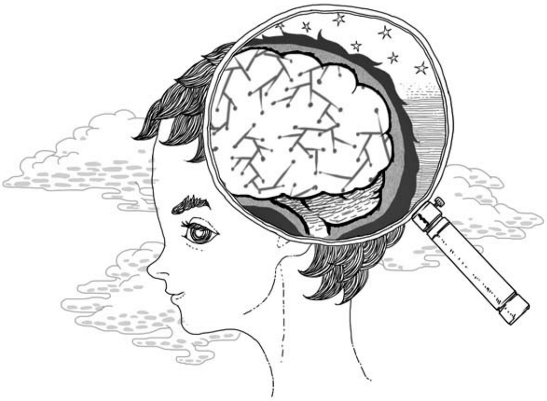
# Examining the validity of the assessment of Gender Identity Disorder

Diagnosis, self-reported psychological distress and strategy adjustment

*Muirne Caitlin Shonagh Paap*



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

Institute of Clinical Medicine, University of Oslo, Norway

*Series of dissertations submitted to the*
*Faculty of Medicine, University of Oslo*
*No. 1073*

ISBN 978-82-8264-014-5

Illustration on title page by Yo Masuda.

Cover: Inger Sandved Anfinsen.
Printed in Norway: AIT Oslo AS.

Produced in co-operation with Unipub.

***To my parents René Stahn and Mitzi Paap***

## Table of contents

## Acknowledgements

This work was conducted at the Department for Neuropsychiatry and Psychosomatic Medicine, Oslo University Hospital, Rikshospitalet, Oslo and at the Institute of Clinical Medicine, University of Oslo, Norway. The research was supported by grant funding from the South-Eastern Norway Regional Health Authority and the Norwegian Research Council.

I thank my research supervisors, Ira Haraldsen and Ulrik Malt, for their constant support and undiminishing faith in my capabilities. I am particularly grateful to Ira Haraldsen for introducing me to the world of scientific research. Ira, your ambition and passion for science have been truly inspiring. Your guidance throughout the writing process has proven invaluable and has contributed substantially to the quality of my work. Most of all, I would like to thank you for your generosity and hospitality and for inviting me to come work with you in this wonderful and beautiful country. I really appreciated your warm welcome at the airport, when I arrived in Oslo for the first time.

I would like to extend my gratitude to several of my dearest colleagues. Torhild Garen, Åse Fløistad, Solveig Pedersen, Heidi Johnsen, Bjørg Høgnes, Torbjørn Elvsåshagen, Ingrid Funderud, Swavek Wojniusz, Frøydis Hellem, Thomas Mengshoel, and Kjersti Gulbrandsen, thank you for making me feel welcome, for the inspiring discussions, for your friendship and emotional support.

I thank Sigmund Karterud and Geir Pedersen for a very fruitful and pleasant collaboration, and for entrusting their database to me. I also thank Griet de Cuypere, Hertha Richter-Appelt and Peggy Cohen-Kettenis, who co-founded the ENIGI-project with Ira Haraldsen, for entrusting their data to me, as well as for their valuable theoretical input during the writing of two of the papers. I thank Baudewijntje Kreukels, Susanne Cerwenka, and Timo Nieder for inspiring discussions and several pleasant evenings.

A special thanks goes to Rob R. Meijer, University of Groningen, for agreeing to co-supervise the writing of two of my papers. Unfortunately, Item Response Theory has yet to be discovered by Norwegian psychometricians, and for this reason I contacted Professor Meijer, who is a Dutch expert in the field. Rob, your support means a lot to me.

Although he may think it sounds a bit corny, I would like to express my gratitude to my beloved *samboer* Jan van Bebber. Jan, thank you for the constructive discussions, your uncensored feedback, your faith in me, and your valuable contribution to one of the papers. It

is yet another test we passed, having ’survived’ working together. Thank you for supporting me in my decision to move to Norway for several years, I know it was not easy for you.

I thank my best friend, Anja M. Jansen, for proof-reading several of my papers, and for brainstorming with me. Anja, I know our friendship will last forever, wherever we are in the world. I extend my thanks to several of my dearest friends, for being such good friends during the last couple of years (either by making my stay in Norway a pleasant one, or by not making me feel too guilty for leaving the Netherlands!): Mattijs Vreeling, Lone Nassar, Nils Sønderland, Daniel Bright, Rikard Tvedby, Xi Zhao, Reynolds Cameron, Aron Paap, Nienke Tolsma, Alice Spruit and Sonja Vanessa Schmitz.

Last but not least, I would like to thank my parents. I thank my mother, Mitzi Paap, for double-checking my spelling, and use of grammar. Mum, thank you for your undying faith in me, from the moment I was born. Thank you for being such a wonderful role-model, for never exerting pressure or pushing me, but for always providing me with inspiration. I thank my father, René Stahn, for always making me feel loved. Dad, thank you for being there when I need you most, for our wonderful philosophical discussions, for being the best dad I could wish for. Thank you for supporting the choices I make. Mum and dad, I would not have come this far without your love, faith and support. Thank you for installing in me a firm sense of self-worth. I cannot thank you enough. I dedicate this thesis to you.

## Preface

In the 1950s, Lee Cronbach (1954) described the immense differences between the ‘countries’ of Clinicia and Psychokometrika. Inhabitants of these countries spoke different languages, adhered to a different set of rules, and had different personality traits. In his paper, Cronbach stressed that, in spite of these differences, it was essential that the two tribes join hands to further scientific knowledge. Many clinicians, statisticians and researchers have followed his advice, and the gap seems to have diminished to some extent. However, it is my observation that the countries of Clinicia and Psychometrika are divided by vast waters still. I chose to train myself in both languages, so as to be an interpreter as well as a bridge builder. The result of my efforts is this thesis which is based on four articles I wrote, to which experts from both fields contributed. The four articles share the same take-home message: the utility, generality, validity and interpretation of diagnostic criteria and test scores should not be taken for granted, but should be continually and carefully monitored. This should logically result in revisions of the criteria/tests themselves and/or their application and interpretation. There are two topics that dominate this thesis, one of which is of clinical and one of statistical nature: Gender Identity Disorder (GID) and Item Response Theory (IRT), respectively.

## List of papers

**I** *Paap, M. C. S.*, Kreukels, B. P. C., Cohen Kettenis, P. T., Richter-Appelt, H., de Cuypere, G. and Haraldsen, I. R. Assessing the Utility of Diagnostic Criteria: A Multi-Site Study on Gender Identity Disorder. The Journal of Sexual Medicine, no. doi: 10.1111/j.1743-6109.2010.02066.x

**II** *Paap, M. C. S.*, Meijer, R. R., van Bebber, J., Pedersen, G., Karterud, S., Hellem, F. and Haraldsen, I. R. A study of the dimensionality and measurement precision of the SCL-90-R using Item Response Theory. Submitted.

**III** *Paap, M. C. S.*, Meijer, R. R., Cohen Kettenis, P. T., Richter-Appelt, H., de Cuypere, G., Kreukels, B., Pedersen, G., Karterud, S., Malt, U. F. and Haraldsen, I. R. Why the factorial structure of the SCL-90-R is unstable: comparing patient groups with different levels of psychological distress using Mokken Scale Analysis. Submitted.

**IV** *Paap, M. C. S.* and Haraldsen, I. R., (2010). Sex-based differences in answering strategy and the influence of cross-sex hormones. Psychiatry Research, 175, 266-270.

## Abbreviations

| | |
|---|---|
| 1-PL | 1 parameter logistic model |
| 2-PL | 2 parameter logistic model |
| 3-PL | 3 parameter logistic model |
| *a* | estimator of the discrimination parameter (parametric IRT) |
| AISP | Algorithm for Item Selection |
| ANOVA | Analysis of Variance |
| Anx | Anxiety (SCL-90-R subscale) |
| AO | Arithmetic Operations |
| *b* | estimator of the difficulty parameter (parametric IRT) |
| C | Control |
| *c* | user-specified constant in MSP5.0 |
| CTT | Classical Test Theory |
| Dep | Depression (SCL-90-R subscale) |
| DIF | differential item functioning |
| DMM | Double Monotonicity Model |
| DSM | Diagnostic Statistical Manual of Mental Disorders |
| ENIGI | European Network for the Investigation of Gender Incongruence |
| ETS | Educational Testing Service |
| FA | Factor Analysis |
| FtM | Female-to-Male |
| GI | Gender Incongruence |
| GID | Gender Identity Disorder |
| GLM | Genderalized Linear Model |
| GRM | Graded Response Model |
| GSI | Global Severity Index (sumscore on SCL-90-R) |
| *H* | scalability-coefficient (used in MSA) |
| Hos | Hostility (SCL-90-R subscale) |
| ICC | Item Characteristic Curve |
| ICD | International Classification of Diseases |
| IIO | invariant item ordering |
| Int | Interpersonal Sensitivity (SCL-90-R subscale) |
| IRF | Item Response Function |
| IRSF | Item Step Response Function |
| IRT | Item Response Theory |
| *m* | number of item categories |
| MHM | Monotone Homogeneity Model |
| MSA | Mokken Scale Analysis |
| MSP5.0 | Mokken Scaling for Polytomous items version 5.0 |
| MtF | Male-to-Female |
| *n* | sample size |
| Obs | Obsessive-Compulsive (SCL-90-R subscale) |
| Par | Paranoid Ideation (SCL-90-R subscale) |
| PCA | Principal Component Analysis |

| | |
|---|---|
| Pho | Phobic Anxiety (SCL-90-R subscale) |
| Psy | Psychoticism (SCL-90-R subscale) |
| P-value | the probability of obtaining a test statistic at least as extreme as the one that was actually observed |
| RLE | Real Life Experience |
| SCL-90-R | Symptom Checklist-90-Revised |
| Som | Somatization (SCL-90-R subscale) |
| SPSS15.0 | Statistical Package for the Social Sciences version 15.0 |
| SRS | Sexual Reassignment Surgery |
| WHO | World Health Organization |
| $X_+$ | total score |
| AA | Arithmetic Aptitude |
| $\alpha$ | level of significance |
| $\theta$ | latent trait |

## Abstract

*Aim* This thesis examines the validity of the assessment of Gender Identity Disorder (GID). More specifically, it scrutinises the utility and generality of the diagnosis itself by investigating whether the symptoms underlying the diagnostic criteria for the diagnosis of GID are interpreted in the same way in four European GID clinics. It also examines whether the level of Gender Incongruence (GI) differs among the clinics and sexes. Second, it scrutinises the dimensionality of the SCL-90-R, a measure used in the diagnostic phase in the four aforementioned GID clinics; this is done in three patient groups: patients referred for personality disorder (PD), depressed patients without either a PD or GID, and individuals referred for GID. Finally, it investigates whether cross-sex hormone therapy in GID patients has an effect on the answering strategy they employ on a math test that is known to show sex differences.

*Results* No differences were found among the four clinics, with respect to the way the symptoms were interpreted. For three of the four clinics, a one scale solution was found. In Amsterdam, two scales were found: severity/persistence and onset/duration. In Ghent and Oslo, higher levels of GI were reported for GID patients than in Hamburg and Amsterdam. The dimensionality of the SCL-90-R was shown to be unstable; our results indicated the dimensionality of the SCL-90-R at least depended on (1) the reported level of psychological distress; (2) sex. Finally, our results indicated that GID males differed from control males with respect to adjustment in answering strategy: control males adapted their strategy over time, resulting in more guessing and more correct answers, whereas this adaptation was not seen in GID males.

*Conclusion* The diagnostic criteria were interpreted in a similar manner in the four clinics. However, the distinction made in Amsterdam between onset/duration on the one hand and severity/persistence on the other hand may lead to differences in diagnostic decisions among the clinics. We recommend that severity and duration be taken into account in the next version of the DSM. Our results suggest that the dimensionality of the SCL-90-R is not stable. We suggest subscale scores should be used with care in patient groups reporting little distress, such as GID patients. Finally, we conclude that even though previous studies have shown that cross-sex hormone treatment does not influence cognitive performance as such, it may still influence other cognitive factors, such as answering strategy and adjustment.

## 1. Introduction

Transsexualism is characterised by a discrepancy between biological sex and gender identification, in spite of hormonal levels that are normal with respect to the biological sex. It is a phenomenon that has been described since antiquity (Heath, 2006). However, cultural attitudes toward transsexualism have varied greatly throughout history. In Western societies, it has long been labelled as a mental disorder by the medical profession; however, in the last two decades, there has been much debate about this, since many transsexuals are not reported to show impairment and distress but are in fact high-functioning individuals (Meyer-Bahlburg, 2010).

Throughout the 20th century, aversion therapies, hormone ’reinforcement’, psychoanalytic therapy and even electroconvulsive shock treatments were employed in an effort to ’cure’ the patient (Benjamin, 1967; Bancroft and Marks, 1968; Callahan and Leitenberg, 1973; Cohen-Kettenis and Kuiper, 1984; Gurney, 2010). In the past few decades, Sexual Reassignment Surgery (SRS) preceded by cross-sex hormone treatment, as a treatment for transsexualism, has been gaining ground, and in many countries transsexuals are now being diagnosed and treated by specialists. Generally, having the diagnosis ‘transsexualism’ (WHO, 1992) or ‘Gender Identity Disorder’ (APA, 1994) is a prerequisite for hormonal and surgical treatment. The DSM-IV criteria that must be fulfilled to receive the diagnosis ‘Gender Identity Disorder’ (GID) in childhood or adulthood are:

**A.** A strong persistent cross-gender identification (not merely a desire for any perceived cultural advantages of being the other sex).

In children, the disturbance is manifested by four (or more) of the following:

- Repeatedly stated desire to be, or insistence that he or she is, the other sex.

- In boys, preference for cross-dressing or simulating female attire; In girls, insistence on wearing only stereotypically masculine clothing.

- Strong and persistent preferences for cross-sex roles in make believe play or persistent fantasies of being the other sex.

- Intense desire to participate in the stereotypical games and pastimes of the other sex.

- Strong preference for playmates of the other sex.

In adolescents and adults, the disturbance is manifested by symptoms such as a stated desire to be the other sex, frequent passing as the other sex, desire to live or be treated as the other sex, or the conviction that he or she has the typical feelings and reactions of the other sex.

**B.** Persistent discomfort with his or her sex or sense of inappropriateness in the gender role of that sex.

In children, the disturbance is manifested by any of the following:

In boys, the assertion that his penis or testes are disgusting or will disappear, or the assertion that it would be better not to have a penis, or aversion toward rough-and-tumble play and rejection of male stereotypical toys, games, and activities.

In girls, rejection of urinating in a sitting position, assertion that she has or will grow a penis, or assertion that she does not want to grow breasts or menstruate, or marked aversion toward normative feminine clothing.

In adolescents and adults, the disturbance is manifested by symptoms such as preoccupation with getting rid of primary and secondary sex characteristics (e.g., request for hormones, surgery, or other procedures to physically alter sexual characteristics to simulate the other sex) or belief that he or she was born the wrong sex.

**C.** The disturbance is not concurrent with physical intersex condition.

**D.** The disturbance causes clinically significant distress or impairment in social, occupational, or other important areas of functioning.

The diagnostic code is based on current age: 302.6 for Gender Identity Disorder in Children and 302.85 for Gender Identity Disorder in Adolescents or Adults. Sexual orientation is used as a specifier for sexually mature individuals: attracted to males, attracted to females, attracted to both, attracted to neither.

### *1.1 Diagnostic phase and Treatment*

The GID clinic (Seksjon for Transsexualisme) at Oslo University Hospital-Rikshospitalet has been evaluating and treating adult patients with GID since 1967. The gender clinic evaluates *all* Norwegian gender reassignment applicants. Yearly, 50-80 adult applicants are referred. During the diagnostic phase, the patient is evaluated by two or more independent senior psychiatrists or psychologists. The mean duration for the diagnostic procedure is approximately twelve months, with monthly visits during this period. After the diagnostic work is finished, the applicant is discussed by a multidisciplinary team, and the members of the team jointly decide whether the applicant is eligible for treatment.

The treatment programme that is currently offered in Norway is in accordance with the Harry Benjamin International Gender Dysphoria Association recommendations, and consists of hormonal therapy and sex reassignment surgery (SRS) (Meyer III, et al., 2002). Before the start of treatment, the patient has to initiate the so-called ‘Real Life Experience’ (RLE); this entails experimenting with the desired gender role in daily life, and finally making the switch fully. Changing the gender role often has an enormous impact on personal and social life; it is of huge importance that the patient is aware of the consequences, and is able to make an ’informed’ decision before embarking on the treatment. The RLE typically precedes cross-sex hormone treatment (at least 3 months). In practice, many patients have already started the real-life experience when they come to the clinic. During hormone treatment the individual is seen every three months by an endocrinologist and by a mental health clinician (psychologist, psychiatrist, or psychiatric nurse). If there are no contraindications after one year of hormone treatment, the individual will be referred for surgery. Psychological follow-up evaluations are offered every sixth month until the last surgery, and three structured follow-up sessions are available up to five years after the last surgery.

### *1.2 Comparability of research findings*

Scientific interest in the phenomenon of transsexualism or Gender Identity Disorder (GID) has increased in recent years, which is reflected in a growing body of international research on this patient group, especially by specialists working in multidisciplinary gender teams (Herman-Jeglinska, et al., 2002; Haraldsen, et al., 2005; De Cuypere, et al., 2007; Gomez-Gil, et al., 2008; Okabe, et al., 2008; Sommer, et al., 2008; Vujovic, et al., 2008). This increase of international publications is of huge importance, since it enables us to critically assess possible cultural factors that interact with the diagnostic process. Furthermore, since transsexualism is such a rare phenomenon, and a so-called ‘gold standard’ against which the diagnosis could be evaluated is lacking, it is of utmost importance that reliable information be published by as many clinics as possible (Kraemer, et al., 2007). This way, a large enough sample can be obtained to yield reliable statistics; providing both the clinical and scientific community with more in-depth, up-to-date and reliable information about the disorder.

So far, international research has shown differences in sex ratio, comorbidity and socio-demographic variables (see Gomez-Gil et al., 2008). Differences between subgroups have also been published; among the investigated grouping variables are biological sex (Kockott and Fahrner, 1988; Herman-Jeglinska et al., 2002; Smith, et al., 2005), onset, and sexual orientation (Blanchard, et al., 1987). The published results have been far from homogeneous.

One major factor stands in the way of performing a ‘study of studies’ (meta-analysis): the lack of comparability of the data between the publishing clinics and countries (Kraemer et al., 2007). Presently, it is practically impossible to diagnose transsexualism on the basis of objective criteria due to a lack of psychometrically sound psychological instruments to measure the disorder (Cohen-Kettenis and Gooren, 1999). Thus, the next-best choice is a diagnosis set by at least one experienced clinician. Indeed, almost all publications state that the disorder was diagnosed according to the latest version of the DSM or International Classification of Diseases (ICD); however, no specifics are given. It is impossible then to know whether consensus about a diagnosis would be reached by two clinicians of different clinics. Unfortunately, the criteria as stated in the DSM and ICD still leave much room for interpretation, and for that reason the reliability of the diagnosis is questionable.

The question about comparability makes interpretation of differences that are found among countries difficult. Are the differences that were reported ‘real’ differences, or were they caused by differences in the diagnostic process and the resulting labelling of patients? The latter could pose a serious problem. Some clinicians may use ‘transsexualism’ and ‘Gender Identity Disorder’ inter-changeably, whereas others may use a more conservative approach where they see ‘true transsexualism’ as a sub-group of GID.[1] The degree to which clinicians take into account information about the sexuality of the patient and onset of the disorder when setting a diagnosis, may also vary. This unspoken, and in some cases maybe even unconscious, labelling makes comparability of patient groups almost impossible.

[1] In the ICD-10, transsexualism is still listed as a diagnosis; however, since its introduction in the DSM-III, the diagnosis of transsexualism has broadened, to eventually become what it is today in the DSM-IV: ‘Gender Identity Disorder’ (GID). The current GID diagnosis encompasses three disorders, which were listed as seperate diagnoses in the DSM-III: transsexualism, Gender Identity Disorder of Childhood and Gender Identity Disorder of Adolescence or Adulthood, Nontranssexual Type.

In 2006, the heads of the GID clinics in Oslo (Norway), Amsterdam (the Netherlands), Ghent (Belgium) and Hamburg (Germany) decided to form a research collaboration called the European Network for the Investigation of Gender Incongruence (ENIGI) (Kreukels, et al., 2010) in order to increase diagnostic transparency and comparability between countries. The main aim was to investigate potential differences in diagnostic 'habits' or interpretation of the classification rules as provided by DSM-IV and ICD-10. The four clinics that are part of the ENIGI initiative use the same diagnostic protocol and assessment battery, enabling more reliable comparisons between the countries. In *Paper I*, which is based on data from the ENIGI, the validity of the DSM-IV diagnosis GID is investigated by examining whether the symptoms underlying the core criteria (A and B) are interpreted in a similar fashion in the four countries.

### *1.3 Psychological distress*

Psychological distress plays a key role in the diagnosis of GID. First, the applicant has to experience severe and persistent distress or discomfort about his or her assigned sex (criterion A). Second, there must be evidence of distress in the clinical, occupational, social or another area of functioning (criterion D). Third, psychological distress pertaining to another disorder than GID should not be too severe, since severe comorbidity could imply that the Gender Dysphoria is actually caused by a different disorder (e.g. Schizophrenia), or interfere with the diagnostic process or treatment (clinical 'rule').

Historically, transsexuals have often been looked upon as severely disturbed persons (Lothstein, 1984). More recent studies have shown, however, that transsexuals show psychological functioning within the non-clinical range (Haraldsen and Dahl, 2000; Seikowski, et al., 2008; Gomez-Gil, et al., 2009). This finding satisfies the aforementioned clinical 'rule' (that severe comorbidity should not be present because it could interfere with the diagnostic process/treatment), but it might be in conflict with the diagnosis a-specific D criterion (that there must be evidence of distress). In fact, there has recently been much debate about the usefulness of the D-criterion in setting the diagnosis of GID, recently. Cohen-Kettenis and Pfäfflin (2010) argue that impairment of functioning is not necessarily associated with gender dysphoria, because many applicants that strongly desire sex reassignment in fact are employed, have relationships, and function well socially.

One of the most frequently used self-report questionnaires to assess psychological distress is the Symptom Checklist Revised (SCL-90-R) (Derogatis, 1994). This questionnaire is also incorporated in the assessment battery used in the diagnostic phase by the ENIGI. The 90 items were designed to cover nine different subscales (factors) of psychological distress: somatization (Som), interpersonal sensitivity (Int), depression (Dep), anxiety (Anx), phobic anxiety (Pho), obsession-compulsion (Obs), hostility (Hos), paranoid ideation (Par), and psychoticism (Psy). Each item is scored on a scale ranging from 0 ('not at all') through 4 ('extremely'). In addition, the Global Severity Index (GSI) can be calculated by taking the mean item score across all 90 items. Studies have consistently shown high inter-correlations between the subscales (Dinning and Evans, 1977; Brophy, et al., 1988; Hafkenscheid, 1993; Schmitz, et al., 2000; Olsen, et al., 2004; Arrindell, et al., 2006), but there has been disagreement about whether the high correlations cast doubt on the multidimensionality of the instrument or not. In *Paper II*, the scale structure of the SCL-90-R is investigated and improved upon, based on a large group of patients referred for personality disorders. In *Paper III*, it is investigated whether the scale structure found in *Paper II* is generalisable to patients with GID as well as patients with depression.

### *1.4 Effects of cross-sex hormone treatment*

The physical effects of cross-sex hormone treatment (such as lowering of the voice and beard growth in female-to-males and breast growth in male-to-females) are well-established. The psychological or cognitive effects have been subject to study as well, but the outcomes are less straightforward and homogenous (Resnick, et al., 1986; Van Goozen, et al., 1995; Slabbekoorn, et al., 1999; Malouf, et al., 2006; Puts, et al., 2008). However, the most recent studies of GID patients have consistently found that GID patients showed a pattern of cognitive performance similar to their biological sex, in spite of current hormonal treatment (Van Goozen, et al., 2002; Haraldsen, et al., 2003; Haraldsen et al., 2005). This is an interesting finding, because research has repeatedly demonstrated gender differences in certain areas of cognition such as language skills, mathematical skills and mental rotation abilities (Torres, et al., 2006). It seems that these differences are established in an earlier stage of life, and cannot be influenced by exposure to sex-hormones later in life. This might imply a so-called organising effect of sex-hormones on cognitive abilities (as opposed to an activating one). Does this imply that the brain or psyche cannot become more 'male' or 'female' as an effect of cross-sex hormone therapy? In *paper IV*, a slightly different angle on sex differences on cognition is taken, and it is investigated whether the response style of GID patients on neuropsychological tests pertaining to mathematics is immune to the effects of cross-sex hormone therapy.

## 2. Aims and research questions

In the daily bustle of clinical practise, the validity of diagnoses and tests are often taken for granted. Many clinicians are interested in research, and do participate in it or at least read up on the latest findings in their field; nevertheless, psychometric studies about for example test validity are often regarded as stuff for statisticians. After all, the clinicians use tests that *were* ”validated” at some point, and are widely used by highly esteemed colleagues. The main aim of the research on which this thesis is based, was to scrutinise that which is taken for granted by many. The starting point was addressing the following fundamental question:

I. How valid or generalisable is the diagnosis of GID itself?

More specifically, is the diagnosis itself usable, and is it interpreted in the same way by clinicians in different clinics or countries? After having determined this, it was investigated whether the SCL-90-R, which is a measure of psychological distress that is widely used all over the world, including in studies reporting about GID, is a valid and useful measure to use. This was evaluated by addressing the following two questions:

II. Can the factorial structure of the SCL-90-R be replicated in a study based on a large sample of disturbed patients, using a theory-driven Item Response Theory approach?

III. Is the scale solution found in the large sample of disturbed patients equally valid for depressed patients and patients with GID?

Recent findings suggest that cross-sex hormone treatment does not impact the overall cognitive performance of GID patients, neither in natal males nor natal females. The purpose of this study was to investigate whether this also holds for answering strategies employed by males and females with or without GID:

IV. Does the answering strategy of GID patients change as an effect of cross-sex hormone treatment on a math test that is known to show gender differences?

## 3. Materials and Methods

Detailed descriptions can be found in the original papers. I will present an overview of the samples, tests and statistics which were used in this study. One particular statistical method will be discussed in greater detail: Item Response Theory (IRT); for the reason that this is still a relatively unknown method in Norway, even though its application in medical and psychiatric settings is becoming increasingly popular internationally.

### *3.1 Design and Participants*

A cross-sectional design was used in the first three papers, and in the fourth a longitudinal design with three measurement-occasions was used. Different samples were used in each paper. In *Paper I*, the sample consisted of new applicants that applied to the GID clinics participating in the ENIGI (Ghent, Hamburg, Amsterdam, Oslo) between January 2007 and March 2009 (n=214, 42% natal female, mean age = 32 ± 12 years). In order to be included in the study, the applicants had to be at least 16 years of age at their first visit, and had to have completed the diagnostic assessment. The total sample used in *Paper II*, comprised 3078 patients (72% female, mean age = 35 ± 9 years) admitted to 14 different day hospitals participating in the Norwegian Network of Personality-Focused Treatment Programs. In *Paper III*, three samples were included. The first corresponds with the sample used in *Paper II*, the second was a sample of new applicants (n=410, 36% natal female, mean age = 32 ± 11 years) that were seen at the four clinics participating in the ENIGI between January 2007 and December 2009, and the third was a sample of depressed patients (n=223, 60% female, mean age = 43 ± 13 years) treated at the Department for Neuropsychiatry and Psychosomatic Medicine at Oslo University Hospital. In *Paper IV*, two samples were used: one consisting of patients that had been referred to the GID clinic in Oslo between January 1996 and December

1998 and had received the diagnosis Gender Identity Disorder (n=33, 64% natal female, mean age =25 ± 6 years), and one consisting of controls (n=29, 52% female, mean age =27 ± 11 years). The control group members were either high school graduates, military recruits from the armed forces, college students or employees of the University of Oslo. They were recruited by advertisement. All participants were tested on three occasions: baseline (T1), 3 months (T2), and 12 months (T3), respectively, after the GID patients had started with hormone treatment. All C females were tested during the first 2 weeks of their menstrual cycle.

### *3.2 Instruments*

In *Paper I*, the validity of the DSM-IV (APA, 1994) diagnosis of Gender Identity Disorder itself was investigated. To make a detailed comparison among the four clinics possible, we operationalised the diagnosis and this resulted in a scoring sheet which existed of 23 items. These items consisted of a combination of a symptom and an 'aspect'. The aspects were: severity, onset, duration, frequency, persistence. The aspects that were applicable to the given symptom were used. For example, it is noted in the DSM-IV that one of the symptoms of the A-criterion is 'a stated desire to be the other sex'; we measured this using four items: 'how strong', 'how persistent', 'since when', and 'how long'. Each item was scored dichotomously with categories 'moderately/mildly' (0) and 'very strong' (1).

In *Paper II* and *III*, the scale structure of the SCL-90-R was scrutinised. The instrument consists of 9 scales that were designed to measure one symptom dimension each (comprising a total of 83 items), and includes 7 additional items. The additional items are only used to calculate the Global Severity Index (GSI; range 0-4), which is calculated by taking the average on all 90 items; the GSI is widely used as a global index for psychological distress. The predefined scales are: somatization (Som), interpersonal sensitivity (Int), depression (Dep), anxiety (Anx), phobic anxiety (Pho), obsession-compulsion (Obs), hostility (Hos), paranoid ideation (Par), and psychoticism (Psy). Each item is scored on a scale ranging from 0 (‘not at all’) through 4 (‘extremely’).

*Paper IV* was based on data collected for a previous study (Haraldsen et al., 2005). A full description of all cognitive tests used in the neuropsychological testing battery can be found there. The two subtests used in *Paper* IV were taken from the factor ‘Reasoning, general’ which is included in the officially distributed “Kit of factor-referenced cognitive tests” by ETS [Educational Testing Service (www.ets.org), (Ekstrom, et al., 1976)]. The factor is based on three subtests, of which we used ‘arithmetic aptitude’ (AA) and ‘arithmetic operations’ (AO), each consisting of 15 items. In the first subtest (AA), the participant has to calculate the answer and select it from 5 alternative answers. In AO, the participant selects the correct arithmetic operation required for a given result (e.g. addition, subtraction).

### *3.3 Statistics*

#### *3.3.1 Item Response theory: Papers I, II, III*

In papers *I*, *II* and *III*, Item Response Theory was used to analyse the data. In all three papers, a form of nonparametric IRT was used (Mokken Scale Analysis; MSA), and in *Paper II* a form of parametric IRT was used as well (Graded Response Model; GRM). In general, parametric IRT is better known than nonparametric IRT; and texts that explain nonparametric IRT often start with referring to parametric models, and discuss where nonparametric ones differ from the parametric ones. However, in my view, it is more logical to start with the nonparametric model. The reason is that the parametric models could be seen as special versions of the nonparametric model, imposing more stringent conditions on the Item

Response Function (IRF). In the following paragraphs I will, therefore, first describe the Mokken Models, and then continue to the parametric Graded Response Model. I will also discuss some differences between Principal Component Analysis (PCA) and MSA.

#### *3.3.1.1 Nonparametric IRT*

One of the most frequently used nonparametric IRT techniques is MSA (Mokken, 1971). Among the many advantages of MSA is that its outcomes are much easier to comprehend than those of parametric models for the inexperienced user. MSA can be performed in Mokken Scaling for Polytomous items (MSP5 for Windows; Molenaar and Sijtsma, 2000), which is a very user-friendly program. Other possibilities also exist, for example performing MSA in the popular free software package *R* (van der Ark, 2007).

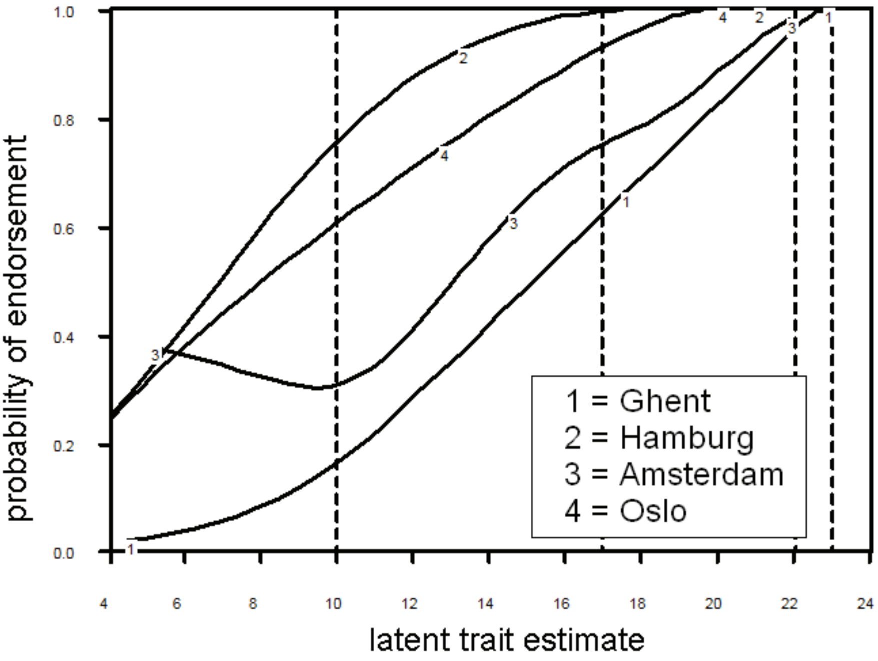
When using MSA in psychiatric or medical research, the data file usually consists of answers that patients gave to a large number of questions (items), and the goal is to detect the underlying dimensional structure of the data. Often, as is the case with PCA, MSA can be used as a tool for *data-reduction*. Traditionally, PCA has been very popular for this purpose in the medical field. Unfortunately, in spite of its popularity, the method is often applied inappropriately. First of all, PCA should strictly speaking be based on tetrachoric or polychoric correlations when the variables (items/questions) are of ordinal or dichotomous nature (which is usually the case!). In MSA, this problem is solved by using coefficients (*H*) that 'correct' interitem covariances for the maximum covariance, given the discrete item-score distributions (Michielsen, et al., 2004; van Abswoude, et al., 2004; Wismeijer, et al., 2008). Second, it is often overlooked that there is a distinction between investigating the factorial structure and proposing a scale solution. PCA is suited for dimensionality analysis, but it is not a measurement model implying useful scale properties; furthermore, PCA always results in as many components as there are items, whether or not these components (scales) are useful (Wismeijer et al., 2008). MSA, on the other hand, is a technique that was specifically designed to discover the underlying dimensional structure of dichotomous or ordinal data. In addition, it provides the user with scales that adhere to a set of criteria, and this allows the resulting scales to be used immediately as a ’safe’ means for rank-ordering the patients on the underlying trait. Another advantage worth mentioning is that the user can influence the analyses on many levels, allowing researchers to make use of their expert knowledge of the content of the items or of the construct being measured. I will return to this issue later; I would like to first discuss the Item Response Function (IRF) and its key role in both nonparametric and parametric IRT.

The basic unit in any IRT model is the IRF (also known as the Item Characteristic Curve, ICC). In case of dichotomous items, the IRF depicts the relationship between the latent trait $\theta$ (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 nonparametric Mokken’s Monotone Homogeneity Model (which I will elaborate on later), the only demand regarding the shape of the IRFs is that the IRFs be monotone non-decreasing (monotonicity). This means that an increase in $\theta$-level never corresponds with a decreased probability of endorsing the item, which is illustrated by the figure on the next page. The figure depicts four IRFs for the item ‘strong conviction that he or she has the typical feelings of the other sex’: each line represents the IRF for one of the four countries participating in the ENIGI. In this case, the latent trait that is estimated is Gender Dysphoria (x-axis). It can be seen that three of the four lines fulfil

[2] In *Paper I*, dichotomous data were analysed. However, in *Paper II* and *III*, the data were polytomous (multiple answering categories). 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 1, we are left with (*m*-1) ISRFs for each item. In our case, there were 5 answering categories, hence the number of ISRFs per item are 4.

the criterion that the IRFs should be monotonely nondecreasing, but the IRF of Amsterdam does not. So for three of the four clinics, one can conclude that the higher the estimated Gender Dysphoria score, the higher the probability that a patient will have a 'very strong conviction that he or she has the typical feelings of the other sex'.

**Figure** The item response functions (IRFs) of the four clinics for item *A4_st* ('strong conviction that he or she has the typical feelings of the other sex'). Item Response Theory (IRT) allows for different IRFs to be created for different groups and to be placed on a common scale. The IRFs show that patients in Ghent and Amsterdam need to score higher than patients in Hamburg and Oslo on the latent trait estimate for this item to be endorsed.



The aforementioned 'Monotone Homogeneity Model' (MHM) was the first model proposed by Mokken (1971). It is based on three assumptions, one of which has already been described (monotonicity). 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). This model was used in *Paper II* and *III*.

In addition to the MHM, Mokken (1971; 1997) also proposed the model of double monotonicity (DMM), in which the assumption nonintersection (also known as invariant item ordering, IIO) is added to the MHM assumptions. The DMM allows for the ordering of respondents, *as well as items* on the underlying scale. When the DMM holds, it also implies the same ordering of items in all subgroups, and, therefore, allows for the investigation of differential item functioning (DIF) or item bias in subgroups (Sijtsma and Molenaar, 2002). In terms of IRFs, the IRFs of different subgroups (in the previously mentioned example the subgroups were the countries) are not allowed to cross under the DMM. This was the model used in *Paper I*, which enabled us to study whether symptoms (items) were interpreted in the same way in the four clinics.

As mentioned previously, MSA makes use of the *H*-coefficient (Molenaar, 1997), which is based on coefficients that 'correct' interitem covariances for the maximum covariance given the discrete item-score distributions. It implies that the coefficients used in MSA are not artificially diminished due to a difference in popularity[3]. In the dichotomous case, if one item is endorsed very frequently, and another item very infrequently, the product-moment correlation is low by definition. When calculating the *H*-coefficient, the pair-wise covariances are modified in such a way that two items with very different popularity can still have a high pair-wise *H*-value. *H*-coefficients can be calculated between item-pairs ($H_{ij}$), on item-level ($H_i$) and on scale-level ($H$). $H_i$ is based on $H_{ij}$, and expresses the degree to which an item is related to other items in the scale: a high $H_i$ value means that the item distinguishes

[3] synonym for probability

well between people with relatively low latent trait values and people with relatively high latent trait values. $H$ is based on $H_i$ and expresses the degree to which the total score ($X_+$) accurately orders persons on the latent trait scale ($\theta$). 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).

In MSP5.0, it is possible to carry out either an exploratory analysis ('SEARCH') or a confirmative one ('TEST'). I mentioned previously that the researcher has several options to influence the analysis. First, when carrying out exploratory analyses in MSP5.0, one can either 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. In *Paper II*, the focus was on finding a scale solution that was well-founded in clinical theory/experience. Hence, we chose to provide the program with 2 start items for each scale. In our opinion, these two items best reflected the construct the scale was aiming to measure, of all 90 items. The algorithm that MSP5.0 uses to build one or more scales is called Algorithm for Item Selection (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.

Another way the researcher can influence the analyses, is by choosing a *c*-value. This is the 'user-specified' constant: the $H$-values of the total scales and $H_i$-value of the item (at the time it enters the analysis) should be equal to or be higher than this value. 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, 2010). Usually, *c* is set at 0.3, but this is by no means obligatory (Sijtsma and Molenaar, 2002). One reason to change this value from the default is when one wants to determine whether one's data are uni- or multidimensional, such as we wanted to do in *Paper II* and *III*. Following Sijtsma and Molenaar (2002), we  ran the AISP repeatedly for increasing values of *c*. The resulting sequence of outcomes indicates whether the data-set is unidimensional or multidimensional. Sijtsma and Molenaar (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.

#### *3.3.1.2 Parametric IRT*

Nonparametric IRT is very useful for detecting the underlying dimensional structure of a data-set consisting of dichotomous or polytomous items, as well as for investigating invariant item ordering. However, some questions cannot be answered by using MSA. One of those questions was asked in *Paper II*, namely, whether the scales could reliably distinguish patients from each other across different values of the latent trait scale. This is referred to as measurement precision (and is related to the concept of reliability).

Parametric IRT models differ from nonparametric ones in that they assume a specified form for the IRF. 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). The form of the IRF is determined by the estimated parameters. The number of parameters (one, two or three) being estimated depends on the model that is chosen. The so-called item-difficulty parameter ($\beta$) is always estimated. This parameter is defined as the score on the latent trait (x-axis) for which the probability (y-axis) is exactly 0.5 that the item (or in that category or higher in case of polytomous items) is endorsed. In the Rasch-model, also known as the one-parameter logistic (1-PL) model, only the item difficulty is estimated. The 2-parameter logistic (2-PL) model extends the 1-PL model by also estimating an item discrimination parameter ($\alpha$). The higher the discrimination parameter, the steeper the slope of the IRF. The parameter indicates how well the item can distinguish between persons with a high score on the latent trait and those with a low score. A higher $\alpha$ indicates that the item can discriminate well between persons with different scores on the latent trait. The equivalent of the $\alpha$ in CTT is the corrected item-total point-biserial correlation (Hays, et al., 2000; DeMars, 2010), and in MSA it is the $H_i$ coefficient; like the $H_i$ coefficient, $\alpha$ reflects the degree to which the item is related to the latent trait (Egberink and Meijer, 2010). One can extend the model further to a 3-parameter model by adding a guessing parameter, which adjusts for the impact of chance on the observed scores.

The model used in *Paper II* is the Graded Response Model (GRM). This is an extension of the dichotomous 2-PL model. The GRM can be used when item responses are of an ordered categorical nature. As in the dichotomous 2-PL model, both the item difficulty as well as the discrimination are estimated per item. But in the polytomous case, one also has to deal with item steps, and thus Item Step Response Functions (ISRFs; see footnote 1). Under the GRM, the discrimination parameter is held constant for all ISRFs belonging to one item, but the location parameter is specific for the ISRF (and thus the number of location parameters for one item is equal $m$-1, the number of ISRFs for one item). In general, items with a high $a$ (estimator of $\alpha$) contribute most information. The value of the $b$ coefficient (estimator of $\beta$) can be interpreted as the point on the $\theta$-scale at which the probability equals 50% of responding in category $m$ or higher. If the $b$'s for one item are close together, this indicates that the patient is not able to distinguish well between the response categories.

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 $\theta$ (Embretson and Reise, 2000; Meijer, et al., 2010). This means that the reliability is not a single estimate such as in MSA or CTT, but depends on the value of $\theta$ (Egberink and Meijer, 2010). The information curve depicts the measurement precision conditionally on $\theta$. Information curves can be generated for each item separately (item information function), as well as for the whole scale (test information function). In *Paper II*, the item and test information functions were used to evaluate the subscales found in the exploratory (MSA) data analyses.

### *3.3.2 Repeated measures: Paper IV*

In *Paper IV*, a repeated measures analysis of variance (GLM repeated measures, SPSS 15.0) was applied. In this model, the number of unanswered items (averaged over the two subtests) was the dependent variable, and time (baseline, 3 months, 12 months) served as the within-subject factor. The reported significance values for the repeated measures ANOVAs were based on the Huynh-Feldt estimator of Epsilon (Huynh and Feldt, 1976), which is to be preferred to the more conservative Greenhouse-Geisser estimator when the estimated Epsilon is above .70 (Stevens, 2002). To correct for capitalisation of chance, the Bonferroni-Holm procedure was used (Shaffer, 1995), which can always be used instead of the classical Bonferroni procedure (Shaffer, 1995; Ekenstierna, 2004), and which is less conservative and therefore more powerful than the simple Bonferroni procedure. When using the Bonferonni-Holm procedure, the *P*-values of the pair-wise comparisons are ordered from low to high, and then $\alpha$ is divided by the total number of comparisons ($K$) for the lowest *P*-value , by $K-1$ for the second lowest *P*-value and so on. When the first non-significant effect in this list of ordered *P*-values is encountered, one stops the procedure and looks no further. This means that for the lowest *P*-value, $\alpha$ is corrected in the same way as when we would use the Bonferonni procedure. But for all other *P*-values, the Bonferonni-Holm procedure results in a larger $\alpha$ and is consequently more lenient than the Bonferonni procedure.

## 4. Summary of papers and results

### *4.1 Paper I*

Paap, M. C. S., Kreukels, B. P. C., Cohen Kettenis, P. T., Richter-Appelt, H., de Cuypere, G. and Haraldsen, I. R. (2010). **Assessing the Utility of Diagnostic Criteria: A Multi-Site Study on Gender Identity Disorder**. Journal of Sexual Medicine, no. doi: 10.1111/j.1743-6109.2010.02066.x

This study presents results from data gathered within the framework of the ENIGI. All new applicants who were seen between January 2007 and March 2009 at the four GID clinics, were at least 16 years of age at their first visit, and had completed the diagnostic assessment (n=214, mean age = 32 ± 12.2 years) were included. Operationalisation and quantification of the core criteria A and B resulted in a twenty-three-item score sheet which was filled out by the participating clinicians after they had made a diagnosis.

The aims of this paper were:

- To investigate whether the symptoms underlying the diagnostic criteria for the diagnosis of GID are interpreted in a similar manner in the four clinics participating in the ENIGI.
- To examine whether the score on the GI scale differs among the clinics, and between the sexes for *(1)* the total group, and *(2)* the patients diagnosed with GID only.

When all data were analysed jointly, only one strong unidimensional scale emerged; ‘the general GI scale’. Neither the checks for monotonicity nor the checks for Invariant Item Ordering (IIO) revealed any deviations when the entire data set was analysed, indicating that the DMM held for the general GI scale. When the data was analysed separately for each country, a one-scale solution was found for three of the four clinics. For Amsterdam, however, two scales emerged from the analysis: one that included the ‘onset’ and ‘duration’ items (‘Amst 1’), and one that included the ‘severity’ and ‘persistence’ items (‘Amst 2’).

When all data were divided into two groups based on birth sex, a one-scale solution was found for both the Female-to-Male (FtM) group and the Male-to-Female (MtF) group. Only two items violated the assumption of equal item ordering in subgroups that is implied by the DMM, when comparing the clinics: ‘strong conviction that he or she has the typical feelings of the other sex’ (item *A4_st*) and ‘persistent conviction that he or she has the typical feelings of the other sex’ (item *A4_pe*). One item violated the DMM model when comparing the sexes: ‘strong belief to be born the wrong sex’ (item *B2_st*).

When considering the data of all applicants, regardless of diagnosis, it was found that the medians of Hamburg, Amsterdam, and Oslo were highly comparable. Ghent’s median was significantly higher than those of the other clinics. Comparison of the sexes revealed that FtMs showed a significantly higher median than MtFs. When only the data of applicants diagnosed with GID were considered, all medians (except the Amsterdam and FtM medians) had a higher value than those for the whole group of applicants. The largest increase was seen for Oslo, accompanied by a decrease in spread. The difference in medians between the sexes diminished, but remained statistically significant.

In the face of our results, we concluded that it might be helpful for clinicians if the severity and duration of symptoms would be taken into account in the next version of the DSM. The distinction between A and B criteria was not supported by our findings and might have to be reconsidered.

### *4.2 Paper II*

Paap, M. C. S., Meijer, R. R., van Bebber, J., Pedersen, G., Karterud, S., Hellem, F. and Haraldsen, I. R., (submitted). **A study of the dimensionality and measurement precision of the SCL-90-R using Item Response Theory.**

This study was conducted using a data-set comprising patients referred for a personality disorder. The total sample consisted of 3078 patients (72% women, mean age = 35 ± 9) admitted to 14 different day hospitals participating in the Norwegian Network of Personality-Focused Treatment Programs. The patients were severely disturbed, and exhibited severe comorbidity.

The aims of this paper were:

- To examine the properties of the existing scale structure of the SCL-90-R in this group of severely disturbed patients.
- To investigate whether a more optimal scale solution could be found, using a theory-driven nonparametric IRT approach.
- To ascertain whether the new scale solution is equally valid for two subgroups of patients: those with and those without a final personality disorder diagnosis.
- To assess the measurement precision of the new subscales.

The H-values of the existing scales were within an acceptable range. However, the exploratory analyses showed that the scale solution could be improved upon. A final scale solution of seven scales was proposed, which was found to be equally valid for both subgroups. In total, 60 of the 90 items were kept. The new scales were: 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. Our finding that measurement precision is dependent on the estimated level of distress should be taken into account when interpreting change scores (treatment effects).

### *4.3 Paper III*

Paap, M. C. S., Meijer, R. R., Cohen Kettenis, P. T., Richter-Appelt, H., de Cuypere, G., Kreukels, B., Pedersen, G., Karterud, S., Malt, U. F. and Haraldsen, I. R., (submitted). **Why the factorial structure of the SCL-90-R is unstable: comparing patient groups with different levels of psychological distress using Mokken Scale Analysis.**

In this study, three samples were used: a sample of severely disturbed patients (n=3078) admitted to 14 different day hospitals participating in the Norwegian Network of Personality-Focused Treatment Programs, a sample of patients with Gender Incongruence (GI; n=410) that were seen at 4 different Gender Identity Disorder clinics participating in the European Network for Investigation of Gender Incongruence and a sample of depressed patients (n=223) treated at the Department for Neuropsychiatry and Psychosomatic Medicine at Oslo University Hospital. The first of the samples was used in *Paper II* as well.

The aims of this study were:

- To answer the following question: is the dimensionality of the SCL-90-R sensitive to the level of psychological distress reported by the patient?
- To investigate the effect of variance in total scores on the SCL-90-R on dimensionality.

A unidimensional pattern of findings was 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. We did not find an effect of variance in total score on the dimensionality. Our analyses suggest that *(1)* differences in variance of SCL-90-R scores are unlikely to have a big impact on the dimensionality, and *(2)* subscale scores in patient groups with low self-reported level of distress, such as GI patients, may be unreliable. Future studies are needed to investigate in what way the main diagnosis and degree of comorbidity impacts the dimensional structure.

### *4.4 Paper IV*

Paap, M. C. S. and Haraldsen, I. R., (2010). **Sex-based differences in answering strategy and the influence of cross-sex hormones.** Psychiatry Research, 175, 266-270.

In this study, somatically healthy male and female GID patients (*n*=33, 21 females, 12 males) were tested at three measurement points: before hormonal treatment, 3 months and 12 months after the start of treatment. Their performance was compared to that of untreated healthy subjects without GID (*n*=29, 15 females, 14 males). The control group existed of high school graduates, military recruits from the armed forces, college students and employees of the University of Oslo. They were recruited by advertisement. The patient group consisted of somatically healthy individuals diagnosed with GID who consecutively sought sex reassignment surgery (SRS) in Norway from 1996 to 1998. The data used for this study were also part of a previously reported study (Haraldsen et al., 2005).

The aim of this paper was:

- To investigate whether hormonal treatment has an impact on the answering strategy that men and women use when being administered a mathematical test.

The results showed that men and women did not differ in the answering strategy used at baseline, in contrast to previous reported findings which indicated that men guessed more than women on mathematical tests. When being retested, however, the guessing tendency of the control males increased when being retested, which was not the case for the control females. The sex differences that were found in this study might impact the calculation of scores based on standardised multiple choice tests, especially arithmetic subtests, and when the interpretation of these scores. This could be particularly relevant when retesting the participant.  The Female-to-Male GID patients resembled the control males, in that they guessed more at each time point; however, this trend was not as outspoken as for the control males. The Male-to-Female GID patients did not adjust their answering strategy at all when being retested. Even though cognitive performance as such may not be influenced by cross-sex hormone treatment, the treatment may still influence other psychological traits, such as answering strategy and adjustment.

## 5. General discussion

### *5.1 Utility and generality of diagnostic criteria*

In *Paper I*, Mokken Scale Analysis (MSA) was used to evaluate whether the DSM-IV diagnostic criteria for GID were used in a similar fashion in the Gender Identity clinics in Ghent (Belgium), Hamburg (Germany), Oslo (Norway) and Amsterdam (the Netherlands). In addition, it was investigated whether the criteria were used differently when diagnosing natal males (MtF) and females (FtM). To make the comparisons possible, the diagnostic criteria were operationalised and quantified on item-level, and an item-analysis and a scale-analysis were conducted. This was followed by comparing the average total score among clinics and between sexes.

Many authors have stressed the advantage of Item Response Theory (IRT) analyses for detecting possible item bias (Doolittle and Cleary, 1987; Santor, et al., 1994; Hartung and Widiger, 1998; Embretson and Reise, 2000; Gierl, et al., 2003; Reise, et al., 2005; Jane, et al., 2007; Uebelacker, et al., 2009; Weinstock, et al., 2009), since it accounts for the potential confounding effect of the value on the latent trait (here: Gender Dysphoria) when evaluating group differences. Most of them had parametric IRT in mind when raising this point. In this study, we illustrated the usefulness of Nonparametric IRT (MSA) for detecting potential item bias, as well as for scale-analysis purposes.

#### *5.1.1 The 'general GI scale'*

Our results indicated that most criteria were free from cultural and gender bias, and that the GID criteria were largely interpreted in the same way in the four clinics participating in this study. However, clinicians participating in the study had trouble interpreting the sub criterion 'conviction that he or she has the typical feelings of the other sex', which was expressed in differential item functioning for two items pertaining to this criterion. When analysing all data regardless of subgroup-membership, only one scale emerged, comprising the diagnosis-specific criteria A and B (the ‘general GI scale’). This one-scale solution was also found for three of the clinics when the data were analysed by clinic, and for both sexes when the data were analysed by sex.

### *5.1.2 A two-scale solution for Amsterdam*

In Amsterdam, a two-scale solution was found: one scale consisted of all duration and onset items, and the other scale consisted of all strength and persistence items. This difference in scale solutions is of high clinical importance. It could lead to differences in diagnostic decisions among clinics: in Amsterdam, an applicant could still receive the diagnosis when symptoms are very severe and persistent but of relatively recent onset, whereas this seems less likely to happen in the other clinics. A possible explanation could be that Dutch patients present themselves differently than other patients. It could, however, also mean that Dutch clinicians have a different way of diagnosing than clinicians in other countries.

### *5.1.3 Cultural and sex differences in levels of Gender Dysphoria*

We found that patients diagnosed with GID in Ghent or Oslo have higher Gender Dysphoria scores than those in Amsterdam and Hamburg. This could point towards differences in diagnostic thresholds; maybe patients need to have more severe symptoms in Ghent and Oslo in order to receive the diagnosis of GID than in Hamburg and Amsterdam. However, the Gender Dysphoria score for all applicants regardless of diagnosis was much higher in Ghent than in Oslo. This could point towards differences in pre-selection between the clinics in Ghent and Oslo. Therefore, it is unclear whether the high threshold in Ghent is attributable to a referral bias of applicants in Flanders, or whether it reflects a systematic difference in judgment between the two groups of clinicians in Ghent and Oslo.

The finding that the Gender Dysphoria score for GID patients in Oslo was so much higher than the score for all applicants regardless of diagnosis, in combination with the observation that only 44.1% of the total patient group received the diagnosis (versus 83.3%–97.6% in the other clinics), could point towards a more 'conservative' view of GID in Oslo; and the low spread in scores for applicants diagnosed with GID in Oslo could reflect a narrower interpretation of the GID criteria than in the other clinics. However, the percentages cannot be compared directly among the clinics. In Oslo all applicants went through the first part of the diagnostic phase (6 months) and as a consequence the diagnostic scoring sheet is filled out for almost all of them. In the other clinics, some applicants were referred elsewhere or dropped out of the diagnostic process at an earlier stage. As a result, no diagnostic data are available for those patients.

Our results show that there were more MtF applicants than FtMs (and the number we found may even be an underestimation, since the DIA was not filled out for the applicants that were turned away within the first six months in three of the four countries). However, a larger percentage of FtMs received the GID diagnosis and FtMs had, on average, a higher score on the Gender Dysphoria scale. We only found one item to be gender biased on the basis of our analyses 'strong belief to be born the wrong sex'. Given the same average score, FtMs had a higher probability of having endorsed this item than MtFs.

### *5.2 An examination of the validity and utility of the SCL-90-R*

Our examination of the validity and utility of the SCL-90-R was done in two steps. First, a clinically meaningful scale solution was sought, that would improve on the current 9 scale structure. This was done using a large, severely disturbed patient group. Second, it was investigated whether the new scale solution was usable for depressed and GI patients, as well.

#### *5.2.1 A clinically meaningful scale solution*

When we were planning *Paper II*, we asked ourselves two questions: (1) Is the SCL-90-R unidimensional or multidimensional? (2) If we find that there is room for improvement, what procedure do we follow in order to provide the readers with meaningful recommendations? We reached the following conclusion: the starting point of the search for answers should be a clinical one. Conveniently, this is possible in the Mokken scaling software package MSP5.0 (Molenaar and Sijtsma, 2000) we used for investigating the dimensionality of the data. When carrying out an exploratory analysis, the user can identify two starting items. In order to improve the scales in a clinically meaningful way, two items were chosen that best reflected the syndrome the subscale aimed to measure. By taking this angle, this study distinguishes itself from studies using exploratory analyses, in which clinical meaning and interpretability is typically assessed *after* the analyses have been performed.

Even though the confirmatory analysis indicated that the predefined scale solution was usable, 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. Furthermore, our analyses indicated that Derogatis'

Paranoid Ideation and Interpersonal Sensitivity scales could be combined into one scale which we labelled ‘Distrust’. Most of the excluded items were dropped, because they did not cluster well with any of the scales. However, a few of the items were dropped for the opposite reason: they clustered with several of the subscales.

Our parametric IRT analyses showed that most of the new 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 could not be estimated reliably and the same holds for the patients located on the high end. This implicates that the scales may not detect a clinically meaningful decrease in symptoms as an effect of therapy for patients with very high initial levels of distress. This finding is in contrast with many other clinical studies, which have showed 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., 2010), showing most reliable measurement for average to moderately high scores.

### *5.2.2 An explanation for the dimensional instability of the SCL-90-R*

The original 9-scale solution (Derogatis, 1994) remains controversial to this day (Holi, et al., 1998; Vassend and Skrondal, 1999; Schmitz et al., 2000; Olsen et al., 2004; Arrindell et al., 2006; Elliott, et al., 2006; Hafkenscheid, et al., 2007; Paap, et al., submitted). In *Paper III*, we attempted to explain the inconsistent findings in the literature; at the same time we wanted to investigate whether the SCL-90-R was usable for GID patients in any form.

Our results showed that the SCL-90-R was unidimensional, when analysing the data from the Gender Incongruence sample. This would seem to be in conflict with the findings of *Paper II*, which lent support for the multidimensionality of the SCL-90-R. However, we believe that these conflicting results may be related to the difference in the reported psychological distress in the two samples. Another variable that influenced the outcomes in our study was birth sex. This was only the case for the patients in the depression sample, however; the depressed males demonstrated a dimensional structure that was highly similar to that of the severely disturbed group, whereas the depressed females resembled the GI patients, in that they interpreted the SCL-90-R 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, it demonstrates that finding factorial invariance for sex in one patient group is not necessarily generalisable 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. Other researchers have speculated that the lack of factorial invariance of the SCL-90-R may be explained by a different variance in scale scores in different samples. Our analyses suggest that differences in variance of SCL-90-R scores are unlikely to have a big impact on the dimensionality.

Our results suggest that subscale scores may be unreliable in patient groups with low self-reported level of distress, such as GI patients, but total scores (GSI) can be reliably used. Whether the low level of reported psychological distress reflects the real level of psychological distresses experienced by the patient, is another question. It is conceivable, for instance, that GI patients feel the need to portray themselves as more stable than other patients, and that the reported scores, as a consequence, are an under-estimation of their experienced distress. They might feel that people in their surroundings (family, friends, but also professionals from the GI clinic) will accept their 'condition' more readily if they seem to be doing 'fine' otherwise. It is a well known problem that many GI patients 'adopt' biographical stories of others (so-called 'identity work'; Schrock and Reid, 2006), because they think these contain all the elements the clinician wants to hear (and will grant them access to SRS). As mentioned previously, it is indeed true that (severe) comorbidity is seen as a contra-indication by many clinicians. GID patients being less willing to see themselves as psychiatric patients than, for example, patients with a personality disorder or depression might also lead to GID patients downplaying their psychological problems.

In a paper which is currently in progress, we indeed found evidence in the brains of GID patients that they were extremely stressed. In this study, atrophy of the hippocampus and cerebellum was shown, which is a pattern also commonly seen in patients with Post Traumatic Stress Disorder. These findings indicate that a self-report questionnaire such as the SCL-90-R is probably an insufficient (and maybe altogether invalid) measure for psychological distress in this patient group.

### *5.3 Cross-sex hormone treatment cognition: a different approach*

Research has shown that men are inclined to take more risks than women. Ben-Shakhar and Sinai (1991) showed that this also applied to the type of strategy used by males and females when taking a standardised test: females showed a higher level of omission rates than males, implying a more conservative answering strategy for the females versus a bolder one for the males. In this study, we wanted to see if men would guess more readily than women, and if this difference would be immune to cross-sex hormone treatment in GID patients.

#### *5.3.1 No sex differences at baseline*

In our study, we did not find a significant difference in answering strategy between control males and females at baseline. This is in conflict with previous research findings that showed differences in risk-taking behaviour (in a testing situation that was both novel and measuring math ability) between males and females (Ben-Shakhar and Sinai, 1991; McGillicuddy-De Lisi and De Lisi, 2002) as well as the ‘novelty versus familiarity hypothesis’ (Kimball, 1989).

These studies were not conducted in Scandinavia, however. In the majority of the world's countries, the stereotype exists that men are better at maths than women, in spite of their equal abilities in the class-room. When this negative stereotype is activated in some way, it has a detrimental effect on the math performance of women, as studies have shown (Inzlicht and Ben-Zeev, 2000; Keller and Dauenheimer, 2003). It has been proposed that gender differences in general may be smaller in the Scandinavian countries; since there is a strong cultural emphasis on gender equality (see Eriksson and Lindholm, 2007). This attitude of 'gender equality' would also translate into weaker stereotypes about sex-differences in math-ability; this was illustrated by Brandell et al. (2005) who conducted a survey among grammar school students in Sweden, and found that only a subgroup of the students (males who had chosen the 'natural sciences' programme) felt that mathematics was a 'typically male' subject. As a consequence, Scandinavian women may not be subjected to stereotype threat as frequently as women from other countries (Inzlicht and Ben-Zeev, 2000; Keller and Dauenheimer, 2003). Indeed, in a Swedish study by Wester and Henriksson (2000), no support was found for the idea that males are more inclined to guess than females. This finding, as well as the lack of differences at baseline in our study may be interpreted as an influence of culture on feelings of self-confidence. It could be argued, that 'even' males need a confidence boost when a testing situation is very new to them (c.q. first measurement). It follows, then, that in Norway males do not experience  the 'advantage' of the stereotype that males excel at mathematics; Presumably, they do not get the confidence boost that might otherwise have produced a bold strategy.

### *5.3.2 Retesting the controls*

Interestingly, our findings did indicate that Norwegian males become more confident when being retested. We found that males adjusted their answering strategy in a bold direction over time, whereas females did not. The decline in null answers for the males was accompanied by a slight increase in false answers and a larger increase in correct answers. This seems to suggest that the increasingly daring strategy of the males pays off; their adjustment is beneficial to their total score on the tests. We propose that this finding indicates that men feel most confident when in a testing situation that is out of the ordinary, but not completely new to them. Women seem to be immune to this effect, and stick to a relatively conservative answering strategy. We recommend examiners, as well as researchers and testing psychologists, to take this sex difference in guessing strategy (and the change herein) into account when calculating scores based on standardized multiple choice tests, especially when it contains arithmetic subtests, and when interpreting these scores. This could be particularly important when retesting the participant.

### *5.3.3 Retesting the GID patients after start of cross-sex hormone treatment*

GID females did not show a pattern that differed from C females. Surprisingly, male GID patients undergoing cross-sex hormone treatment did not adjust their guessing strategy at all; after twelve months of cross-sex hormone treatment, the male GID patients guessed significantly less frequently than the C males. This is an important finding because it indicates that even though it has been shown that cross-sex hormone treatment does not result in an absolute change in cognitive performance, hormonal treatment may still have an important impact on other psychological traits that indirectly impact performance or an adjustment herein, and which have been shown to differ between men and women, such as risk taking behaviour and feelings of self-confidence. In fact, one study that did focus both on psychological traits as well as cognitive performance, showed that the trait ‘anger proneness’ changed as an effect of cross-sex hormone treatment (Van Goozen et al., 1995).

## 6. Conclusions, implications and recommendations

The main conclusions of this thesis are:

- The symptoms pertaining to the DSM-IV diagnosis of Gender Identity Disorder were interpreted largely in the same way in the Gender Identity clinics in Ghent, Hamburg, Oslo and Amsterdam. The diagnosis seems valid and generalisable.
- A subdivision of diagnosis-specific symptoms in an A and B criterion is superfluous.
- The scale solution of the SCL-90-R proposed by Derogatis is not optimal. We proposed a superior scale solution, using 60 items and resulting in 7 scales which can be used in clinical practice.
- Most of the new scales discriminated reliably between patients with *moderately low* scores to *moderately high scores*. Thus, the measurement precision depended on the level of distress measured by the given subscale.
- The dimensionality of the SCL-90-R was not found to be invariant for sex nor differences in the level of self-reported distress as measured by the GSI.
- 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.
- Norwegian healthy males not receiving hormone treatment showed an adjustment in answering strategy on a math test when retested: they guessed more at T2 and T3.
- GID males receiving hormone treatment did not show an adjustment in guessing tendency over time.

### *6.1 Clinical implications*

The Diagnostic and Statistical Manuals of Mental Disorders (DSMs) have been used as a guideline for the diagnosis of psychiatric disorders for decades in the United States and in other parts of the world. In the absence of definitive standards in diagnosing GID, the DSM is the next best thing to a 'gold standard': it summarises what is known about disorders from clinical experience and past research, and what has been learned about the disorders from current research (Kraemer et al., 2007). However, in contrast to a 'gold standard', the DSM is subject to change. Over the years, shifts have been made from criteria that were based on clinical consensus only towards diagnostic criteria that had to be descriptive, explicit, and rule-driven, so that the diagnostic assessments could be conducted more reliably; and from solely relying on the opinions of 'experts' to using empirical evidence as a diagnostic basis (Wilson, 1993; Kraemer et al., 2007). These shifts reflect the ongoing efforts in the field of psychiatry to "lift itself up by its bootstraps" (Kraemer et al., 2007): researchers use diagnostic rules based on clinical experience to define their research population and then draw a sample reflecting that population; and clinicians need researchers to help refine existing diagnostic criteria, based on the most recent and reliable empirical evidence.

Since the DSMs are used in place of a 'gold standard' in both clinical work and research, it is of the utmost importance that the reliability, validity, generality, and utility of DSM diagnoses and their underlying criteria be investigated continuously. Indeed, many articles published over the years have been directed toward investigating the reliability and validity of DSM-based diagnoses. Studies on the reliability of several psychiatric diagnoses revealed that diagnostic disagreement could be due to a number of factors, such as differences in symptom interpretation, threshold severity, misinterpretation of the DSM-rules, interviewer error, change in applicant status or applicant report, presence or absence of comorbidity, and presence or absence of behavioural symptoms (Chorpita, et al., 1998).
In 2006, the heads of the GID clinics in Oslo (Norway), Amsterdam (the Netherlands), Ghent (Belgium) and Hamburg (Germany) decided to form a research collaboration, with the aim to investigate potential differences in diagnostic 'habits' or interpretation of the classification

rules as provided by DSM-IV and ICD-10. The reason for this was that studies involving patients with GID were inconsistent with regard to outcomes, and difficult to compare due to vague descriptions of the diagnostic process. One of the first differences that we discovered was that not all clinics employed the same classification (ICD or DSM). Since we aimed at comparability of diagnostic decisions, we decided to focus on one classification: DSM-IV. DSM-IV is the classification most commonly used in studies reporting about patients with GID. The fact that the participating clinics had not been using the same classification underlines that "the majority of the current diagnostic criteria are still provisional" (Jablensky, 2009) and, in our view, highlights the necessity of multi-site studies. Currently, a new version of the DSM is in the pipeline. Based on our findings, we would like to give the GID working group several recommendations. First, it might be helpful for clinicians if the severity and duration of symptoms be taken into account in the next version of the DSM. Second, the distinction between A and B criteria was not supported by our findings, and might have to be reconsidered. Third, clinicians who participated in our study had trouble interpreting the sub criterion 'conviction that he or she has the typical feelings of the other sex', which was expressed in differential item functioning for two items pertaining to this criterion. This could be a reason to remove or rewrite this criterion in the next DSM.

### *6.2 Recommendations for future research*

#### *6.2.1 Paper I*

This study was foremost aimed at describing the way diagnoses are made in four European GID clinics, so as to create more transparency and shed light on our clinical decision-making. The next step would be reaching a cross-cultural consensus of how aspects such as onset and severity of symptoms should be weighed when reaching a diagnosis. To enable such a process, worldwide data-collection that takes severity and duration of the GID symptoms into account is needed.

IRT is very useful for assessing whether the discriminative validity of items vary between males and females at the same level of GID, but the absence of item bias does not imply that the criteria themselves are equally valid for both sexes (Hartung and Widiger, 1998). It might be conceivable that GID (as any other disorder) expresses itself slightly differently in males and females, and that this is the cause of differences found in scores as well as prevalence/incidence. Future studies directed at assessing male-female differences with respect to the strength of the relationship between diagnostic criteria and external validators, such as treatment outcome, would be particularly useful in elucidating this issue (Hartung and Widiger, 1998).

### *6.2.2 Paper II & III*

We found that sex and level of psychological distress 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 generalisable to other patient groups or whether it is typical for depressed patients with moderate levels of psychological distress.

An MRI-study we recently conducted involving adolescents and young adults diagnosed with GID indicated that these patients were much more distressed than their SCL-90-R-scores indicated. This illustrates the importance of combining sound statistical techniques to calculate the reliability and internal validity of a scale with clinical studies to establish its construct validity. We hope more studies will focus on at combining different sources of information when investigating the validity of self-report inventories. Future studies are also needed to provide more information about the exact role of stress in GID. It is, for example, an interesting question whether the brain atrophy we found are a consequence of GID or whether they played a role in its onset.

### *6.2.3 Paper IV*

Our results were quite surprising with regard to the latest research findings on cognitive performance in GID patients, which interpreted the finding that cross-sex hormone treatment is not accompanied by a change in overall cognitive performance in GID patients as evidence for a resemblance of male and female GID patients to participants of the same biological sex (Van Goozen et al., 2002; Haraldsen et al., 2005; Sommer et al., 2008). Our study indicates, to the contrary, that the lack of change in performance of GID males receiving cross-sex hormone treatment may be seen as evidence for them *not* resembling their biological sex, since we found that C males guessed more when being retested; an adjustment in strategy which benefited their performance. Pinpointing which traits are subject to change as a result of hormonal treatment would be useful for psychologists who can then prepare the patients for these changes as well as guide the patients through them. We hope that future research will further elucidate the potential influence of cross-sex hormone treatment on personality traits.

The nature of our design did not permit us to draw firm conclusions as to whether it is the condition (GID) itself or the hormonal treatment that caused the observed differences between GID and C males. A study that would include a control group of GID patients not yet receiving treatment could further clarify this issue.

## References

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

Arrindell, W. A., Barelds, D. P., Janssen, I. C., Buwalda, F. M. and 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, 377-391.

Bancroft, J. and Marks, I., (1968). Electric aversion therapy of sexual deviations. Proceedings of the Royal Society of Medicine, 61, 796-799.

Ben-Shakhar, G. and Sinai, Y., (1991). Gender differences in multiple-choice tests: The role of differential guessing tendencies. Journal of Educational Measurement, 28, 23-35.

Benjamin, H., (1967). Transvestism and Transsexualism in the Male and Female. The Journal of Sex Research, 3, 107-127.

Blanchard, R., Clemmensen, L. H. and Steiner, B. W., (1987). Heterosexual and homosexual gender dysphoria. Arch Sex Behav, 16, 139-152.

Brandell, G., Larsson, S., Nyström, P., Palbom, A., Staberg, E.-M. and Sundqvist, C., (2005). Kön och matematik. (Reprints in Mathemetical Sciences, 2005:20). Lund, Sweden: Centre for Mathematical Sciences, Lund University.

Brophy, C. J., Norvell, N. K. and 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.

Callahan, E. J. and Leitenberg, H., (1973). Aversion therapy for sexual deviation: Contingent shock and covert sensitization. Journal of Abnormal Psychology, 81, 60-73.

Chorpita, B. F., Brown, T. A. and Barlow, D. H., (1998). Diagnostic reliability of the DSM-III-R anxiety disorders. Mediating effects of patient and diagnostician characteristics. Behavior Modification, 22, 307-320.

Cohen-Kettenis, P. and Kuiper, B., (1984). Transexuality and psychotherapy. Tijdschrift voor Psychotherapie, 10, 153-166.

Cohen-Kettenis, P. and Pfäfflin, F., (2010). The DSM Diagnostic Criteria for Gender Identity Disorder in Adolescents and Adults. Archives of Sexual Behavior, 39, 499-513.

Cohen-Kettenis, P. T. and Gooren, L. J., (1999). Transsexualism: a review of etiology, diagnosis and treatment. J Psychosom Res, 46, 315-333.

Cronbach, L. J., (1954). Report on a psychometric mission to clinicia. Psychometrika, 19, 263-270.

De Cuypere, G., Van Hemelrijck, M., Michel, A., Carael, B., Heylens, G., Rubens, R., Hoebeke, P. and Monstrey, S., (2007). Prevalence and demography of transsexualism in Belgium. Eur Psychiatry, 22, 137-141.

DeMars, C., (2010). Item Resonse Theory. New York: Oxford University Press.

Derogatis, L. R., (1994). SCL-90-R: Administration, scoring and procedures manual. Minneapolis, MN: National Computer Systems.

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

Doolittle, A. E. and Cleary, T. A., (1987). Gender-based differential item performance in mathematics achievement items. Journal of Educational Measurement, 24, 157-166.

Egberink, I. J. L. and Meijer, R. R., (2010). An IRT analysis of Harter's Self-Perception Profile for Children (SPPC) or why strong clinical scales should be distrusted. Assessment (in press).

Ekenstierna, M., (2004). Multiple comparison procedures based on marginal p-values. http://www.math.uu.se/research/pub/Ekenstierna.pdf. Uppsala: Uppsala University.

Ekstrom, R. B., French, J. W., Harman, H. H. and Dermen, D., (1976). Kit of factor-referenced cognitive tests. Princeton, NJ: Educational Testing Service.

Elliott, R., Fox, C. M., Beltyukova, S. A., Stone, G. E., Gunderson, J. and 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.

Embretson, S. E. and Reise, S., (2000). Item response theory for psychologists. Mahwah, NJ: Erlbaum.

Eriksson, K. and Lindholm, T., (2007). Making gender matter: the role of gender-based expectancies and gender identification on women's and men's math performance in Sweden. Scandinavian Journal of Psychology, 48, 329-338.

Gierl, M. J., Bisanz, J., Bisanz, G. L. and Boughton, K. A., (2003). Identifying content and cognitive skills that produce gender differences in mathematics: A demonstration of the multidimensionality-based DIF analysis paradigm. Journal of Educational Measurement, 40, 281-306.

Gomez-Gil, E., Trilla, A., Salamero, M., Godas, T. and Valdes, M., (2008). Sociodemographic, Clinical, and Psychiatric Characteristics of Transsexuals from Spain. Arch Sex Behav, 38, 378-392.

Gomez-Gil, E., Trilla, A., Salamero, M., Godas, T. and Valdes, M., (2009). Sociodemographic, Clinical, and Psychiatric Characteristics of Transsexuals from Spain. Arch Sex Behav, 38, 378-392.

Gurney, K. W., (2010). Transsexualism - attitudes in general practice. Australian Family Physician, 39, 183.

Hafkenscheid, A., (1993). Psychometric evaluation of the symptom checklist (SCL-90) in psychiatric inpatients. Personality and Individual Differences, 14, 751-756.

Hafkenscheid, A., Maassen, G. and 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. and Dahl, A. A., (2000). Symptom profiles of gender dysphoric patients of transsexual type compared to patients with personality disorders and healthy adults. Acta Psychiatr Scand, 102, 276-281.

Haraldsen, I. R., Opjordsmoen, S., Egeland, T. and Finset, A., (2003). Sex-sensitive cognitive performance in untreated patients with early onset gender identity disorder. Psychoneuroendocrinology, 28, 906-915.

Haraldsen, I. R., Egeland, T., Haug, E., Finset, A. and Opjordsmoen, S., (2005). Cross-sex hormone treatment does not change sex-sensitive cognitive performance in gender identity disorder patients. Psychiatry Research, 137, 161-174.

Hartung, C. M. and Widiger, T. A., (1998). Gender differences in the diagnosis of mental disorders: conclusions and controversies of the DSM-IV. Psychological Bulletin, 123, 260-278.

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

Heath, R. A., (2006). The Praeger handbook of transsexuality: changing gender to match mindset. Westport, CT: Praeger Publishers.

Herman-Jeglinska, A., Grabowska, A. and Dulko, S., (2002). Masculinity, femininity, and transsexualism. Arch Sex Behav, 31, 527-534.

Holi, M. M., Sammallahti, P. R. and Aalberg, V. A., (1998). A Finnish validation study of the SCL-90. Acta Psychiatrica Scandinavica, 97, 42-46.

Huynh, H. and Feldt, L. S., (1976). Estimation of the Box Correction for Degrees of Freedom from Sample Data in Randomized Block and Split-Plot Designs. Journal of Educational and Behavioral Statistics, 1, 69-82.

Inzlicht, M. and Ben-Zeev, T., (2000). A threatening intellectual environment: why females are susceptible to experiencing problem-solving deficits in the presence of males. Psychological Science, 11, 365-371.

Jablensky, A., (2009). A meta-commentary on the proposal for a meta-structure for DSM-V and ICD-11. Psychological Medicine, 39, 2099-2103.

Jane, J. S., Oltmanns, T. F., South, S. C. and Turkheimer, E., (2007). Gender bias in diagnostic criteria for personality disorders: an item response theory analysis. Journal of Abnormal Psychology, 116, 166-175.

Keller, J. and Dauenheimer, D., (2003). Stereotype threat in the classroom: dejection mediates the disrupting threat effect on women's math performance. Personality and Social Psychology Bulletin, 29, 371-381.

Kimball, M. M., (1989). A new perspective on women's math achievement. Psychological Bulletin, 105, 198-214.

Kockott, G. and Fahrner, E. M., (1988). Male-to-female and female-to-male transsexuals: a comparison. Arch Sex Behav, 17, 539-546.

Kraemer, H. C., Shrout, P. E. and 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.

Kreukels, B. P. C., Haraldsen, I. R., De Cuypere, G., Richter-Appelt, H., Gijs, L. and Cohen Kettenis, P. T., (2010). A European Network for the Investigation of Gender Incongruence: The ENIGI initiative. European Psychiatry, doi: 10.1016/j.eurpsy.2010.04.009.

Lothstein, L. M., (1984). Psychological testing with transsexuals: a 30-year review. Journal of Personality Assessment, 48, 500-507.

Malouf, M. A., Migeon, C. J., Carson, K. A., Petrucci, L. and Wisniewski, A. B., (2006). Cognitive outcome in adult women affected by congenital adrenal hyperplasia due to 21-hydroxylase deficiency. Hormone Research, 65, 142-150.

McGillicuddy-De Lisi, A. V. and De Lisi, R., (2002). Biology, society, and behavior: the development of sex differences in cognition. Westport, Conn.: Ablex Publishing.

Meijer, R. R. and 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. and van Bruggen, V., (2010). An evaluation of the brief symptom inventory-18 using item response theory or Which items are most strongly related to psychological distress? Psychological Assessment (accepted).

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.

Meyer III, W. J., Bockting, W. O., Cohen-Kettenis, P. T., Coleman, E., DiCeglie, D., Devore, H., Gooren, L., Hage, J. J., Kirk, S., Kuiper, B., Laub, D., Lawrence, A., Menard, Y., Patton, J., Shaefer, L., Webb, A., Christine, C. and Monstrey, S., (2002). The standards of care for gender identity disorders, 6th version. Journal of Psychology & Human Sexuality, 13, 1-30.

Michielsen, H. J., De Vries, J., Van Heck, G. L., Van de Vijver, F. J. R. and Sijtsma, K., (2004). Examination of the Dimensionality of Fatigue: The Construction of the Fatigue Assessment Scale (FAS). European Journal of Psychological Assessment, 20, 39-48.

Mokken, R. J., (1971). A theory and procedure of scale analysis. The Hague: Mouton.

Mokken, R. J. (1997). Nonparametric models for dichotomous responses. In W. J. van der Linden and R. K. Hambleton (Eds.), Handbook of modern item response theory (351-367). New York: Springer.

Molenaar, I. W. (1997). Nonparametric models for polytomous responses. In W. J. van der Linden and R. K. Hambleton (Eds.), Handbook of modern item response theory (369-380). New York: Springer.

Molenaar, I. W. and Sijtsma, K., (2000). MSP5 for Windows. Groningen, The Netherlands: iecProGAMMA.

Okabe, N., Sato, T., Matsumoto, Y., Ido, Y., Terada, S. and Kuroda, S., (2008). Clinical characteristics of patients with gender identity disorder at a Japanese gender identity disorder clinic. Psychiatry Res, 157, 315-318.

Olsen, L. R., Mortensen, E. L. and 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.

Puts, D. A., McDaniel, M. A., Jordan, C. L. and Breedlove, S. M., (2008). Spatial ability and prenatal androgens: meta-analyses of congenital adrenal hyperplasia and digit ratio (2D:4D) studies. Archives of Sexual Behavior, 37, 100-111.

Paap, M. C. S., Meijer, R. R., van Bebber, J., Pedersen, G., Karterud, S., Hellem, F. M. and Haraldsen, I. R., (submitted). A study of the dimensionality and measurement precision of the SCL-90-R using Item Response Theory.

Reise, S. P., Ainsworth, A. T. and Haviland, M. G., (2005). Item Response Theory. Current Directions in Psychological Science, 14, 95-101.

Reise, S. P. and Waller, N. G., (2009). Item Response Theory and Clinical Measurement. Annual Review of Clinical Psychology, 5, 27-48.

Resnick, S. M., Berenbaum, S. A., Gottesman, I. I. and Bouchard, T. J., (1986). Early hormonal influences on cognitive functioning in congenital adrenal hyperplasia. Developmental Psychology, 22, 191-198.

Santor, D. A., Ramsay, J. O. and Zuroff, D. C., (1994). Nonparametric item analyses of the beck depression inventory: evaluating gender item bias and response option weights. Psychological Assessment, 6, 255-270.

Schmitz, N., Hartkamp, N., Kiuse, J., Franke, G. H., Reister, G. and Tress, W., (2000). The Symptom Check-List-90-R (SCL-90-R): A German validation study. Quality of Life Research, 9, 185-193.

Schrock, D. and Reid, L., (2006). Transsexuals' Sexual Stories. Arch Sex Behav, 35, 75-86.

Seikowski, K., Gollek, S., Harth, W. and Reinhardt, M., (2008). Borderline-Persönlichkeit und Transsexualität [Borderline personality disorder and transsexualism]. Psychiatrische Praxis, 35, 135–141.

Shaffer, J. P., (1995). Multiple hypothesis testing. Annual Review of Psychology, 46, 561-584.

Sijtsma, K. and 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.

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

Sijtsma, K., Emons, W. H., Bouwmeester, S., Nyklicek, I. and 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.

Slabbekoorn, D., van Goozen, S. H., Megens, J., Gooren, L. J. and Cohen-Kettenis, P. T., (1999). Activating effects of cross-sex hormones on cognitive functioning: a study of short-term and long-term hormone effects in transsexuals. Psychoneuroendocrinology, 24, 423-447.

Smith, Y. L., Van Goozen, S. H., Kuiper, A. J. and Cohen-Kettenis, P. T., (2005). Sex reassignment: outcomes and predictors of treatment for adolescent and adult transsexuals. Psychol Med, 35, 89-99.

Sommer, I. E. C., Cohen-Kettenis, P. T., van Raalten, T., vd Veer, A. J., Ramsey, L. E., Gooren, L. J. G., Kahn, R. S. and Ramsey, N. F., (2008). Effects of cross-sex hormones on cerebral activation during language and mental rotation: An fMRI study in transsexuals. European Neuropsychopharmacology, 18, 215-221.

Stevens, J. P., (2002). Applied Multivariate Statistics For The Social Sciences. Mahwah, New Jearsey: Lawrence Erlbaum Associates.

Torres, A., Gomez-Gil, E., Vidal, A., Puig, O., Boget, T. and Salamero, M., (2006). [Gender differences in cognitive functions and influence of sex hormones]. Actas Españolas de Psiquiatría, 34, 408-415.

Uebelacker, L. A., Strong, D., Weinstock, L. M. and Miller, I. W., (2009). Use of item response theory to understand differential functioning of DSM-IV major depression symptoms by race, ethnicity and gender. Psychological Medicine, 39, 591-601.

van Abswoude, A. A. H., van der Ark, L. A. and Sijtsma, K., (2004). A Comparative Study of Test Data Dimensionality Assessment Procedures Under Nonparametric IRT Models. Applied Psychological Measurement, 28, 3-24.

van der Ark, L. A., (2007). Mokken scale analysis in R. Journal of Statistical Software, 20, 1-19.

Van Goozen, S. H., Cohen-Kettenis, P. T., Gooren, L. J., Frijda, N. H. and Van de Poll, N. E., (1995). Gender differences in behaviour: activating effects of cross-sex hormones. Psychoneuroendocrinology, 20, 343-363.

Van Goozen, S. H., Slabbekoorn, D., Gooren, L. J., Sanders, G. and Cohen-Kettenis, P. T., (2002). Organizing and activating effects of sex hormones in homosexual transsexuals. Behavioral Neuroscience, 116, 982-988.

Vassend, O. and 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.

Vujovic, S., Popovic, S., Sbutega-Milosevic, G., Djordjevic, M. and Gooren, L., (2008). Transsexualism in Serbia: A Twenty-Year Follow-Up Study. J Sex Med, 6, 1018-1023.

Weinstock, L. M., Strong, D., Uebelacker, L. A. and Miller, I. W., (2009). Differential item functioning of DSM-IV depressive symptoms in individuals with a history of mania versus those without: an item response theory analysis. Bipolar Disorders, 11, 289-297.

Wester, A. and Henriksson, W., (2000). The interaction between item format and gender differences in mathematics performance based on TIMSS data. Studies in Educational Evaluation, 26, 79-90.

World Health Organization, (1992). The ICD–10 Classification of Mental and Behavioral Disorders: Clinical Descriptions and Diagnostic Guidelines. Geneva.

Wilson, M., (1993). DSM-III and the transformation of American psychiatry: a history. American Journal of Psychiatry, 150, 399-410.

Winters, K., (2006). Gender Dissonance: Diagnostic Reform of Gender Identity Disorder for Adults. Journal of Psychology & Human Sexuality, 17, 71-89.

Wismeijer, A. A., Sijtsma, K., van Assen, M. A. and 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.

ERRATA

Page 10, ”Algorithm for Item Selection” should read”Automated Item Selection Procedure”.

Page 19, line 2: the second “recently” in this sentence should be omitted.

Page 28, last line: the colon should be replaced with a full stop.

Page 29, ”Algorithm for Item Selection” should read ”Automated Item Selection Procedure”.

Page 31, fifth line from bottom: it should read “is equal to *m*-1” instead of “is equal *m*-1”.

Page 38, fifth line from bottom: the word “when” should be removed.

Page 44, line 11: it should read “which have shown” instead of “which have showed”.

Page 45, line 18: it should read “distress” instead of “distresses”.

Page 46, line 11: the word “and” should be inserted between “treatment” and “cognition”.

Page 52, third line from bottom: the word “at” should be removed.

Page 53, line 1: the word “atrophy” should be changed to “atrophies”.

I

# II

# 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] and Ira R. Haraldsen[1]

1 Department of Neuropsychiatry and Psychosomatic Medicine, Oslo University Hospital, Norway

2 Institute of Clinical Medicine, University of Oslo, Norway

3 Department of Psychometrics and Statistical Techniques, Faculty of Behavioural and Social Sciences, University of Groningen, the Netherlands

4 Meurs HRM, Woerden, the Netherlands

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

## Abstract

We used item response theory (IRT) to (a) investigate the dimensionality of the 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 Programs. 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.

**Keywords**: item response theory, validity, personality disorder, questionnaire evaluation

## Introduction

The Symptom Checklist-90-Revised (SCL-90-R) (Derogatis, 1994) is a popular psychological screening instrument**,** which is both used to obtain an estimation of the general symptom level (Global Severity Index) as well as a more specific subscale profile. The 90 items were designed to cover nine different subscales (factors) of psychological distress: somatization (Som), interpersonal sensitivity (Int), depression (Dep), anxiety (Anx), phobic anxiety (Pho), obsession-compulsion (Obs), hostility (Hos), paranoid ideation (Par), and psychoticism (Psy). Each item is scored on a scale ranging from 0 ('not at all') through 4 ('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; Hafkenscheid, 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., 2004; Arrindell, et al., 2004; 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 negative affectivity).

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 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 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 0 (‘not at all’) through 4 (‘extremely’), the patients did not effectively discriminate between 2 (‘moderately’) and 3 (‘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, 2010). A nonparametric 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:

(a) creating clinically meaningful scales by entering two items as a starting pair around which the exploratory analysis builds the scale (nonparametric IRT)

(b) investigating item-functioning given the estimated score on the latent trait (for example depression; both nonparametric and parametric IRT)

(c) 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 Programs (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 clinical disorder group (CD), 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 personality disorder group (PD). 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).

The majority of the patients were women (72% in both groups) and the mean age was 35 years in both groups (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 sociodemographic 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, anonymised 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 Symptom Checklist 90-Revised (SCL-90-R: Derogatis, 1994). The instrument encompasses nine symptom subscales (comprising a total of 83 items) as well as 7 additional items. The mean score on all 90 items (including the 7 additional items) is referred to as the Global Severity Index (GSI) and is widely used as a global index for psychological distress. All patients were diagnosed by means of the Mini International Neuropsychiatric Interview (M.I.N.I.) (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) for axis II disorders. We refer to Pedersen and Karterud (2004) for more information regarding the diagnostic procedure.

*Investigating dimensionality: Nonparametric Item Response Theory (NIRT)*

To investigate the dimensionality of the SCL-90-R, Mokken's Monotone Homogeneity Model (MHM) was used (Mokken, 1971; Mokken, 1997). This is a nonparametric item response theory (NIRT) model, which is based on the assumptions of unidimensionality, local independence, and monotonicity (Sijtsma and Molenaar, 2002). This model was tested 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 ($H$). $H_{ij}$ 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 $H_i$s are based on the $H_{ij}$s, 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. $H$ is based on the $H_i$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 $.3 \leq H < 0.4$, good if $.4 \leq H < .5$, and strong if $H \geq .5$ (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 analyzed separately. In addition, the GSI was analyzed to investigate the unidimensionality of the SCL-90-R. Then, exploratory analyses (option ‘SEARCH normal’ in MSP5) were performed. 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. 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 Algorithm for Item Selection (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, 2010). 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 group.

*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 analyzing 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.

The basic unit in any IRT model is the item response function (IRF; also known as the item characteristic curve). In case of dichotomous items, the IRF depicts the relationship between the latent trait ($\theta$) and the probability of the item being endorsed. In case of polytomous items, the IRF is defined as 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 1, we are left with (*m*-1) ISRFs for each item. An important difference between the parametric GRM and the nonparametric MHM described in the previous paragraph concerns the assumptions underlying the shape of the item step response functions (ISRFs). Under a nonparametric model such as the MHM, the only demand is that the ISRFs be monotonely non-decreasing. This means that a higher $\theta$-level corresponds with a higher 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, 2010). 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 $m$-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 $\theta$-scale at which the probability equals 50% of responding in category $m$ or higher. If the $b$'s 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 $\theta$. There is an ORC for each item category $m$, and at each value of $\theta$ the sum of the $m$ probabilities is equal to 1 (Partchev, 2004). Fig. 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 $\theta$-values the 'not at all' option is most likely to be chosen, for slightly higher $\theta$-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 $\theta$ (Embretson and Reise, 2000; Meijer, et al., 2010). This means that the reliability is not a single estimate such as in Mokken scaling or classical test theory, but depends on the value of $\theta$ (Egberink and Meijer, 2010). The information curve depicts the measurement precision conditionally on $\theta$. 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*

Missing data occurred for 1064 of 277020 cells (0.004%). 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. 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 (difference equal to 0.7) and paranoid ideation (difference equal 0.6) scales.

Table 2 and 3 show the correlations between the subscales of the SCL-90-R as well as some other psychometric properties, for the C 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 the lowest mean correlation (0.33 and 0.37, respectively). When comparing table 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 group.

### *Dimensionality of the SCL-90-R*

#### Confirmatory analysis

The *H*-value for the Global Scale Index, which comprises all 90 items, was lower than .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 $H_i$<.3, for the PD group 7. 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). 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.

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 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 7 proposed scales can be found in Table 4.

*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 analyzed items, and Fig. 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 (1) $\theta$ is normally distributed, with the mean equal to zero and a standard deviation of one and (2) $\theta = 0$ corresponds to the mean for the total group on the subscale being analyzed. Inspection of the *b* parameters for the Dep+ scale showed that most of the items are located left of the mean $\theta$, indicating that most of the items are uninformative about individual differences at the range of the $\theta$ 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 $\theta$ 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, is the SCL-90-R primarily a unidimensional or multidimensional instrument? 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.

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. 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 indicates that the correlations between the subscales are stable over time (and generalisable). 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'. Most of the excluded items were dropped, because they did not cluster well with any of the scales. However, a few of the items were dropped for the opposite reason: they clustered with many of the subscales. Item 16 (hearing voices) is an interesting example. Based on the definitions of the DSM-IV, we 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.

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 showed 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., 2010), showing most reliable measurement for average to moderately high scores. This implicates 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 negative affectivity) 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 9-factor models and the 10-factor model they tested showed a poor fit. However, Arrindell et al. (2004) 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., 2004; Arrindell et al., 2006). Interestingly, these confirmatory analyses have shown almost equal support for the Dutch 8-factor model, Derogatis' 9-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 true dimensionality structure of the data (Sijtsma and Molenaar, 2002).

In conclusion, this study has produced seven new scales that can be used in clinical practice which allow for more reliable discrimination between patients than the old scales. Additionally, it has been shown that the measurement precision is dependent on the estimated level of distress. This should be taken into account when interpreting change scores (treatment effects). Finally, this study has clearly illustrated the advantages of IRT, and we propose that our analytic strategy be preferred over more traditional methods, such as exploratory and confirmatory factor analysis, when investigating the scalability of the items of the SCL-90-R.

## Acknowledgements

The authors thank Andries van der Ark of Tilburg University for advising on methods of imputation. We also thank the patients and staff from the following treatment units in the Norwegian Network of Personality-Focused Treatment Programs for their contribution to this study: 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

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

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

Arrindell, W.A. and Ettema, H. (1981). Dimensionele structuur, betrouwbaarheid en validiteit van de Nederlandse bewerking van de Symptom Checklist (SCL-90). *Nederlands Tijdschrift Voor de Psychologie*, **36**, 77-108.

Arrindell, W.A., Boomsma, A., Ettema, H. and Stewart, R. (2004). Verdere steun voor het multidimensionale karakter van de SCL-90-R. *De Psycholoog*, **39**, 195-201.

Arrindell, W.A., Boomsma, A., Ettema, H. and Stewart, R. (2004). Nog meer steun voor het multidimensionale karakter van de SCL-90-R. *De Psycholoog*, **39**, 368-371.

Arrindell, W.A., Barelds, D.P., Janssen, I.C., Buwalda, F.M. and van der Ende, J. (2006). Invariance of SCL-90-R dimensions of symptom distress in patients with peri partum pelvic pain (PPPP) syndrome. *Br J Clin Psychol*, **45**, 377-391.

Bernaards, C.A. and 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. and Kiluk, D.J. (1988). An examination of the factor structure and convergent and discriminant validity of the SCL-90R in an outpatient clinic population. *J Pers Assess*, **52**, 334-340.

Cronbach, L.J. (1954). Report on a psychometric mission to clinicia. *Psychometrika*, **19**, 263-270.

Cyr, J.J., McKenna-Foley, J.M. and Peacock, E. (1985). Factor structure of the SCL-90-R: is there one? *J Pers Assess*, **49**, 571-578.

Derogatis, L.R. (1994). *SCL-90-R: Administration, scoring and procedures manual.* Minneapolis, MN: National Computer Systems.

Dinning, W.D. and Evans, R.G. (1977). Discriminant and convergent validity of the SCL-90 in psychiatric inpatients. *J Pers Assess*, **41**, 304-310.

Egberink, I.J.L. and Meijer, R.R. (2010). An IRT analysis of Harter's Self-Perception Profile for Children (SPPC) or why strong clinical scales should be distrusted. *Assessment*, (in press).

Elliott, R., Fox, C.M., Beltyukova, S.A., Stone, G.E., Gunderson, J. and 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.

Embretson, S.E. and Reise, S. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum.

Emons, W.H.M., Meijer, R.R. and Denollet, J. (2007). Negative affectivity and social inhibition in cardiovascular disease: Evaluating type-D personality and its assessment using item response theory. *J Psychosom Res*, **63**, 27-39.

First, M.B., Spitzer, R.L., Gibbon, M. and Williams, J.B. (1995). The structured clinical interview for DSM-III-r personality disorders (SCID-II): Part I. Description. *J Personal Disord*, **9**, 83-91.

Hafkenscheid, A. (1993). Psychometric evaluation of the symptom checklist (SCL-90) in psychiatric inpatients. *Personality and Individual Differences*, **14**, 751-756.

Hafkenscheid, A. (2004). Hoe multidimensionaal is de Symptom Checklist (SCL-90) nu eigenlijk? *De Psycholoog*, **39**, 191-194.

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

Holi, M.M., Sammallahti, P.R. and Aalberg, V.A. (1998). A Finnish validation study of the SCL-90. *Acta Psychiatr Scand*, **97**, 42-46.

Jenner, J.A., Rutten, S., Beuckens, J., Boonstra, N. and Sytema, S. (2008). Positive and useful auditory vocal hallucinations: prevalence, characteristics, attributions, and implications for treatment. *Acta Psychiatr Scand*, **118**, 238-245.

Karterud, S., Pedersen, G., Friis, S., Urnes, Ø., Irion, T., Brabrand, J., Falkum, L.R. and Leirvåg, H. (1998). The Norwegian Network of Psychotherapeutic Day Hospitals. *Therapeutic Communities*, **19**, 15-28.

Karterud, S., Pedersen, G., Bjordal, E., Brabrand, J., Friis, S., Haaseth, O., Haavaldsen, G., Irion, T., Leirvag, H., Torum, E. and Urnes, O. (2003). Day treatment of patients with personality disorders: experiences from a Norwegian treatment research network. *J Pers Disord*, **17**, 243-62.

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

Meijer, R.R. and 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. and van Bruggen, V. (2010). An evaluation of the brief symptom inventory-18 using item response theory or Which items are most strongly related to psychological distress? (submitted).

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* (eds van der Linden, W.J. and Hambleton, R.K.), 351-367, New York: Springer.

Molenaar, I.W. (1997). Nonparametric models for polytomous responses. In: *Handbook of modern item response theory* (eds van der Linden, W.J. and Hambleton, R.K.), 369-380, New York: Springer.

Molenaar, I.W. and Sijtsma, K. (2000). *MSP5 for Windows*. Groningen, The Netherlands: iecProGAMMA.

Olsen, L.R., Mortensen, E.L. and Bech, P. (2004). The SCL-90 and SCL-90R versions validated by item response models in a Danish community sample. *Acta Psychiatr Scand*, **110**, 225-229.

Partchev, I. (2004). *A visual guide to item response theory*. www.metheval.uni-jena.de/irt/VisualIRT.pdf. Jena: Friedrich-Schiller-Universität Jena.

Pedersen, G. and Karterud, S. (2004). Is SCL-90R helpful for the clinician in assessing DSM-IV symptom disorders? *Acta Psychiatr Scand*, **110**, 215-224.

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

Reise, S.P., Ainsworth, A.T. and Haviland, M.G. (2005). Item Response Theory. *Current Directions in Psychological Science*, **14**, 95-101.

Reise, S.P. and Waller, N.G. (2009). Item Response Theory and Clinical Measurement. *Annual Review of Clinical Psychology*, **5**, 27-48.

Samejima, F. (1996). The graded response model. In: *Handbook of modern item response theory* (eds van der Linden, W.J. and Hambleton, R.K.), 85-100, New York: Springer.

Schmitz, N., Hartkamp, N., Kiuse, J., Franke, G.H., Reister, G. and Tress, W. (2000). The Symptom Check-List-90-R (SCL-90-R): A German validation study. *Qual Life Res*, **9**, 185-193.

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

Sijtsma, K. and 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.

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

Sijtsma, K. and 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.* Chicago: SPSS Inc.

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

van Ginkel, J.R. and 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. and Skrondal, A. (1999). The problem of structural indeterminacy in multidimensional symptom report instruments. The case of SCL-90-R. *Behav Res Ther*, **37**, 685-701.

Wismeijer, A.A., Sijtsma, K., van Assen, M.A. and Vingerhoets, A.J. (2008). A comparative study of the dimensionality of the self-concealment scale using principal components analysis and Mokken scale analysis. *J Pers Assess*, **90**, 323-334.

**Table 1** Mean scores for the 90 items, the 9 subscales and the GSI for the CD and PD groups separately

| ***Scale / item (nr.)*** | ***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 | | | |
| | | | *Paranoid ideation* | *0.8* | *1.4* |
| *Hostility* | *0.5* | *0.8* | Others are to blame (8) | 0.8 | 1.2 |
| Easily annoyed (11) | 1.6 | 2.0 | Most people can’t be trusted (18) | 0.8 | 1.6 |
| Temper outbursts (24) | 0.3 | 0.7 | Feeling watched (43) | 0.8 | 1.6 |
| Urges to harm someone (63) | 0.1 | 0.5 | Having beliefs that others do not share (68) | 0.6 | 1.0 |
| Urges to break things (67) | 0.4 | 0.7 | Not getting enough credit (76) | 1.1 | 1.6 |
| Arguing frequently (74) | 0.3 | 0.7 | People will take advantage (83) | 0.7 | 1.4 |
| Shouting/throwing (81) | 0.2 | 0.4 | | | |
| | | | *Psychoticism* | *0.6* | *0.9* |
| *Additional items* | | | Someone can control your thoughts (7) | 0.2 | 0.4 |
| Poor appetite (19) | 1.0 | 1.3 | Hearing voices (16) | 0.1 | 0.1 |
| Overeating (60) | 1.1 | 1.4 | Others knowing your private thoughts (35) | 0.3 | 0.5 |
| Trouble falling asleep (44) | 2.0 | 2.3 | Thoughts not your own (62) | 0.2 | 0.5 |
| Awakening early (64) | 1.5 | 1.4 | Feeling lonely with others (77) | 1.6 | 2.2 |
| Restless sleep (66) | 2.2 | 2.4 | Thoughts about sex that bother you a lot (84) | 0.2 | 0.4 |
| Thoughts of death (59) | 1.2 | 1.6 | You should be punished for your sins (85) | 0.3 | 0.6 |
| Feelings of guilt (89) | 1.9 | 2.4 | Something is wrong with your body (87) | 1.0 | 1.3 |
| | | | Never feeling close to another person (88) | 1.1 | 1.6 |
| ***Total scale (GSI)*** | ***1.3*** | ***1.6*** | Something is wrong with your mind (90) | 0.8 | 1.4 |

**Table 2** Correlations on the SCL-90-R subscales, Cronbach's alpha (α) and *H*-values based on the confirmatory NIRT analysis (Clinical Disorder group)

| | Som | Obs | Int | Anx | Pho | Dep | Hos | Par | Psy |
|---|---|---|---|---|---|---|---|---|---|
| Somatization (Som) | 1 | .55 | .43 | . 69 | .44 | . 56 | .31 | .35 | .40 |
| Obsessive-compulsive (Obs) | | 1 | .57 | .57 | .32 | .74 | .36 | .48 | .52 |
| Interpersonal sensitivity (Int) | | | 1 | .53 | .45 | . 66 | .38 | .67 | .61 |
| Anxiety (Anx) | | | | 1 | .56 | .65 | .39 | .43 | .54 |
| Phobic Anxiety (Pho) | | | | | 1 | .30 | .13 | .22 | .25 |
| Depression (Dep) | | | | | | 1 | .35 | .49 | .61 |
| Hostility (Hos) | | | | | | | 1 | .46 | .40 |
| Paranoid ideation (Par) | | | | | | | | 1 | .64 |
| Psychoticism (Psy) | | | | | | | | | 1 |
| Mean correlation | .46 | .51 | .54 | .54 | .33 | .54 | .35 | .47 | .50 |
| α | .86 | .83 | .81 | .85 | .85 | .87 | .72 | .72 | .69 |
| *H* | .36 | .38 | .35 | .41 | .49 | .39 | .42 | .32 | .26 |

GSI: Cronbach's alpha = .96, H = .24

**Table 3** Correlations on the SCL-90-R subscales, Cronbach's alpha (α) and H-values based on the confirmatory NIRT analysis (Personality Disorder group)

| | Som | Obs | Int | Anx | Pho | Dep | Hos | Par | Psy |
|---|---|---|---|---|---|---|---|---|---|
| Somatization (Som) | 1 | .57 | .43 | .70 | .54 | .56 | .31 | .42 | .47 |
| Obsessive-compulsive (Obs) | | 1 | .59 | .66 | .49 | .71 | .38 | .55 | .60 |
| Interpersonal sensitivity (Int) | | | 1 | .59 | .55 | .68 | .36 | .68 | .63 |
| Anxiety (Anx) | | | | 1 | .67 | .68 | .37 | .56 | .62 |
| Phobic Anxiety (Pho) | | | | | 1 | .47 | .25 | .43 | .44 |
| Depression (Dep) | | | | | | 1 | .35 | .56 | .61 |
| Hostility (Hos) | | | | | | | 1 | .48 | .44 |
| Paranoid ideation (Par) | | | | | | | | 1 | .67 |
| Psychoticism (Psy) | | | | | | | | | 1 |
| Mean correlation | .50 | .57 | .56 | .61 | .48 | .58 | .37 | .54 | .56 |
| α | .88 | .83 | .83 | .86 | .85 | .86 | .80 | .75 | .76 |
| *H* | .39 | .36 | .38 | .43 | .49 | .36 | .46 | .36 | .32 |

GSI: Cronbach's alpha = .96, H = .27

**Table 4** Properties of the seven proposed subscales based on the Nonparametric IRT analyses

| | | ***CD*** | ***PD*** | ***Total Group*** | | |
|---|---|---|---|---|---|---|
| *Subscale* | *Items* | *Mean (SD)* | *Mean (SD)* | *Mean (SD)* | *H* (range $H_i$s) | $\alpha^*$ |
| 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 (.83) | 2.5 (.81) | 2.4 (.84) | .45 (.39-.54) | .93 |
| Ag | Phob: 13, 25, 47, 50, 70, 82; Int: 73; Anx: 23, 39, 57, 72; Som: 48 | 1.1 (.83) | 1.5 (.94) | 1.4 (.93) | .47 (.40-.52) | .90 |
| Phy | Som: 4, 27, 42, 52, 56, 58 | 1.6 (.95) | 1.8 (.99) | 1.8 (.99) | .45 (.39-.51) | .81 |
| Obs- | Obs: 3, 38, 45, 46, 65; Anx: 86 | 1.1 (.73) | 1.5 (.84) | 1.5 (.84) | .38 (.33-.45) | .74 |
| Hos | Hos: 11, 24, 63, 67, 74, 81 | 0.5 (.50) | 0.8 (.77) | 1.3 (.86) | .47 (.40-.52) | .80 |
| Dis | Para: 18, 43, 83; Int: 36, 37, 61, 69 | 1.0 (.74) | 1.7 (.94) | 1.6 (.95) | .48 (.40-.52) | .85 |
| Psy-** | Psy: 7, 35, 62, 85, 90; Extra: 89 | 0.6 (.53) | 1.0 (.72) | 0.9 (.70) | .40 (.36-.42) | .72 |

* Cronbach's alpha

**For the CD group, two smaller PSY- clusters were found, the first consisting of items 7, 35 and 62 ($H$ = .43) and the second of 85, 89 and 90 ($H$ = .45)

**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 (.06) | –2.98 (.14) | –1.64 (.07) | –0.67 (.04) | 0.94 (.05) |
| 14 | 1.40 (.06) | –1.87 (.08) | –0.85 (.05) | –0.08 (.04) | 1.22 (.06) |
| 15 | 1.15 (.06) | 0.37 (.05) | 1.37 (.07) | 2.22 (.11) | 3.44 (.18) |
| 26 | 1.49 (.06) | –2.03 (.09) | –0.85 (.05) | –0.10 (.04) | 1.15 (.05) |
| 28 | 1.66 (.06) | –2.40 (.10) | –1.24 (.05) | –0.46 (.04) | 0.78 (.04) |
| 29 | 1.66 (.06) | –1.75 (.07) | –0.82 (.04) | –0.13 (.04) | 0.90 (.07) |
| 30 | 2.80 (.09) | –2.01 (.06) | –0.99 (.03) | –0.36 (.03) | 0.63 (.03) |
| 31 | 1.97 (.07) | –2.35 (.09) | –1.26 (.05) | –0.60 (.04) | 0.51 (.03) |
| 32 | 1.60 (.06) | –1.45 (.06) | –0.46 (.04) | 0.35 (.04) | 1.55 (.06) |
| 33 | 1.82 (.06) | –1.52 (.06) | –0.57 (.04) | 0.16 (.03) | 1.28 (.05) |
| 34 | 1.62 (.06) | –2.02 (.08) | –0.97 (.05) | –0.22 (.04) | 1.00 (.05) |
| 41 | 1.78 (.06) | –1.70 (.07) | –0.72 (.04) | –0.07 (.03) | 0.95 (.04) |
| 54 | 1.94 (.07) | –2.01 (.07) | –0.92 (.04) | –0.28 (.03) | 0.74 (.04) |
| 55 | 1.49 (.06) | –2.42 (.10) | –1.28 (.06) | –0.38 (.04) | 0.92 (.05) |
| 71 | 1.82 (.06) | –1.58 (.06) | –0.57 (.04) | 0.11 (.03) | 1.17 (.05) |
| 77 | 1.64 (.06) | –1.47 (.06) | –0.40 (.04) | 0.29 (.04) | 1.45 (.06) |
| 79 | 2.04 (.07) | –1.31 (.05) | –0.49 (.03) | 0.06 (.03) | 1.00 (.04) |
| *Agoraphobia* | | | | | |
| 13 | 2.57 (.10) | –0.03 (.03) | 0.61 (.03) | 1.10 (.04) | 1.79 (.06) |
| 23 | 1.69 (.07) | –0.70 (.04) | 0.05 (.04) | 0.71 (.04) | 1.81 (.07) |
| 25 | 2.24 (.09) | 0.04 (.03) | 0.67 (.03) | 1.10 (.04) | 1.72 (.06) |
| 39 | 1.34 (.06) | –0.58 (.05) | 0.30 (.04) | 0.97 (.06) | 2.05 (.10) |
| 47 | 2.62 (.09) | –0.18 (.03) | 0.32 (.03) | 0.73 (.03) | 1.27 (.04) |
| 48 | 1.48 (.06) | –0.30 (.04) | 0.49 (.04) | 1.20 (.06) | 2.34 (.10) |
| 50 | 1.95 (.07) | –0.82 (.04) | –0.13 (.03) | 0.42 (.03) | 1.32 (.05) |
| 57 | 1.05 (.05) | –2.85 (.14) | –1.45 (.08) | –0.41 (.05) | 1.28 (.08) |
| 70 | 2.25 (.08) | –0.76 (.03) | –0.06 (.03) | 0.45 (.03) | 1.30 (.05) |
| 72 | 1.79 (.07) | –0.60 (.04) | 0.10 (.03) | 0.76 (.04) | 1.74 (.07) |
| 73 | 1.77 (.07) | –0.40 (.04) | 0.30 (.04) | 0.86 (.04) | 1.69 (.07) |
| 82 | 1.71 (.08) | 0.51 (.04) | 1.12 (.05) | 1.55 (.06) | 2.24 (.10) |
| *Physical complaints* | | | | | |
| 4 | 1.36 (.05) | –1.51 (.07) | –0.37 (.04) | 0.50 (.04) | 1.97 (.08) |
| 27 | 1.17 (.06) | –0.72 (.06) | 0.07 (.05) | 0.71 (.05) | 1.84 (.09) |
| 42 | 1.66 (.06) | –1.26 (.05) | –0.55 (.04) | 0.00 (.03) | 0.88 (.03) |
| 52 | 1.46 (.06) | –0.26 (.04) | 0.57 (.04) | 1.22 (.06) | 2.40 (.10) |
| 56 | 2.80 (.08) | –0.90 (.03) | –0.16 (.02) | 0.45 (.03) | 1.30 (.04) |
| 58 | 2.62 (.08) | –0.84 (.03) | –0.10 (.03) | 0.47 (.03) | 1.38 (.04) |
| *Obsessive-Compulsive–* | | | | | |
| 3 | 0.95 (.05) | –2.26 (.12) | –1.00 (.07) | 0.07 (.05) | 1.88 (.11) |
| 38 | 1.84 (.06) | –0.25 (.03) | 0.57 (.03) | 1.27 (.05) | 2.15 (.08) |
| 45 | 2.64 (.08) | –0.59 (.03) | 0.20 (.03) | 0.74 (.03) | 1.57 (.04) |
| 46 | 1.22 (.05) | –2.17 (.10) | –0.80 (.05) | 0.11 (.04) | 1.54 (.08) |
| 65 | 1.46 (.07) | 0.66 (.04) | 1.29 (.06) | 1.72 (.07) | 2.45 (.11) |
| 86 | 1.02 (.06) | 0.33 (.05) | 1.20 (.07) | 1.98 (.11) | 3.30 (.18) |
| *Hostility* | | | | | |
| 11 | 1.70 (.06) | –1.55 (.06) | –0.39 (.04) | 0.44 (.04) | 1.59 (.06) |
| 24 | 2.83 (.11) | 0.57 (.03) | 1.12 (.03) | 1.61 (.04) | 2.15 (.07) |
| 63 | 1.67 (.09) | 1.07 (.05) | 1.79 (.07) | 2.36 (.10) | 3.22 (.17) |
| 67 | 1.77 (.08) | 0.57 (.04) | 1.25 (.05) | 1.81 (.07) | 2.70 (.11) |
| 74 | 1.58 (.07) | 0.58 (.04) | 1.48 (.06) | 2.19 (.09) | 3.20 (.16) |
| 81 | 2.79 (.13) | 0.93 (.03) | 1.47 (.04) | 1.99 (.06) | 2.61 (.10) |
| *Distrust* | | | | | |
| 18 | 1.64 (.06) | –0.67 (.04) | 0.33 (.04) | 0.99 (.05) | 2.03 (.08) |
| 36 | 1.67 (.06) | –1.36 (.05) | –0.19 (.04) | 0.65 (.04) | 1.84 (.07) |
| 37 | 2.34 (.07) | –0.49 (.03) | 0.33 (.03) | 0.99 (.04) | 1.91 (.06) |

|  |  |  |  |  |  |
|---|---|---|---|---|---|
| 43 | 2.41 (.08) | –0.50 (.03) | 0.26 (.03) | 0.84 (.03) | 1.64 (.05) |
| 61 | 1.99 (.06) | –1.28 (.05) | –0.36 (.03) | 0.18 (.03) | 1.09 (.04) |
| 69 | 1.26 (.05) | –1.43 (.07) | –0.14 (.04) | 0.76 (.05) | 1.96 (.09) |
| 83 | 1.60 (.06) | –0.39 (.04) | 0.46 (.04) | 1.14 (.05) | 2.19 (.09) |
| *Psychoticism–* |  |  |  |  |  |
| 7 | 2.79 (.13) | 1.02 (.03) | 1.49 (.04) | 1.90 (.06) | 2.65 (.10) |
| 35 | 2.19 (.09) | 0.82 (.03) | 1.40 (.05) | 1.93 (.06) | 2.72 (.11) |
| 62 | 2.08 (.10) | 1.04 (.04) | 1.55 (.05) | 2.06 (.07) | 2.74 (.11) |
| 85 | 1.31 (.07) | 1.01 (.05) | 1.73 (.08) | 2.31 (.11) | 3.30 (.18) |
| 89 | 1.00 (.05) | –2.10 (.11) | –0.93 (.07) | 0.01 (.05) | 1.40 (.08) |
| 90 | 1.35 (.06) | –0.41 (.04) | 0.52 (.04) | 1.28 (.06) | 2.42 (.11) |

**Fig. 1** (A) Option Response Curve for item 89 from the Psy scale with *a* = 1.00. (B) Option Response Curve for item 30 from the Dep scale with *a* = 2.80.

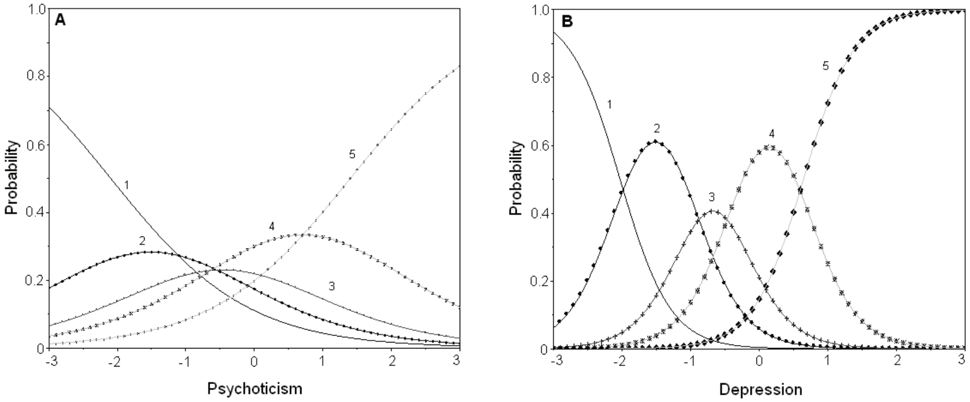
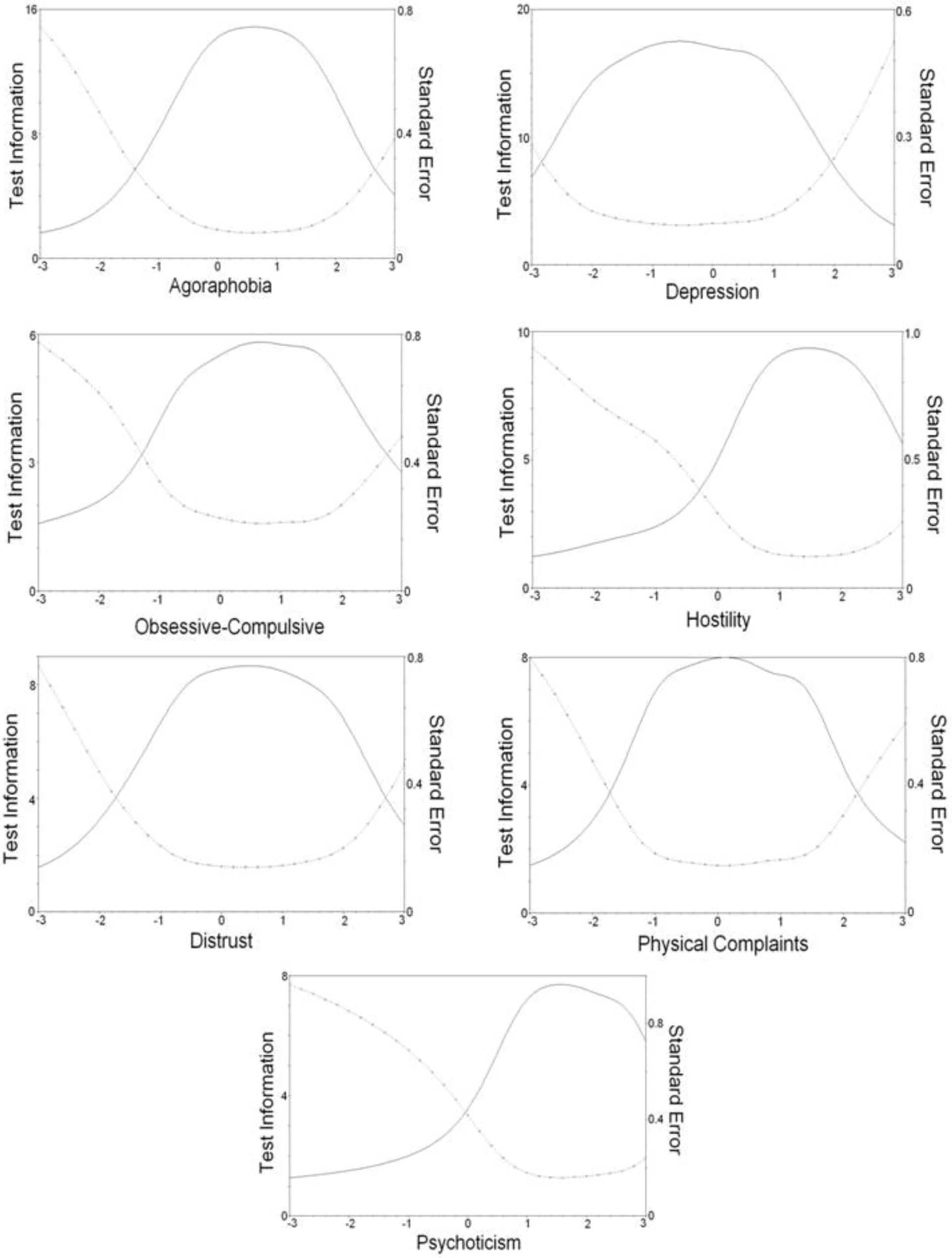


**Fig. 2** Test Information Functions for the seven new subscales, with estimated trait score on the horizontal axis, test information (solid line) on the left vertical axis and standard error of measurement (dotted line) on the right vertical axis



# 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[1,2], Rob R. Meijer[3], Peggy T. Cohen-Kettenis[4], Hertha Richter-Appelt[5], Griet de Cuypere[6], Baudewijntje P. C. Kreukels[4], Geir Pedersen[7], Sigmund Karterud[2,7], Ulrik F. Malt[1], Ira R. Haraldsen[1]

1 Department of Neuropsychiatry and Psychosomatic Medicine, Oslo University Hospital, Norway

2 Institute of Clinical Medicine, University of Oslo, Norway

3 Department of Psychometrics and Statistical Techniques, Faculty of Behavioural and Social Sciences, University of Groningen, the Netherlands

4 Department of Medical Psychology, VU University Medical Center, Amsterdam, The Netherlands

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

6 Department of Sexology and Gender Problems, University Hospital Gent, Belgium

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

## Abstract

*Objective:* 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.

*Methods:* Three samples were included: a sample of severely psychiatrically disturbed patients (n=3078) admitted to 14 different day hospitals participating in the Norwegian Network of Personality-Focused Treatment Programs, a sample of persons with Gender Incongruence (GI; n=410) that were seen at 4 different Gender Identity clinics participating in the European Network for Investigation of Gender Incongruence and a sample of depressed patients (n=223) treated at the Department for Neuropsychiatry and Psychosomatic Medicine at Oslo University Hospital. Mokken Scale Analysis was used to investigate the dimensionality of the SCL-90-R.

*Results:* A unidimensional pattern of findings was 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.

*Conclusion:* 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, such as persons with GI.

**Keywords:** item response theory, validity, personality disorder, questionnaire evaluation, Gender Identity Disorder, depression

## Introduction

The Symptom Checklist-90-Revised (SCL-90-R) [1] 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?’[2]. More than two decades have passed since the publication of that paper; however, the debate has still not abated, as recent publications have demonstrated [3], [4], [5], [6]. On the one hand, there is a group of researchers that firmly believe in the multidimensionality of the instrument [5], [7], [8], 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 [6], [9]. In a recent paper, Paap et al. [10] 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. 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 [10].

In the current study, we investigate whether the findings in the study by Paap et al. can be generalised 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 the one hand, and one's assigned gender and/or one's congenital primary and secondary sex characteristics on the other hand [11], [12]. Following Kreukels et al., we use GI when referring to patients who have not yet been diagnosed with GID [13] or transsexualism [14]. 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 [15] 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) [16], 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. [10] and Meijer et al. [17], Mokken Scale Analysis [18] was used to analyse the data.

## Materials and methods

### *Participants*

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 [19], treated in the period from January 1993 through July 2007. This sample was also used in the study by Paap et al. [10]. 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. [10] and Karterud et al. [20] for sociodemographic and diagnostic details. Patients admitted before 1996 were diagnosed according to the DSM-III-R [21] and patients admitted from 1996 onwards according to the DSM-IV [13]. To 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) into two subgroups. 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.

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' (ENIGI) initiative [12]. 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 [22]. The ENIGI 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 ENIGI 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.

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 [13] 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. [23] 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.

*Assessment*

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: 1]. 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.

*Investigating dimensionality: Mokken Scale Analysis (MSA)*

To investigate the dimensionality of the SCL-90-R, Mokken's Monotone Homogeneity Model (MHM) was used [18], [24]. A scale fulfilling the criteria of the 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). It implies an ordering of *respondents* on an underlying unidimensional scale using the unweighted sum of item scores [25], [26], [27], [28]. MSA was applied using the software package Mokken Scale Analysis for Polytomous items (MSP5.0) [29].

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 [30]. 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 [25]. 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 $H_i$s and expresses the degree to which the total score accurately orders persons on the latent trait scale [27]. A scale is considered acceptable if $.3 \leq H < 0.4$, good if $.4 \leq H < .5$, and strong if $H \geq .5$ [18], [27].

The algorithm that MSP5.0 uses to build one or more scales is called Algorithm for Item Selection (AISP). In the fully automated version ('SEARCH' in MSP.0), the AISP starts by selecting the item pair which has the largest positive $H_{ij}$ of all item pairs. Subsequently, the AISP 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 [18], [27], [31]. The SEARCH-procedure is highly useful for investigating the dimensionality of a questionnaire. Following Sijtsma and Molenaar [27], 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 [27; 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.

## Results

*Missing data: two-way imputation*

Less than 1% of the data were missing in each of the data-sets. Following Paap et al.[10], we used Two-Way imputation [32], 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 [33]. This method is easy to implement using SPSS [34], using the syntax provided by van Ginkel and van der Ark [35].

*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 $\text{PD}_{\text{high}}$ , followed by the $\text{PD}_{\text{low}}$ group and depression sample, and finally the GI sample. Sex differences in mean GSI scores were small (0.1 for each group). Table 2, 3, 4 and 5 show the correlations between the subscales of the SCL-90-R, mean subscale scores with SD, and Cronbach's alpha, for the $\text{PD}_{\text{high}}$, $\text{PD}_{\text{low}}$, 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 $\text{PD}_{\text{low}}$ group to 0.35 for the anxiety (Anx) and psychoticism (Psy) scales in the $\text{PD}_{\text{high}}$ group. The hostility (Hos) and Pho scales had the lowest mean correlations. When comparing Table 2 and 3, it can be seen that the correlations and SD's for the Pho and Psy scales show the largest difference (correlations: 0.07; SD's: 0.31). However, the difference in correlations and SD's for the obsessive-compulsive (Obs) and interpersonal sensitivity (Int) was very small (correlations: 0.01; SD'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 SD'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 the 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 .40 (Hos) to .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.

*Dimensionality of the SCL-90-R*

Four (groups) x eight (different values of *c*) = 32 exploratory analyses were performed using the SEARCH-procedure. A summary of the findings can be found 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 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.

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 Figure 1. To explore potential explanations for these differences, we performed t-tests to ascertain whether there were any sex differences in mean score on particular subscales. We found that the females in the depression sample scored significantly ($\alpha = 0.05$) higher than the males on Som ($T$ = -3.03, $P$ = 0.003), Int ($T$ = -1.99, $P$ = 0.047), Pho ($T$ = -2.05, $P$ = 0.041) and Dep ($T$ = -2.09, $P$ = 0.038). In spite of the difference in subscale scores for the Pho scale, the percentage of females diagnosed with agoraphobia was 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). 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.

## Discussion

Studies reporting on the dimensionality of the SCL-90-R have had very diverse outcomes. To this day, the original 9-scale solution [1] remains controversial [3], [4], [5], [6], [10], [36], [37], [38]. 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 analysing the data from the Gender Incongruence sample, which was characterised by a low level of psychological distress. In contrast, Paap et al. [10] 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.

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 [3], [36], [37]. 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. [10] 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, in spite of the diminished variance. Therefore it is unlikely that the largely 'unidimensional findings' reported by others using samples characterised by low levels of psychological distress can be merely explained by a lack of variance in SCL-90-R scores.

To test the generalisability of our findings, we investigated the dimensionality in a third sample, consisting of depressed outpatients. This sample was characterised 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 generalisable to an other 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 Skrondal [36], who demonstrated factorial invariance for sex, and Olsen et al. [3], who showed that there were two items in the SCL-90-R that were sex biased ('having to do things slowly' and 'crying easily'). At present, we can only speculate as to why we found an effect of sex in the depressed group only. A potential explanation could be that depressed women have more general psychological complaints, as an effect of their depression. This might also explain their elevated mean scores on four of the subscales. However, it could also be that men and women on the whole have a different subjective experience of depression. Alternatively, the effect of sex on dimensionality might be characteristic for patients with intermediate levels of psychological distress.

*Conclusion and recommendations*

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 generalisable 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. In patient groups with high levels of psychopathology, such as patients with personality disorders, we propose that using the seven scales proposed by Paap et al. [10] may possibly be the best option.

## Acknowledgements

We  thank Jan van Bebber for imputating the missing data, and Xi Zhao, 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

1. Derogatis LR. SCL-90-R: Administration, scoring and procedures manual. Minneapolis, MN: National Computer Systems 1994.

2. Cyr JJ, McKenna-Foley JM, Peacock E. Factor structure of the SCL-90-R: is there one? J Pers Assess 1985;49:571-578.

3. Olsen LR, Mortensen EL, Bech P. The SCL-90 and SCL-90R versions validated by item response models in a Danish community sample. Acta Psychiatr Scand 2004;110:225-229.

4. Elliott R, Fox CM, Beltyukova SA, Stone GE, Gunderson J, Zhang X. Deconstructing therapy outcome measurement with rasch analysis of a measure of general clinical distress: The Symptom Checklist-90-Revised. Psychological Assessment 2006;18:359-372.

5. Arrindell WA, Barelds DP, Janssen IC, Buwalda FM, van der Ende J. Invariance of SCL-90-R dimensions of symptom distress in patients with peri partum pelvic pain (PPPP) syndrome. Br J Clin Psychol 2006;45:377-391.

6. Hafkenscheid A, Maassen G, Veeninga A. The dimensions of the Dutch SCL-90: more than one, but how many? Netherlands Journal of Psychology 2007;63:25-30.

7. Arrindell WA, Boomsma A, Ettema H, Stewart R. Verdere steun voor het multidimensionale karakter van de SCL-90-R [Further support for the multidimensional nature of the SCL-90-R]. De Psycholoog 2004;39:195-201.

8. Arrindell WA, Boomsma A, Ettema H, Stewart R. 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 2004;39:368-371.

9. Hafkenscheid A. Hoe multidimensionaal is de Symptom Checklist (SCL-90) nu eigenlijk? [How multidimensional is the Symptom Checklist (SCL-90) really?]. De Psycholoog 2004;39:191-194.

10. Paap MCS, Meijer RR, van Bebber J, Pedersen G, Karterud S, Hellem FM, Haraldsen IR. A study of the dimensionality and measurement precision of the SCL-90-R using Item Response Theory. submitted;

11. Meyer-Bahlburg H. From Mental Disorder to Iatrogenic Hypogonadism: Dilemmas in Conceptualizing Gender Identity Variants as Psychiatric Conditions. Arch Sex Behav 2010;39:461-476.

12. Kreukels BPC, Haraldsen IR, De Cuypere G, Richter-Appelt H, Gijs L, Cohen Kettenis PT. A European Network for the Investigation of Gender Incongruence: The ENIGI initiative. Eur Psychiatry 2010;doi: 10.1016/j.eurpsy.2010.04.009:

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

14. World Health Organization. The ICD–10 Classification of Mental and Behavioral Disorders: Clinical Descriptions and Diagnostic Guidelines. Geneva 1992.

15. Haraldsen IR, Dahl AA. Symptom profiles of gender dysphoric patients of transsexual type compared to patients with personality disorders and healthy adults. Acta Psychiatr Scand 2000;102:276-81.

16. Leinonen E, Niemi H. The influence of educational information on depressed outpatients treated with escitalopram: A semi-naturalistic study. Nord J Psychiat 2007;61:109-114.

17. Meijer RR, de Vries RM, van Bruggen V. An evaluation of the brief symptom inventory-18 using item response theory or Which items are most strongly related to psychological distress? Psychological Assessment 2010;accepted:

18. Mokken RJ. A theory and procedure of scale analysis. The Hague: Mouton 1971.

19. Karterud S, Pedersen G, Friis S, Urnes Ø, Irion T, Brabrand J, Falkum LR, Leirvåg H. The Norwegian Network of Psychotherapeutic Day Hospitals. Therapeutic Communities 1998;19:15-28.

20. Karterud S, Pedersen G, Bjordal E, Brabrand J, Friis S, Haaseth O, Haavaldsen G, Irion T, Leirvag H, Torum E, Urnes O. Day treatment of patients with personality disorders: experiences from a Norwegian treatment research network. J Pers Disord 2003;17:243-262.

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

22. Paap MCS, Kreukels BPC, Cohen Kettenis PT, Richter-Appelt H, De Cuypere G, Haraldsen IR. Assessing the Utility of Diagnostic Criteria: A Multi-Site Study on Gender Identity Disorder. J Sex Med 2010 [accepted];

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

24. Mokken RJ. Nonparametric models for dichotomous responses. In: van der Linden WJ, Hambleton RK, eds. Handbook of modern item response theory. New York: Springer 1997:351-367.

25. Wismeijer AA, Sijtsma K, van Assen MA, Vingerhoets AJ. A comparative study of the dimensionality of the self-concealment scale using principal components analysis and Mokken scale analysis. J Pers Assess 2008;90:323-334.

26. Sijtsma K, Emons WH, Bouwmeester S, Nyklicek I, Roorda LD. Nonparametric IRT analysis of Quality-of-Life Scales and its application to the World Health Organization Quality-of-Life Scale (WHOQOL-Bref). Qual Life Res 2008;17:275-90.

27. Sijtsma K, Molenaar IW. Introduction to Nonparametric Item Response Theory. Thousand Oaks: Sage Publications 2002.

28. Meijer RR, Baneke JJ. Analyzing psychopathology items: a case for nonparametric item response theory modeling. Psychological Methods 2004;9:354-368.

29. Molenaar IW, Sijtsma K. MSP5 for Windows. Groningen, The Netherlands: iecProGAMMA 2000.

30. Molenaar IW. Nonparametric models for polytomous responses. In: van der Linden WJ, Hambleton RK, eds. Handbook of modern item response theory. New York: Springer 1997:369-380.

31. Egberink IJL, Meijer RR. An IRT analysis of Harter's Self-Perception Profile for Children (SPPC) or why strong clinical scales should be distrusted. Assessment 2010; (in press).

32. Bernaards CA, Sijtsma K. Influence of simple imputation and EM methods on factor analysis when item nonresponse in questionnaire data is nonignorable. Multivar Behav Res 2000;35:321-364.

33. Sijtsma K, van der Ark LA. Investigation and treatment of missing item scores in test and questionnaire data. Multivar Behav Res 2003;38:505-528.

34. SPSS. SPSS for Windows, Rel. 16.0.1. Chicago: SPSS Inc. 2007.

35. van Ginkel JR, van der Ark LA. SPSS syntax for missing value imputation in test and questionnaire data. Applied Psychological Measurement 2005;29:152-153.

36. Vassend O, Skrondal A. The problem of structural indeterminacy in multidimensional symptom report instruments. The case of SCL-90-R. Behav Res Ther 1999;37:685-701.

37. Holi MM, Sammallahti PR, Aalberg VA. A Finnish validation study of the SCL-90. Acta Psychiatr Scand 1998;97:42-46.

38. Schmitz N, Hartkamp N, Kiuse J, Franke GH, Reister G, Tress W. The Symptom Check-List-90-R (SCL-90-R): A German validation study. Qual Life Res 2000;9:185-193.

**Table 1** Descriptive statistics for the four samples

| | **PDhigh** | | **PDlow** | | **GI** | | **Depression** | |
|---|---|---|---|---|---|---|---|---|
| | *males* | *females* | *males* | *females* | *males* | *females* | *males* | *females* |
| ***N*** | 386 | 1164 | 459 | 1069 | 264 | 146 | 94 | 129 |
| **Mean age ± SD** | 37 ± 9 | 34 ± 9 | 37 ± 9 | 35 ± 9 | 35 ± 12 | 27 ± 10 | 47 ± 14 | 44 ± 13 |
| **Mean GSI ± SD** | 2.0 ± .38 | 2.1 ± .40 | 1.0 ± .34 | 1.1 ± .34 | .5 ± .46 | .6 ± .54 | 1.2 ± .50 | 1.3 ± .62 |

**Table 2** Correlations on the SCL-90-R subscales, mean subscale scores with SD, Cronbach's alpha (α), $\mathrm{PD}_{\mathrm{high}}$ group

| | Som | Obs | Int | Anx | Pho | Dep | Hos | Par | Psy |
|---|---|---|---|---|---|---|---|---|---|
| Somatization (Som) | 1 | .31 | .08 | .51 | .31 | .25 | .11 | .11 | .18 |
| Obsessive-compulsive (Obs) | | 1 | .25 | .38 | .22 | .38 | .20 | .28 | .37 |
| Interpersonal sensitivity (Int) | | | 1 | .23 | .34 | .38 | .19 | .54 | .43 |
| Anxiety (Anx) | | | | 1 | .46 | .36 | .20 | .29 | .38 |
| Phobic Anxiety (Pho) | | | | | 1 | .12 | .05 | .16 | .17 |
| Depression (Dep) | | | | | | 1 | .15 | .28 | .36 |
| Hostility (Hos) | | | | | | | 1 | .38 | .33 |
| Paranoid ideation (Par) | | | | | | | | 1 | .55 |
| Psychoticism (Psy) | | | | | | | | | 1 |
| Mean correlation | .23 | .30 | .31 | .35 | .23 | .29 | .20 | .32 | .35 |
| Mean subscale score | 2.2 | 2.4 | 2.4 | 2.3 | 1.8 | 2.7 | 1.0 | 1.8 | 1.2 |
| SD | .75 | .60 | .68 | .64 | .96 | .50 | .81 | .80 | .60 |
| α | .81 | .68 | .72 | .72 | .81 | .67 | .79 | .67 | .69 |

**Table 3** Correlations on the SCL-90-R subscales, mean subscale scores with SD, Cronbach's alpha (α), PD$_{low}$ group

| | Som | Obs | Int | Anx | Pho | Dep | Hos | Par | Psy |
|---|---|---|---|---|---|---|---|---|---|
| Somatization (Som) | 1 | .23 | .03 | .43 | .22 | .23 | .09 | .02 | .09 |
| Obsessive-compulsive (Obs) | | 1 | .40 | .31 | .11 | .57 | .20 | .30 | .33 |
| Interpersonal sensitivity (Int) | | | 1 | .26 | .24 | .48 | .23 | .51 | .44 |
| Anxiety (Anx) | | | | 1 | .44 | .37 | .12 | .16 | .28 |
| Phobic Anxiety (Pho) | | | | | 1 | .08 | .02 | .08 | .08 |
| Depression (Dep) | | | | | | 1 | .15 | .26 | .41 |
| Hostility (Hos) | | | | | | | 1 | .33 | .18 |
| Paranoid ideation (Par) | | | | | | | | 1 | .42 |
| Psychoticism (Psy) | | | | | | | | | 1 |
| Mean correlation | .17 | .31 | .32 | .30 | .16 | .32 | .17 | .26 | .28 |
| Mean subscale score | 1.1 | 1.4 | 1.3 | 1.1 | .7 | 1.6 | .5 | .7 | .5 |
| SD | .62 | .61 | .64 | .56 | .65 | .63 | .45 | .57 | .33 |
| α | .78 | .73 | .73 | .73 | .76 | .79 | .65 | .61 | .47 |

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

| | Som | Obs | Int | Anx | Pho | Dep | Hos | Par | Psy |
|---|---|---|---|---|---|---|---|---|---|
| Somatization (Som) | 1 | .69 | .60 | .75 | .60 | .64 | .56 | .55 | .56 |
| Obsessive-compulsive (Obs) | | 1 | .76 | .78 | .61 | .80 | .68 | .68 | .71 |
| Interpersonal sensitivity (Int) | | | 1 | .74 | .69 | .81 | .67 | .80 | .76 |
| Anxiety (Anx) | | | | 1 | .68 | .77 | .61 | .64 | .68 |
| Phobic Anxiety (Pho) | | | | | 1 | .56 | .49 | .55 | .52 |
| Depression (Dep) | | | | | | 1 | .64 | .68 | .73 |
| Hostility (Hos) | | | | | | | 1 | .64 | .61 |
| Paranoid ideation (Par) | | | | | | | | 1 | .75 |
| Psychoticism (Psy) | | | | | | | | | 1 |
| Mean correlation | .62 | .71 | .73 | .71 | .59 | .70 | .61 | .66 | .67 |
| Mean subscale score | .4 | .7 | .7 | .5 | .3 | .9 | .4 | .5 | .4 |
| SD | .49 | .61 | .69 | .54 | .56 | .74 | .51 | .61 | .44 |
| α | .85 | .85 | .87 | .87 | .84 | .90 | .79 | .78 | .72 |

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

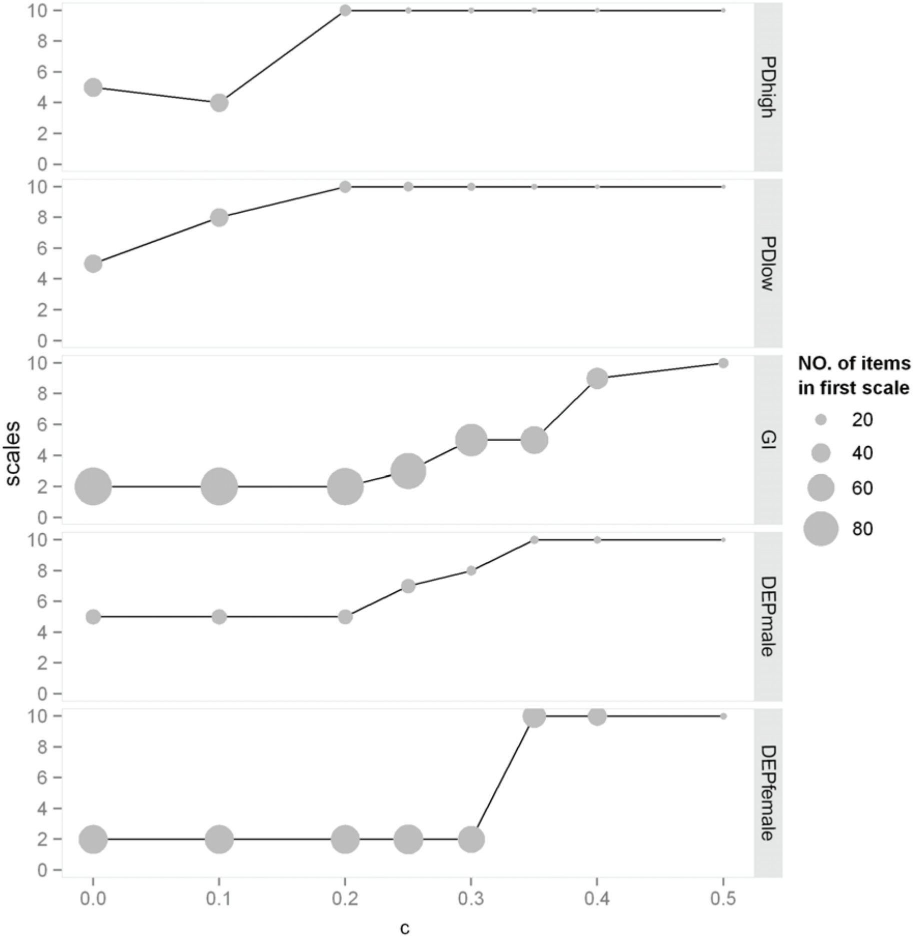
| | Som | Obs | Int | Anx | Pho | Dep | Hos | Par | Psy |
|---|---|---|---|---|---|---|---|---|---|
| Somatization (Som) | 1 | .45 | .32 | .53 | .43 | .38 | .24 | .22 | .33 |
| Obsessive-compulsive (Obs) | | 1 | .63 | .65 | .51 | .72 | .44 | .48 | .58 |
| Interpersonal sensitivity (Int) | | | 1 | .57 | .67 | .69 | .48 | .73 | .60 |
| Anxiety (Anx) | | | | 1 | .64 | .71 | .40 | .42 | .57 |
| Phobic Anxiety (Pho) | | | | | 1 | .60 | .25 | .47 | .45 |
| Depression (Dep) | | | | | | 1 | .40 | .48 | .57 |
| Hostility (Hos) | | | | | | | 1 | .53 | .48 |
| Paranoid ideation (Par) | | | | | | | | 1 | .69 |
| Psychoticism (Psy) | | | | | | | | | 1 |
| Mean correlation | .36 | .56 | .59 | .56 | .50 | .57 | .40 | .50 | .53 |
| Mean subscale score | 1.5 | 1.6 | 1.1 | 1.3 | 0.8 | 1.8 | 0.5 | 0.6 | 0.6 |
| SD | .85 | .83 | .83 | .82 | .86 | .83 | .63 | .68 | .48 |
| α | .86 | .85 | .85 | .86 | .84 | .87 | .82 | .77 | .69 |

**Table 6** Number (No.) of scales, number of items in the first scale and number of excluded items for 8 levels of *c*, reported seperately 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** |
|---|---|---|---|---|---|---|---|---|
| ***PD*$_{high}$** | | | | | | | | |
| 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* | 0 | 7 | 7 | 21 | 39 | 54 | 59 | 67 |
| ***PD*$_{low}$** | | | | | | | | |
| 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* | 1 | 1 | 16 | 32 | 39 | 54 | 59 | 69 |
| ***GI*** | | | | | | | | |
| 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* | 1 | 1 | 2 | 2 | 7 | 13 | 16 | 42 |
| ***Depression*** | | | | | | | | |
| 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* | 0 | 0 | 4 | 5 | 8 | 10 | 24 | 43 |

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

**Figure 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 seperate pannels for the different groups. The GI sample and 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.


PDhigh
PDlow
GI
DEPmale
DEPfemale
scales
c
NO. of items in first scale