Validation of instruments for diagnosing depression and measuring stress in general practice

PhD dissertation

Marie Germund Nielsen

Health
Aarhus University
Research Unit for General Practice & Section for General Medical Practice
Department of Public Health
2018
Erratum

Corrections to M.G. Nielsen (2018), PhD dissertation entitled “Validation of instruments for diagnosing depression and measuring stress in general practice”, Faculty of Health, Aarhus University.

Two references have been corrected in the electronic version of the dissertation:

- **Figure 2.8. Item characteristic curve of dichotomised item (p. 41)**
  This was originally incorrectly referred to as “103. RUMM Laboratory. Interpreting RUMM2030. Part 1: Dichotomous data; 2009.”

  The correct reference is instead:


- **Figure 2.3. A ruler with an ordinal scale on the top and an interval scale at the bottom (p. 31)**

  The correct reference is instead:

  “Brodersen J. Measuring psychosocial consequences of false-positive screening results - breast cancer as an example. Copenhagen: Faculty of Health Sciences, University of Copenhagen. 2006.”
PhD student:
MSc in Public Health, Marie Germund Nielsen, Research Unit for General Practice, Department of Public Health, Aarhus, Denmark

Supervisors:
Professor, Kaj Sparle Christensen, PhD, MD, Research Unit for General Practice, Department of Public Health, Aarhus, Denmark

Statistician, Eva Ørnbøl, MSc, Research Clinic for Functional Disorders and Psychosomatics, Aarhus University Hospital

Professor, Mogens Vestergaard, PhD, MD, Research Unit for General Practice and Section for General Medical Practice, Department of Public Health, Aarhus University.

Professor, Per Bech, PhD, MD, Psychiatric Research Unit, Psychiatric Centre North Zealand, Copenhagen University Hospital

Assessment committee:
Professor, Christopher Dowrick, BA, MSc, MD, FRCGP, Institute of Psychology Health and Society, Psychological Sciences, University of Liverpool

Associate Professor, Tine Nielsen, PhD, MSc Psychology, Psychological Testing & Psychometrics, Department of Psychology, University of Copenhagen

Associate professor, Søren Dinesen Østergaard, PhD, MD, Department of Clinical Medicine, Aarhus University (Chairman of the committee)

Financial support:
The Lundbeck Foundation (grant number: R155-2012-11280), the Foundation for Research in Primary Care at Central Denmark Region, the Oticon Foundation, Danish Ministry of Higher Education and Science, Danish Agency for Science, Technology and Innovation, Multipractice Study Committee of the Danish College of General Practitioners.

Acknowledgements

This dissertation was carried out during my time as a PhD student at the Research Unit for General Practice at Aarhus University. My profound thanks go to everyone at the research unit, my family and friends, who helped me to accomplish this PhD project. First, I would like to thank my supervisors, Kaj Sparle Christensen, Eva Ørnbøl, Mogens Vestergaard and Per Bech for all their guidance, support and kind supervision. Thank you, Kaj, for your patience, but also for helping me to become more and more independent as a researcher. Thank you, Eva, for discussing Rasch analysis with me. You always took the time to explain the complicated theories in depth. Thank you, Mogens, for your always positive comments and attitude. Thank you, Per, for introducing me to the great history of psychometric testing and for letting me learn from all your vast experience in this field.

A warm and kind atmosphere characterises the research environment at the Research Unit for General Practice. This has helped me tremendously in my development as a researcher.

I want to thank Hanne Beyer for her help developing ‘Sundhedsmappen.dk’. I want to thank the dedicated student workers on my project; Pia Walfried, Fie Hjorth Hansen and Anita Bruun. A special thanks goes to Lone Niedziella for her availability for linguistic advice and for her support in the process of finalizing my dissertation.

My thanks go to the research network MEPRICA, which has been inspiring for me to be part of. Especially, I would like to thank my great colleagues; Anders Prior, Anette Riisgaard Ribe, Maiken Ina Siegismund Kjærgaard, Anita Tønder Nielsen, Marlene Rosengaard Møller and Henrik Søndergaard Pedersen. Anette, you have been a role model for me in how to be productive and setting high goals, while still being a kind and helpful colleague. In addition, special thanks go to my friend Anne Bo for being a mentor during my process, to my colleague Karen Busk Nørøxe for her caring mind and coffee talks and to my colleague Therese Koops Grønborg for her support and methodological discussions.

During my research stay as a visiting scholar at the Department of Psychiatry and Behavioural Science, University of Washington, Seattle, I formed new and delightful research collaborations. Thank you, Joan Russo and Dimitry Davydow, for making this stay possible. Furthermore, I would like to thank all the general practitioners and patients participating in my study for their time and patience.

My heartfelt appreciation goes to my mother and father for always believing in me and for taking care of my son whenever needed. My deepest gratitude goes to my beloved husband, Andy, for being a positive and clever person in my life and for always patiently supporting and encouraging me. My precious son, Karl Johan, I want to thank you for bringing love and happiness into my life every day.

Marie Germund Nielsen August 2017
Preface ............................................................................................................................ 9

Outline of the dissertation ............................................................................................... 10

The three papers of the dissertation ............................................................................. 11

List of abbreviations ....................................................................................................... 12

Chapter 1: Introduction and Aims .................................................................................. 13

Chapter 2: Background and theoretical framework ..................................................... 17

1.1 Mental health in primary care ................................................................................... 18

1.2.1 Patient-reported outcomes .................................................................................. 19

1.2.2 Perceived stress in general practice ...................................................................... 19

1.2.3 Perceived Stress Scale ....................................................................................... 21

1.2.4 Depression in general practice ........................................................................... 23

1.2.5 Major Depression Inventory .............................................................................. 25

Chapter 3: Subjects and methods .................................................................................. 49

1.1 The process of the PhD study .................................................................................. 50

1.2 Data sources ............................................................................................................. 51

1.2.1 The Danish National Health Survey of 2010 (Paper I) ..................................... 51

1.2.2 Sundhedsmappen.dk (Paper II-III) ...................................................................... 51

1.2.2.1 Recruitment procedure for Sundhedsmappen.dk (Paper II-III) ....................... 52

1.2.3 Munich-Composite International Diagnostic Interview (Paper III) .................... 53
PREFACE
This PhD dissertation is based on the project ‘Improving instruments for diagnosing depression and measuring stress in general practice’. The project was carried out during my time as a research fellow in the research network ‘Mental Health in Primary Care (MEPRICA)’ at the Research Unit for General Practice, Department of Public Health, Aarhus University.

Chapter 1 contains Introduction and Aims. The background for initiating this study, including the measurement of stress and diagnosing of depression, and description of methods and statistical analyses applied to measure validity is provided in Chapter 2 Background and theoretical framework. Chapter 3. Subjects and methods describes the data sources and the study populations. Chapter 4 outlines the main results of the three studies. Chapters 5–7 present the three papers of the dissertation. Chapter 8 and 9 offer a general discussion of the methods used and the presented results. Chapter 10 offer a supplementary discussion which further enlighten specific aspects of the dissertation. Chapter 11 presents the main conclusions. Chapter 12 describes the perspectives raised by the present research and offers ideas for future research. English summary and Danish summary of the dissertation are found in Chapter 13 and 14 respectively. Finally the references, which are used throughout the study, are listed. As an aid to the reader, the Appendix provides study information material and the applied measurement instruments (in Danish).


LIST OF ABBREVIATIONS

ASS  Anxiety Symptom Scheme
AUC  Area under the curve
CDR  Central Denmark Region
CI   Confidence interval
CTT  Classical test theory
DIF  Differential item functioning
DMHMM Double monotone homogeneity model of Mokken
DNHS Danish National Health Survey
DSM-V Diagnostic and Statistical Manual of Mental Disorders, 4th Ed.
ICC  Item Characteristic Curve
ICD-10 International classification of disease-10
ICP  Item Category Probability Curve
IRT  Item Response Theory
M-CIDI Munich-Composite International Diagnostic Interview
MDI  10-item Major Depression Inventory
NPV  Negative predictive value
OECD Organisation for Economic Co-operation and Development
PCM  Partial credit model
PPV  Positive predictive value
PRO Patient-Reported Outcome
PSE  Present State Examination
PSS  10-item Perceived Stress Scale
ROC  Receiver Operating Curve
RSM  Rating scale model
RUMM Rasch Unidimensional Measurement Model-2030
SCAN Schedules for Clinical Assessment in Neuropsychiatry
SEN  Sensitivity
SPE  Specificity
Stress Psychological stress
Sentinel GPs GPs participating in an internet-based data collection system (Sentinel Data Capture).

Latent variable The construct, latent trait, latent ability and latent variable are words for the same meaning in the dissertation
CHAPTER 1: INTRODUCTION AND AIMS
This chapter offers a short introduction to the dissertation and a presentation of the specific aims for the dissertation.

**Introduction**

More and more individuals seem to suffer from psychological stress (stress) in modern society. Stress is not considered an illness in itself, but merely a risk factor for developing depression and anxiety disorders. The World Health Organization (WHO) estimates that more than 300 million people of all ages suffer from depression worldwide and depression is currently the leading cause of disability.

Depression and stress are, together with anxiety, the most frequently encountered mental health problems in primary care and these problems are often either unrecognised or over-diagnosed worldwide. This also includes measuring stress and diagnosing depression. Incomplete recognition and misclassification of mental health problems are consequently also a well-known issue in general practice.

A group of patients with symptoms of stress do not meet the accepted criteria for a diagnosis of depression in general practice. These patients are still exposed to increased morbidity and mortality. Although the frequency of stress-related consultations is high, no psychometric instrument is recommended for measuring stress in general practice. The Perceived Stress Scale (PSS) is a widely used instrument for measuring perceived stress. However, the psychometric properties of the instrument have not been validated by modern test theory and Rasch analysis.

An accurate measurement or diagnosis is a prerequisite for appropriate treatment. And although the application of psychometric instruments for identification and monitoring depression in general practice is increasing in Denmark, the GPs need to be provided with better diagnostic tools. A validation is needed of the diagnostic instrument which is most often applied by GPs on clinical suspicion of depression (i.e. presence of two of three core symptoms of depression). The Danish clinical guidelines recommend using the Major Depression Inventory (MDI), but no previous studies have assessed the diagnostic accuracy of the MDI when used on clinical suspicion of depression in a general practice setting. Additionally, despite the recent technological breakthroughs, only paper and pencil versions of the MDI have so far been used.
The dissertation aims to explore the measurement of mental disorders in primary care by investigating the validity of the two selected psychometric tests, the internationally widely used PSS and the MDI which is widely used in Danish general practice.

**Aims**

1) To study the construct validity of the PSS among adults in the general population in the Central Denmark Region.

2) To study the construct validity of the web-based diagnostic instrument MDI when used on clinical suspicion of depression in patients in Danish general practice.

3) To study the criterion validity of the web-based MDI for diagnosing depression on clinical suspicion of depression in patients in Danish general practice.
CHAPTER 2:

BACKGROUND AND THEORETICAL FRAMEWORK
This chapter offers an introduction to the background and the theoretical framework for this dissertation on psychometric instruments for measuring stress (PSS) and depression (MDI). The chapter outlines the methods applied to assess the construct validity of the PSS and the MDI and the background for performing the criterion validation of the MDI.

### 2.1 MENTAL HEALTH IN PRIMARY CARE

The publicly funded health care system in Denmark provides all Danish citizens with free access to general practice, outpatient clinics and hospital care. Nearly all Danish citizens are registered with a GP. A typical GP has a list of approximately 1600 persons, and the GPs treat both physical and mental disease.

A distinction is often made between severe and long-term mental health disorders (e.g. schizophrenia) and common mental health disorders (e.g. depression or anxiety). Common mental health problems are often diagnosed and treated in general practice. Danish general practitioners (GPs) handle about 40 million patient contacts per year. Approximately 9% of all contacts primarily concern mental health problems; 46% relate to depression and 24% to stress. This makes depression and stress, together with anxiety, the most frequently encountered mental health problems in primary care.

Incomplete acknowledgement and misclassification of mental disorders is a well-known problem in general practice. In a meta-analysis of 19 studies, Mitchell, Vaze & Rao found a diagnostic sensitivity of approximately 50.1% (CI: 41.3%-59.0%) and a specificity of 81.3% (CI: 74.5%-87.3%) for unassisted diagnoses of depression performed by GPs. Additionally, 60% of the identified cases may be erroneously diagnosed due to the positive predictive value (PPV) of 42% (CI: 39.6%-44.3%). Due to the relatively low specificity and the low prevalence, this causes that many are expected to have a false positive diagnosis. This suggests that only half of the true depressive cases are diagnosed. To improve the diagnosis of mental disorders in general practice, valid diagnostic tools are needed. An accurate diagnosis is a prerequisite for appropriate treatment. In daily clinical practice, the case finding is based on the GP’s suspicion of psychiatric illness, and a high prevalence can thus be expected.
2.1.1 Patient-reported outcomes

Patient-reported outcome (PRO) data reporting on the patient’s health conditions have the potential to improve the quality of health care and can be integrated into clinical practice. PRO data can be used to understand the impact of a disease on the people who have it and what they value most in terms of symptom relief. PRO data are most often collected via standardized questionnaires designed to assess underlying constructs that are not directly measurable, such as depression. In psychometric theory, such underlying constructs are referred to as latent traits or latent variables. Individual items (i.e. a question and its response categories) are grouped into one or more domains, depending on the concept they represent. Patient-reported outcome measures (PROMs), such as psychometric tests are increasingly used in clinical practice.

2.1.2 Perceived stress in general practice

Psychological stress is a feeling of strain and pressure. It is a well-established risk factor for developing mental and physical health problems. Cohen et al. state that studies of psychological stress usually focus either on the occurrence of environmental events that are demanding for one’s ability to cope or on individual responses to events that are indicative of this overload, such as perceived stress. Perceived stress can be seen as an expression of psychological stress.

Although a stress condition carries a substantial burden, it is merely considered a ‘risk factor’. No diagnosis code for stress exists in the 10th version of the International Classification of Disease (ICD-10) or in the International Classification of Primary Care (ICPC). In the ICPC system, GPs may use the P02 for an acute stress reaction/adjustment disorder. The P02 corresponds to the F43 in the ICD-10 for a reaction to severe stress and adjustment disorders. The prevalence of stress has been widely assessed using the PSS. The proportion of people with a high stress level was 29.5% in 2010 in Denmark (a PSS score of ≥ 15 for men and a PSS score of ≥ 17 for women), whereas the proportion with a high perceived stress was 21% in a Danish population-based study from 2013. In primary care, the prevalence of stress among patients is generally high; 60–80% of visits may have a stress-related component. As a consequence, the indirect costs of poor mental health are substantial.
health are significant. Employees with poor mental health, such as stress, are less productive at work and are more likely to take sick leave. Perceived stress has been found to have considerable negative impact on a person’s physical health status. Elevated perceived stress levels are associated with an increased risk of preventable hospitalisations and poor short-term prognosis. A study by Russ et al. found a dose-response association between psychological distress and an increased risk of mortality. In our research group, we have also suggested that perceived stress contributes significantly to higher mortality rates in people with multimorbidity. Several studies have found a link between stress and depression due to release of hormones in a physiological stress response. These findings underline the need for a validated psychometric instrument for measuring perceived stress, and no instrument specifically targeting stress has so far been recommended in Danish general practice.
### 2.1.3 Perceived Stress Scale

A candidate for measuring stress in general practice is the PSS. This scale assesses the degree to which situations are perceived as stressful by the respondent.\(^{21,29}\) The original 14-item scale (PSS-14) was developed in 1983 by Sheldon Cohen.\(^{29}\) It was later revised and reduced into a 10-item and a 4-item version.\(^{21}\) In this dissertation, the 10-item PSS (PSS) will be studied (see 2.1. for English version and Appendix 3 for Danish version).

The PSS was developed on the basis of Lazarus’ assumptions that stress is both an objective operational state and a contextual condition of life that is an expression of the individual’s ability to act in relation to the both negative as positive challenges.\(^{30}\) The PSS also assesses the degree to which external demands seem to exceed the individual's perceived ability to cope.\(^{21,29}\)

The PSS asks, for example, if the respondent has found life uncontrollable during the past month. Each item is rated on a five-point Likert scale (from 0=never to 4=very often). The total score is generated by summarising responses after reverse scoring of items. Scores range from 0 to 40; high scores indicate high perceived stress.\(^{29}\)

#### Table 2.1. The Perceived Stress Scale: Cohen's original English version\(^{29}\)

<table>
<thead>
<tr>
<th></th>
<th>0 = Never</th>
<th>1 = Almost Never</th>
<th>2 = Sometimes</th>
<th>3 = Fairly Often</th>
<th>4 = Very Often</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. In the last month, how often have you been upset because of something that happened unexpectedly?</td>
<td></td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>2. In the last month, how often have you felt that you were unable to control the important things in your life?</td>
<td></td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>3. In the last month, how often have you felt nervous and “stressed”?</td>
<td></td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>4. In the last month, how often have you felt confident about your ability to handle your personal problems?</td>
<td></td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>5. In the last month, how often have you felt that things were going your way?</td>
<td></td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>6. In the last month, how often have you found that you could not cope with all the things that you had to do?</td>
<td></td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>7. In the last month, how often have you been able to control irritations in your life?</td>
<td></td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>8. In the last month, how often have you felt that you were on top of things?</td>
<td></td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>9. In the last month, how often have you been angered because of things that were outside of your control?</td>
<td></td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>10. In the last month, how often have you felt difficulties were piling up so high that you could not overcome them?</td>
<td></td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
</tbody>
</table>
The PSS was originally defined as a single construct, as the distinction between the two different dimensions in terms of the positively and negatively scored items was considered irrelevant. However, exploratory factor analysis (EFA) later indicated that a two-dimensional structure was more dominant. A confirmatory factor analysis (CFA) by Andreou et al. confirmed that the one-dimensional model did not provide acceptable fit, whereas the two-dimensional model showed a better fit for the PSS. A principal component analysis (PCA) supported the existence of two dimensions: one dimension related to perceived stress (measured by six negatively worded items), whereas another related to coping ability and stress resilience (measured by four positively worded items).

The psychometric properties of the different versions of the PSS have been studied in many countries, but mainly with classical test theory (CTT). Cronbach’s alpha (α) is a measure of internal consistency reliability, and a value of >.70 is considered a minimum measure of internal consistency. Several studies conducted in the general population have found that Cronbach’s α for the total scale ranges between 0.75 and 0.91. The PSS seems to have unsolved issues as several studies (based on CTT or IRT models) have indicated problems with unidimensionality. These unsolved issues might have a consequence for the validity of the PSS score as a measure of stress in a population.
2.1.4 Depression in general practice

Depression in general practice is often viewed in terms of the stress-vulnerability model. This model states that getting symptoms is the result of long-lasting vulnerability factors (e.g. genetic risk, early life experience, physical illness and lack of social support) in interaction with exposures to environmental stressors. Depression without mania is sometimes referred to as unipolar because the mood remains at one emotional state or ‘pole’. Unipolar depression is among the most common mental disorders and is a major contributor to the global burden of disease.

The essential features of major depression are the presence of depressed mood, loss of interest and pleasure, together with at least four other accompanying symptoms (e.g. disturbed sleep and poor concentration) for a duration of at least two weeks. Differential diagnostic considerations are bipolar disorders, somatic disorders, medical side effects, psycho-organic syndromes, substance abuse, psychotic disorders, anxiety disorders, somatoform and personality disorders.

Most patients with depression are treated in general practice. There are many reasons why depression goes unrecognised in primary care. Patients may attribute their symptoms to physical illnesses, or they may not realise that they need help. Since 2007, the GPs in Denmark have had the opportunity to refer patients aged 18-37 years (from 1 July 2012, > 18 years) with mild to moderate depression to a psychologist for 12 sessions of psychotherapy with a self-payment of about a third of the total fee. This treatment option is widely used.

Public confidence in the ability of GPs to manage depression has been challenged by current concerns about Selective Serotonin Reuptake Inhibitor (SSRIs), which has received much attention in the media.

Internationally, the Patient Health Questionnaire (PHQ-9) is often used to diagnose depression in the primary care setting. A recent diagnostic meta-analysis of depression in primary care by Mitchell et al. reported that the PHQ-2 is suitable for initial assessment, and that it should be followed by either the PHQ-9 or another similar suitable tool. This was supported in a study by Arrol et al. A study by Sherina et al. found that the additional question “During the past month, have you often been bothered by..?” increased the specificity to 95% (95% CI: 89-98%).
Another widely used instrument is the Beck Depression Inventory-II. The guidelines on depression from the Danish College of General Practitioners recommend screening with three ICD-10 core symptoms of depression: low in spirits, lost interest and lack of energy (see Figure 2.1). The screening result is considered positive if the patient has at least two of the three core symptoms. If two core symptoms of depression are present, the guidelines recommend diagnostic assessment by the Major Depression Inventory (MDI), which is a self-report questionnaire (see Table 2.2 and Appendix 1 for Danish version). The use of psychometric instruments has been increasing in general practice in Denmark (see Appendix 2). The GPs are reimbursed a fee of DKK 130.43 (rate per October 2016), corresponding to EUR ≈17.50, for administering a psychometric test for depression.

Figure 2.1 Core symptoms of depression for diagnosing depression in primary care patients. Adapted from Whooley et al. and the guidelines on unipolar depression by the Danish College of General Practitioners.
2.1.5 Major Depression Inventory

The MDI was developed with reference to both the ICD-10 and the DSM-IV classification systems. The MDI explores the ICD-10 criteria for depression in a questionnaire form. The scale was originally developed for diagnosing depression in general practice in Denmark in collaboration with the WHO. The MDI was released by Bech and

<table>
<thead>
<tr>
<th>Table 2.2. Major Depression Inventory</th>
<th>The following questions ask about how you have been feeling over the past two weeks. Please put a tick in the box, which is closest to how you have been feeling.</th>
</tr>
</thead>
<tbody>
<tr>
<td>How much of the time...</td>
<td>All the time</td>
</tr>
<tr>
<td>1. Have you felt low in spirits or sad?</td>
<td>5</td>
</tr>
<tr>
<td>2. Have you lost interest in your daily activities?</td>
<td>5</td>
</tr>
<tr>
<td>3. Have you felt lacking in energy and strength?</td>
<td>5</td>
</tr>
<tr>
<td>4. Have you felt less self-confident?</td>
<td>5</td>
</tr>
<tr>
<td>5. Have you had a bad conscience or feelings of guilt?</td>
<td>5</td>
</tr>
<tr>
<td>6. Have you felt that life wasn’t worth living?</td>
<td>5</td>
</tr>
<tr>
<td>7. Have you had difficulty in concentrating, e.g. when reading the newspaper or watching television?</td>
<td>5</td>
</tr>
<tr>
<td>8 a. Have you felt very restless?</td>
<td>5</td>
</tr>
<tr>
<td>8 b. Have you felt subdued or slowed down?</td>
<td>5</td>
</tr>
<tr>
<td>*9 a. Have you slept too little?</td>
<td>5</td>
</tr>
<tr>
<td>*9 b. Have you slept too much?</td>
<td>5</td>
</tr>
<tr>
<td>10 a. Have you suffered from reduced appetite?</td>
<td>5</td>
</tr>
<tr>
<td>10 b. Have you suffered from increased appetite?</td>
<td>5</td>
</tr>
</tbody>
</table>
Wermuth in 1998 and has been translated into 20 different languages, including English. The severity of depression is measured over the last two weeks in the form of a Likert scale in which the frequency of each symptom can be indicated from 0 (at no time) to 5 (all the time). According to the Psychiatric Research Unit, where the MDI was originally developed, the MDI can be used in two ways: a) as a diagnostic instrument or b) as a severity scale of depression. A core symptom is present if the score is at least 4, and an associated symptom is present if the score is at least 3. The ICD-10 algorithm is coded as: mild depression (at least two core symptoms and two associated symptoms), moderate depression (at least two core symptoms and four associated symptoms) or severe depression (at least three core symptoms and five associated symptoms). The core items are indicated by items 1-3, while the associated symptoms are indicated by items 4-10.

The MDI has a severity rating score of 0-50. The cut-points for the total score are: no depression (≤ 19), mild depression (20-24), moderate depression (25-29) and severe depression (≥30).

The scoring instructions in the manual for the MDI state that, for items 8, 9 and 10, alternative a or b with the highest score should be considered. Eating and sleeping too much at the same time are not typical signs of depression. Therefore, it is important to distinguish between too little and too much sleep. Studies suggest that up to 30% of the patients who are treated in primary care have atypical depression. In Denmark, GPs are recommended to retest the patient again after two weeks using the MDI before they start up any treatment.

Only few studies have assessed the prevalence of depression with the MDI. Ellervik et al. found a 2.3% prevalence of depression according to the MDI in a general population. Amris et al. found a prevalence of 29% for any depression according to the MDI in a population of chronic pain patients. More often diagnostic interviews are used for assessing the prevalence of depression. In a study by Toft et al. the present state prevalence of depression was 13.5% (11.1–16.3) in a waiting room population when estimated with the SCAN interview for Mood disorders (F30–F39) the prevalence of moderate/severe depression was: 2.7% (1.9–3.8). To the best of my knowledge the
prevalence of depression on clinical suspicion has not previously been established in general practice. A study by Cuijpers et al. investigated the criterion validity of the MDI and found a sensitivity of 66% and a specificity of 65% in an investigation of the presence of major depressive disorder in a consecutive sample of 258 psychiatric outpatients. The area under the curve (AUC) was 0.68. The MDI has been validated using Present State Examination (PSE) in a sample of 43 subjects with a spectrum of depressive symptoms in a psychiatric department. The sensitivity of the MDI algorithms for major depression varied between 86% and 92%, whereas the specificity varied between 82% and 86%. A study by Amris et al. including a Rasch analysis of the MDI found problems with the rating scale properties and lack of unidimensionality in females with chronic widespread pain. Misfit was identified for four items: appetite (two items), sleep and feelings of guilt/bad conscience. A study by Olsen et al., which was based on PCA and Rasch analysis in a sample of non-depressed and moderately depressed patients, supported a unidimensional model, even though the analyses showed misfit for the sleep item. A total Mokken coefficient of (0.52) was found, while the sleep and appetite items showed suboptimal fitting (Loevinger’s coefficients <0.40). Ellervik et al. found that the MDI had adequate measures of scalability, corresponding to a Mokken coefficient of homogeneity above 0.40.
The following sections will describe the technical specifications regarding validity and measurement theory.

2.1.6 Validity of psychometric scales

There are divergent views on the concept of validity. According to the CONsensus-based Standards for the selection of health Measurement INstruments (COSMIN checklist)\(^6\), three domains are considered when assessing the quality of a measurement scale: reliability, validity and responsiveness (see Figure 2.2). Each domain contains one or more measurement properties. Validity contains three measurement properties: content, construct and criterion.\(^6\) Construct validity consists of structural validity, cross-cultural validity and hypothesis testing.\(^6\) This PhD dissertation will focus on the construct and criterion validity of selected instruments. As only selected existing instruments are assessed and validated, this dissertation does not address hypothesis testing or explore the concurrent, cross-cultural or content validity of the investigated scales.

![The COSMIN taxonomy for quality of Health Related-Patient Reported Outcomes (HR-PRO).](https://example.com/cosmin.png)

2.2 CONSTRUCT VALIDATION

The construct validity assesses the degree to which a scale is valid and appears to measure the construct that it was designed to measure. Messick described the construct validity as the overarching concept of validity. Others have stated that assessing the construct validity of an instrument is an ongoing process. Consequently, construct validity is unlikely to be fully achieved.

Construct validation refers to a wide range of approaches used when trying to measure a ‘hypothetical construct’, like depression, rather than something which can be readily observed. In our studies, the primary focus has been on construct validity. We have specifically investigated multi-item scales with several items reflecting the same latent variable. In the following; the construct, latent trait, latent ability and latent variable are words for the same meaning and in the dissertation the “latent variable” will most often be used.
2.2.1 Classical test theory

In psychometric testing, there are essentially two theoretical approaches: traditional methods (known as CTT) and modern test methods (known as latent-trait theory or IRT). CTT focuses on construct validity but also on reliability. A reliable test is one that measures the construct in a consistent, repeatable and reproducible manner. In this dissertation, the focus is only on the internal consistency part of reliability. Internal consistency is a measure based on the correlations between different items on the same test, and it measures whether several items that are intended to measure the same general construct produce similar scores.

CTT assumes that each person has a true score, which would be obtained if there were no errors in the measurement. It is assumed that the observed score equals the true score plus some error. When applying CTT, it is necessary to make an assumption of normally distributed data. CTT assumes that intervals between the values (response categories) are equally spaced, which is a problem since measurements are usually ordinal. On an ordinal scale, the distance between two scores are not always the same (see figure 2.3.), which illustrates that the distance between a score of 3 and a score of 4 on the ordinal scale is not the same as the distance between, e.g. a score of 8 and 9. However, the distances between these scores (or any scores next to each other) are always the same on an interval scale. In clinical practice, the assumption of equal intervals is seldom fulfilled (see Figure 2.3).

CTT analysis often relies on two statistical measures: item difficulty and item discrimination. Item difficulty expresses the proportion or percentage of respondents who affirmed the item. In CTT, item discrimination is determined as an index of the item-total correlation. An essential assumption is unidimensionality, i.e. that each item is a manifest of the underlying construct. CTT focuses primarily on test-level information and is assessed with correlations and primary Factor Analysis (FA). A limitation of CTT is that scale evaluations are sample dependent.
2.2.1.1 Confirmatory factor analysis

In CTT, the number of latent variables or factors that account for the variation or covariation among a set of observed measures is often assessed by factor analysis (FA). A factor is an unobservable variable that influences more than one observed measure and accounts for the correlations among these observed measures. FA is a general approach which can model a number of factors at the same time by using the inter-item correlations and standard deviation (SDs) to estimate the models and carry out statistical ‘goodness-of-fit’ tests. The factor structure models are linear combinations of the observed variables, with the latent variables being estimated by weighted summation.

There are two approaches to FA: an exploratory (EFA) and a confirmatory (CFA). CFA aims to reproduce the observed relationships among a group of indicators with a smaller set of latent variables. However, CFA differs fundamentally from EFA by the number of a priori specifications and restrictions made on the latent variable measurement model.

CFA is used to verify the number of underlying dimensions (factors) of the instrument and the pattern of item-factor relationships (factor loadings). The analysis for the construct validity of the PSS also contained CFA to further study the dimensionality and to allow comparison with existing validation studies of the PSS. A number of indices are available to assess the fit of a model on the basis of categorical data. These are all presented in Paper I (page 25-27). In FA models both the latent variables and the items are assumed to be located on an interval scale. However, through the broader framework of Structural Equation Models (SEM), the latent trait model has gained a foothold. Here, the latent variables are placed on an interval scale, and the items are located on an ordinal scale. These developments of the FA models are particularly seen as a way of establishing measurement invariance.
2.2.2 Item response theory
IRT refers to a framework encompassing a group of models, and the Rasch model applied in this dissertation is one of these models. When using IRT, it is not necessary to make an assumption of normally distributed data as in mainstream CTT. IRT thus represents a group of models which treats the difficulty of each item as information to be incorporated in the model. Several IRT models have been developed with different numbers of parameters. The 1-parameter Rasch model will be described in detail in section 2.3. More often, the 2- and 3-parameter models are used. The 2- and 3-parameter models of IRT accept variations in the shape of the item response curves. The 1-, 2- and 3-parameter models differ, however, in the number of parameters they allow to vary. The 2- and 3-parameter models allow for more deviations by having more parameters. When selecting the model to use, the most stringent model could be chosen to accurately represent the observed data.

The assumptions shared by all IRT models are unidimensionality, monotonicity, local independence and no differential item functioning (DIF). According to Rosenbaum, indirect measurement by a set of item responses is criterion-related construct valid if these assumptions are met. A scale is unidimensional if the items measure only one trait or ability. A single unobservable latent variable lies behind the item performance on the test, referred to as theta. The strict mathematical assumption of unidimensionality in IRT models is rarely fully met in practice. However, even if unidimensionality is recognized as being relative, the construction of a set of items intended to measure a unidimensional latent variable still requires knowledge of the subject matter. Unidimensionality is the basis of IRT and can be explained in terms of the Guttman scaling. A Guttman scale is considered the ideal in measurement theory (see section 2.2.2.1). Multidimensionality is present if items belong to more than one dimension or if items form subgroups where each subgroup belong to different factors/dimensions.

Items are assumed to relate monotonically to one dimension. The item scores are monotonically increasing functions of the latent variable. This property implies that the total score for all items will be positively correlated with all variables, which are known to be positively correlated with the latent variable.
Local independency is defined as the response to any one item being independent on the response to any other item after controlling for the underlying trait. In local independency, the term local refers to the idea that all the variation among responses to an item is accounted for by the latent variable. Therefore, for the same value of the underlying trait, there are no further relationships among responses.69

Differential item functioning (DIF) is a type of item bias that can occur when groups within the sample (e.g. males and females) respond differently to an item despite equal levels of the underlying trait.83 When one subgroup (e.g. females) consistently scores differently on an item across all levels of the trait, this is known as uniform DIF. When uniform DIF is present, the difficulty of the item depends on gender in the same way for all levels of the latent variable.82 When the effect of gender differ across the levels of the trait there is non-uniformity in the differences between the groups this is known as non-uniform DIF.84 Rather than giving a consistent advantage to the reference group, the conditional dependency moves and changes direction at different locations. It is possible to adjust for uniform DIF since the item of interest consistently gives one group an advantage across all levels of ability. While the conditional dependency for non-uniform DIF moves and changes direction at different locations on the continuum.72

2.2.2.1 The Guttman pattern and unidimensionality

A Guttman scale consists of several items of varying difficulty and is rigidly hierarchical as items are ranked according to the difficulty they appear to represent for the respondents. If a respondent can accomplish a difficult task at the upper end of the scale, the respondent must be able to accomplish all of the easier tasks.69 To illustrate the Guttman pattern, an example with dichotomous data (with two possible response categories) is used for simplicity in the rest of section 2.2. Information on the polytomous model (used for items with more than two possible responses) will be given in section 2.3. The Guttman scale is seen as the ideal in terms of evidence of unidimensionality (see Figure 2.4.a).
Validation of instruments for diagnosing depression and measuring stress in general practice

Figure 2.4.a. Guttman pattern.
Persons illustrated by letters and items by numbers

Figure 2.4.b. Non-hierarchical pattern.
An example: Guttman errors for person A on items 3 & 5, for person B on items 4 & 6, for person C on items 4 & 6, for person D on items 1 & 2 and for person F on items 1 & 3.

Figure 2.4.c. Probabilistic Guttman pattern (imperfect pattern).
The above figures 2.4.a., 2.4.b. and 2.4.c. have all been adapted from Streiner.  

The above figures 2.4.a., 2.4.b. and 2.4.c. have all been adapted from Streiner.  

12
Figure 2.4.a. presents a perfect Guttman pattern; person A to F represent six persons completing a questionnaire, and items 1 to 6 represent six items in the scale. The response ‘1’ means that the item is confirmed, and ‘0’ means that the item is not confirmed. For example, if the six items formed a depression scale, person A would be the least depressed person and person F the most depressed person. Item 1 would be measuring the mildest aspects of depression and item 6 the most severe aspects of depression. The unidimensionality of the Guttman scale is of absolute importance because the response pattern is perfect. For example, agreement with item 3 implies agreement with items 1 and 2.

Figure 2.4.b. presents a non-hierarchical pattern with Guttman errors. The scores do not follow any pattern, thereby making it difficult to determine which patient would affirm most items or which item would be affirmed by most patients. Figure 2.4.c. illustrates a probabilistic Guttman pattern. For example, person A did not affirm item 2 but affirmed item 3, although item 2 was affirmed depression in five other persons. This pattern reflects what often happens in real life. There is a discrepancy between the observed and the expected response. This type of probabilistic imperfect response pattern (as seen in Figure 2.4.c.) is allowed in IRT models, but only to a certain degree. Item responses from the observed are expected to be close enough to what is expected under the model. The number of Guttmann errors are statistically insignificant if the model requirements are fulfilled.12

2.2.2.2 Item response theory graph

IRT models, such as the Rasch model, can be presented as an IRT graph (see Figure 2.5). The graphs capture the assumptions of unidimensionality, local independence and absence of DIF, which are assumptions for all IRT models.72 The properties of the model can be discussed from both a statistical and a psychometric point of view. In the following, the focus will only be on the psychometric properties and the visual display, which is a way to illustrate the concept that items depend only on the latent variable.72 Graphical IRT models allow exogenous variables (independent variable that affects a model without being affected by it) to be added to the graph.85 The graphical IRT model consists of items (items 1-4), the latent variable and exogenous variables (covariables 1-3). The items and exogenous variables are assumed to be independent given the latent
variable, e.g. depression (see Figure 2.5). In Figure 2.6., the violation of the assumption about no DIF is illustrated by the red arrow between covariable 2 and item 1. Violation of the assumption of local independency between items is illustrated by a red arrow between item 3 and item 4. The violation of the assumption of unidimensionality is illustrated with an extra dimension that influences items 1 and 2 (red arrow).

**Figure 2.5. IRT graph.** Adapted from “Rasch Models in Health” Edited by Karl Bang Christensen & Svend Kreiner 2013.72

**Figure 2.6. IRT graph with violations of assumptions.** Adapted from “Rasch Models in Health” Edited by Karl Bang Christensen & Svend Kreiner 2013.72
2.2.3 Mokken scale

Mokken scale analysis (Mokken analysis) is a nonparametric IRT model used to study the properties of a set of items in the IRT framework. Mokken’s definition of a scale is simply a mathematical one it does not guarantee validity. The purpose of the Mokken analysis is to test an ordinal measure of a latent variable, e.g. stress. If a given group of items satisfies the criteria of the Mokken model based on the scalability of coefficients, the sum of the responses across items can be used to rank respondents on the latent trait. The Loevinger’s H coefficient summarise the number of errors according to the Guttman scalogram: the number of times an incorrect answer is given to an easy item while a correct answer is given to a difficult item.

Higher values of H mean that more confidence can be held in the ordering of the respondents on theta (ability). Mokken states that there is no empirical basis for the interpretation of the value of H. Though Mokken suggest the following classification of scales; a Loevinger coefficient of 0.30–0.39 is a weak scale, whereas a coefficient of 0.40-0.49 is considered an medium scale and a coefficient above 0.50 is a strong scale.

Mokken analysis can take the form of either the monotone homogeneity (MHMM) or the double monotone homogeneity model of Mokken (DMHMM). The assumptions behind the MHMM are thus unidimensionality, monotonicity and local independence. The assumptions behind the DMHMM are the same, with an additional feature of non-crossing response functions and thus also invariant item ordering. The MHMM is found to be a suitable alternative to more restrictive models for basic testing applications.

The Mokken analysis was conducted with Stata statistical software, version 14 (Stata Corporation, College Station, Tex.). The Stata LoevH command provides the values of the Loevinger’s H coefficients. And the Mokken scale procedure (MSP) algorithm for constructing subscales allows selecting items from a bank of items that satisfy a Mokken scale.
2.3 RASCH ANALYSIS

The dichotomous Rasch model is a one-parameter IRT model and it is the simplest Rasch model. In the Rasch model, the probability of a response to any particular item is a function of the difference between the ability of the person (e.g. the level of depression) and the difficulty (location) of the item (e.g. the level of depression implied by the item). 73, 91 Both of these person and item parameters refer to the same scale of the continuous latent variable (ability). Like the Guttman scale, the Rasch model assumes that the ‘easier’ the item is to endorse, the more likely it will be affirmed; and the more affected the respondent is, the more likely the respondent will affirm a more difficult item compared to a less affected individual. The Rasch model assumes a probabilistic Guttman response pattern (Figure 2.4.c.) as it takes into account that measurement errors exist and that no respondents are 100% logical in their responses. 92

In Figure 2.7, the formula is presented. The following presentation of the dichotomous model is based on the manual for the applied software program Rasch Unidimensional Measurement Model-2030 (RUMM-2030). 92 The polytomous model is applied in our studies, but we have chosen not to describe the advanced statistics behind the polytomous model in this dissertation. Interested readers can find the formula in the Manual for the RUMM-2030 program “Part II: Interpreting Polytomous Data” on page 3. 92 The locations of the items are denoted by $\delta_i, i = 1, 2, ... I$ and those of the persons by $\beta_n, n = 1, 2, ... N$. 
In the Rasch model for dichotomous items, i.e. two ordered categories, there is only one parameter for an item; its location \( \delta_i \). In psychometric literature, this is called a threshold, which is the point at which each of the responses has a 50% equal chance of occurring.

The simplest possible response structure is scored as \( x_{ni} = 0 \) or \( x_{ni} = 1 \).

\[
\begin{align*}
  n &= 1, \ldots, N \quad (N=\text{number of persons}) \\
  i &= 1, \ldots, I \quad (I=\text{number of items})
\end{align*}
\]

In the formula of the dichotomous Rasch model,

\[
P(x_{ni} = x) = \frac{e^{x(\beta_n - \delta_i)} }{ 1 + e^{x(\beta_n - \delta_i)} }
\]

In the formula of the dichotomous Rasch model, \( P(x_{ni} = x) \) is the probability that a person \( n \), who is affected to a certain degree \( \beta_n \), will affirm an item \( i \) with the location \( \delta_i \).

Specifying \( x_{ni} = 0 \) or \( x_{ni} = 1 \) gives:

\[
\begin{align*}
  \text{e.g. response=\text{no}} \quad P(x_{ni} = 0) &= \frac{e^{0(\beta_n - \delta_i)} }{ 1 + e^{0(\beta_n - \delta_i)} } = \frac{1}{1 + e^{0(\beta_n - \delta_i)}} \\
  \text{e.g. response=\text{yes}} \quad P(x_{ni} = 1) &= \frac{e^{1(\beta_n - \delta_i)} }{ 1 + e^{1(\beta_n - \delta_i)} } = \frac{e^{(\beta_n - \delta_i)}}{1 + e^{(\beta_n - \delta_i)}}
\end{align*}
\]

\textbf{Figure 2.7. Formula for the dichotomous Rasch model.}\(^92\)

The probability that a person \( n \) gives a positive answer to an item \( i \) is dependent on the value that the latent variable has for the person in question and for the location of the item \( i \). If a person responds positively to an item for which the severity is exactly the same as the degree to which the person is affected, then \( \beta = \delta \) and \( P(x_{ni} = x) = 0.5 \) (50%). If the severity of the item is less than the degree to which a person is affected \( (\beta > \delta) \), the probability will be greater than 50%. Likewise, when \( (\beta < \delta) \), the probability will be less than 50%.

The polytomous Rasch model accounts for more than two response categories and can take the form of either the Partial Credit Model (PCM), which was introduced by Masters\(^93\), or the Rating Scale Model (RSM), which was introduced by Andrich.\(^94\)

The RSM specifies that a set of items share the same rating scale structure. The PCM specifies that each item has its own rating scale structure, and the thresholds in the PCM may differ between items.\(^95\) Instead of the two response categories, there are now more, which need more estimates to describe item difficulty. The restricted RSM assumes the distance between the thresholds to be equal across items.\(^71\)
2.3.1 Requirements for the Rasch analysis

The Rasch model is considered a valuable reference standard because: it can include data on an ordinal scale, and it provides formal representation of perfect measurement. Besides the assumptions for IRT models (unidimensionality, monotonicity, local independence and no DIF), a measure fitting the Rasch model generally possesses: specific objectivity and sufficiency.\textsuperscript{81, 96-98} Specific objectivity and sufficiency of the raw score are inherent to the Rasch model.

Measurement is \textit{specifically objective} when the comparison of two persons are independent of everything but the two persons and their observed relation. A consequence of the principle of specific objectivity is that the estimated difference in ability between persons is independent of the difficulty of any particular test item used to compare them.\textsuperscript{99}

When an item set fits a Rasch model, the scale will possess what is known as \textit{statistical sufficiency}.\textsuperscript{100} This means that the summary score will yield all relevant information about the person, and there is therefore no additional information than what is contained in the total score. Also the number of positive answers for an item is sufficient to estimate item location.\textsuperscript{83}

The Rasch model will thus be taken as a criterion for the structure of the responses rather than simply a statistical description of the responses from patients.\textsuperscript{101}

To assess the internal consistency, the Person Separation Index (PSI) was investigated. A PSI is an index which is applicable for the scale as a whole. The PSI is based on the estimated locations of the persons who are non-linear transformations of the raw scores. This is equivalent to Cronbach’s alpha, although the logit value (linear person estimate) is used instead of the raw score in the same formulae.\textsuperscript{102} When the persons and items are well aligned, then the two indices are very close in value.

2.3.2 Item Characteristic Curve

The item Characteristic Curve (ICC) is a graphical expression of each item within IRT models, such as the one-parameter Rasch model and the two- and three-parameter logistic models.\textsuperscript{72} Thus, the ICC shows that the probability that the person will affirm the item (y-axis) increases as the person’s ability increases on the latent variable (x-axis) (see Figure 2.8). This is the point at which a person has a fifty percent likelihood of affirming
or not affirming the item (red line). The item threshold is considered the item’s level of location along the latent variable.\textsuperscript{94}

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{item_characteristic_curve.png}
\caption{Item characteristic curve of dichotomised item. Example from Comins J.D. 2012.\textsuperscript{103}}
\end{figure}

2.3.3 Category probability curve

The graph showing the relation between the probability of a given response category as a function of person location is referred to as a category probability curve (CPC) (see Figure 2.9). The threshold corresponds to the location on a latent variable at which it is equally likely that a person will be classified into neighbouring categories and, therefore, will obtain one of two scores. The first threshold of an item is the location on the continuum at which a person is equally likely to obtain a score of 0 or 1 and so on.\textsuperscript{92}

\begin{itemize}
\item \textbf{Item 8 (well-functioning category responses)}
\item Item: 8: ‘Have you felt very restless/slowed down?’
\item Item: 9a/9b: ‘Have you slept too little/Have you slept too much?’
\end{itemize}

A polytomous version of the CPC is illustrated in Figure 2.9 for item 8 and item 9 from the MDI with six response categories. Item 8 from the MDI is an example of a well-functioning category response scale, whereas item 9 from the MDI is an example of disordered category response.
A scale can fulfil the Rasch model without having ordered response categories, as described by Adams et al. the ordering of the Rasch rating model thresholds is not connected to the ordering of the item response categories.105 Andrich states that the rationale for the underlying Guttman structure is that the thresholds of an item are required to be ordered.106 Inspired by Andrich’s work, we chose to investigate the ordering of the thresholds.

2.3.4 Model comparison

2.3.4.1 Classical Test Theory versus Item Response Theory
Whereas CTT focus upon absolute measures such as averages, IRT focuses on probabilities of responses.69 This is because CTT items and person parameters are sample dependent, whereas IRT items and person parameters are sample independent if the data fit the model.107 Original CTT was developed for continuous data, which was one of the differences between CTT and IRT, but due to e.g. computer development, CTT is no longer restricted to continuous data.

2.3.4.2 Confirmatory Factor Analysis compared to Rasch analysis
CFA models rely on one particular item for defining the latent variable, while the Rasch model allows for testing whether an item follows the model or not.108 CFA is a method for detecting the dimensionality of a measurement procedure, and Rasch is a method for measuring persons and items on a single dimension. CFA is not a measurement procedure – it is a statistical procedure for describing relationships among items in terms of how they correspond to a hypothesized factor structure.77 The Rasch model produces estimate measures for persons and items and then uses fit statistics to describe departures from the measurement properties.109

2.3.4.3 Mokken scale analysis compared to Rasch analysis
For item responses fitting the Rasch model, the assumptions underlying the Mokken model of double monotonicity are met.87 Stochl stated that non-parametric IRT, such as the Mokken model, is a natural starting point for Rasch item analysis.87 The major advantage of the non-parametric IRT compared to more commonly used IRTs, such as the Rasch model, is that non-parametric IRT relaxes some of the strong assumptions
about the nonlinear behaviour of response probabilities, and the Mokken model is easy to apply.\textsuperscript{110} Since the Mokken model is considered a good starting point for a Rasch analysis we applied it as an initial step in our studies.\textsuperscript{90}

2.3.5 RUMM-2030
Both our datasets for Paper I and for Paper II were based on polytomous data, and the Rasch analysis was performed using the statistical software programme RUMM-2030 (RUMM) for Windows \textsuperscript{792}, which was developed by David Andrich, originator of the Rasch-Andrich rating scale model.\textsuperscript{72} RUMM can perform analyses with different numbers of response categories and different parameterisations. The estimation in Rumm is based on pairwise conditional maximum likelihood estimation.\textsuperscript{72} The creation of projects, the running of analyses and the performance of reanalysis are all steps that are facilitated in RUMM.\textsuperscript{92} The RUMM can produce a wide range of multicolor charts. Another advantage is that RUMM provides many statistical details (including the statistical key findings that form the basis for conclusions) that are related to normal distributions in the ANOVA framework.

2.3.6 Rasch analysis strategy
The Rasch analysis was performed by applying the procedure described by Pallant & Tennant in 2007.\textsuperscript{84} The assumptions behind unidimensionality, monotonicity, local independence and DIF have been described in detail in section 2.2.3. Assessing the internal construct validity involves investigating the overall scale fit, the individual item fit and the response structure.\textsuperscript{84, 102} The fundamental aspects of Rasch analysis are described more in detail in Paper I and Paper II. The following steps were applied for the Rasch analysis of the PSS and the MDI.
First, the mathematical derivation of the Rasch model was chosen. The likelihood ratio statistical test was performed in RUMM to determine whether to assess the RSM or the PCM.\textsuperscript{102} The PCM was applied in both Rasch studies conducted in the dissertation.
The fit of data to the Rasch model with the original dataset was then investigated. A significant chi-square result indicates that the hierarchical ordering of the items varies across the trait, which compromises the required property of invariance.\textsuperscript{102} When
assessing polytomous scales, the category structure is examined by investigating the CPC for each item for disordered thresholds. Targeting of items and persons was investigated by examining the locations of the items and persons. Comparison of the mean location score (obtained for persons with the value zero set for the items) provides an indication of how well targeted the items are for people in the sample.

If any items presented misfit, the analysis explored, on a general level, ways to account for any misfit found within the scale. The iterations of the analyses involved item removal and collapsing of non-fitting response categories. This process was performed by first excluding the most misfitting item, and this was done one item at a time. If the model was still not fulfilled after deleting the misfitting items from the analysis, collapsing of response categories for misfitting response categories was performed. In the PSS study we only collapsed the response categories for the misfitting items. While in the MDI study only two items were satisfied with six response categories, which argued for changing the response categories consistently for all items. In both studies if more than one item presented misfitting response categories, we choose to collapse the response categories in the same way for all items with misfitting response categories.

There are different ways to test the assumption of unidimensionality. Rasch programs usually provide a principle component analysis (PCA) of the residuals.

2.3.6.1 DIF analysis for the Major Depression Inventory
For the construct validation of the MDI, DIF analysis was performed for gender, age, work status and education as part of Paper I.

2.3.6.2 Additional DIF analysis for the Perceived Stress Scale
After completion of Paper I, additional analyses were performed for DIF for the PSS to investigate whether these analyses would have altered the results. These analyses were first conducted in the original sample and subsequently in ten different random subsamples of 500 to take into account the large sample size of the dataset. The following exogenous variables were chosen for the analysis of DIF in the ten random samples; age, native language, long term illness, anxiety and years in school. These variables were chosen due to their availability in the dataset and since we hypothesised that one or more items in the scale performed differently in some of these subgroups.
2.4 CRITERION VALIDATION

Criterion validity and the accuracy of a test are considered to determine whether a disease is truly present or not; an indication of the truth is often referred to as reference standard. Criterion validity can be divided into concurrent validity and predictive validity. We studied concurrent validity by comparing a questionnaire test with a diagnostic interview. The Standards for Reporting of Diagnostic Accuracy Studies (STARD) guidelines were followed in our study together with the COSMIN checklist for criterion validation studies.

2.4.1 Diagnostic interviews for depression
Several diagnostic interviews for assessing depression exist, such as the Structured Clinical Interview (SCID), the Mini-International Neuropsychiatric Interview (MINI) and Schedules for Clinical Assessment in Neuropsychiatry (SCAN), also known as PSE. The WHO developed the Composite International Diagnostic Interview (CIDI), which is intended for both clinical and research purposes. The World Mental Health CIDI (WHO-CIDI) consists of 41 sections. The first is an introductory screening and lifetime review section. There are also 22 diagnostic sections for assessing disorders, for example mood disorders and anxiety disorders. The Munich-Composite International Diagnostic Interview (M-CIDI) is a comprehensive, fully-structured, standardised diagnostic computer-assisted interview designed for trained certified lay interviewers for assessment of mental disorders according to the ICD-10 or DSM-IV criteria. Compared to the standard WHO-CIDI (version 1.2), the M-CIDI also allows for an evaluation of additional DSM-IV diagnoses. The M-CIDI was used as a telephone interview that served as a reference standard to measure the prevalence of ICD-10 depression in the criterion validation of the MDI.

The CIDI has been shown to have good test-retest and inter-rater reliability. In a study by Jordanova et al. among primary care attendees, the CIDI was found to be a highly valid assessment tool for common mental disorders when compared to the SCAN interview. However, the CIDI tended to overdiagnose; the prevalence of any depressive episode or disorder was 18.1% compared with 7.6% with the SCAN interview. In a study by Brugha et al. comparing the SCAN interview and the CIDI interview, the kappa coefficient for any
depression was found to be 0.39; this means that only fair concordance was found.\textsuperscript{117} In a study by Andrews et al., the inter-rater reliability of the CIDI was found to be perfect (overall intra-class kappa = 1.0), whereas the SCAN was found to be good (overall intra-class kappa = 0.67). This supports our choice of using the M-CIDI in our study.\textsuperscript{118}

2.4.2 The accuracy of a test result
The MDI was compared with the M-CIDI to study the accuracy of the test and assess the criterion validity of the MDI.

A diagnostic test has two purposes: finding persons with disease (sensitivity) and identifying healthy persons (specificity)\textsuperscript{121}. As no psychometric test is perfect, the sensitivity and specificity are never both 100%. For example, persons who present with some symptoms of depression and who test positive for having depression may not have a depression (false positive). Likewise, persons with normal test results may have a depression (false negative) (see Table 2.3). The PPV is the probability of disease in a patient with a positive test result, and the negative predictive value (NPV) of a diagnostic test is the probability of not having the disease when the test result is negative. These outcomes are important for the choice of action and from decision-making and socioeconomic perspectives.\textsuperscript{121}

‘A’ symbolises persons with a disease identified by a psychometric test.

Table 2.3. Possible outcomes of medical testing

<table>
<thead>
<tr>
<th>Persons with disease</th>
<th>Healthy persons</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive test</td>
<td>A (TP)</td>
</tr>
<tr>
<td></td>
<td>B (FP)</td>
</tr>
<tr>
<td></td>
<td>A + B</td>
</tr>
<tr>
<td>Negative test</td>
<td>C (FN)</td>
</tr>
<tr>
<td></td>
<td>D (TN)</td>
</tr>
<tr>
<td></td>
<td>C + D</td>
</tr>
<tr>
<td></td>
<td>A + C</td>
</tr>
<tr>
<td></td>
<td>B + D</td>
</tr>
<tr>
<td></td>
<td>N = A + B + C + D</td>
</tr>
</tbody>
</table>

Sensitivity: $A/(A+C)$  
Specificity: $D/(B+D)$  
PPV: $A/(A+B)$  
NPV: $D/(C+D)$  
TP: True positive finding, FP: False positive finding,  
FN: False negative finding, TN: True negative finding.\textsuperscript{121}

‘C’ is the number of persons with disease missed by the test. ‘B’ is the number of healthy persons with an abnormal test result. Finally, ‘D’ symbolises healthy persons with a negative result. A high PPV of a test is found in a highly prevalent population, whereas a low PPV of a test is always found in a population with low prevalence despite
a high specificity.\textsuperscript{122} The performance of a test also depends on the sensitivity and the specificity of the test and on the prevalence of the disease. The criterion validation procedure also included receiver operating curve (ROC) statistics for further exploring the MDI sum score.

2.4.2.1 Additional analysis for the Major Depression Inventory

Additional analysis was performed for the MDI without items 9 and 10 to take the findings of Paper II into consideration. The analysis compared a ROC curve for the MDI without item 9 and 10 to a ROC curve with all ten items of the MDI. Furthermore, a ROC curve without all the somatic items was developed and compared to the ROC curve with all ten items of the MDI.
CHAPTER 3:

SUBJECTS AND METHODS
The following chapter will describe the choice of methods and provide a presentation of the methodology used. A further detailed description of methods is given in the three different papers in Chapters 5-7.

### 3.1 THE PROCESS OF THE PHD STUDY

The dissertation was based on three studies with two different themes and study populations. To illustrate the different processes of conducting these studies, a figure was made to provide an overview of the required tasks, see Figure 3.1.

**Figure 3.1. Process of the PhD study**
3.2 DATA SOURCES

3.2.1 The Danish National Health Survey of 2010 (Paper I)

The PSS formed part of a battery of self-report questionnaires on physical and mental health in the Danish National Health Survey (DNHS), which was conducted in 2010. The DNHS was based on five random subsamples: one from each of the five Danish regions. In this study, we used the population-based sample of 52,400 persons from the Central Denmark Region (CDR). All randomly selected individuals received an introduction letter, which briefly described the purpose of the voluntary survey and invited the recipient to complete an enclosed paper questionnaire. Data were collected from February to April 2010, and non-respondents received up to three postal reminders. In total, 34,168 (65.2%) completed and returned the questionnaire in CDR. See Appendix 3 for the Danish version of the PSS used in the DNHS.

3.2.2 Sundhedsmappen.dk (Paper II-III)

We developed the web page Sundhedsmappen.dk [Healthfolder.dk] (in Danish) to collect web-based versions of the MDI from Danish general practices to our study. The site is an online system intended to support the diagnosis and monitoring of depression, anxiety disorders and monitoring of home blood pressure. The site provided access to psychometric tests to be filled out by the patient, as recommended in the clinical guidelines drawn up by the Danish College of General Practitioners. The tests were the MDI and the Anxiety Symptom Scale (ASS) for diagnosis and monitoring of depression and anxiety disorders. The idea with Sundhedsmappen.dk was to create an online system that gathered psychometric tests and other diagnostic instruments for use in general practice in a joint electronic library that both GP and patient could access (see Appendix 4 for instruction sheet for GPs). A video demonstration of how to use Sundhedsmappen.dk, was conducted and uploaded on the login site for Sundhedsmappen.dk.

The patient could fill out the first test in collaboration with the GP during a consultation, and then the patient could fill out the test at home through direct access to the system (see Appendix 5 for screenshot from Sundhedsmappen.dk). We provided each of the GPs with an Ipad, which could be used to fill out the web-based MDI with the patients. The
Ipad was provided on a loan basis running six months at a time. The Ipad was intended to make the use of Sundhedsmappen.dk easier. The program generated diagnoses and sum scores according to the MDI based on entered data. The results were automatically transferred via a secure connection to the GP’s electronic system in the form of laboratory results. The GP had the opportunity to follow the patient’s progress over time by activating ‘reminders’. The patient then automatically received an invitation every two weeks to go to Sundhedsmappen.dk on his/her own computer, mobile phone or tablet and enter new values for the last two weeks. Once two normal sum scores on the MDI had been acquired two weeks apart, the patient was considered healthy. The GPs had access to updated statistics showing the average time for the patients to achieve normal sum scores.124

3.2.2.1 Recruitment procedure for Sundhedsmappen.dk (Paper II-III)
We included 22 practices in the study. The recruitment of the GPs was conducted in the following way: invitation sent by postal mail to 700 practices (see Appendix 5 for invitation letter), presentation at the Danish national conference for GPs ‘Lægedage’, newsfeed on the Danish site ‘Praksis.dk’, article in the Danish trade magazine ‘Practicus’ and invitation sent by e-mail to network practices. The GPs invited by postal mail were invited because they were already participating in an internet-based data collection system (Sentinel Data Capture).

When the GP had indicated interest in the validation study, the project leader visited the GP and gave detailed information about the study. The GPs received DKK 122.57 (≈ EUR 16.50) per included patient. Reminders about the study were sent by postal mail to the GPs and thus who did not recruit any patients after signing up for the study were contacted by phone to resolve any questions. On clinical suspicion of depression (i.e. presence of two of three core symptoms of depression), the GP asked consecutive patients to complete the web-based version of the MDI at Sundhedsmappen.dk on a tablet PC or desktop computer in the clinic. The data were then securely saved at our database. The GP handed out an information brochure about the study to patients who had signed up (see Appendix 7 for patient information brochure). In Danish, the study was called ‘Bedre diagnostik af stress, depression og angst’ [Better diagnosis of stress, depression and anxiety]. The patient information brochure should inform the patients about the study and make them aware that they would be contacted by phone for the
interview within the following two weeks. When the patients were contacted by phone, they were given a brief description of the study and the overall purpose of validating the questionnaires used in general practice. If the patient agreed to participate, the interview could take place straightaway or could be scheduled for the next couple of days.

3.2.3 Munich-Composite International Diagnostic Interview (Paper III)
For the MDI, we examined both the construct and the criterion validity. The criterion validity was assessed using the M-CIDI as the reference standard. The interviewers in our study were M-CIDI certified and blinded to the patient’s MDI score. We used the Norwegian version of the computerised M-CIDI as this version of the M-CIDI was easy to adapt into Danish as Norwegian and Danish are fairly similar languages. A manual for the Norwegian version was followed when performing the interview. The following sections of the M-CIDI was conducted; demography, depression, anxiety, Post-traumatic stress disorder (PTSD) and Obsessive-Compulsive Disorder (OCD). The lay interviewers were all students at the final year of their Master’s programme in Psychology. After consulting the experienced M-CIDI instructors from Norway, we chose psychologist students instead of medical students or psychiatrists. Psychology students were expected to have knowledge about dealing with people with mental issues and the choice of words, but at the same time not being trained in diagnosing. The lay interviewers received a certification that consisted of a two-day course with lecturing and cases in conducting interviews. After a couple of months the interviewers was recertified with a one-day course.
Additional an interview guide was developed for conducting the M-CIDI interviews for this study (see Appendix 8 for interview guide). The interview guide contained an introductory text and a detailed description of possible responses if the patient answered ‘yes’ to questions concerning thoughts of death or suicide. If the patient seemed much tormented during the interview and answered ‘yes’ concerning thoughts of dying, the patient was encouraged to contact the GP and discuss this further. The interview guide also included telephone numbers for psychiatric helplines. Ethical considerations were discussed in the research group when planning the telephone interviews. If necessary, the interviewer consulted a senior GP about an interview. Also it was discussed in the research group that there is no evidence that enquiry about suicidal thoughts increases a person’s risk. In reality many patients are relieved to be able to talk about suicidal thoughts. When contacting the patients the aim of the study was also announced cautiously and the name of the patient was confirmed, to make sure that no family member were informed about the focus of the interview. In the research group we also discussed the decision that no information was given to the GP about the outcome of the interview. Since the GP already have had the clinical suspicion of depression when included the patient in the study, we knew that the GP had an eye on the patient. Phone numbers and text messages with patients was deleted regularly to secure data security. The interviewer was instructed to interrupt the interview if the patient was very tormented by depression and gently talk the patient down.
3.3 STUDY POPULATION

3.3.1 Construct validity of the Perceived Stress Scale (Paper I)

Eligible persons for the construct validity of the PSS was 32,374. To achieve a more homogeneous population, the non-Danish or native language missing (N=1794) was excluded. Collected data were split into two random samples: a development dataset and a validation dataset. The development dataset was created to modify the PSS to achieve better fit with the Rasch model, whereas the validation dataset was created to test the modified version of the PSS. Persons with missing values on any of the PSS items (N = 1,156) were excluded. Likewise, extreme values in RUMM were excluded (N = 22). To assess the influence of a large sample size, we examined ten subsamples (each of 500 cases); these were randomly extracted from the development dataset (see flowchart Figure 3.3).
3.3.2 Construct validity of the Major Depression Inventory (Paper II)

The study population for investigating the construct validity of the MDI consisted of 245 consecutive persons who were recruited through the GPs using Sundhedsmappen.dk. Three non-Danish speakers were excluded and four persons was excluded since they were younger than 18.
3.3.3 Criterion validity of the Major Depression Inventory (Paper III)

A total of 246 persons was assessed eligible for the Criterion validity study. Out of these; 31 declined to participate, while no contact was received by phone for 56 persons. A total of 152 patients were enrolled for the M-CIDI interview. They were recruited through the GPs using Sundhedsmappen.dk. Out of the 152 patients, two were excluded because of faulty CIDI output. A total of 18 patients was excluded as the time between the interview and the MDI completion was too long. The study population for investigating the criterion validity of the MDI then consisted of 132 consecutive persons who completed the interview. See flowchart in Figure 3.4.

According to the ICD-10 applied in M-CIDI, the diagnostic codes for depression are: F32.0 Mild depressive episode, F32.1 Moderate depressive episode and F32.2 Severe depressive episode without psychotic symptoms.126 F34.1 Dysthymia is a chronic depression of mood, which does not currently fulfil the criteria for recurrent depressive disorder with mild (F33.0) or moderate (F33.1) severity.
Figure 3.4. Flow of patients for studying the criterion validity of the MDI.¹²⁷
3.4 APPROVAL AND ETHICS

The construct validity study of the PSS was based on anonymised data and was approved by the Danish Data Protection Agency (file no. 1-16-02-571-13).

For the construct and criterion validity study of the MDI, oral informed consent was obtained from participants, and no written informed consent was required according to Danish law. Permission to conduct the study was granted by the Committee of Multipractice Studies in General Practice and by the Danish Data Protection Agency (file no. 2013-41-1756). For the interviewed participants oral consent was obtained again before the interview was conducted.
CHAPTER 4:

RESULTS IN SUMMARY
This chapter offers a brief summary of the results from each paper in the dissertation. More detailed descriptions of the results are presented in Chapters 5-7.

4.1 PAPER I

This paper investigates the construct validity of the PSS by performing a Rasch analysis of a sample from the DNHS. The dataset included 32,374 persons with information on gender, age and responses to the PSS.

In the CFA analyses, the fit of data significantly improved for the two-dimensional model ($\chi^2(35) = 23,841, p \leq 0.0000001$) versus ($\chi^2(34) = 4,610, p \leq 0.000001$). The problems are mainly centred on the positively worded items, specifically item 4. The RMSEA fit index showed reasonable fit, and both the CFI and the TFI showed good fit. The two dimensions are closely related (0.687) which confirmed the two-dimensional model of the PSS.

The Mokken analysis for the MHMM revealed that the PSS had medium Mokken scale properties in terms of total Loevinger’s H coefficient (0.46) and the Loevinger’s H coefficient was above 0.4 for all items: 1 (0.49), 2 (0.51), 3 (0.49), 4 (0.31), 5 (0.43), 6 (0.47), 7 (0.44), 8 (0.51), 9 (0.40), 10 (0.56), except for item 4. For the DMHMM the H coefficient classified the scale as having weak Mokken scale properties. The MSP command for the PSS revealed one scale.

This large population-based study revealed that the unrestricted PCM contained more information than the RSM. The Danish version of the PSS did not fit the Rasch model for the ten items and the five response categories: overall scale fit ($\chi^2$ item trait interaction term) ($\chi^2(90) = 5248, p \leq 0.0000001$). The response categories functioned as intended after collapsing the response categories ‘often’, ‘sometimes’ and ‘rarely’ for items 4 and 7. However, still no fit to the model was achieved with this analysis ($\chi^2(90) = 2798, p \leq 0.0000001$). Overall no unidimensional scale emerged after wide modifications of the PSS. We did not test for local dependency.

4.1.1 Additional analysis

As an additional analysis DIF analysis for Paper I was conducted. DIF was investigated for the following exogenous variables gender, age, native language, long term illness, anxiety and years in school. The uniform DIF analysis for the total sample (N= 13 307) is presented in Table 4.1. The uniform DIF analyses was also conducted in ten different random subsamples with 500 persons in each sample, see Table 4.2.
Table 4.1 Differential item function for 10-item Perceived Stress Scale (Total sample) (Uniform DIF)

<table>
<thead>
<tr>
<th>Gender</th>
<th>Age groups</th>
<th>Native language</th>
<th>Long-term illness</th>
<th>Anxiety</th>
<th>Years in school</th>
</tr>
</thead>
<tbody>
<tr>
<td>Item</td>
<td>F</td>
<td>P</td>
<td>Item</td>
<td>F</td>
<td>P</td>
</tr>
<tr>
<td>3</td>
<td>123.113</td>
<td>0.000043</td>
<td>3 264.838</td>
<td>0.00076</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>109.997</td>
<td>0.000000</td>
<td>4 157.015</td>
<td>0.00175</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>106.582</td>
<td>0.000000</td>
<td>7 76.191</td>
<td>0.000000</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>101.362</td>
<td>0.000302</td>
<td>1 67.044</td>
<td>0.000117</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>73.220</td>
<td>0.000000</td>
<td>2 50.450</td>
<td>0.000117</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>26.378</td>
<td>0.000014</td>
<td>5 27.051</td>
<td>0.000006</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>15.268</td>
<td>0.000028</td>
<td>9 18.638</td>
<td>0.000000</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>18.490</td>
<td>0.000000</td>
<td>8 29.364</td>
<td>0.000340</td>
<td></td>
</tr>
</tbody>
</table>

aAge groups: 1: 15-30 years, 2: 31-60 years 3: 61->70 years.
bNative language: “What is your native language?”: 1: Danish 2: Other.
cLong-term illness: “Do you have any long-term illness, long-term rehabilitation of injury, disability or other long-term illness?”: 1: Yes. 2: No.
dAnxiety: “Transient mental illness (E.g. mild depression or anxiety)?”: 1: No, I have never had that. 2: Yes, I have now. 3: Yes, have had it before.
eYears in school: “What school education do you have?”: 1: Still in school. 2: 7 years or less. 3: 8-9 years. 4: 10-11 years. 5: High school. 6: Other. High school graduation.
**Table 4.2 Differential item function for 10-item Perceived Stress Scale (Random subsamples) (Uniform)**

<table>
<thead>
<tr>
<th>Subsample</th>
<th>Gender</th>
<th>Age groups(^a)</th>
<th>Native language(^b)</th>
<th>Long-term illness(^c)</th>
<th>Anxiety(^d)</th>
<th>Years in school(^e)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Item</td>
<td>(F)</td>
<td>(P)</td>
<td>Item</td>
<td>(F)</td>
<td>(P)</td>
</tr>
<tr>
<td>1</td>
<td>9</td>
<td>7.074</td>
<td>0.000934</td>
<td>3</td>
<td>5.383</td>
<td>0.000306</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>14.493</td>
<td>0.000151</td>
<td>7</td>
<td>12.166</td>
<td>0.000151</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>11.598</td>
<td>0.000007</td>
<td>3</td>
<td>14.374</td>
<td>0.000178</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>13.301</td>
<td>0.000000</td>
<td>3</td>
<td>8.027</td>
<td>0.000383</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>7.047</td>
<td>0.000958</td>
<td>1</td>
<td>13.016</td>
<td>0.000000</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>8.827</td>
<td>0.000479</td>
<td>2</td>
<td>7.768</td>
<td>0.000479</td>
</tr>
<tr>
<td>3</td>
<td>10</td>
<td>19.141</td>
<td>0.000017</td>
<td>4</td>
<td>11.413</td>
<td>0.000020</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>10.958</td>
<td>0.001006</td>
<td>3</td>
<td>19.141</td>
<td>0.000017</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>10.826</td>
<td>0.000108</td>
<td>4</td>
<td>10.826</td>
<td>0.000108</td>
</tr>
<tr>
<td></td>
<td>9</td>
<td>10.219</td>
<td>0.001475</td>
<td>3</td>
<td>10.817</td>
<td>0.00019</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>6.591</td>
<td>0.001494</td>
<td>4</td>
<td>10.817</td>
<td>0.00019</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>7.058</td>
<td>0.000957</td>
<td>5</td>
<td>10.252</td>
<td>0.000053</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>10.219</td>
<td>0.001475</td>
<td>3</td>
<td>7.058</td>
<td>0.000957</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>9.232</td>
<td>0.000124</td>
<td>4</td>
<td>7.058</td>
<td>0.000957</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>7.007</td>
<td>0.001014</td>
<td>5</td>
<td>6.798</td>
<td>0.001227</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>6.591</td>
<td>0.001494</td>
<td>6</td>
<td>6.798</td>
<td>0.001227</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>6.565</td>
<td>0.00153</td>
<td>7</td>
<td>6.798</td>
<td>0.001227</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>13.098</td>
<td>0.000000</td>
<td>4</td>
<td>9.205</td>
<td>0.000126</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>12.166</td>
<td>0.000151</td>
<td>3</td>
<td>17.848</td>
<td>0.000034</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>11.423</td>
<td>0.000782</td>
<td>4</td>
<td>20.583</td>
<td>0.000000</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>11.255</td>
<td>0.000865</td>
<td>3</td>
<td>11.255</td>
<td>0.000000</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>8.717</td>
<td>0.000183</td>
<td>4</td>
<td>8.717</td>
<td>0.000183</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>7.827</td>
<td>0.000491</td>
<td>5</td>
<td>8.717</td>
<td>0.000183</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>9.232</td>
<td>0.000124</td>
<td>6</td>
<td>7.428</td>
<td>0.000067</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>11.255</td>
<td>0.000865</td>
<td>7</td>
<td>11.255</td>
<td>0.000865</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>8.717</td>
<td>0.000183</td>
<td>8</td>
<td>8.717</td>
<td>0.000183</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>8.857</td>
<td>0.000215</td>
<td>9</td>
<td>8.857</td>
<td>0.000215</td>
</tr>
</tbody>
</table>

\(^a\)Age groups: 1: 15-30 years, 2: 31-60 years 3: 61->70 years.

\(^b\)Native language: “What is your native language?”: 1: Danish 2: Other.

\(^c\)Long-term illness: “Do you have any long-term illness, long-term rehabilitation of injury, disability or other long-term illness?”: 1: Yes. 2: No.

\(^d\)Anxiety: “Transient mental illness (E.g. mild depression or anxiety)”?: 1: No, I have never had that. 2: Yes, I have now. 3: Yes, have had it before.

\(^e\)Years in school: “What school education do you have?”: 1: Still in school. 2: 7 years or less. 3: 8-9 years. 4: 10-11 years. 5: High school. 6: Other. High school graduation.
In summary, the DIF analysis in table 4.1 & 4.2 showed evidence of uniform DIF for item 3 relative to age, anxiety and years in school. Also evidence of uniform DIF was shown for item 10 relative to gender, long-term illness and anxiety. Item 3 looks a bit worrisome, and to some extent also item 10. This calls for further investigation of item 3. Item 3 is; “How often have you felt nervous and “stressed”?”, which may raise the question whether older persons interprets the word “stressed” as a more serious condition/state than the young persons. See chapter 10, section 10.4 for further discussion of the DIF results.
The construct validity of the MDI scale was investigated using Rasch analysis in Paper II. In the study sample (N = 245), the MDI score was in the range from 3 to 48. The MDI sum score mean standard deviation (SD) was 29.2 (9.0) for the total of 245 persons. The mean age (SD) was 37.1 (12.9) years; 97 (39.6%) were males and 148 (60.4%) were females. Further characteristics were only available for the interviewed persons (n=132). More women than men were included in this study, which is in alignment with a general practice population but not twice as many as women as men as found in a former study by Kjøller.128 Our subjects were aged 18–80 years (only five were above age 65 years). Therefore, our results may not apply to young adolescents or elderly patients.

Initially the results of the Mokken analysis showed that items 9 and 10 had a Loevinger coefficient below 0.30 and that items 2, 4, 5, 6 and 7 had a Loevinger coefficient below 0.40. The total coefficient for all ten items was classified as having weak Mokken scale properties (0.36). The MSP identified two sub-groups of items: a) items 1, 2, 3, 4, 5, 6, 7 and 8 and b) item 9 and 10. When item 9 and item 10 were excluded from the analyses, the MDI scale was classified having medium Mokken scale properties (Loevinger coefficient total = 0.44).

For the Rasch analysis of the MDI, we concluded that the unrestricted model was the appropriate to apply due to the results of the likelihood ratio statistic test. Initial analysis suggested misfit within the scale, which was indicated by a significant overall scale fit $\chi^2$ item trait interaction term ($\chi^2 (30) = 61.15, p = 0.0007$). The person separation reliability was acceptable (0.82) for the initial model. Our Rasch analysis showed misfit for the sleep and appetite items (items 9 and 10). The response categories were disordered for eight of the ten items. The Rasch analyses revealed two dimensions, but the MDI showed fit to one dimension if items 9 and 10 were excluded and resoring all items (from the original six response categories to four response categories), overall fit ($\chi^2(24) = 17, p = 0.8464$) and ordered thresholds were then found for all items.

The scale appeared to be well targeted to this clinical sample. No significant DIF was observed for gender, age, work status and education.
4.3 PAPER III

The criterion validity of the MDI compared to the M-CIDI interview was investigated in Paper III. 22 Danish general practices recruited 246 patients. Even though the patients had agreed to participate at the consultation with their GP, 56 persons did not answer their phone and 31 persons declined to participate in the interview when asked on the phone. In total (n=7) were not meeting the inclusion criteria and (n=2) were excluded due to no correct CIDI output. In total, 18 persons were excluded from the analysis due to the timeframe of a maximum of two weeks, between the MDI and the M-CIDI interview. Of the M-CIDI interviewed persons (n=132), 116 (87.9%) were diagnosed with depression, and 16 (12.1%) had no depression. According to the interview 116 had a depression; 8 (6%) had mild depression, 23 (17.33%) had a moderate depression and 85 (64.4%) had a severe depression. According to the ICD-10 algorithm of the MDI (n=132), 78 (59%) were diagnosed with depression, and 54 (40.9%) had no depression. Of the 78 with a depression; 14 (9.9%) had mild depression, 22 (17.4%) had a moderate depression and 42 (31.8%) had a severe depression. According to the MDI 24 patients reported no core symptoms, 29 patients reported one core symptoms, 31 patients had two core symptoms and 48 patients had three core symptoms. These results might be explained by the way the GPs are asking about the core symptoms which is less restrictive and without a specific timeframe than the questions in the MDI.

The sensitivity of the MDI for any depression was 62.1% (95% CI: 52.6-70.9), the specificity was 62.5% (95% CI: 35.4-84.8), the PPV was 92.3% (95% CI: 84.0-97.1), and the NPV was 18.5% (95% CI: 9.3-31.4). The sensitivity for severe depression was 41.2% (95% CI: 30.6-52.4), the specificity was 85.1% (95% CI: 71.7-93.8), the PPV was 83.3% (95% CI: 68.6-93.0), and the NPV was 44.4% (95% CI: 34.0-55.3).

The ROC showed an AUC of 66 (95% CI: 0.52-0.81) for any depression and 0.72 (95% CI: 0.63-0.81) for severe depression.

4.3.1 Additional analysis

The results from the analyses of the construct validity in Paper II (the 8-item version of the MDI fitting the Rasch model) was investigated in Paper III, by performing ROC curves without items 9 and 10 which were misfitting items in the Rasch analysis.
Figure 4.1. ROC curves for any depression with all ten items and without items 9 and 10.

In figure 4.1, the results of the ROC curve without item 9 and 10 are compared to the result of the ROC for all ten items for any depression. There are no substantial differences between the AUC.

Figure 4.2 ROC curves all ten items and without items 9 and 10, for severe depression

In figure 4.2, the ROC curve for all ten items is compared to the ROC curve without item 9 and 10 for severe depression, and no substantial difference between the AUC.

Figure 4.3. ROC curves for any depression with all ten items and without items 8, 9 & 10.
In figure 4.3, the results of the ROC curve without the three somatic items; 8, 9 and 10 are compared to the result of the ROC for all ten items for any depression. There are no substantial difference between the AUC.

A test was conducted in Stata to investigate for any differences between the ROC with all ten items, without items 9 and 10 and without item 8, 9 and 10. The command “roccomp” was applied for this analysis. This analysis showed no differences between the two ROC curves. For any depression the results were p=0.23>0.05 and for severe depression p=0.80>0.05.

Figure 4.4. ROC curves for severe depression with all ten items and without items 8, 9 & 10.
CHAPTER 8:

DISCUSSION OF METHODS
This chapter presents a critical discussion of the applied methods and links these with the presented results.

The chapter consists of four main sections. The first section discusses the study populations, data sources and methods applied in Paper I. The second section describes the Rasch model, which was applied in both Paper I and Paper II. The third section discusses the study populations and methods applied in Paper II and Paper III. Finally, the fourth section discusses the diagnostic interview used in Paper III.

### 8.1 Construct Validity of the Perceived Stress Scale (Paper I)

#### 8.1.1 Sampling of respondents and generalizability of findings

The respondents included in the DNHS were random (mutually exclusive) subsamples from the general population in the CDR.\(^{123}\) The participation rate in the DNHS was 56%. The survey respondents might have perceived different levels of stress (more or less) than the general population. However, even if this is the case, application of the Rasch analysis is unlikely to have limited the generalisability of the findings for the PSS as the Rasch analysis is not sample dependent due to the incorporated principle of specific objectivity.

#### 8.1.2 Danish version of the Perceived Stress Scale

The Danish version of the PSS used in the DNHS survey had a different layout than the original English version of the PSS. The headline differed in the Danish version since it was translated into “Everyday stress” [Dagligdagens stress] instead of “Perceived Stress Scale”. A potential limitation was the translation procedure used for the Danish version of the PSS, which was applied in the DNHS. The translation procedure defined as; “The scale was translated into Danish and the consensus version was translated back into American English, and accepted by the developer of the scale”.\(^{129}\) It is uncertain whether the translation was done in line with the general recommendations drawn up by the WHO for translation and adaptation instruments: forward translation, expert panel back-translation, pre-testing and cognitive interviewing, final version.\(^{130}\) If the recommendations have not been followed this might cause that the PSS does not perform in the same way in the Danish version as in the English version. When comparing the English version (source text) and the applied Danish version of the PSS (target text),
one of the main differences between these two versions is the use of the time frame ‘in the last month’. In the original English version, this wording is stated for each item separately. In the Danish translation, it is used only once in the introductory text. The response behaviour of the participants may have been affected by this as they might not have been aware of the time frame when answering each item, and this risk is likely to have increased as the respondent progressed through the questionnaire. This could imply that the respondent may have responded on the basis of a different (possibly shorter) time frame which might overestimate the perceived stress score. Another relevant difference might be the use of quotation marks around the word “stressed” in item 3 in the English source text; these quotation marks have been left out in the Danish target text. The use of quotation marks might imply that the word should be taken with reservations. Therefore, the understanding of this item and thereby the response behaviour might be different in the Danish version because the respondents might interpret (or think they are intended to interpret) “stressed” more literally.

According to the generalizability of our results of the PSS, this is a potential limitation of our study that the Danish version of the PSS differs from the English version in this respect several Danish versions of the PSS are available. At the time when the study was initiated, the version used in the DNHS was the version approved by Sheldon Cohen. However, a consensus version was recently presented in a Danish study, which is now the formally approved Danish version according to Cohen’s homepage. In this consensus version the timeframe of asking for “in the last month” has been placed before each item as in agreement with the English version of the PSS. Additionally, the wording of the problematic item 4 has been changed from ‘Hvor ofte har du følt at du var i stand til at klare dine personlige problemer’ [..How often have you felt that you were able to handle your personal problems] into ‘Hvor ofte inden for den sidste måned har du følt dig sikker på dine evner til at klare dine personlige problemer’ [..How often within the last month have you felt confident about your ability to handle your personal problems]. It would be interesting to investigate how this change may affect the responses.

8.1.3 Confirmatory factor analysis

The PSS was analysed with CFA to study the fit of data for a two-dimensional model. We could also have used EFA to investigate the PSS, but this was outside the scope of this
dissertation since the EFA is often used to reduce a set of items to a small number of summary scale scores.82

8.1.4 Mokken scale analysis
As a starting point, a Mokken analysis was completed as an initial step before the item analysis for the Rasch model was performed. This was in line with the recommendations by Koning et al., who generally advise to start an item analysis with the most liberal models, such as the Mokken model, and then continue with a more restrictive model, such as Rasch model.88

8.2 RASCH ANALYSIS (PAPER I-II)
The use of the Rasch model is one of the strengths of our construct validation studies because it facilitates disclosure of measurement problems that may not be easily detected by CTT, e.g. ranking of persons, response categories and items.133 Both person- and test-level statistics are sample dependent in CTT, which may result in different psychometric properties when based on different samples.134 In some sense, measurement by Rasch items is an ideal type of measurement because measurements are valid, objective and sufficient.82 In IRT models, both the item and the person scores are established as single values on the same latent variable.69 If the data fit with the Rasch model, it is considered a measure of specific objectivity, which indicates that persons and items can be placed on the same scale. The Rasch model is the only IRT model, which contains specific objectivity, which is also the reason why we chose to use this model. The 2- and 3-parameter IRT versions accept variations in the shape of the item response curves. Such models are more useful for finding a model that suits the data than testing item and person fit.
Using the RSMs were considered during the analysis in our studies but abandoned because of lack of fit between the model and the observed item responses.
Another IRT model, which could also have been used for the studies conducted in this thesis, are the graded response model (GRM). The GRM is less restrictive than the Rasch model. Hence, the discrimination varies across items.69 Discrimination is the ability of a test to separate subjects into high and low levels of ability.69 Another model is the
graphical log-linear Rasch Model (GLLRM) which provides the possibility of modelling uniform DIF and uniform local dependency, but this model was not available within the RUMM. Analyses by GLLRMs can be used to suggest improvements that may lead to a scale with better psychometric qualities, as suggested in Nielsen and Kreiner’s scale improvement strategy. Such a comprehensive scale improvement strategy were outside the scope of this dissertation.

We leaned towards the paradigm where all response categories should be rescored in the same way rather than striving for a model that fits. We aimed for a scale, which fits the Rasch model as a whole. In the traditional paradigm, one choose a model over another if it accounts better for the data and the emphasis is on finding a model that best characterizes the given data. In the Rasch paradigm, the emphasis is placed on identifying and studying anomalies in the data disclosed by the Rasch model. Our analytical process involved exploration of ways to account for any misfit found within the scales and the contributions of each item. First, the statistical properties were investigated and then item removal and collapse of non-fitting response categories were conducted. This was followed by another investigation of the statistical properties and using the same procedures until a fit of the model was achieved.

8.2.1 Software programs for Rasch analysis

Before selecting the most appropriate software program for conducting the Rasch analysis, a scale validation course on a range of software programs at Copenhagen University was attended by the author of this dissertation. Afterward course completion, several meetings were held with colleagues at Aarhus University and the Danish School of Education who had thorough experience with Rasch analysis and could help which program to choose. Furthermore, literature was searched and studied. The RUMM program was chosen for the analyses in the dissertation as all of the analyses could then be conducted in the same program. To learn and work with one program was more appealing than getting superficial knowledge of two or more programs.

A potential disadvantage of using the RUMM program is that the chi-square based fit statistics differs from the infit and outfit mean-square statistics output by Winsteps, as also reported in many measurement journals. Besides Winstep, several software programs for Rasch analysis provide estimates of items, person parameters and fit
statistics\textsuperscript{72}, such as Conquest developed by Wu.\textsuperscript{137} Another software program is Diagram, which supports item analysis by graphical and log-linear Rasch models.\textsuperscript{138} Software procedures such as NLMIXED in SAS and the SEM in Stata 14 can also perform Rasch analysis.\textsuperscript{72} Another option was the R Project for Statistical Computing, the eRm package provides tools for assessing goodness of fit.\textsuperscript{72} Several updates have recently been made in these programs; these were not available at the time when the writer of this dissertation had to choose a program for conducting the Rasch analysis.

8.2.2 Sample size and Rasch analysis
Large sample sizes generally constitute a problem for significance tests based on chi-square statistics as small differences are often reported to indicate a statistically significant misfit between the data and the model.\textsuperscript{102} Therefore, the large sample size for our validation of the PSS was problematic. A study by Hagell et al. suggests that a sample size around $N = 250$ to $N = 500$ for Rasch analysis conducted with RUMM may provide a good balance for the statistical interpretation of the RUMM fit statistics.\textsuperscript{139} To assess the influence of the large sample size in our study, we examined ten subsamples (each of $n = 500$) randomly extracted from the development dataset. For the PSS study, we could also have analysed the dataset with missing values as this is possible in RUMM.

8.2.3 Differential item functioning
For the Rasch analysis, an underlying property is the absence of DIF (see Figure 2.6, page 36). For the dissertation we additionally investigated DIF for the PSS scale. Also we included the former excluded data to be able to conduct DIF analysis for ethnicity, Danish or non-Danish native language. The DIF analysis showed evidence of DIF for item 3 relative to age, anxiety and years in school. Also evidence of DIF was shown for item 10 relative to gender, long-term illness and anxiety. The results are discussed in accordance with existing literature in chapter 9, and the results are further discussed in chapter 10.

8.3 CONSTRUCT- & CRITERION VALIDITY OF THE MDI (PAPER II-III)
In the following, the study population, the methodological design and the data sources will be discussed in terms of the construct and criterion validity of the MDI.
8.3.1. Sampling of general practitioners

GP s were self-selected for participation in the validation studies of the MDI and were invited by postal mail. We included 22 practices from different Danish counties. The study was launched during a period of much public debate about data collection for research and quality assurance in general practice. The debate centred around the legal rights to access the data. Many GPs found that only the patient and the GP should share the collected data. As some GPs were critical about data collection in general, this external factor challenged our recruitment of GPs, the implementation of Sundhedsmappen.dk in general practice and the recruitment of patients for interviews. We originally planned to begin our study in the spring of 2013, but a labour conflict between Danish Regions and the Organisation of General Practitioners in Denmark (PLO) postponed the initiation of our study.

To encourage recruitment, introduce the study and answer any questions regarding the inclusion criteria (to recruit only patients with two or three core symptoms of depression), the project leader visited participating GPs to provide them with detailed information about the study. The recruitment criteria was in alignment with the guidelines for depression from the Danish College of General Practitioners. The information given at these visits seems essential for inclusion of potentially depressed patients in the study which is important for our results. The GPs could also contact the project leader by phone or email if they had any questions about the project.

GP reminders were sent by postal mail, and the GPs were contacted by phone if they did not recruit any patients for the study. The 22 GPs in our study had an average of 4393 patients (in the range 916-14,615). Ten of the general practices were situated outside one of the four largest cities in Denmark and only four were located in the countryside. Due to the overrepresentation of GPs from large cities, the included GPs might not be representative for all GPs in Denmark. Other factors such as gender, age and experience of the GP may have affected their choice when including patients on clinical suspicion of depression. We have no information on whether included GPs in our study differ from other GPs, for example in terms of taking greater interest in mental health problems than other GPs. If the participating GPs had greater interest in mental health it might have affected our results as they have more focus on depression.
8.3.2 Sampling of patients for Sundhedsmappen.dk

The inclusion criteria for the study were based on the clinical guideline for depression in Danish general practice. On clinical suspicion of depression, the GP asked consecutive patients to complete the web-based version of the MDI. The mean number of patients per practice was: 10.78 (range: 1-38). We have no knowledge of whether the GP included all 'eligible' patients with suspected depression, or if they (consciously or unconsciously) selected patients with specific characteristics. However, as we instructed the GPs to use the MDI only on clinical suspicion of depression (i.e. presence of two of three core symptoms of depression), as also recommended in daily clinical practice, any potential selection bias in our study would have been minimised. Some GPs could be less likely to include patients with low socioeconomic status and severe comorbidities because they did not wish to burden them further. Before recruiting the patient, the GP might also have considered the severity of the depression and thereby the patient’s ability to participate in a long interview. We did not collect information from the participating GPs regarding the reasons for patients not to enter the study, which is a limitation of our study. These information’s might have been useful for conducting sub-group analyses. Underrepresentation of certain groups of patients, such as patients with low educational level and low household income, may have affected the generalizability of our findings to all patients in Danish general practice. The high prevalence of depression is partly a natural consequence of the sampling procedure as GPs were instructed to recruit only patients with two or three core symptoms of depression. It might also be due to GPs oversampled patients with an already known depression and not only included new cases or new episodes of depression. We did not collect information regarding presented symptoms from the GP. At the time of the interview, the patients could already have received treatment for their depression. In our study, there is a maximum time period of 14 days between the MDI test and the M-CIDI interview. We acknowledge that patients may have changed in score over time, but we aimed to minimize the time frame from the MDI to the M-CIDI interview. The only demographic information available for patients entering the study was gender and age.

Another point for discussion is the time period (from 14 August 2013 to 19 February 2016) required for recruiting a relatively small number of patients (i.e. 2.5 years for 246 patients). This long period may have implied a risk that they were highly selected which
are in line with the high prevalence stated above and the suspicion that GPs have primarily chosen patients motivated to attend the study. Which might have resulted in that more young patients with depression was included in our study since the GPs might have selected those they thought were more likely not to be distressed about using a web based version of the MDI. Still, the study had several steps of recruitment. First, we had to recruit the GPs who then needed to recruit the patients on clinical suspicion of depression. After this, we had to get in contact with the patients by phone; first to recruit the patients for the interviews and later to perform the interviews. To ensure a sufficient sample size, we decided to keep on including until we reached a sufficient number of patients ($N = 132$) to ensure valid results.

8.3.3 Sampling of patients for interview

A study by Ekholm et al. found that survey response rates have been declining over the past decades, and surveys using telephone interviewing have been particularly declining. Furthermore, the non-response rate was high among persons with low socioeconomic position.140 This could also be the case in our study. To investigate this further, we should have obtained the social security number of all participants to be able to study this through information in the registries. We did not collect information of the patients who did not participate in the study besides their MDI. Neither did we collect information from the telephone contact with patients regarding their reasons for not wanting to participate. We did not want to burden the patients asking for additional information when they had already declined to participate. Due to the findings of no difference in MDI sum scores between those who participated in the interview and those who did not, it indicates that there has been a non-differential drop-out.

Insufficient information from GP to patient may have led to non-participation in the interview. Unfortunately, some of the patients did not answer their phone at the scheduled appointment. We are aware that the depressed patients participating in our study might not have had enough energy to take part in a phone interview and may, therefore, have neglected the phone calls. Some of those who declined to participate in the interview might have regard themselves as not depressed and therefore did not think of themselves as relevant for the study. We sent a text message to make the non-
answering patients aware of who was calling without revealing that the interview concerned depression.

8.3.4 Web-based version of the Major Depression Inventory
8.3.4.1 Implementation of Sundhedsmappen.dk
The web-based MDI test has shown good results in relation to the completeness of data, required physician time and clarification of the patient’s symptoms.\(^{141}\) Despite these benefits, Sundhedsmappen.dk (including the web-based version of the MDI) was only partially implemented by the participating GPs. We are aware that Sundhedsmappen.dk is a new procedure, which may have caused some technical issues for some GPs. This could also be reflected in the variation of recruited patients per GP (range: 1-38).

Even though Sundhedsmappen.dk was linked to the laboratory system used by the GPs, the individual GP still had to use a separate log in to Sundhedsmappen.dk with a username and password only for this system. Furthermore, the GP needed to set up the patient in the system. It would have been an advantage if the GP could have used the same login information as for the applied medical system and if the patient could have used NemID (easy identification), which is held by every citizen in Denmark. This would have reduced the inconvenience of an additional login for the GP and patient. The best solution would be if it was fully implemented in the GPs already existing IT system. There might also be other unknown barriers towards using such a web-based solution which recently have been addressed in a qualitative study in our research group.

8.3.4.2 Paper-based versus web-based Major Depression Inventory
Another issue concerning Sundhedsmappen.dk was that it was designed to be used on PC, smartphone or tablet. To allow this, each item of the MDI had to be presented on a separate screen display during completion (see Appendix 4). This approach is different from the paper version, where all ten items are displayed on the same page. We are unaware if this could have affected the responses and thus the validity of the responses as the patient could only read one item at a time and thus could not compare the responses for different items in the questionnaire. However, a former study by Toepoel et al. showed that more items on a screen had a negative influence on the evaluation of the test.\(^{142}\)
A strength of using Sundhedsmappen.dk was that we received information about an included patient right away, which generally made it possible for us to conduct the M-CIDI interview shortly after the patient had filled out the MDI. Another strength is that there were no missing for any of the items of the MDI.
8.4 CRITERION VALIDITY OF THE MAJOR DEPRESSION INVENTORY (PAPER III)

8.4.1 M-CIDI as reference standard

Our main reasons for choosing the M-CIDI as our reference standard were, that it is developed by the WHO, is internationally acknowledged, fully-structured and can be applied by lay interviews as a telephone interview. We chose not to use the SCAN interview because this required a psychiatrist to perform the interview and is not available as a telephone interview. This would have been costly and might have delayed the time frame between the test and the interview. Another interview which could have been applied was the Mini-International Neuropsychiatric Interview (MINI interview); this could also have been conducted by phone by a lay interviewer. But it is not so international acknowledged as the M-CIDI.

Five different interviewers performed the interviews and conducted 64, 15, 35, 35 and 27 interviews each. We selected the M-CIDI interview because the interviewers were certified, recertified and supervised by the project leader, which ensured high quality of the interviews. Data from the M-CIDI telephone interviews were recorded according to the standard manual.

The anxiety diagnoses from the M-CIDI were handled as ‘no depression’ diagnoses. For 17 patients, the M-CIDI generated the ICD-10 diagnosis F06.32, which is an organic depressive episode. By reviewing their responses to the M-CIDI interview, these patients were recoded according to their individual degree of core symptoms and accompanying symptoms. If a patient with had a depression diagnosis according to the ICD-10 algorithm, the patient was categorised as having a present depression. In order to investigate the impact of recoding, we performed a sensitivity analysis while excluding the 17 patients. These analyses showed no substantial changes in our results.

8.4.2 M-CIDI compared to other interviews

When comparing M-CIDI with another diagnostic interview (SCAN), CIDI tended to over-diagnose with respect to SCAN with a prevalence of 18.1% compared with 7.6% for any depressive episode or disorder. This could be a limitation of using the M-CIDI in our study. The concordance (interrater reliability) for depressive disorders was fair (kappa = 54%), the sensitivity was 100% and specificity of 88%. If so, our study (Paper III) might
underestimate the sensitivity of the MDI. However Brugha et al. stated that a consistent pattern of false positives was seen for all CIDI diagnoses when set against the SCAN calibration data. They reported a sensitivity of 50% (95%CI: 12–88%) and a specificity of 87% (95%CI: 81–91%). The kappa coefficient for any depression between CIDI and SCAN showed only a fair concordance. As mentioned earlier in a study by Andrews et al., the inter-rater reliability of the CIDI was found to be perfect. So the concordance between the interviewers was better for CIDI than SCAN, which is result of M-CIDI being a fully structured interview. The high concordance is a strength of applying the M-CIDI interview since the lower sensitivity of the CIDI noted in our study and those of others may be due to the fact that the CIDI is a fully structured interview with precisely worded questions that cannot be rephrased.117, 143, 145

Due to the previous findings of high agreement between SCAN and MDI and the findings of CIDI over diagnosing according to SCAN, the MDI may actually be considered as the reference standard in our study instead of the M-CIDI. We are unaware if this is the case, but it might be an important reflection regarding interpretation of the results from our study.

8.4.3 Telephone interview versus face-to-face interview

Using a telephone interview for research focusing on depression has previously been found a proper and valid method. To our knowledge, no study has compared the M-CIDI administered by phone and by face-to-face interview, which is a limitation of the M-CIDI applied as a telephone interview. However, several studies have compared the SCID performed as a telephone and as a face-to-face interview, such as the study by Allen et al. who conclude that a substantial proportion of patients may actually prefer telephone interviews. A previous study of the SCID interview found diagnostic telephone interview for depression to be valid when compared to face-to-face assessments. Their results encourage the use of diagnostic interviews by telephone for depression, particularly when face-to-face interviews are found to be impracticable. We also believe that due to M-CIDI being a fully structured interview, it minimizes the differences between the face-to-face and telephone interview. The M-CIDI seems acceptable for the respondents and is efficient considering the required time and the ease of administration. Furthermore, it has been shown that telephone interviews constitute a
valid mode of performing clinical assessment. During the interviews in our study, some of the patients were very distressed and expressed a need for someone to talk to. Some patients found it helpful to talk about past time periods in their lives and mentioned that the interview was the first time they had the opportunity to talk about earlier periods of their lives. The interviews each took on average 1–1.5 hours to perform. Some patients expressed afterwards that the long duration made it difficult to stay focused throughout the entire interview.

Some barriers may also arise when conducting the interviews by phone. One study has shown that technology-based assessment encourages self-disclosure, whereas another study found that the technology discourages self-disclosure. In our study, we experienced that the participating patients were satisfied with having the interview done by phone; the phone interview did not require them to show up in person and could be performed at the time that best suited the individual patient.
CHAPTER 9:

DISCUSSION OF RESULTS
This chapter offers a discussion of the results and a comparison with existing literature for each of the three papers in the dissertation.

9.1 CONSTRUCT VALIDITY OF THE PERCEIVED STRESS SCALE (PAPER I)

Our findings that the PSS should be considered a two-dimensional rather than a unidimensional scale are in line with previous CTT studies. Several previous studies have evaluated the PSS using CTT and factor analysis, but few studies have examined the measurement properties of the PSS in a general population. A Greek community-based study found CFA results similar to ours as they found acceptable fit for the two-dimensional PSS model.32

As a natural starting point for item analysis we conducted the Mokken analysis,87 which suggested medium Mokken scale properties of the PSS in our study. We were unable to identify any former studies performing Mokken analysis of the PSS. However, a Mokken coefficient of homogeneity of 0.44 for the PSS scale was reported in “Clinical psychometrics”.91

However the PSS did not fit the Rasch model. We could not establish unidimensionality, even though the data showed improved fit to the Rasch model and the CFA analyses for the two dimensions with the positively formulated items (4, 5, 7 & 8) and the negatively formulated items (1, 2, 3, 6, 9 & 10). Item 4 indicated the largest misfit, and items 4 and 7 displayed disordered thresholds.

To further elaborate the two dimensions of the PSS, additional analyses (Mokken, CFA and Rasch analyses) were conducted to examine if the effects of gender could explain the two dimensions. The data did not fit the model in any gender subgroups. We concluded that gender differences did not seem to influence our findings since the two dimensions appear to be dominant for both men and women.

Our findings are in agreement with the former division of items into a negative dimension (items 1, 2, 3, 6, 9 and 10) and a positive dimension (items 4, 5, 7 and 8). A study by Dougherty et al. from 2017 investigating the PSS used in patients with age-related macular degeneration showed unidimensionality when item 4 was removed and assessed by Rasch person separation statistics.152 A Rasch study by Medvedev showed that the best model fit was achieved by combining locally dependent items into three subtest sets: items 1,2,9, items 4,5,7,8 and items 3,6,10. The authors suggest that the
The precision of the PSS could be improved by applying an ordinal-to-linear conversion table. To the best of our knowledge, these two studies are the only studies besides our study that have validated the PSS by Rasch analysis. Additionally, we identified only one other study using IRT for assessing the psychometric properties of PSS. A study by Taylor et al. investigated the PSS by a different factor analysis and IRT model, the GRM. Their analysis also revealed the same two dimensions and named them: helplessness (six negatively worded items) and self-efficacy (four positively worded). A PCA by Golden-Kreutz et al. found ‘close fit’ for the two-dimensional model when assessing stress in a cancer population and labelled the two dimensions ‘perceived stress’ (negatively worded items) and ‘counter-stress’ (positively worded items). In an EFA and a CFA by Ezzati et al. conducted in a community-based sample of older adults, a two-factor structure of PSS was established. A PCA by Örücü and Demir (2009) found gender differences in the PSS, which were in agreement with a study by Gitchel et al., finding women generally reported higher levels of perceived stress than men in the positively worded items. Whereas a study by Barbosa-Leiker et al. indicated that stress and counter-stress were measured equivalently in men and women as in accordance with our findings of no gender differences explaining the two dimensions of the scale.

A recent CFA study by Denovan et al. found a closer fit to a bifactor model consisting of perceived stress Total, Distress and Coping, than to a one-, two or three factor model. They suggest that although individual stress and coping factors exist, the PSS is driven by a single underlying dimension of perceived stress. Which are new interesting perspectives on the PSS scale. Golden-Kreutz et al. call for future research to make recommendations about the use of these dimensions as subscales. Taylor et al. call into question that the multidimensionality of the PSS is ignored since multidimensionality can leave interpretations ambiguous. The result of our study also support that there is a need for investigating the application and use of the two dimensions of the PSS. Therefore it can be discussed whether the interpretation of the PSS as a unidimensional scale can be trusted.
Additional analysis

For the dissertation, we additionally investigated the PSS for DIF and our findings were in agreement with previous studies investigating DIF of the PSS.\textsuperscript{152, 153, 159} We found DIF for item 3 for age, anxiety and years of school. As discussed earlier, the meaning of item 3 “felt nervous and “stressed”?” might be understood differently in the younger and older age groups. Also, for item 10 “felt difficulties were piling up so high that you could not overcome them?”, we saw some DIF for age and years of school. We found misfit for item 10 in our original analysis, but not as large a misfit as for item 4. The misfit for item 10 in our original analysis might be explained by DIF. The DIF which we investigated for in the additional analysis did not explain the findings of the largest misfit for the positively worded item 4 of the PSS. We identified four previous studies investigating DIF. Our findings of DIF of item 3 and 10 was in agreement with an earlier study by Cole et al. of a sample of US adults who found that women reported higher perceived stress for items 3, 6, 7, 8 and 10 than men. Additionally, items 3 and 4 functioned differently by ethnicity. For items 3 and 9, participants with higher education reported increased perceived stress, while lower educated participants seemed to report increased perceived stress for items 4 and 8.\textsuperscript{159} The study by Dougherty et al. showed some evidence of DIF by age as younger participants found it easier to endorse item 3 and more difficult to endorse item 8 (‘I felt on top of things’) than participants above the median age. None of the PSS items exhibited evidence of significant DIF by gender.\textsuperscript{152} Medvedev et al. found no significant DIF for gender, age, ethnic groups and education level.\textsuperscript{153} Neither did Taylor et al. found any evidence of DIF for gender.\textsuperscript{154} For our PSS study the exclusion of ethnicity was made to achieve a more homogeneous population and to avoid any problems with understanding the items. A previous PSS study by Sharp et al. exploring literacy by DIF identified items 2, 3, and 4 as problematic. Respondents with the same levels on perceived stress but with low literacy scoring were less likely to agree with these three items as those with high literacy scoring. Additionally, a readability analysis showed that item 4 required a reading level corresponding to ten years of school (10\textsuperscript{th} grade).\textsuperscript{160}

In summary, the PSS is not found to be unidimensional, and there seems to be comprehensive problems with the positively formulated items in the PSS scale. The PSS has primarily been investigated with the use of CTT, whereas only sparse IRT has been conducted, also in regard to DIF.
9.2 CONSTRUCT VALIDITY OF THE MAJOR DEPRESSION INVENTORY (PAPER II)

The Mokken analysis of the MDI showed weak Mokken scale properties (a homogeneity coefficient of 0.36). Two items in the MDI (sleep and appetite) gave homogeneity coefficients below 0.30 when a DMHMM was applied. Our findings are in contrast to those of Ellervik et al., who found that the MDI showed adequate measures of scalability as they reported a Mokken homogeneity coefficient above 0.40. The study by Olsen et al. found a total Mokken coefficient of 0.52 even though the physical items (sleep and appetite) had coefficients < 0.40.

The construct validation of the MDI did demonstrated misfit to item 9 and 10 and no overall fit to the Rasch model. Disordered response categories were demonstrated for eight of the ten items. After modifying the original six-point scoring system to a four-point system for all items, ordered response categories were achieved. We found acceptable fit with four response categories if items 9 and 10 were omitted. Even though we know that no fit to the Rasch model was achieved with six response categories, we do not know how five response categories with different wording would perform why this needs to be investigated.

The misfit of items 9 and 10 in our study is in agreement with the findings from the Rasch analysis reported in the study by Amris et al. on females with chronic widespread pain as they also reported misfit of four items. This suggests lack of unidimensionality and thereby indicates that the MDI does not measure a single construct. The findings by Amris et al. and our findings are in contrast with the findings by Olsen et al., who found the MDI to be unidimensional in a Rasch study of 91 patients.

The results of our study of the MDI regarding misfitting items measuring appetite and sleep are equivalent to those obtained for former studies of other depression scales. In a former study by Lerdal et al. investigating the Beck Depression Inventory (BDI) by Rasch analysis, item 16 (changes in sleep) and item 18 (changes in appetite) did not demonstrate acceptable goodness-of-fit to the Rasch model. Which was supported by Siegert et al. In a study by Horton et al. of the PHQ-9 in a primary care setting, dependency issues were identified between items 3 and 4 (‘Trouble falling or staying asleep, or sleeping too much’ and ‘Feeling tired or having little energy’). Bech et al.
studied the monotonicity of the Hamilton Depression Scale (HAM-6) and found that item 6 (insomnia, delayed) and item 12 (somatic symptoms) delayed when validating using experienced psychiatrist.\textsuperscript{164} A Rasch analysis by Bech et al. found that general somatic symptoms’ of the HAM-6 was less sensitive to differentiate between the highest degrees of depression.\textsuperscript{165}

No significant DIF was observed for gender, age, work extent in the last 12 months, current work status and education level. Our data only allowed us to investigate DIF for patients who participated in the M-CIDI interview. Even though we did not find any evidence of DIF, it would be of great importance to test for DIF with respect to use of medicine and mental and physical comorbidity. The identified problems with the somatic items of the MDI could be caused by DIF, which is a limitation of our study. This could be further explored by testing DIF for medication use and for mental and physical comorbidity, which might have influenced the responses on the items focusing on appetite, sleep and tiredness. If DIF is observed, it is difficult to ascertain whether any differences in the prevalence of symptoms measured between groups represent a true difference or whether the difference is due to item bias.\textsuperscript{166} This could lead to undertreatment or overtreatment of a particular group of patients. The assumption of specific objectivity for the Rasch analysis implies that the scale is not sample dependent.\textsuperscript{99} Still, if misfit of a single item is found, it is important to investigate DIF for different socioeconomic factors. We were not able to identify any former studies investigating DIF for the MDI scale. A study of DIF in PHQ-9 by Cameron et al., which was based on a clinical primary care sample of 895 individuals aged 55 years or older, the respondents exhibited greater tendency to endorse the PHQ-9 items regarding anhedonia and low mood than respondents under age 55 years. Likewise, older adults showed lesser tendency to endorse the item on ‘Feeling tired and having little energy’.\textsuperscript{167} Which imply the importance of further investigation for DIF of the MDI.

In summary, the Rasch analysis of the MDI revealed problems with the somatic items measuring sleep and appetite. Future studies need to address the wording of the modified response categories and further investigate DIF.\textsuperscript{104}
9.3 CRITERION VALIDITY OF MAJOR DEPRESSION INVENTORY (PAPER III)

For the patients who were not interviewed, information on age, gender and sum score of the MDI was available. The mean age for interviewed patients was higher (38.6 years), whereas the mean age for non-interviewed patients was lower (34.8 years). Younger patients tend to hold on to their job after they are diagnosed with depression, while the older patients tend to go on sick leave or retire and therefore might have better time to participate in the telephone interview. The gender distribution was comparable among those who were not interviewed (males: 35 (37.2), females: 59 (62.8)) and those who were interviewed (males: 63 (41.4), females: 89 (58.6)). The sum score was almost the same for these groups, which could indicate that there was no difference in the severity of depression between those interviewed (28.9) and those not interviewed (29.3).

We found a high prevalence of depression in our study sample consisting of patients suspected of clinical depression by their GP: 87.9% for any depression and 64.4% for severe depression. The high prevalence was expected as the participating GPs were instructed to include only patients suspected of depression. Additionally, findings from previous studies on depression in general practice found that GPs recognize the disease in 56–75% of cases and another study of depression screening of primary care patients found high-risk prevalence of depression. Our study is the first to establish the prevalence of depression on clinical suspicion of depression.

The MDI demonstrated a sensitivity of 62% for any depression, corresponding to a false negative rate of 38%, and a low sensitivity of 41% for severe depression, consistent to a false negative rate of 59%.

The MDI demonstrated a modest specificity of 63% for any depression, consistent to a false positive rate of 37%. The specificity of the MDI test was relatively high (85.1%) for severe depression, corresponding to a false positive rate of only 14.9%. Findings by Mitchell et al., who reported that the PHQ-2 is suitable for initial assessment followed by the PHQ-9. This was supported by the studies by Arrol et al. and Sherina et al. These studies supports our aim of evaluating the MDI when used on clinical suspicion of depression (to recruit only patients with two or three core symptoms of depression).
To the best of our knowledge, no previous studies have tested the MDI in a primary care setting. In a study by Cuijpers et al., the DSM-IV criteria of the MDI was compared to a regular diagnostic interview by experienced psychiatrists. The psychiatric outpatient population in their study was also highly prevalent. Our findings are in line with their findings (with a sensitivity of 66% and a specificity of 65%) and the AUC of 0.68 was similar to the AUC of 0.66 in our study.56

In the study by Bech et al., the sensitivity of the MDI algorithms for major depression varied between 86% and 92%, whereas the specificity varied between 82% and 86%. 53 These findings differ from our results. This is possibly because their study was performed with a SCAN interview, which in a study by Jordanova had a lower prevalence of depression than the CIDI interview.120 A possible explanation for the low sensitivity of the MDI could be that the M-CIDI interview is too sensitive, which is in line with the findings by Jordanova et al. and Brugha et al.117, 120

An issue of great debate in the Danish media has been whether GPs tend to diagnose too many healthy (non-depressed) people with depression as a consequence of their use of the MDI. The discussion has primarily focused on whether transient stress or adjustment disorders are incorrectly diagnosed as depression.44 Using the MDI on clinical suspicion of depression reduces the problem of false positive depression diagnoses and overdiagnosing of depression in general practice. Our results suggest that the use of the MDI may result in under diagnosing depression if compared with the M-CIDI. If the patient is depressed according to MDI, it is valid and the MDI captures more of the severely depressed. Since the prevalence of depression on clinical suspicion in general practice has not previously been established, the generalisation of our results of the MDI to other studies, such as studies conducted on a waiting room population, is not possible.

Additional analysis

For the dissertation, we additionally investigated the findings from Paper II of misfit for item 9 and 10 in the MDI. To further test how this result would impact our findings of the criterion validity of the MDI we performed the ROC analysis without item 9 and 10. These analysis showed no substantial changes to our result of the criterion validity of the MDI. As previously mentioned studies have shown that somatic symptoms are important when diagnosing depression in general practice as the clinical picture is dominated by
these symptoms.\textsuperscript{170, 171} From a clinical perspective, it is therefore not advisable to exclude item 9 and 10 from the MDI.

In summary, it is of great importance to be cautious when interpreting psychometric tests. Additionally, it is crucial to discuss how the sensitivity and specificity of a test should be interpreted in daily clinical practice. Future research is needed to cross-validate our findings and further explore the validity of the MDI.
CHAPTER 10:

SUPPLEMENTARY DISCUSSION
The following chapter will provide a supplementary discussion to further enlighten specific aspects of the dissertation.

10.1 APPLIED METHODS

10.1.1 Selection of applied Item Response Theory methods

Before initiating the study, it was not possible to make a thoroughly defined analysis strategy. The initial selection of the IRT methods used in the dissertation is more a result of the process of the PhD study. Rasch analysis is generally considered the reference standard within modern psychometric testing. Moreover, the Rasch model had not previously been applied for the PSS. On this background, combined with discussions in the research group in which most of the supervisors knew about the Rasch and the Mokken model, the models were selected as appropriate for analysis of the construct validity of both the MDI and the PSS. Other options could have been to use different IRT models. When we had decided on the software programme for conducting the Rasch analysis, the applied types of analysis was primarily based on the availability in the RUMM programme. Other Rasch model programmes than the RUMM could also have been chosen.

10.1.1.1 Mokken scale analysis

The Mokken scale analysis is relevant to conduct in order to provide further information about the construct validity of the PSS and the MDI. We found it a more thorough approach to present both methods for the data as this would illuminate the subject matter more carefully than if ending the analysis after the Rasch analysis. Also it enables comparison of the results with other studies which have applied the Mokken scale analysis. Some readers of our studies might not be familiar with the Rasch analysis and may prefer interpreting the results of a Mokken scale analysis, which was also a reason for conducting the analysis.

A recent paper by Sijtsma et al. suggests evaluation of dimensionality with Mokken scale analysis and points out several advantages. Sijtsma argues that if the assumptions of the DMHMM are fulfilled, the results of the Mokken and Rasch analysis are closely related as the DMHMM implies an invariant item ordering and adds the assumption of non-intersecting item response functions. Sijtsma et al. state that considering only
scalability coefficients provides an incomplete picture of the analysis. Applying the detailed procedure by Sijtsma would have allowed us to take full advantage of the properties of the Mokken scale analysis. Sijtsma et al. suggest a three-stage model: 1) to recode scores of items that are negatively worded, treat unacceptable scores as missing values and identify if any item scores are outliers, 2) to establish if the scales satisfy the model of monotone homogeneity and the model of double monotonicity and 3) to determine the scale properties identified in the second stage.

We were interested in investigating if the results of the Mokken scale analysis were in accordance with the results of the Rasch analysis. However, we acknowledge that we should have had more confidence in the Rasch analysis than in the Mokken scale analysis in relation to the final conclusions. The results in Paper 1 led us to claim that the Mokken scale analysis identified a unidimensional model. However, the fit of the scale was insufficient for the DMHMM as item 4 presented misfit. We now interpret the results as the PSS having weak Mokken scale properties, which is stated in the discussion of the findings in the dissertation. In the overall conclusion in Paper 1, we relied more on the results of the Rasch and CFA analyses, and we concluded that our results revealed statistical support for a two-dimensional scale.

In Paper 2, we applied a more cautious interpretation of the Mokken scale analysis, and we did not find that we needed to downgrade our finding that items 9 and 10 constitute a separate scale than the other items investigated in the dissertation.

10.1.2 RUMM-2030
The Rasch analyses carried out in Study 1 and Study 2 were performed in the RUMM-2030 (RUMM) programme. The main restrictions with the RUMM concern the fit statistics and the interaction with the sample size. For fit statistics, the distribution of the residuals is known, and this should be conditional on the actual scores that have been observed; RUMM calculates these residuals (unconditionally) based on a theoretical estimate of a person’s score rather than the actual observed value. Unconditional item fit statistics become unreliable for $n \geq 500$. The sample size issue concerns potential ‘false’ misfit: If the sample size gets too big, the fit statistics becomes overpowered, and everything may appear to look misfitting although it is not. This makes the interpretation difficult.
In Study 1, the sample size was very large, which caused some challenges when conducting the Rasch analysis in RUMM. Our sample size procedure with ten random subsamples of 500 seems to be a suitable strategy.

In Study 2, the sample size was only just sufficient. Hagell & Westergren assume that a sample size of 250-500 appear to provide a good balance for the statistical interpretation of the fit statistics, and our sample size also follows the guideline for sample size recommended by Linacre.

10.1.2.2 The likelihood ratio test
As stated by Tennant & Conaghan, the first step in the process of conducting a Rasch analysis is deciding which mathematical deviation of the model should be chosen. In RUMM, likelihood ratio statistics help decide which polytomous version of the model to use. The PCM specifies that each item has its own rating scale structure, and the thresholds in the PCM may differ between items. The restricted RSM assumes the distance between the thresholds to be equal across items.

A more experienced researcher would probably have searched for a fitting model and then have tested this model against the RSM. As all members of the research group were first-time users of the Rasch model in RUMM, we had no reason to question the before mentioned guidance of analysis described by Tennant & Conaghan. We did not consider the large sample size in the likelihood ratio test for Study 1, which is a limitation of our analysis.

10.2 OVERALL ANALYSIS PLAN

10.2.1 Order of conducted analyses
In Paper 1, the analyses were conducted in the following order: 1) CFA analysis, 2) Rasch analysis and 3) Mokken scale analysis. In Paper 2, the order was: 1) Rasch analysis and 2) Mokken scale analysis.

However, due to the many steps of conducting the Rasch analysis, each step of the analysis was not completed before the Mokken scale analysis was applied; this is why the order of the analysis was not as fixed as in the above-mentioned order.

Initially in our studies, the Mokken scale analysis was considered a supplementary analysis to the Rasch analysis, while it was conducted after the Rasch analysis in both
Study 1 and 2. This order was taken because of the strategy chosen in the research group in the early stages of the studies. At that time we regarded the Mokken scale analysis as an additional analysis to Rasch analysis. Later on in our studies, it became apparent that the Mokken scale analysis is often used in the step before the Rasch analysis.\(^{87}\) I now consider this latter approach a more appropriate approach, although I am aware that Mokken scale analysis and Rasch analysis are closely related as also stated by Sijtsma.\(^{173}\)

In the introduction part of the dissertation in section 2.2, the analysis is presented in the following order: CFA, Mokken scale analysis and Rasch analysis.

### 10.2.2 Clarification of followed plan

At the beginning of our analytical work, we decided to investigate if the data fitted the Rasch model and (if not) if we could make the data fit to the model with only a few changes of the scales rather than searching for other IRT models that might have fitted the data better than the selected model. We were inspired by the viewpoints and the approach described by Pallant & Tennant 2007 and Conaghan & Tennant 2007.\(^{84,\;102}\)

According to their approach, the items for a new or an existing scale are tested against the expectations of this measurement model.\(^{84,\;102}\) Although we partially followed the approach by Pallant & Tennant, we now realize that several different decisions could have been made when conducting our analysis.

Both the PSS and the MDI were already well-established scales when we conducted our studies. The content validity of the scales was an important aspect to us because we did not exclusively investigate the scales to achieve fit; we also wanted the scales to have high reliability and to be useful in clinical practice.

In our analytical plan, we aimed to investigate whether the PSS and the MDI revealed fit to a unidimensional Rasch model. We explored ways to account for any misfits found in the scales, but our overall analysis plan for the construct validity studies was conservative as we planned only to exploratory remove few of the items from the MDI and PSS scales. We did not want to continue removing items as we could not be sure that the construct for which the scale had originally been developed would still be reflected.

The iterations of the analyses involved item removal and collapsing of non-fitting response categories. This process was performed by excluding the most misfitting item (one item at a time), and the statistical properties were then tested again. If the model
still not received fit to the Rasch model, we collapsed response categories for items with misfitting response categories.\textsuperscript{36, 104}

10.3 EFFECT OF THE APPLIED ANALYSIS ON THE RESULTS

The original analytic plan, which was followed in the performed Rasch analysis in Study 1 and 2, may have affected the reported results in various ways. These will be reviewed as a merged discussion of methods and results.

10.3.1 Specific steps and choices in Study 1

In study 1, we initially tested the fit to the Rasch model for the total sample. We then chose to make ten random subsamples, each of 500 respondents, to take into account the problems with significance tests in large samples. We did not have a predefined strategy concerning how to conclude on the basis of the results from the ten subsamples. We initially investigated if the items were the main reason for the misfit to the Rasch model. As part of a purely exploratory approach, the most misfitting item was deleted first. The two items demonstrating the largest misfits were then deleted one at a time, but this procedure revealed no improved fit to the Rasch model for the remaining items.\textsuperscript{36}

To test whether extreme persons with large residuals may have affected the estimates, a modification was performed, but the data did still not fit the Rasch model. To investigate whether disordered thresholds could explain the misfit to the Rasch model, we collapsed the response categories. We chose to collapse the response categories in the same way for every item with misfitting response categories, and they were only collapsed for the misfitting items. After rescoring all response categories with the same collapsing process for all items, we obtained a lower degree of statistical perfection, but we also got a scale that was easier to interpret and use.\textsuperscript{177}

Although the global test of fit improved with the collapsed response categories for items 4 and 7 with disordered thresholds, the analysis did still not demonstrate fit to the model with all ten items. Moreover, the data with the collapsed response categories did not fit the model with nine or eight items after deletion of item 10 (with the largest misfit) and item 5 (with the second largest misfit).

DIF was not reported in the manuscript for Study 1. Therefore, we cannot reject that DIF could explain the misfit of the PSS to the Rasch model. If the uniform DIF analysis had
been further studied as part of the original analysis plan, the findings could have been further explored by resolving the items that revealed uniform DIF (see section 10.4 for additional analysis). We could also have included a test for uniform DIF in the two dimensions (see section 10.4.3).36

Local dependency was not reported for the PSS items, which is a limitation of our study. Dependency between two or more items can be investigated by combining these items to produce one single and more extended item, which is called a subtest in the RUMM.92 However, we did not perform subtests to account for possible dependency and investigate if the fit improved. It could also have been appropriate to test for local dependency within the suggested two dimensions of the scale.

We investigated the model fit after splitting the scale into the two suggested dimensions as found in previous research.31, 32 The Rasch analysis of the six negatively worded items did not show fit to the model, nor did the analysis of the four positively worded items. It would have been appropriate to report the analysis of the two dimensions for each of the ten random subsamples of 500 to take into account the large sample size.36

10.3.2 Specific steps and choices in Study 2

When conducting Study 2, we did not experience challenges regarding sample size. Therefore, we did not make random subsamples. We developed a more clearly defined analysis plan as we had already dealt with many of the key decisions in Study 1, that are necessary when handling real-life data in the Rasch model.

First, we tested the fit to the Rasch model for the total sample. Second, we investigated if items were the main reason for the misfit to the Rasch model by excluding misfitting items. Excluding item 9 did not improve the overall model fit, while excluding both items 9 and 10 (with the two largest item fit residuals) did improve the overall model fit. We also tried to exclude all three somatic items (items 8, 9 and 10). However, as the fit did not improve, we chose to keep item 8 in the scale.104

RUMM identifies extreme persons. We did not investigate whether exclusion of persons with large residuals would have improved the fit of the MDI to the Rasch model.

To investigate whether disordered thresholds could explain the misfit to the Rasch model, the next step was to collapse response categories. Only two items displayed ordered thresholds with six response categories, which could indicate a general problem with
response categories. One response category was first rescored for all items (including the two with ordered thresholds), then two response categories, and ordered thresholds were then presented for all items even though no fit to the Rasch model was achieved. The most misfitting item was then excluded one item at a time with the collapsed response categories. When rescoring all items (four response categories) and excluding items 9 and 10, overall fit and ordered thresholds were achieved for the remaining items.104

DIF was investigated for gender, age, work status and education, and no significant uniform or non-uniform DIF was observed. We suggest that further DIF analysis is needed to explore whether the misfit of items 9 and 10 can be explained by medication use or by mental or physical comorbidity before deleting the items from the scale.

We also tested for local dependency, and the results suggest that the MDI contains items that are characterized by response dependency (1 & 8, 4 & 5, and 9 & 10). If we had made a subtest to consider the response dependency for items 9 and 10, we could have tested if we achieved fit. If fit was then achieved, items 9 and 10 should not have been excluded.104

### 10.3.3 Investigating the ordering of thresholds

Inspired by Andrich’s work, we chose to investigate the ordering of the thresholds in Study 1 and 2.106 However, we are aware of the ongoing discussion of the significance of investigating disordered thresholds. García-Pérez et al. (2017) state that disordered crossings should not be considered as a criterion for removal of items or for collapsing of response categories. Even when the respondents use some categories a little, they can still make distinctions that matter to practitioners.178 The fact that we chose to investigate thresholds in Study 1 and 2 might have affected our results by making the scales weaker since collapsing categories may reduce the item information function.178

### 10.4 ADDITIONAL DIF ANALYSIS STUDY 1

In the following, the additional DIF analysis conducted for Study 1 is further discussed. Even though more DIF is claimed in the total sample, we find that the combination of the results from the total sample and the random subsamples is the most appropriate since the DIF in the total sample will be random and spurious.139 The reason is that the total
Supplementary discussion

sample most likely will signal a lot of DIF that is irrelevant in practice. Supplementary data for the number of individuals in the different subgroups of each subsample are presented in Appendix 9 in the dissertation (see Table 9.1 “Description of the ten random subsamples for the conducted DIF analysis”). We found that the six exogenous variables investigated for DIF seemed reasonably well distributed in the ten different subsamples.

10.4.1 Non-uniform DIF results
In the total sample (N= 13 307), non-uniform DIF was present for: gender (item 7), age group (item 5-10), native language (item 5) and years in school (item 1-5, 7 and 10) (data not shown in table).

In one subsample, non-uniform DIF was present for age in item 7. For native language, non-uniform DIF was present for item 4 in two subsamples and for item 5 and 7 in one subsample each. For anxiety, non-uniform DIF was present for item 3 in one subsample. For years in school, non-uniform DIF was present for item 10 in two subsamples and for item 4 in one subsample (data not shown in the table). According to the results from the subsamples, there seem to be little evidence of non-uniform DIF for the PSS.

10.4.2 Uniform DIF results
The uniform DIF analysis for the total sample (N= 13 307) is presented in Table 4.1 “Differential item function for 10-item Perceived Stress Scale (Total sample) (Uniform)” in Chapter 4. We wanted to be able to assess whether sample size problems alone were the reason why the uniform DIF appeared to be present in the total sample, and the results and the test statistics of the ten random subsamples are presented in Table 4.2 “Differential item function for 10-item Perceived Stress Scale (Random subsamples) (Uniform)” in Chapter 4.

For gender in the total sample, items 1-3 and 7-10 presented uniform DIF. Item 3 presented uniform DIF in two subsamples (women found it easier to endorse item 3), and items 4 and 7 each presented uniform DIF in one subsample. As only two subsamples presented uniform DIF, we do not suggest further investigation.

For age in the total sample, items 1-9 revealed uniform DIF. In eight subsamples, younger participants found it easier to endorse item 3. Item 4 revealed uniform DIF in four subsamples, and item 1 revealed uniform DIF in two subsamples. Item 5, 7 and 8 revealed
uniform DIF in one subsample each. To further investigate uniform DIF, item 3 should first be accounted for and hereafter maybe item 4.

For native language, items 3, 6, 7 and 9 showed evidence of uniform DIF in the total sample. In one subsample, participants with another native language than Danish found it more difficult to endorse item 3. Items 7 and 9 also revealed uniform DIF in one subsample. However, we would not suggest further investigation for two reasons: only few persons had another native language than Danish, and only one subsample presented uniform DIF.

For anxiety, all ten items in the total sample revealed uniform DIF. In the subsamples, those who previously or presently had transient mental illness (e.g. mild depression or anxiety) found it more difficult to endorse item 3 in five out of ten subsamples. Item 4 revealed uniform DIF in four subsamples, and item 2 in two subsamples. Items 1, 6, 9 and 10 also revealed uniform DIF in one subsample. Further analysis might investigate item 3 and afterwards maybe item 4.

For years in school, all ten items revealed uniform DIF in the total sample. In the subsamples, those who had few years in school found it more difficult to endorse item 3 in nine subsamples. Also item 4 revealed uniform DIF in three subsamples and item 9 in two subsamples. Items 1 and 7 revealed uniform DIF in one subsample. If uniform DIF should be investigated further for years in school, item 3 should first be examined and afterwards perhaps item 4.

For long-term illness in the total sample, items 2, 4, 6, 7, 9 and 10 showed evidence of uniform DIF. In the subsamples, those who had not experienced long-term illness found it more difficult to endorse item 6 in four subsamples. Furthermore, item 10 revealed uniform DIF in one subsample. If uniform DIF is to be investigated further, one could investigate item 6.

In the subsamples, there is overall consistency in the reporting of which items present uniform DIF. In conclusion, item 3 indicates uniform DIF for age (8 out of 10), years in school (9 out of 10) and probably anxiety (3 out of 10). Regarding item 4, the most uniform DIF was available for age in the subsamples (4 out of 10). Thus, the tendency of DIF for item 4 could be discussed. We suggest that uniform DIF for item 3 is investigated before resolving the DIF for item 4.
10.4.3 The consequences of the investigated uniform DIF
Our analysis revealed uniform DIF as the results showed that several subgroups might respond differently to specifically item 3 (“How often have you felt nervous and ‘stressed’?”), especially older and less educated persons. The uniform DIF that appears when measuring perceived stress with the PSS could suggest that persons with a shorter education find it more difficult to endorse item 3 than persons with a longer education. The uniform DIF for age (as mentioned in Chapter 4) may raise the question whether older persons tend to interpret the word “stressed” as a more serious condition than younger persons do. If we look generally at the results of uniform DIF, we might identify a group of older persons with low education who may have experienced long-term illness. This indicates that there might be a subgroup of persons in whom we do not measure stress with the available indirect variables. This subgroup may experience stress differently than others, which might have biased the results of the PSS study.

10.4.4 The consequences of the methodological approach for the DIF analysis
To further investigate the uniform DIF of the PSS, we could have performed separate uniform DIF analyses of the two subscales (the four positively worded items and the six negatively worded items). We acknowledge that performing separate DIF analyses for the two identified dimensions would have been a more appropriate approach to the analysis. However, as this was outside the initial aim of our study, we chose not to further investigate the two dimensions separately before we knew of the consequences of the applied wording in the reversed items. This is also in accordance with the approach taken by Christensen et al., who state that, in case of misfitting items, the researcher must take a second look at both the subject matter arguments that lead to the development of the item, the contents, the formulations of questions and response categories.72

In the study by Taylor et al., DIF analysis was conducted on the two subscales of the PSS, but these showed no evidence of uniform or non-uniform DIF for any of the items.154 Furthermore, the size of the uniform DIF could be explored further by investigating the difference between respondents with the same score on the remaining items.181 The estimates of the item parameters can be compared between different groups and the principles of test equating are often applied.180
10.4.5 Accounting for DIF
The substantial DIF identified for the PSS needs further investigation in future studies. Andrich et al. state that two main challenges exist within DIF analysis. First, there is a need to distinguish between real and artificial DIF items because real DIF in one item may induce artificial DIF in other items. Second, resolving an item that shows evidence of DIF (into multiple items, one for each group) may improve the fit of the data to the model for measurement, but the content validity of the scale may worsen. Removing an item might decrease both reliability and person separation, and resolving uniform DIF is thus preferable.\textsuperscript{180} The next step for investigating the PSS further could be to take DIF into account by splitting the items. In RUMM, uniform DIF can be resolved by splitting of items into separate sample-specific items.\textsuperscript{92, 182}

10.5 FINDINGS OF THE PSS

10.5.1 The dimensionality of the PSS
We acknowledge that our findings do not support the PSS with all ten items to be a unidimensional and valid scale, and the interpretation of the PSS results should be cautious. Our CFA analysis revealed that the two dimensions of the PSS are closely related (correlation: 0.687). No prior research has been made into the performance of the two subscales that predict outcomes such as mortality or multimorbidity. Therefore, we suggest that the work with revising the wording of the four positive items of the PSS is continued in order to maintain as many items as possible, before splitting the PSS into two dimensions. Further investigation is also in line with the suggestions made in other studies on this topic.\textsuperscript{36}

Items 3 and 4 seem to constitute a major problem. Our finding that item 4 was the item representing the largest misfit is in agreement with the findings by Dougherty et al.\textsuperscript{152} However, in contrast to their analysis, our analysis did not reach unidimensionality, not even after removing item 4 from the analysis. The recent findings of DIF for item 4 might explain the misfit. We believe that the DIF for item 3 needs to be resolved before it can be established if the DIF for item 4 is only an artefact. Medvedev et al. found local dependency between items 1, 2 and 9; 4, 5, 7 and 8; 6, 10 and 3. They found the best fit when using three subtest sets of items.\textsuperscript{153} One of the suggested subtest of items covers all items from the positive dimension of the PSS.
A CFA by Ezzati et al. suggests that the two factors should be considered as weakly correlated yet independent factors because these two factors show different correlations with other variables (i.e. memory domain). The two factors are suggested to be used as separate indicators of stress in future studies of elderly adults. Other studies have shown that positive versus negative wording of similar items may influence the outcome. We also suggest a concurrent validity study investigating if the PSS or one of the two subscales correlate best with other stress measures (e.g. p-cortisol measures). For example, a study by Walvekar et al. found a positive correlation between serum cortisol and the PSS.

10.5.2 The prevalence of stress measured with the PSS

As the construct is not unidimensional, it is difficult to interpret the prevalence of PSS based on a common sum score. However, a recent CFA by Denovan et al. found closer fit to a bi-factor model consisting of three latent factors (i.e. a general perceived stress factor, Distress and Coping) than to a one-, two- or three-factor model. They suggest that the PSS is driven by a single underlying dimension. Their findings indicate that prevalence of stress estimated with PSS in its current format is useful. As no standard cut-off values were predefined, PSS sum scores are often categorized into quintiles; scores in the highest quintile (≥ 15–17) are generally considered abnormal. However, as the PSS consists of two dimensions, one should look at quartiles for both subscales. This might be problematic if a person is in the highest quartile for one subscale, but in the lowest for the other subscale.

As no reference standard is available for measuring stress, it is difficult to evaluate the quintile approach as well as the cut point (≥ 15–17).

10.6 CONSEQUENCES FOR THE USE OF THE MDI

We chose to further test how the misfit for items 9 and 10 could have impacted our findings concerning the criterion validation of the MDI, and we performed the ROC analysis without items 9 and 10. If we had chosen to remove the appetite and sleep items in Study 3 due to misfit in Study 2, a new division of the cut-points should have been
made as the score range is changed in the 8-item version of the MDI. The new cut-points should have been divided into no, moderate and severe depression and should have been determined according to the score on the ROC curve because the total score will now sum up to only 40. However, we chose not to delete the two items on appetite and sleep in Study 3. This decision was taken because, when diagnosing depression in general practice, somatic symptoms are important for maintaining content validity.170

10.6.1 Contribution of the MDI studies
The construct validation study is the first study to conduct a comprehensive Rasch analysis of the MDI in a general practice population. Our study is also the first to test for DIF, and we find it encouraging that no DIF was found for the investigated exogenous variables. Our finding that the MDI was not a unidimensional scale is in accordance with an earlier Rasch study by Amris et al.65, but it is in controversy with a study by Olsen et al. that supported a unidimensional model, even though they showed misfit for the sleep item.55 The Rasch analysis by Olsen et al. was based on only 91 patients and less comprehensive analyses than those conducted in our study. The MDI should be further explored for DIF to investigate if the misfit found for the appetite and sleep items in our study could be explained by DIF related to medication use or comorbidity.104

Our criterion validation study was the first to be conducted in general practice, where the MDI is often applied. Our finding that the MDI is a conservative instrument for diagnosing depression is encouraging in terms of preventing over-diagnosis (false positive diagnosis) of depression in general practice. However, our results suggest that the MDI did not identify all patients with depression when the M-CIDI is used as reference standard.127 We found a lower diagnostic precision of the MDI than found in the earlier study by Bech et al. using the SCAN interview among patients from psychiatric departments as their study reported a sensitivity for moderate to severe depression of 86% and a specificity of 86%.53 Our findings are more in agreement with the earlier findings of Cuijpers et al., who reported a sensitivity of 66% and a specificity of 65% compared to a regular diagnostic interview by a psychiatrist.56 However, our reference standard, the M-CIDI, might have been too sensitive, and our results should be seen in this light. If our analyses had been based on a more conservative reference standard, e.g. the PSE, our results may have been more in agreement with the results reported by Bech et al.53
CHAPTER 11:

CONCLUSIONS
MAIN CONCLUSIONS FOR AIM I-III

I: Our population-based study showed scalability problems in the applied Danish version of the PSS. The CFA analysis confirmed the same grouping of items as found in previous studies. Initially, the Mokken scale analysis revealed that the PSS was classified has having medium Mokken scale properties. A better fit to the Rasch model was obtained by splitting the items into the two dimensions suggested in previous research. Overall, our results revealed statistical support for a two-dimensional scale as the PSS did not fit the Rasch model as a unidimensional scale. Additional analysis showed DIF for items 3 and 4.

II: Our study showed scalability problems with the Danish version of the MDI when applied in a clinical sample from general practice. Initially the Mokken scale analysis classified the total scale as a weak scale. When items 9 and 10 were excluded from the Rasch analysis a better overall statistical fit revealed. Disordered response categories was demonstrated in the majority of items, which might be solved by collapsing response categories. A revised unidimensional eight-item scale of the MDI without items 9 and 10 and with four response categories showed a better fit to the Rasch model. No significant DIF was observed for gender, age, work status and education. DIF needs to be investigated further for medicine use, physical and mental comorbidity.

III: The MDI was found to be a conservative instrument for diagnosing depression in a clinical setting according to the ICD-10 criteria compared to an M-CIDI interview. In contrast to general concerns, the MDI does not seem to overdiagnose depression in general practice. Especially severe depression is not overdiagnosed. This is also important because severe depression is often treated for long periods with antidepressant medications, and public discussions have often focused on the risk of overdiagnosing depression when using MDI in general practice. However, it is essential to be aware that GPs might risk underdiagnosing depression if they rely only on the MDI. Our findings are encouraging as the MDI depression scale appears to be a reasonably valid tool for diagnosing depression on clinical suspicion. This is an important finding as depression is a common disorder that significantly contributes to the morbidity in many persons seen in general practice. Additional analysis revealed no significant difference in the AUC for the ROC curve without item 9 and 10, which had shown misfit in Paper II.
CHAPTER 12:

PERSPECTIVES AND FUTURE RESEARCH
12.1 Increasing global burden of mental problems

In the past decade, more and more individuals experience psychosocial stress on a daily basis. Heavy workloads, job insecurity, and financial difficulties are circumstances that can result in chronically increased stress.\(^{185}\) In line with this, the WHO and the Organisation for Economic Co-operation and Development (OECD) state that the global burden of mental health problems is increasing.\(^{40, 186}\) According to the WHO, the number of people living with depression increased by more than 18% between 2005 and 2015.\(^{187}\) Effective diagnosis and coordination by GPs should be secured.\(^{186}\)

Katon et al. found that twenty-four percent of high utilizers of medical care (the top 10%) have been found to suffer from current major depression. Depressed patients made an average of 15 visits and 15 telephone calls to a medical clinic during a single year.\(^{188}\)

The global burden of stress and depression are emphasizing the importance of our validation studies of psychometric instruments, since recognition and diagnosing are the first steps for appropriate treatment.

12.2 Measuring stress by the Perceived Stress Scale

To be able to measure perceived stress in general practice, further studies of a revised version of the PSS are required to assess its validity and reliability because the Rasch model was not fulfilled in the studied population.

According to our findings of DIF for item 3 and 10 of the PSS further research need to investigate this in depth. This might have implications of the findings of the prevalence of perceived stress when applying the PSS. Guidelines on how to manage DIF will be required.

Additionally, it might be a problem that the PSS consists of both positively and negatively worded items. A study by Salazar stated that the combination of positively and negatively worded items seriously affected the internal consistency of the scales.\(^{189}\)

To further explore the PSS, we suggest a rephrasing of the four positively worded items into negatively worded items, as the two dimensions could be artefacts formed by opposing formulations of the two sets of items. We have collected data from 1153 young persons from the CDR on these two different versions of the PSS. The analyses are currently ongoing. An evaluation of a revised version of the PSS with all negatively formulated items could be done by conducting qualitative interviews to ensure comprehensibility,
coverage and relevance. Followed by a new Rasch testing this could be another way to approach this issue.\textsuperscript{36} Moreover, the quality of the translation process of the Danish version of the PSS could be further examined. Or the Danish consensus version of the PSS could be investigated.

The large correlation between the positive and negative dimensions found in our study call for further interpretation of the concept of perceived stress as opposed to counter-stress.\textsuperscript{36} These two dimensions influence the content validity and the overall concept of perceived stress. Additionally, the predictive validity of the two dimensions should be examined by exploring which of the two dimensions predict more stress-related sick-leave, hospitalisations or co-morbidity.

The PSS was developed back in 1983. Modern society and our way of living have undergone much change since then, and we seem to experience more stress-related illness today. In consideration of this development, it would be interesting to explore whether perceived stress can still be measured by the same wording of items as used in the original version. Further studies of the PSS could be the beginning of developing guidelines for GPs on how to measure and treat stress in general practice.

12.3 Diagnosing depression by the Major Depression Inventory

The MDI was developed in 1989, and hence the content validity of the MDI could be further explored. For example, it could be investigated whether the wording of item 7 on concentration is still up-to-date as it is asking on reading newspapers and watching TV. More research on the construct of the MDI is needed to further explore whether the misfit of items 9 and 10 can be explained by medication use, mental or physical comorbidity (especially anxiety). Furthermore, the wording of the suggested collapsed response categories needs to be investigated in detail. Likewise, we need to explore how this wording could affect the thresholds of the response categories and whether five or four response categories are best.\textsuperscript{104}

Due to our findings of misfit for items 9 and 10, it might also be interesting to test the MDI without the somatic items. Several studies have shown that somatic symptoms are important due to the clinical picture is dominated by these symptoms, such as lack of energy and general aches and pains, in approximately two-thirds of patients with depression.\textsuperscript{170, 171} Many patients frequently attribute their depression as being due to somatic symptoms. This makes it even more complex to diagnose depression and the
somatic symptoms might also be effected by DIF more easily, which is why this should be investigated further.

It would also be interesting to examine whether the MDI is sensitive to change over time, for example, in terms of treatment or mental comorbidity.

Further research is required to cross-validate our findings of the criterion validity of the MDI, e.g. using the SCAN interview. It will also be interesting to compare paper-pencil and web-based versions of the MDI to test for modus effects (e.g. using DIF analysis).

Cultural aspects of depression also play a role and make the diagnosis of depression even harder to establish. Some items might have a different meaning in another cultural background. Therefore, it is important to investigate the Danish version of the MDI used in persons speaking a non-Danish language.

**12.4 Web-based psychometric testing and use of PRO**

Since we developed Sundhedsmappen.dk, a new web-based system has come into use in Danish general practice called Web-Patient and Web-Req. PRO in general practice is a project funded by the Danish Ministry of Health to broaden the use of PROs in the laboratory requisition system Web-Req and Web-Patient. The system contains many different treatment options, including psychometric tests targeting stress, anxiety and depression.

This system makes it possible for the patient to fill out the MDI or the ASS at home, and the GP then receives the patient’s current score in his or her laboratory system. In a current study we are conducting qualitative interviews with 12 GPs to investigate barriers and attitudes towards using web-based psychometric testing in general practice. These new opportunities will be exciting to follow in the future.
CHAPTER 13:

ENGLISH SUMMARY
Introduction: Stress and depression are the most frequently encountered mental health problems in the general population and in primary care. This PhD dissertation focuses on measurement of mental disorders in general practice. It specifically investigates the construct validity of the Perceived Stress Scale (PSS) and the construct and criterion validity of the Major Depression Inventory (MDI). No instrument is recommended for measuring stress in Danish general practice. The PSS is widely used internationally, but the underlying construct validity has never been investigated by modern item response theory. The Danish College of General Practitioners recommend using the MDI, but no previous studies have assessed the validity and the diagnostic accuracy when used on clinical suspicion of depression in general practice.

Aims: This PhD project aims to investigate the construct validity of the PSS when used among adults in the Central Denmark Region, the construct validity of the web-based MDI when used in Danish general practice, and the criterion validity of the web-based MDI for diagnosing adults with depression on clinical suspicion in Danish general practice.

Methods: The PSS formed part of the Danish National Health Survey in 2010. These data were used to assess the construct validity of the PSS by Confirmatory Factor Analysis, Mokken and Rasch analyses. A web-based version of the MDI was validated for diagnostic use on clinical suspicion of depression in general practice. We compared the criterion validity of the MDI by using the Munich-Composite International Diagnostic Interview by phone reference.

Results: The PSS demonstrated medium Mokken scale properties. According to the Rasch analysis the PSS showed a better fit to a two-dimensional model. A CFA confirmed the two dimensions. The MDI demonstrated weak Mokken scale properties and the Mokken scale procedure revealed two subgroups of items. The MDI demonstrated better fit to the Rasch model with collapsed response categories and exclusion of items 9 and 10. The MDI demonstrated a sensitivity of 62% for any depression and 41% for severe depression and a specificity of 62% for any depression and of 85% for severe depression.

Conclusion and perspectives
The PSS and the MDI both demonstrated scalability problems. However, the MDI still seems to be a valid instrument for diagnosing depression when used on clinical suspicion. Our findings suggest that both of the investigated scales have the potential to become even more valid tools if further adjusted and tested according to our findings. Such adjustments may lead to improved measurement, diagnosis and treatment of both stress and depression in primary care in the future.
CHAPTER 14:

DANSK RESUMÉ

Formål: Formålet med denne ph.d.-afhandling var at undersøge konstruktionsvaliditeten af PSS blandt voksne i Region Midtjylland, konstruktionsvaliditeten af en web-baseret udgave af MDI ved anvendelse i dansk almen praksis samt kriteri validiteten af en web-baseret udgave af MDI i dansk almen praksis.


Resultater: PSS påviste middel Mokken-skala-egenskaber. PSS passede bedre med en todimensionel skala ifølge Rasch-analyse. En konfirmatorisk faktoranalyse bekræftede de to dimensioner. MDI påviste svage Mokken-skala-egenskaber og Mokken skala proceduren påviste to subgrupper af spørgsmål. MDI passede bedre med Rasch-modellen efter sammenlægning af svarkategorier og udeladelse af spørgsmål 9 og 10. MDI demonstrerede en sensitivitet på 62% for enhver depression og 41% for svær depression og en specificitet på 62% for enhver depression og 85% for svær depression.

Konklusion og perspektiver: PSS og MDI demonstrerede begge problemer med skalerbarheden. MDI synes dog stadig at være et validt diagnosticeringsværktøj, når det anvendes ved klinisk mistanke om depression. Vore resultater peger på, at begge undersøgte spørgeskemaer vil kunne måle endnu mere validt, hvis de bliver yderligere justeret og afprøvet på baggrund af vores resultater. Sådanne tilpasninger kan medføre forbedret måling, diagnosticering og behandling af både stress og depression i almen praksis i fremtiden.
REFERENCES
Validation of instruments for diagnosing depression and measuring stress in general practice

10. Videnscenter for Brugerinddragelse (ViBIS) [The Danish Knowledge Center for User Involvement in Health Care (ViBIS)]. Program PRO - Anvendelse af PRO-data i kvalitetsudviklingen af det danske sundhedssystem- anbefalinger og vidensgrundlag [Programme PRO- Use of PRO data in the quality development of the Danish health system- recommendations and knowledge]. 2016.


44. Danmarks Radio [Danish Broadcasting Corporation]. *DR1 Dokumentar: De raske syge* [Danish Broadcasting Coporation Documentary: The healthy ill]. [TV documentary.]. Copenhagen, Denmark.: Danish Broadcasting Corporation; 2016.


50. The Danish College of General Practitioners. Unipolar depression-diagnostik og behandling [Unipolar depression-diagnostics and treatment]. 2010;2(1).


63. Sundhedssstyrelsen [Danish Health Authority]. Faglige retningslinjer for henvisning til psykolog. For patienter med let til moderat depression eller let til moderat angst [Professional guidelines for referral to a psychologist. For patients with mild to moderate depression or mild to moderate anxiety]. 2012(1).


76. Brodersen J. Measuring psychosocial consequences of false-positive screening results - breast cancer as an example. Copenhagen: Faculty of Health Sciences, University of Copenhagen; 2006.


95. Wright BD. Rating Scale Model (RSM) or Partial Credit Model (PCM)? .... Rasch Measurement Transactions. 1998;12(3):641-642.


103. Comins J.D. Measuring Symptoms, Function, and Psychosocial Consequences in Patients with Anterior Cruciate Ligament Rupture. Development and Validation of a Patient-related Outcome Questionnaire. Copenhagen: Faculty of Health and Medical Sciences, University of Copenhagen; 2012.


116. Wittchen H, Pfister H. *DIA-X-Interviews (M-CIDI): Manual für Screening-Verfahren und Interview; Interviewheft Längsschnittuntersuchung (DIA-X-Lifetime); Ergänzungsheft (DIA-X-Lifetime); Interviewheft Querschnittuntersuchung (DIA-X 12 Monate); Ergänzungsheft (DIA-X 12 Monate) [DIA-X interviews (M-CIDI): manual for screening procedures and interview; Interviewheft longitudinal section analysis (DIA-X-Lifetime); Supplementary booklet (DIA-X-Lifetime); Interviewheft cross-section study (DIA-X 12 months); Supplementary booklet (DIA-X 12 months)].* Frankfurt, Germany: Swets & Zeitlinger; 1997.


APPENDIX
### APPENDIX 1: MAJOR DEPRESSION INVENTORY

**MDI: Items in Danish**

<table>
<thead>
<tr>
<th>Hvor meget af tiden inden for de sidste 14 dage</th>
<th>Hele tiden</th>
<th>Det meste af tiden</th>
<th>Lidt over halvdelen af tiden</th>
<th>Lidt under halvdelen af tiden</th>
<th>Lidt af tiden</th>
<th>På intet tidspunkt</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Har du følt dig trist til mode, ked af det?</td>
<td>5 4 3 2 1 0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2 Har du mangelte interesse for dine daglige gøremål?</td>
<td>5 4 3 2 1 0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3 Har du følt at du manglede energi og kræfter</td>
<td>5 4 3 2 1 0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4 Har du haft mindre selvtilfælde?</td>
<td>5 4 3 2 1 0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5 Har du haft dårlig samvittighed eller skyldfedelse</td>
<td>5 4 3 2 1 0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6 Har du følt, at livet ikke var værd at leve?</td>
<td>5 4 3 2 1 0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7 Har du haft besvær med at koncentrere sig, f.eks. at læse aviser eller følge med i fjernsyn?</td>
<td>5 4 3 2 1 0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8a Har du følt dig rastløs?</td>
<td>5 4 3 2 1 0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8b Har du følt dig mere stille?</td>
<td>5 4 3 2 1 0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9a Har du sovet for lidt?</td>
<td>5 4 3 2 1 0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9b Har du sovet for meget?</td>
<td>5 4 3 2 1 0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10a Har du haft nedsat appetit?</td>
<td>5 4 3 2 1 0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10b Har du haft øget appetit?</td>
<td>5 4 3 2 1 0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Total score**

---

Validation of instruments for diagnosing depression and measuring stress in general practice
### MDI: Scoring key (in English)

At the top the diagnostic demarcation line is indicated. The total score of the 10 items is filled in below.

<table>
<thead>
<tr>
<th>Item</th>
<th>Question</th>
<th>All the time</th>
<th>Most of the time</th>
<th>Slightly more than half the time</th>
<th>Slightly less than half the time</th>
<th>Some of the time</th>
<th>At no time</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Have you felt low in spirits or sad?</td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>Have you lost interest in your daily activities?</td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>Have you felt lacking in energy and strength?</td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>Have you felt less self-confident?</td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>Have you had a bad conscience or feelings of guilt?</td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>Have you felt that life wasn't worth living?</td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>Have you had difficulty in concentrating, e.g., when reading the newspaper or watching TV?</td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>8a</td>
<td>Have you felt restless?</td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>8b</td>
<td>Have you felt subdued or slowed down?</td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>9a</td>
<td>Have you been sleeping too little?</td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>9b</td>
<td>Have you been sleeping too much?</td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>10a</td>
<td>Have you suffered from reduced appetite?</td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>10b</td>
<td>Have you suffered from increased appetite?</td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

Total score (item 1 – 10) 

Diagnosis ICD-10 ____________________ DSM-IV ____________________
MDI: Scoring instructions (in English)

Major Depression Inventory (MDI): A depression questionnaire with a dual function

MDI: Scoring instructions

The questionnaire consists of the ten symptoms contained in the World Health Organization WHO's depression demarcation. WHO employs the last two weeks as the period of time in which to assess whether each symptom has been present for more than half the time. These symptoms are mainly subjective, therefore it is natural to ask the patient to complete the questionnaire, allowing the patient to tick each symptom. A higher number signifies a more constant presence of the symptom in question. Remember to fill in patient name and the date.

The patient's completed questionnaire is scored using the scoring key. MDI (Major Depression Inventory) has a dual function, as it is scored both as an instrument of severity (A) similar to the Hamilton Depression Scale, and (B) as a diagnostic tool.

(A) If MDI is used as a rating scale in the same way as the Hamilton scales, then the sum of the ten questions indicates the degree of depression. For items 8, 9 and 10, with two answer categories for each (a) and (b), the highest score is used. The theoretical score range is thus from 0 (no depression) to 50 (maximum depression).

- Mild depression: MDI total score from 21 to 25
- Moderate depression: MDI total score from 26 to 30
- Severe depression: MDI total score of 31 or higher

(B) MDI as a diagnostic tool: the vertical line (the diagnostic demarcation line) is used as indicated above. The three top symptoms which reflect the core symptoms of the WHO/ICD-10 diagnosis of depressions must have been present during the last two weeks for most of the time. The accompanying symptoms in the remaining seven MDI items must have been present during the last two weeks for more than half of the time.

The ICD-10 algorithm:

- Mild depression: 2 core symptoms and 2 accompanying symptoms
- Moderate depression: 2 core symptoms and 4 accompanying symptoms
- Severe depression: 3 core symptoms and 5 accompanying symptoms

MDI can also be employed when diagnosing DSM-IV major depression. According to DSM-IV only nine symptoms are used, as the DSM-IV item 4 is included in item 5. Thus the item with the highest score is used here.

The DSM-IV algorithm:

- 5 out of the 9 symptoms should be present. Of these one should be one of the two first items, according to DSM-IV these are core symptoms.
- A more precise major depression diagnosis depends on the answer to item 9 (a) or (b) and to item 10 (a) or (b).

Major depression without inverse neurovegetative symptoms: a score on 9a and 10a.
Major depression with inverse neurovegetative symptoms: a score on 9b and 10b.
APPENDIX 2: PSYCHOMETRIC TESTS IN DENMARK

GP use of psychometric tests in Denmark

Number

Years

0 50000 100000 150000 200000 250000 300000

APPENDIX 3: PERCEIVED STRESS SCALE FROM DANISH NATIONAL HEALTH SURVEY

Dagligdagens stress

17. Spørgsmålene drejer sig om din oplevelse af belastende eller stressende situationer inden for den seneste måned.

<table>
<thead>
<tr>
<th>(Sæt et X i hver linje)</th>
<th>Aldrig</th>
<th>Næsten aldrig</th>
<th>En gang imellem</th>
<th>Ofte</th>
<th>Meget ofte</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hvor ofte er du blevet bragt ud af ligevægt over noget, der skete uventet?</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hvor ofte har du følt, at du var ude af stand til at kontrollere de vigtige ting i dit liv?</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hvor ofte har du følt dig nervøs og stresset?</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hvor ofte har du følt, at du var i stand til at klare dine personlige problemer?</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hvor ofte har du følt, at tilværelsen formede sig efter dit hoved?</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hvor ofte har du oplevet, at du ikke kunne overkomme alle de ting, du skulle?</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hvor ofte har du været i stand til at håndtere dagligdagens irritationer?</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hvor ofte har du følt, at du havde styr på tingene?</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hvor ofte er du blevet vred på grund af ting, du ikke var herre over?</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
APPENDIX 4: ITEM 9 FROM MAJOR DEPRESSION INVENTORY FROM SUNDHEDSMAPPEN.DK

Gemme de sidste 2 uger, hvor stor en del af tiden: 
Har du følt dig trist til mode, ked af det?

- Hela tiden
- Del måde af tiden
- Litt over halvdelen af tiden
- Litt under halvdelen af tiden
- Litt af tiden
- På nogen tidspunkt
APPENDIX 5: INVITATION LETTER FOR GENERAL PRACTICE

INVITATION TIL AT DELTAGE I "BEDRE DIAGNOSTIK AF STRESS, DEPRESSION OG ANGST"

Hvert år anvender praktiserende læger i Danmark cirka 200.000 psykometriske tests, først og fremmest Major Depression Inventory (MDI) og Angst-Symptom-Spørgeskemaet (ASS). Alligevel mangler vi et sammelt overblik over effektiviteten af denne metode. Derfor skal de psykometriske tests for depression og angstfølelse nu valideres.

Vi henvender os til din praksis, fordi i alieneede anvender Sundhedsmappen, hvor alle valideringdata bliver samlet. Vi håber, at også i vil deltage i valideringen af de webbaserede udgaver af MDI og ASS i almen praksis.

Hvad får I ud af at deltage?

- I modtager honorering (øverst til et modul på 10 min. à 122,57 kr.) for hver inkluderet patient.
- I får en spad Mini stillet til rådighed i et halvt år - her kan patienterne udfylde de diagnosticke skemaer.
- Jeres deltageri bidrager til at sikre en mere precis diagnosticering af depression og angstfølelse.
- Jeres deltageri bidrager til forskning, som er til gavn for både patienter og praktiserende læger.

Hvordan gør I?

Hvis I vil være med, skal I blot fortsætte med at inkludere konsekutive patienter til Sundhedsmappen.dk, hvis I har mistanke om en klinisk depression eller angstfølelse hos patienten. Derudover skal I informere inkluderede patienter om, at hun/hun bliver ringet op af en certificeret interviewere fra forskningsenheten ved Aarhus Universitet med henblik på at gennemføre et interview (30-40 minutter) over telefonen. I kan udføre vedligeholdte informationspakke til patienten. Pjesen opsummerer formål og vilkår for deltageri i "Bedre diagnosik af stress, depression og angst".

Hvordan tilmelder I praksis?

Det er afgørende for undersøgelsens resultater, at så mange som muligt bidrager med data. Derfor har vi brug for jeres deltageri. I tilmelder praksis til valideringsprojektet ved at sende en e-mail til projektkoordinator Mette Mortensen metm@acp.au.dk og oplyser praksisnavn, adresse og telefonnummer, ringe på 873 16 78 13 eller fax vedhæftede tilmeldingsbladet. I er også velkommen til at kontakte Mette Mortensen, hvis I har spørgsmål eller kommentarer til projektet.

Indsamlede data vil blive brugt i anonymiseret form i forskningsprojekter. Projektet er godkendt af DSAAMs Multifaktsudvalg (MFP 11-013) og Datastyrelsen (J.nr. 2013-44-1756).

Vi håber, at du og din praksis har kyst til at deltage i valideringen af MDI og ASS.

Med venlig hilsen

Mette Mortensen
Projektkoordinator, ph.d.-studierende

Mogens Vestergaard
Professor, speciallæge i almen medicin

Kaj Sparre Christensen
Seniorforsker, praktiserende læge

Forskningsenheden for Almen Praksis, Institut for Folkesundhed, Aarhus Universitet, Bartholomäus Allé 2, 8000 Aarhus C
APPENDIX 6: INSTRUCTIONS FOR GENERAL PRACTICE

Hvad er Sundhedsmappen.dk?
Sundhedsmappen.dk er et gratis værktoy til praksis. Patienterne udfylder skemaerne i klinikken sammen med lægen eller darthjemme. Resultatet kommer ind i lægesystemet som laboratorieresvar. Det er dig som læge, der visiterer patienten til sundhedsmappen.dk. Login og påmindelser kommer automatisk til patienten som en mail eller en sms.

Hvordan logger jeg på Sundhedsmappen.dk?

Hvad kan jeg i Sundhedsmappen.dk?
• oprette en ny patient
• udfylde et skema sammen med patienten
• se alle udfyldte skemaer og en grafisk oversigt for den enkelte patient
• se statistikker for din praksis – også i forhold til andre praksis
• resultater sendes automatisk til dit lægesystem med lupap-i-koder (laboratorieresvar)

Hvilke skemaer kan jeg bruge i Sundhedsmappen.dk?
Der findes skemaer til måling af depression (MDII), angststilstand (ASS) og hjemmeblodtryk. For MDII og ASS er det scoren, der sendes, og for hjemmeblodtryk sendes beregnet gennemsnit. Gennemsnittet er beregnet efter DSAM’s vejledning. Du kan altid logge på www.sundhedsmappen.dk og se hele det udfyldte skema.
Undersøgelsen er godkendt af Datalysynet og Multipraksisudvalget under Dansk Selskab for Almen Medicin.

Hvis du har spørgsmål?
...så kontakt
Projektkoordinator Marie Mortensen
Forskningsenheden for Almen Praksis, Institut for Folkesundhed, Aarhus Universitet.
e-mail: mm@ph.au.dk • telefon: 07 16 78 23

AARHUS UNIVERSITY
FORSKNINGSENHEDEN
AF ALMEN PRAKSIS
Aarhus

BEDRE DIAGNOSTIK AF
STRESS, DEPRESSION OG ANGST
- en del af SUNDHEDSMAPPEN.DK
**APPENDIX 7: FLYER FOR PATIENTS (PAGE 2 & 3)**

Bedre diagnostik af stress, depression og angst.

Depression og angst er blandt de mest udbredte psykiske sygdomme i Danmark. Symptomerne kommer til udtryk på mange forskellige måder, og stress kan forvirre tilstanden.

En præcisk diagnose er det første skridt mod den rette behandling. Din læge kan tæmme dine symptomer ved hjælp af diagnostiske skemaer og i Sundhedsmappen.dk foregår det elektronisk.

For at sikre at skemaerne måler præcist, vil vi sammenligne resultaterne for skemaer og personlige interview. Det forøger som en del af forskningsprojektet "Bedre diagnostik af stress, depression og angst" ved Aarhus Universitet.

**Hvordan bliver jeg interviewet?**

- Når du har udfyldt de første spørgeskemaer hos lægen, vil du inden for en uge blive ringet op for et telefoninterview med en projektmedarbejder.
- Interviewet handler om hvordan du har det mentalt og psykisk.
- Interviewet tager cirka 30-40 minutter og bliver gennemført af en professionel interviewere.

**Hvordan bliver jeg tilmeldt?**

- Du bliver tilmeldt Sundhedsmappen.dk hos din egen læge, hvor I sammen besvarer de første skemaer.
- Dine svar sendes automatisk via en sikker forbindelse til din læge.
- Det er frivilligt at deltage. Du kan når som helst melde dig fra.

**Hvad er Sundhedsmappen.dk?**

- Et online system til at måle stress, depression og angst.
- Et væsentligt til at følge din egen helbredsel udvikling.

**Hvad får jeg ud af det deltage?**

- Som deltager i Sundhedsmappen dk kan både du og din læge følge dit helbredsel udvikling. Det betyder, at du kan få en mere målrettet behandling.
- Dine anonyme svar bidrager til forskning i bedre diagnostik og behandling af stress, depression og angst.

**Hvad skal jeg gøre?**

- Du får automatisk en SMS eller e-mail med brugernavn og adgangskode til www.sundhedsmappen.dk. Derudover får du en påmindelse, når der er tid til at udfylde skemaet.
- Som deltager udfylder du hver 14. dag et online spørgeskema på www.sundhedsmappen.dk
- Du får adgang ved at indtaste dit tilhørende brugernavn og din adgangskode.
**APPENDIX 8: INTERVIEW GUIDE**

<table>
<thead>
<tr>
<th>CIDI-interview</th>
<th>Intro-tekst</th>
<th>17-10-2013</th>
</tr>
</thead>
</table>

### FORTROLIGHED OG ANONYMITET:


Alle de oplysninger, du giver os, bliver behandlet strengt fortroligt, og din identitet forbliver ukendt for andre. Oplysningerne vil derfor heller ikke blive videregivet til din egen læge.

### UNDERSØGELSENS FRIVILLIGE KARAKTER:

Det er frivilligt at deltage i undersøgelsen og du kan altid springe fra uden at fortælle hvorfor. Interviewet spænder meget bredt ind til mange forskellige symptomer på både angst og depression, derfor kan der være spørgsmål som ikke har relevans for dig.

Det var de indledende oplysninger nu vil jeg gå i gang med interviewet.

<table>
<thead>
<tr>
<th>Q3. Det er naturligvis helt i orden, at du siger nej, men vil du fortælle mig, hvorfor du ikke ønsker at blive interviewet, bare så vi kan få en idé om, hvorfor folk ikke ønsker at deltage?</th>
</tr>
</thead>
<tbody>
<tr>
<td>01. Det vil være ubehageligt/forældreægnet svært for mig at deltage</td>
</tr>
<tr>
<td>02. Jeg har ikke tid</td>
</tr>
<tr>
<td>03. Undersøgelsen tager for lang tid</td>
</tr>
<tr>
<td>04. Jeg deltaget ikke i interviewsundersøgelsen</td>
</tr>
<tr>
<td>05. Jeg er ikke interesseredt</td>
</tr>
<tr>
<td>06. Finsker ikke at besvare spørgsmålet</td>
</tr>
<tr>
<td>07. Andet,</td>
</tr>
</tbody>
</table>

---
TILBØJ RESPONDENTER, SOM PÅ ET ELLER ANDET TIDSPUNKT VIRKER MEGET KEDE AF DET, ELLER SOM ER BEKYMREDE ELLER ER SELVMORDSTREBET.

Husk værden graden af forpinthed, i forhold graden af hvor selvmordsstræbt patienten er.

1) Har du snakket med din læge om disse selvmordsplaner?

2) Vil du tage kontakt til din læge igen?

3) Må vi tage kontakt til din egen læge? (Hvis pt. ikke kan tage sige sammen til selv at kontakte lægen er det en god ide at spørge om dette).

4) Kontrakten kan vi lave en aftale om at du kontakter din egen læge og taler med din læge om disse tanker?

5) Spørge ind til hvad det er, der holder personens tilbage fra at begå selvmord? (fx hensyn til familie og pårørende) og bestyrke og gengælge dette.

6) Søke sig at anrede tager sig af personen.

7) Rådgiv om akutte hjælpemuligheder.

8) Altid give patienten håb. Jeg ved at du vil få det bedre.

9) Runde sit ansvaret: Ved du hvad jeg er faktisk bekymret for dig. Når du fortæller mig det her, bliver jeg nødt til at mærke på det...

VIGTIGST noter hvis patienten har virket meget selvmordsstræbt skriv det ind i excel-arket. Skriv også ind hvis du har opfordret patienten til at søge læge.

Afskuel interviewet med en opsamling som man har hørt at patienten er ked af det og virker forpint.

Tænk på det som almindelig "hjælpepligt", hvis man oplever et menneske er i krise.

Vær især opmærksom på hvis der kommer noget op i interviewet som pt. gør opmærksom på at det er første gang pt. har snakket med nogen om dette.


Psychiatri skadestue Region Nord: Tlf. 97 64 17 00. Døgnræben Psykiatrisk telefonrådgivning: Tlf. 98 13 42 02.

Psychiatri skadestue Region Midt: Psykiatrisk Rådgivnings telefon 78 47 04 70.


Psychiatri skadestue Region Hovedstaden: Hillerød: 48 29 32 88. København NV.: 35 31 23 00. Brøndby: 38 64 55 02. Glostrup Tlf nr.: 38 64 06 71. Frederiksberg Tlf nr.: 38 16 37 36. Ballerup Tlf nr.: 38 64 50 03. Amager tlf nr.: 38 64 16 50 00.)
Ev. Center for selvomordforhøjede. Livelineses rådgivere på 70 201 201, alle ugerne dage mellem 11.-23.

Ev. information om projektet:
- Hvad får pt. ud af at være med i interviewet?
- Patienten får mind tål.
- Patienten er med til at sikre nogle gode målerekiskaber.
## APPENDIX 9 DESCRIPTION OF THE TEN RANDOM SUBSAMPLES FOR THE DIF ANALYSIS

### Table 9.1 Description of the ten random subsamples for the conducted DIF analysis

<table>
<thead>
<tr>
<th></th>
<th>Sub 1</th>
<th>Sub 2</th>
<th>Sub 3</th>
<th>Sub 4</th>
<th>Sub 5</th>
<th>Sub 6</th>
<th>Sub 7</th>
<th>Sub 8</th>
<th>Sub 9</th>
<th>Sub 10</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td></td>
<td>(%)</td>
<td>(%)</td>
<td>(%)</td>
<td>(%)</td>
<td>(%)</td>
<td>(%)</td>
<td>(%)</td>
<td>(%)</td>
<td>(%)</td>
<td>(%)</td>
</tr>
<tr>
<td><strong>Total sample</strong></td>
<td>488</td>
<td>495</td>
<td>488</td>
<td>493</td>
<td>491</td>
<td>492</td>
<td>495</td>
<td>496</td>
<td>496</td>
<td>493</td>
</tr>
<tr>
<td><strong>Gender</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Men</td>
<td>233</td>
<td>249</td>
<td>237</td>
<td>215</td>
<td>236</td>
<td>252</td>
<td>250</td>
<td>234</td>
<td>222</td>
<td>231</td>
</tr>
<tr>
<td>(47.7)</td>
<td>(50.3)</td>
<td>(48.6)</td>
<td>(43.6)</td>
<td>(48.1)</td>
<td>(51.2)</td>
<td>(50.5)</td>
<td>(47.2)</td>
<td>(44.8)</td>
<td>(46.8)</td>
<td></td>
</tr>
<tr>
<td>Women</td>
<td>255</td>
<td>246</td>
<td>251</td>
<td>278</td>
<td>255</td>
<td>240</td>
<td>245</td>
<td>262</td>
<td>274</td>
<td>262</td>
</tr>
<tr>
<td>(52.3)</td>
<td>(49.7)</td>
<td>(51.4)</td>
<td>(61.9)</td>
<td>(48.8)</td>
<td>(52.4)</td>
<td>(52.8)</td>
<td>(49.5)</td>
<td>(55.2)</td>
<td>(53.2)</td>
<td></td>
</tr>
<tr>
<td><strong>Age groups</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15-30 years</td>
<td>92</td>
<td>81</td>
<td>93</td>
<td>88</td>
<td>93</td>
<td>83</td>
<td>84</td>
<td>84</td>
<td>84</td>
<td>79</td>
</tr>
<tr>
<td>(18.9)</td>
<td>(16.4)</td>
<td>(19.1)</td>
<td>(17.8)</td>
<td>(16.9)</td>
<td>(16.7)</td>
<td>(16.7)</td>
<td>(15.4)</td>
<td>(16.9)</td>
<td>(16.9)</td>
<td>(16.0)</td>
</tr>
<tr>
<td>31-60 years</td>
<td>276</td>
<td>279</td>
<td>257</td>
<td>281</td>
<td>283</td>
<td>285</td>
<td>291</td>
<td>255</td>
<td>278</td>
<td>286</td>
</tr>
<tr>
<td>(56.6)</td>
<td>(56.4)</td>
<td>(52.7)</td>
<td>(57.0)</td>
<td>(57.6)</td>
<td>(57.9)</td>
<td>(58.8)</td>
<td>(51.4)</td>
<td>(56.1)</td>
<td>(58.0)</td>
<td></td>
</tr>
<tr>
<td>61-&gt;70 years</td>
<td>120</td>
<td>135</td>
<td>138</td>
<td>124</td>
<td>125</td>
<td>124</td>
<td>128</td>
<td>157</td>
<td>134</td>
<td>128</td>
</tr>
<tr>
<td>(24.6)</td>
<td>(27.3)</td>
<td>(28.3)</td>
<td>(25.2)</td>
<td>(25.5)</td>
<td>(25.2)</td>
<td>(25.9)</td>
<td>(31.7)</td>
<td>(27.1)</td>
<td>(26.0)</td>
<td></td>
</tr>
<tr>
<td><strong>Native language:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Danish</td>
<td>466</td>
<td>478</td>
<td>459</td>
<td>483</td>
<td>477</td>
<td>473</td>
<td>478</td>
<td>478</td>
<td>478</td>
<td>477</td>
</tr>
<tr>
<td>(95.5)</td>
<td>(96.6)</td>
<td>(94.1)</td>
<td>(98.0)</td>
<td>(97.1)</td>
<td>(96.1)</td>
<td>(96.6)</td>
<td>(94.8)</td>
<td>(94.8)</td>
<td>(96.8)</td>
<td></td>
</tr>
<tr>
<td>Other</td>
<td>22</td>
<td>17</td>
<td>29</td>
<td>17</td>
<td>14</td>
<td>19</td>
<td>17</td>
<td>18</td>
<td>18</td>
<td>16</td>
</tr>
<tr>
<td>(4.5)</td>
<td>(5.4)</td>
<td>(5.9)</td>
<td>(5.9)</td>
<td>(6.2)</td>
<td>(6.9)</td>
<td>(5.1)</td>
<td>(5.5)</td>
<td>(5.5)</td>
<td>(3.6)</td>
<td>(3.2)</td>
</tr>
<tr>
<td><strong>Long term illness</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>154</td>
<td>161</td>
<td>150</td>
<td>161</td>
<td>173</td>
<td>157</td>
<td>159</td>
<td>161</td>
<td>165</td>
<td>159</td>
</tr>
<tr>
<td>(31.6)</td>
<td>(32.5)</td>
<td>(30.7)</td>
<td>(32.7)</td>
<td>(35.2)</td>
<td>(31.9)</td>
<td>(32.1)</td>
<td>(32.5)</td>
<td>(33.3)</td>
<td>(32.3)</td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>334</td>
<td>334</td>
<td>338</td>
<td>332</td>
<td>318</td>
<td>335</td>
<td>336</td>
<td>335</td>
<td>331</td>
<td>334</td>
</tr>
<tr>
<td>(68.4)</td>
<td>(67.5)</td>
<td>(69.3)</td>
<td>(67.3)</td>
<td>(64.8)</td>
<td>(68.1)</td>
<td>(67.9)</td>
<td>(67.5)</td>
<td>(66.7)</td>
<td>(67.7)</td>
<td></td>
</tr>
<tr>
<td><strong>Anxiety:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No, I have never had that</td>
<td>377</td>
<td>398</td>
<td>390</td>
<td>391</td>
<td>390</td>
<td>405</td>
<td>397</td>
<td>389</td>
<td>394</td>
<td>390</td>
</tr>
<tr>
<td>(77.3)</td>
<td>(80.4)</td>
<td>(79.9)</td>
<td>(79.3)</td>
<td>(79.4)</td>
<td>(82.3)</td>
<td>(80.2)</td>
<td>(78.4)</td>
<td>(79.4)</td>
<td>(79.1)</td>
<td></td>
</tr>
<tr>
<td>Yes, I have now</td>
<td>41</td>
<td>31</td>
<td>40</td>
<td>32</td>
<td>37</td>
<td>37</td>
<td>27</td>
<td>36</td>
<td>35</td>
<td>25</td>
</tr>
<tr>
<td>(8.4)</td>
<td>(6.3)</td>
<td>(8.2)</td>
<td>(6.5)</td>
<td>(7.5)</td>
<td>(7.5)</td>
<td>(5.5)</td>
<td>(7.3)</td>
<td>(7.1)</td>
<td>(5.1)</td>
<td></td>
</tr>
<tr>
<td>Yes, I have had it before</td>
<td>70</td>
<td>66</td>
<td>58</td>
<td>70</td>
<td>64</td>
<td>50</td>
<td>71</td>
<td>71</td>
<td>67</td>
<td>85</td>
</tr>
<tr>
<td>(14.3)</td>
<td>(13.3)</td>
<td>(11.9)</td>
<td>(14.2)</td>
<td>(13.0)</td>
<td>(10.2)</td>
<td>(14.3)</td>
<td>(14.3)</td>
<td>(13.5)</td>
<td>(17.2)</td>
<td></td>
</tr>
<tr>
<td><strong>Years in school:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Still in school</td>
<td>4</td>
<td>8</td>
<td>6</td>
<td>2</td>
<td>8</td>
<td>7</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>8</td>
</tr>
<tr>
<td>(0.8)</td>
<td>(1.6)</td>
<td>(1.2)</td>
<td>(0.4)</td>
<td>(1.6)</td>
<td>(1.4)</td>
<td>(0.8)</td>
<td>(0.8)</td>
<td>(0.8)</td>
<td>(1.6)</td>
<td></td>
</tr>
<tr>
<td>7 years or less</td>
<td>57</td>
<td>65</td>
<td>53</td>
<td>47</td>
<td>62</td>
<td>56</td>
<td>55</td>
<td>64</td>
<td>62</td>
<td>55</td>
</tr>
<tr>
<td>(11.7)</td>
<td>(13.1)</td>
<td>(10.9)</td>
<td>(9.5)</td>
<td>(12.6)</td>
<td>(11.4)</td>
<td>(11.1)</td>
<td>(12.9)</td>
<td>(12.5)</td>
<td>(11.2)</td>
<td></td>
</tr>
<tr>
<td>8-9 years</td>
<td>56</td>
<td>72</td>
<td>73</td>
<td>91</td>
<td>80</td>
<td>83</td>
<td>76</td>
<td>70</td>
<td>64</td>
<td>68</td>
</tr>
<tr>
<td>(11.5)</td>
<td>(14.5)</td>
<td>(15.3)</td>
<td>(18.5)</td>
<td>(16.3)</td>
<td>(16.9)</td>
<td>(15.4)</td>
<td>(14.1)</td>
<td>(12.9)</td>
<td>(13.8)</td>
<td></td>
</tr>
<tr>
<td>10-11 years</td>
<td>157</td>
<td>129</td>
<td>136</td>
<td>153</td>
<td>139</td>
<td>144</td>
<td>136</td>
<td>138</td>
<td>138</td>
<td>138</td>
</tr>
<tr>
<td>(32.2)</td>
<td>(26.1)</td>
<td>(27.9)</td>
<td>(31.0)</td>
<td>(28.3)</td>
<td>(29.3)</td>
<td>(27.5)</td>
<td>(33.5)</td>
<td>(27.8)</td>
<td>(28.0)</td>
<td></td>
</tr>
<tr>
<td>High school</td>
<td>186</td>
<td>197</td>
<td>189</td>
<td>175</td>
<td>173</td>
<td>173</td>
<td>191</td>
<td>165</td>
<td>205</td>
<td>202</td>
</tr>
<tr>
<td>(38.1)</td>
<td>(39.8)</td>
<td>(38.7)</td>
<td>(35.5)</td>
<td>(35.2)</td>
<td>(35.2)</td>
<td>(38.6)</td>
<td>(33.3)</td>
<td>(41.3)</td>
<td>(41.0)</td>
<td></td>
</tr>
</tbody>
</table>
Declaration of co-authorship

Full name of the PhD student: Marie Germund Nielsen

This declaration concerns the following article/manuscript:

<table>
<thead>
<tr>
<th>Title:</th>
<th>The construct validity of the Perceived Stress Scale</th>
</tr>
</thead>
<tbody>
<tr>
<td>Authors:</td>
<td>Nielsen MG, Ornbol E, Vestergaard M, Bech P, Larsen FB, Lasgaard M, Christensen KS,</td>
</tr>
</tbody>
</table>

The article/manuscript is: Published ☒ Accepted ☐ Submitted ☐ In preparation ☐


If accepted or submitted, state journal:

Has the article/manuscript previously been used in other PhD or doctoral dissertations?

No ☒ Yes ☐ If yes, give details:

The PhD student has contributed to the elements of this article/manuscript as follows:

<table>
<thead>
<tr>
<th>Element</th>
<th>Extent (A-E)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Formulation/identification of the scientific problem</td>
<td>D</td>
</tr>
<tr>
<td>2. Planning of the experiments and methodology design and development</td>
<td>D</td>
</tr>
<tr>
<td>3. Involvement in the experimental work/clinical studies/data collection</td>
<td>D</td>
</tr>
<tr>
<td>4. Interpretation of the results</td>
<td>D</td>
</tr>
<tr>
<td>5. Writing of the first draft of the manuscript</td>
<td>D</td>
</tr>
<tr>
<td>6. Finalization of the manuscript and submission</td>
<td>D</td>
</tr>
</tbody>
</table>

Signatures of the co-authors

<table>
<thead>
<tr>
<th>Date</th>
<th>Name</th>
<th>Signature</th>
</tr>
</thead>
<tbody>
<tr>
<td>2/1/2017</td>
<td>Kaj Sparle Christensen</td>
<td></td>
</tr>
<tr>
<td>2/2/2011</td>
<td>Mogens Vestergaard</td>
<td></td>
</tr>
<tr>
<td>2/1/2017</td>
<td>Eva Ornbol</td>
<td></td>
</tr>
<tr>
<td>Date</td>
<td>Name</td>
<td></td>
</tr>
<tr>
<td>------</td>
<td>-----------------------------</td>
<td></td>
</tr>
<tr>
<td>3q's 17</td>
<td>Per Bech</td>
<td></td>
</tr>
<tr>
<td>4/4/17</td>
<td>Mathias Lasgaard</td>
<td></td>
</tr>
<tr>
<td>5/4/17</td>
<td>Finn Breinholt Larsen</td>
<td></td>
</tr>
</tbody>
</table>

In case of further co-authors please attach appendix

Date: 5/4/17

Signature of the PhD student: **Ullem. Gerwynd. Nielsen**
Declaration of co-authorship

Full name of the PhD student: Marie Germund Nielsen

This declaration concerns the following article/manuscript:

| Title: | The construct validity of the Major Depression Inventory: a Rasch analysis of a self-rating scale in primary care |
| Authors: | Nielsen MG, Ørnbøl E, Vestergaard M, Bech P, Christensen KS |

The article/manuscript is: Published □ Accepted ☒ Submitted □ In preparation □

If published, state full reference:

If accepted or submitted, state journal: Journal of Psychosomatic Research

Has the article/manuscript previously been used in other PhD or doctoral dissertations?

No ☒ Yes □ If yes, give details:

The PhD student has contributed to the elements of this article/manuscript as follows:

A. No or little contribution
B. Has contributed (10-30 %)
C. Has contributed considerably (40-60 %)
D. Has done most of the work (70-90 %)
E. Has essentially done all the work

<table>
<thead>
<tr>
<th>Element</th>
<th>Extent (A-E)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Formulation/identification of the scientific problem</td>
<td>D</td>
</tr>
<tr>
<td>2. Planning of the experiments and methodology design and development</td>
<td>D</td>
</tr>
<tr>
<td>3. Involvement in the experimental work/clinical studies/data collection</td>
<td>D</td>
</tr>
<tr>
<td>4. Interpretation of the results</td>
<td>D</td>
</tr>
<tr>
<td>5. Writing of the first draft of the manuscript</td>
<td>D</td>
</tr>
<tr>
<td>6. Finalization of the manuscript and submission</td>
<td>D</td>
</tr>
</tbody>
</table>

Signatures of the co-authors

<table>
<thead>
<tr>
<th>Date</th>
<th>Name</th>
<th>Signature</th>
</tr>
</thead>
<tbody>
<tr>
<td>3/5 2017</td>
<td>Kaj Sparle Christensen</td>
<td>[Signature]</td>
</tr>
<tr>
<td>2/5 2017</td>
<td>Ørnbøl</td>
<td>[Signature]</td>
</tr>
<tr>
<td>2/5 2017</td>
<td>Mogens Vestergaard</td>
<td>[Signature]</td>
</tr>
<tr>
<td>3/5 2017</td>
<td>Per Bech</td>
<td>[Signature]</td>
</tr>
</tbody>
</table>

Page 1 of 2
In case of further co-authors please attach appendix.

Date:

5/4/17

Signare of the PhD student
Declaration of co-authorship

Full name of the PhD student: Marie Germund Nielsen

This declaration concerns the following article/manuscript:

| Title: | The criterion validity of the web-based Major Depression Inventory on clinical suspicion of depression in primary care |
| Authors: | Nielsen MG, Ørnbol E, Bech P, Vestergaard M, Christensen KS. |

The article/manuscript is: Published □ Accepted ☒ Submitted □ In preparation □

If published, state full reference:

If accepted or submitted, state journal: Clinical Epidemiology

Has the article/manuscript previously been used in other PhD or doctoral dissertations?

No ☒ Yes □ If yes, give details:

The PhD student has contributed to the elements of this article/manuscript as follows:

A. No or little contribution
B. Has contributed (10-30 %)
C. Has contributed considerably (40-60 %)
D. Has done most of the work (70-90 %)
E. Has essentially done all the work

<table>
<thead>
<tr>
<th>Element</th>
<th>Extent (A-E)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Formulation/identification of the scientific problem</td>
<td>D</td>
</tr>
<tr>
<td>2. Planning of the experiments and methodology design and development</td>
<td>D</td>
</tr>
<tr>
<td>3. Involvement in the experimental work/clinical studies/data collection</td>
<td>D</td>
</tr>
<tr>
<td>4. Interpretation of the results</td>
<td>D</td>
</tr>
<tr>
<td>5. Writing of the first draft of the manuscript</td>
<td>D</td>
</tr>
<tr>
<td>6. Finalization of the manuscript and submission</td>
<td>D</td>
</tr>
</tbody>
</table>

Signatures of the co-authors

<table>
<thead>
<tr>
<th>Date</th>
<th>Name</th>
<th>Signature</th>
</tr>
</thead>
<tbody>
<tr>
<td>21/3/2017</td>
<td>Kaj Sparle Christensen</td>
<td>[Signature]</td>
</tr>
<tr>
<td>21/5/2017</td>
<td>Eva Ørnbol</td>
<td>[Signature]</td>
</tr>
<tr>
<td>25/5/2017</td>
<td>Mogens Vestergaard</td>
<td>[Signature]</td>
</tr>
<tr>
<td>30/6/2017</td>
<td>Per Bech</td>
<td>[Signature]</td>
</tr>
</tbody>
</table>
In case of further co-authors please attach appendix

Date:

5/4/17  Marie Germund Nielsen

Signature of the PhD student