0.04)	0.71 (0.04)	1.81 (0.07)
25	2.24 (0.09)	0.04 (0.03)	0.67 (0.03)	1.10 (0.04)	1.72 (0.06)
39	1.34 (0.06)	−0.58 (0.05)	0.30 (0.04)	0.97 (0.06)	2.05 (0.10)
47	2.62 (0.09)	−0.18 (0.03)	0.32 (0.03)	0.73 (0.03)	1.27 (0.04)
48	1.48 (0.06)	−0.30 (0.04)	0.49 (0.04)	1.20 (0.06)	2.34 (0.10)
50	1.95 (0.07)	−0.82 (0.04)	−0.13 (0.03)	0.42 (0.03)	1.32 (0.05)
57	1.05 (0.05)	−2.85 (0.14)	−1.45 (0.08)	−0.41 (0.05)	1.28 (0.08)
70	2.25 (0.08)	−0.76 (0.03)	−0.06 (0.03)	0.45 (0.03)	1.30 (0.05)
72	1.79 (0.07)	−0.60 (0.04)	0.10 (0.03)	0.76 (0.04)	1.74 (0.07)
73	1.77 (0.07)	−0.40 (0.04)	0.30 (0.04)	0.86 (0.04)	1.69 (0.07)
82	1.71 (0.08)	0.51 (0.04)	1.12 (0.05)	1.55 (0.06)	2.24 (0.10)
Physical complaints					
4	1.36 (0.05)	−1.51 (0.07)	−0.37 (0.04)	0.50 (0.04)	1.97 (0.08)
27	1.17 (0.06)	−0.72 (0.06)	0.07 (0.05)	0.71 (0.05)	1.84 (0.09)
42	1.66 (0.06)	−1.26 (0.05)	−0.55 (0.04)	0.00 (0.03)	0.88 (0.03)
52	1.46 (0.06)	−0.26 (0.04)	0.57 (0.04)	1.22 (0.06)	2.40 (0.10)
56	2.80 (0.08)	−0.90 (0.03)	−0.16 (0.02)	0.45 (0.03)	1.30 (0.04)
58	2.62 (0.08)	−0.84 (0.03)	−0.10 (0.03)	0.47 (0.03)	1.38 (0.04)
Obsessive-compulsive–					
3	0.95 (0.05)	−2.26 (0.12)	−1.00 (0.07)	0.07 (0.05)	1.88 (0.11)
38	1.84 (0.06)	−0.25 (0.03)	0.57 (0.03)	1.27 (0.05)	2.15 (0.08)
45	2.64 (0.08)	−0.59 (0.03)	0.20 (0.03)	0.74 (0.03)	1.57 (0.04)
46	1.22 (0.05)	−2.17 (0.10)	−0.80 (0.05)	0.11 (0.04)	1.54 (0.08)
65	1.46 (0.07)	0.66 (0.04)	1.29 (0.06)	1.72 (0.07)	2.45 (0.11)
86	1.02 (0.06)	0.33 (0.05)	1.20 (0.07)	1.98 (0.11)	3.30 (0.18)
Hostility					
11	1.70 (0.06)	−1.55 (0.06)	−0.39 (0.04)	0.44 (0.04)	1.59 (0.06)
24	2.83 (0.11)	0.57 (0.03)	1.12 (0.03)	1.61 (0.04)	2.15 (0.07)

Table 5. (Continued)

	Slope parameter	Location parameters			
Item	*a*	b_1	b_2	b_3	b_4
63	1.67 (0.09)	1.07 (0.05)	1.79 (0.07)	2.36 (0.10)	3.22 (0.17)
67	1.77 (0.08)	0.57 (0.04)	1.25 (0.05)	1.81 (0.07)	2.70 (0.11)
74	1.58 (0.07)	0.58 (0.04)	1.48 (0.06)	2.19 (0.09)	3.20 (0.16)
81	2.79 (0.13)	0.93 (0.03)	1.47 (0.04)	1.99 (0.06)	2.61 (0.10)
Distrust					
18	1.64 (0.06)	−0.67 (0.04)	0.33 (0.04)	0.99 (0.05)	2.03 (0.08)
36	1.67 (0.06)	−1.36 (0.05)	−0.19 (0.04)	0.65 (0.04)	1.84 (0.07)
37	2.34 (0.07)	−0.49 (0.03)	0.33 (0.03)	0.99 (0.04)	1.91 (0.06)
43	2.41 (0.08)	−0.50 (0.03)	0.26 (0.03)	0.84 (0.03)	1.64 (0.05)
61	1.99 (0.06)	−1.28 (0.05)	−0.36 (0.03)	0.18 (0.03)	1.09 (0.04)
69	1.26 (0.05)	−1.43 (0.07)	−0.14 (0.04)	0.76 (0.05)	1.96 (0.09)
83	1.60 (0.06)	−0.39 (0.04)	0.46 (0.04)	1.14 (0.05)	2.19 (0.09)
Psychoticism–					
7	2.79 (0.13)	1.02 (0.03)	1.49 (0.04)	1.90 (0.06)	2.65 (0.10)
35	2.19 (0.09)	0.82 (0.03)	1.40 (0.05)	1.93 (0.06)	2.72 (0.11)
62	2.08 (0.10)	1.04 (0.04)	1.55 (0.05)	2.06 (0.07)	2.74 (0.11)
85	1.31 (0.07)	1.01 (0.05)	1.73 (0.08)	2.31 (0.11)	3.30 (0.18)
89	1.00 (0.05)	−2.10 (0.11)	−0.93 (0.07)	0.01 (0.05)	1.40 (0.08)
90	1.35 (0.06)	−0.41 (0.04)	0.52 (0.04)	1.28 (0.06)	2.42 (0.11)

In order to improve the scales in a clinically meaningful way, two items were chosen (per subscale) that best reflected the syndrome the subscale aimed to measure. These two items formed the starting pair that formed the foundation on which the scale was built.[3] This approach differentiates our exploratory analyses from other exploratory studies, in which clinical meaning and interpretability is typically assessed *after* the analyses have been performed. Before proceeding with statistical modelling, we examined the correlational pattern among the subscales, and found that it was very similar to that found in previous studies; indeed almost identical to the pattern found by Hafkenscheid (1993) almost 20 years ago. This is an interesting finding, because it may indicate that the correlations between the subscales are stable over time (and generalizable). Like Hafkenscheid and many others, we conducted a confirmatory analysis first. Interestingly, we found that most of the scales performed quite well in psychometrical terms. However, the exploratory analyses showed that the existing scales could be improved upon. Our final scale solution included 60 of the 90 items clustered in seven scales: depression, agoraphobia, physical complaints, obsessive-compulsive, hostility (unchanged), distrust and psychoticism. The enormous overlap between Derogatis' anxiety scale and his depression and phobic anxiety scales led us to conclude that the original anxiety scale was not functioning well as a separate scale. Whether this is caused by a very high "real" correlation between feelings of anxiety (generalized anxiety disorder) and depression/phobic anxiety (agoraphobia), or due to a poor construction of the anxiety scale is a question that is difficult to answer with our data. Furthermore, our analyses indicated that Derogatis' paranoid ideation and interpersonal sensitivity scales could be combined into one scale which we labelled "Distrust".

Sixty items were kept in our final scale solution. This solution was based on IRT analyses as well as clinical considerations. If the analyses showed that an item did not cluster with any scale, we chose to "drop" it (not assign it to any subscale). However, some items clustered with many of the scales. In such situations, we also decided to "drop" the item, because it could not be assigned to one of the scales univocally. Several courses of action are available to deal with items that are not included in any subscale, such as: (i) deleting the item from the questionnaire because it is redundant or does not measure what it was designed to measure, (ii) keeping the item, but not using it in the calculation of a scale score, or (iii) reformulating the item and repeating the analyses. The latter could for example be applicable to item 16 (hearing voices). Based on the definitions of the DSM-IV, we

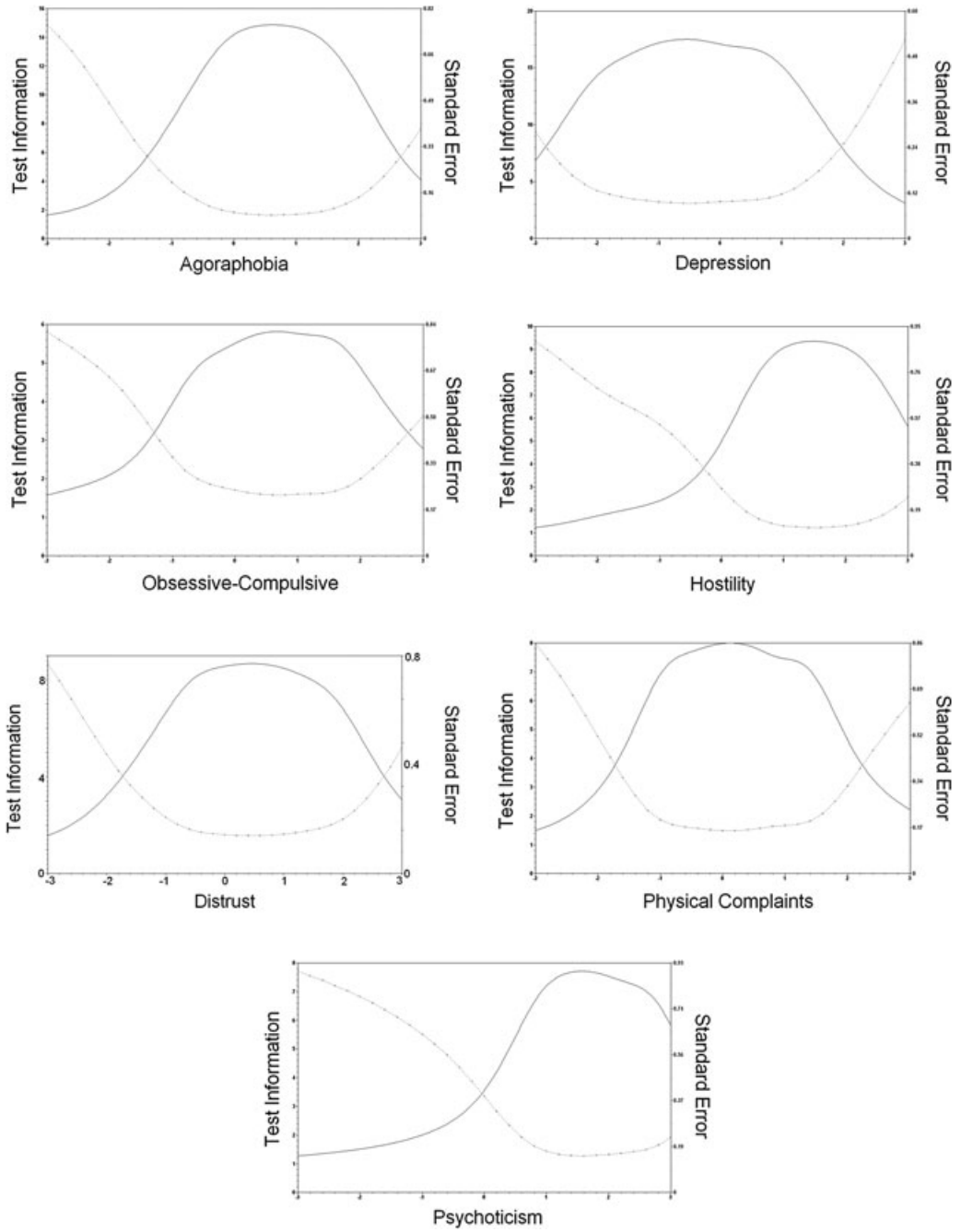

Figure 2. Test information functions for the seven new subscales, with estimated trait score on the horizontal axis, test information (solid line) on the left vertical axis and standard error of measurement (dotted line) on the right vertical axis.

would expect this item to cluster with the psychoticism scale, where it was originally placed by Derogatis. Intriguingly, our analyses showed that this item clustered with the Dep, Anx, Pho and Int scales – though only for the patients with at least one personality disorder. This finding concurs with clinical experience indicating that hearing voices is not necessarily confined to psychotic disorders (Jenner *et al.*, 2008). However, the nature of the psychotic voices is not assessed in the SCL-90-R which might have significant consequences for clinical categorization.

Therefore, we propose that item 16 has to be re-written if it is aimed to tap into psychotic voices only. A new study would be necessary to investigate the most appropriate way to handle the other 29 items that were not included in any of the scales.

Further examination of the seven new scales showed that most of these scales discriminated reliably between patients with moderately low scores to moderately high scores. However, latent trait values of patients that are located on the low end of the scale cannot be estimated reliably and the same holds for the patients located on the high end. This finding is in contrast with many other clinical studies, which have shown that the information (measurement precision/reliability) tends to be highest at the high end of the scale (Reise and Waller, 2009). It is in accordance, however, with the findings of a recent study (Meijer *et al.*, 2011) using data of outpatients, students and prisoners, showing most reliable measurement for average to moderately high scores. This implies that the scales might not detect a clinically meaningful decrease in symptoms as an effect of therapy for patients with very high initial levels of distress.

This study was based on a large sample of severely disturbed patients, with high levels of distress and interpersonal difficulties. The nature of the sample differentiates it from other recent validation studies of the SCL-90-R using IRT, which were either based on a community sample showing little pathology and distress (Olsen *et al.*, 2004), or on small samples of patients with moderate levels of distress (Elliott *et al.*, 2006). Our sample consisting of severely disturbed patients is both a strength and a limitation of our study. It is a limitation, because we were not able to directly compare the results produced by our analytic strategy in this highly distressed group to results in a group with little to moderate distress. It is a strength, because there was a need for validation of the dimensionality of the SCL-90-R in severely distressed patient groups.

When we return to question of dimensionality, we argue that both our findings and the findings of other recent studies offer support for the multidimensionality of the SCL-90-R. However, the conclusions drawn by researchers as to *how many* dimensions there are vary, and seem to depend on several things. First, the results depend on certain sample characteristics. Studies based on low-distress samples have shown support for solutions with only a few factors (Arrindell and Ettema, 1981; Holi *et al.*, 1998; Olsen *et al.*, 2004). This could be a direct result of low variance in these samples. Additionally, structure generating factors (such as NA) have been shown to influence the dimensionality (Vassend and Skrondal, 1999). Second, the researcher's interpretation of the results most likely plays an important role. For example, Schmitz *et al.* (2000) concluded that the nine-factor models and the 10-factor model they tested showed a poor fit. However, Arrindell *et al.* (2004a) reviewed their findings and concluded the opposite. Finally, it might depend on the chosen analytic strategy. Explorative studies have resulted in a range of different factor solutions. In contrast, confirmatory factor analytic studies have found support for Derogatis' factor structure (Arrindell *et al.*, 2004a; Arrindell *et al.*, 2006). Interestingly, these confirmatory analyses have shown almost equal support for the Dutch eight-factor model (Arrindell and Ettema, 1986), Derogatis' nine-factor model, and factor models including higher order factors. Thus, the question arises which model to prefer. In our study, we prevented this dilemma from arising by (a) choosing two core items before hand for each subscale based on clinical theory and (b) running the exploratory Mokken Scale Analysis repeatedly, so that the appropriate lower bound H value was chosen which revealed the underlying dimensionality structure of the data (Sijtsma and Molenaar, 2002).

In conclusion, this study has produced seven new scales that may allow for more reliable discrimination between patients than the old scales. Additionally, our results indicated that the measurement precision may be dependent on the estimated level of distress.

This is of importance to clinicians who are interpreting change scores (treatment effects); they should be aware that confidence intervals around the estimated scale score may vary depending on the *value* of the scale score. More specifically, researchers should be cautious when interpreting scores if the scale scores are in the range with low measurement precision. More research is needed to ascertain whether this scale solution also holds for other patient groups and the general population. We are currently working on a study that explores the dimensionality of the SCL-90-R in two patient groups that report substantially lower overall psychological distress than the sample used in this study.

The DSM-5 is currently under development, and it is being considered whether to enhance the DSM by adding a dimensional adjunct to each of the traditional categorical diagnoses in the DSM (Kraemer, 2007). IRT is an excellent method to create dimensional scales and provides a powerful framework for examining the generality of specific symptoms, and is therefore likely to play an important role in this enhancement (Kraemer, 2007; Kraemer *et al.*, 2007; Paap *et al.*, 2011). One way to create dimensional diagnoses would be to use inventories

which detect a broad spectrum of symptoms, such as the SCL-90-R, and to create an item pool which could subsequently be used to extract dimensional adjuncts for several disorders using the method we described in this paper. We showed that the advantages of IRT are not limited to parametric IRT only, and we recommend that more researchers consider using MSA in future studies involving psychiatric measurement.

Acknowledgements

The authors thank Andries van der Ark of Tilburg University for advising on methods of imputation, and Sonja-Vanessa Schmitz for her assistance in preparing Figure 2. The authors also thank the patients and staff from the following treatment units in the Norwegian Network of Personality-focused Treatment Programmes for their contribution to this study: Department for Personality Psychiatry, Oslo University Hospital; the Group Therapy Unit, Lillestrøm District Psychiatric Centre, Akershus University Hospital; the Unit for Group Therapy, District Psychiatric Centre, 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 Centre; the Unit for Group Therapy, Vestfold Mental Health Care Trust, Tønsberg; the Group Therapy Unit, Alna District Psychiatric Centre, Department of Psychiatry, Aker University Hospital, Oslo; the Årstad Day Unit, Fjell & Årstad District Psychiatric Centre, Bergen; the Bergenhus Day Unit, District Psychiatric Centre, Bergen; the Unit for Group Therapy, Skien District Psychiatric Centre, Telemark Hospital Health Authority; Day Treatment Unit, Furuset District Psychiatric Centre, Aker University Hospital, Oslo; the Group Therapy Unit, Ringerike Psychiatric Centre, Hønefoss; the Outpatient Clinic in Farsund, District Psychiatric Centre, Farsund, and the Unit for Group Therapy, Jessheim District Psychiatric Centre, 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 funding 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.

Declaration of interest statement

The authors have no competing interests.

Notes

1. Of the diagnostic categories investigated in this study, there were no major differences in the criteria sets between the two DSM versions for most disorders; therefore, we do not expect the use of two different versions of the DSM to have any substantial effect on our results.
2. An IRF can still be produced for polytomous data, but is now the sum of the so-called item step response functions (ISRFs). The ISRF could be seen as a special case of the IRF, depicting the probability of answering in category m or higher. Since the probability of answering "at least" in the lowest category is equal to one, we are left with $(m - 1)$ ISRFs for each item. In our case, there were five answering categories, hence the number of ISRFs per item are four.
3. One should be aware that IRT-based item selection procedures tend to result in strong unidimensional scales consisting of items that have similar content, which we already pointed out in Egberink and Meijer (2011). Therefore, arguments that are both content-related and theory-driven should play a key role in scale construction when using IRT modelling.

References

American Psychiatric Association (APA) (1987) *Diagnostic and Statistical Manual of Mental Disorders (3rd edition, revised) (DSM–III-R)*, Washington, DC, APA.

American Psychiatric Association (APA) (1994) *Diagnostic and Statistical Manual of Mental Disorders (4th edition) (DSM–IV)*, Washington, DC, APA.

Arrindell W.A., Barelds 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 peri partum pelvic pain (PPPP) syndrome. *British Journal of Clinical Psychology*, **45**(Pt 3), 377–391.

Arrindell W.A., Boomsma A., Ettema H., Stewart R. (2004a) 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.

Arrindell W.A., Boomsma A., Ettema H., Stewart R. (2004b) 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., Ettema H. (1981) Dimensionele structuur, betrouwbaarheid en validiteit van de Nederlandse bewerking van de Symptom Checklist (SCL-90) [Dimensional structure, reliability and validity of the Dutch edition of the Symptom Checklist (SCL-90)]. *Nederlands Tijdschrift Voor de Psychologie*, **36**, 77–108.

Arrindell W.A., Ettema J.H.M. (1986) Symptom Checklist (SCL-90). *Handleiding bij een multidimensionele psychopathologie-indicator [Symptom Checklist (SCL-90). Manual accompanying a multidimensional psychopathology-indicator]*, Lisse, Swets & Zeitlinger.

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.

Brophy C.J., Norvell N.K., Kiluk D.J. (1988) An examination of the factor structure and convergent and discriminant validity of the SCL-90R in an outpatient clinic population.

Journal of Personality Assessment, **52**, 334–340, DOI: 10.1207/s15327752jpa5202_14

Cronbach L.J. (1954) Report on a psychometric mission to clinicia. *Psychometrika*, **19**, 263–270, DOI: 10.1007/BF02289226

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, DOI: 10.1207/s15327752jpa4906_2

Derogatis L.R. (1994) *SCL-90-R: Administration, Scoring and Procedures Manual*, Minneapolis, *MN, National Computer Systems*.

Dinning W.D., Evans R.G. (1977) Discriminant and convergent validity of the SCL-90 in psychiatric inpatients. *Journal of Personality Assessment*, **41**, 304–310.

Egberink I.J.L., Meijer R.R. (2011) An IRT analysis of Harter's Self-Perception Profile for Children (SPPC) or why strong clinical scales should be distrusted. *Assessment*, **18**, 201–212.

Elliott R., Fox C.M., Beltyukova 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, DOI: 10.1037/1040-3590.18.4.359

Embretson S.E., Reise S. (2000) *Item Response Theory for Psychologists*, Mahwah, NJ, Erlbaum.

Emons W.H.M., Meijer R.R., Denollet J. (2007) Negative affectivity and social inhibition in cardiovascular disease: Evaluating type-D personality and its assessment using item response theory. *Journal of Psychosomatic Research*, **63**, 27–39, DOI:10.1016/j.jpsychores.2007.03.010

First M.B., Spitzer R.L., Gibbon M., Williams J.B. (1995) The structured clinical interview for DSM-III-R personality disorders (SCID-II): Part I. Description. *Journal of Personality Disorders*, **9**, 83–91.

First M.B., Gibbon M., Spitzer R.L., Williams J.B. W., Benjamin L.S. (1997) *The Structured Clinical Interview for DSM-IV Axis II Personality Disorders (SCID-II)*, Washington, DC, American Psychiatric Press.

Hafkenscheid A. (1993) Psychometric evaluation of the symptom checklist (SCL-90) in psychiatric inpatients. *Personality and Individual Differences*, **14**, 751–756, DOI: 10.1016/0191-8869(93)90088-K

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*.

Hays R.D., Morales L.S., Reise S.P. (2000) Item response theory and health outcome measurement in the 21st century. *Medical Care*, **38**, II-28–II-42.

Holi M.M., Sammallahti P.R., Aalberg V.A. (1998) A Finnish validation study of the SCL-90. *Acta Psychiatrica Scandinavica*, **97**, 42–46, DOI: 10.1111/j.1600-0447.1998.tb09961.x

Jenner J.A., Rutten S., Beuckens J., Boonstra N., Sytema S. (2008) Positive and useful auditory vocal hallucinations: prevalence, characteristics, attributions, and implications for treatment. *Acta Psychiatrica Scandinavica*, **118**, 238–245, DOI: 10.1111/j.1600-0447.2008.01226.x

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.

Kraemer H.C. (2007) DSM categories and dimensions in clinical and research contexts. *International Journal of Methods in Psychiatric Research*, **16**, S8–S15, DOI: 10.1002/mpr.211

Kraemer H.C., Shrout P.E., Rubio-Stipec M. (2007) Developing the diagnostic and statistical manual V: What will "statistical" mean in DSM-V? *Social Psychiatry and Psychiatric Epidemiology*, **42**, 259–267, DOI: 10.1007/s00127-007-0163-6

Malt U.F., Retterstøl N., Dahl A.A. (2003) *Lærebok i psykiatri*, Oslo, Gyldendal.

Meijer R.R., Baneke J.J. (2004) Analyzing psychopathology items: A case for nonparametric item response theory modeling. *Psychological Methods*, **9**, 354–368, DOI: 10.1037/1082-989X.9.3.354

Meijer R.R., de Vries R.M., van Bruggen V. (2011) An evaluation of the brief symptom inventory-18 using item response theory or which items are most strongly related to psychological distress?. *Psychological Assessment*, **23**(1), 193–202.

Mokken R.J. (1971) *A Theory and Procedure of Scale Analysis*, The Hague, Mouton.

Mokken R.J. (1997) Nonparametric models for dichotomous responses. In *Handbook of Modern Item Response Theory*, van der Linden W.J., Hambleton R.K. (eds) pp. 351–367, New York, Springer.

Molenaar I.W. (1997) Nonparametric models for polytomous responses. In *Handbook of Modern Item Response Theory*, van der Linden W.J., Hambleton R.K. (eds) pp. 369–380, New York, Springer.

Molenaar I.W., Sijtsma K. (2000) MSP5 for Windows, Groningen, iecProGAMMA.

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, DOI: 10.1111/j.1600-0447.2004.00399.x

Paap M.C.S., Kreukels B.P.C., Cohen-Kettenis P.T., Richter-Appelt H., De Cuypere G., Haraldsen I.R. (2011) Assessing the utility of diagnostic criteria: A multisite study on gender identity disorder. *The Journal of Sexual Medicine*, **8**, 180–190, DOI: 10.1111/j.1743-6109.2010.02066.x

Partchev I. (2004) A visual guide to item response 4theory, Friedrich-Schiller-Universität Jena. http://www.metheval.uni-jena.de/irt/VisualIRT.pdf [Accessed April 2010]

Pedersen G., Karterud S. (2004) Is SCL-90R helpful for the clinician in assessing DSM-IV symptom disorders? *Acta Psychiatrica Scandinavica*, **110**, 215–224, DOI: 10.1111/j.1600-0447.2004.00321.x

Pedersen G., Karterud S. (2010) Using measures from the SCL-90-R to screen for personality disorders. *Personality and Mental Health*, **4**, 121–132, DOI: 10.1002/pmh.122

Reise S.P., Ainsworth A.T., Haviland M.G. (2005) Item response theory. *Current Directions in Psychological Science*, **14**, 95–101, DOI: 10.1111/j.0963-7214.2005.00342.x

Reise S.P., Waller N.G. (2009) Item response theory and clinical measurement. *Annual Review of Clinical Psychology*, **5**, 27–48, DOI: 10.1146/annurev.clinpsy.032408.153553

Samejima F. (1996) The graded response model. In *Handbook of Modern Item Response Theory*, van der Linden W.J., Hambleton R.K. (eds) pp. 85–100, New York, Springer.

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, DOI: 10.1023/A:1008931926181

Sheehan D.V., Lecrubier Y. (1994) Mini International Neuropsychiatric Interview (M.I.N.I.), Tampa, FL/Paris, University of South Florida Institute for Research in Psychiatry/INSERM-Hôpital de la Salpétrière.

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**, 275–90, DOI: 10.1007/s11136-007-9281-6

Sijtsma K., Hemker B.T. (2000) A taxonomy of IRT models for ordering persons and items using simple sum scores. *Journal of Educational and Behavioral Statistics*, **25**, 391–415, DOI: 10.3102/10769986025004391

Sijtsma K., Molenaar I.W. (2002) *Introduction to Nonparametric Item Response Theory*, Thousand Oaks, CA, Sage Publications.

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, DOI: 10.1207/s15327906mbr3804_4

SPSS (2007) *SPSS for Windows*, Rel. 16.0.1, Chicago, IL, SPSS Inc.

Thissen D., Chen W.H., Bock R.D. (2003) *MULTILOG (version 7)*, Lincolnwood, IL, Scientific Software International.

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, DOI: 10.1177/0146621603260688

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, DOI: 10.1016/S0005-7967(98)00182-X

Wismeijer A.A., Sijtsma K., van Assen M.A., Vingerhoets A.J. (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, DOI: 10.1080/00223890802107875