Psychometric analysis of the Patient Health Questionnaire in Danish patients with an implantable cardioverter defibrillator (The DEFIB-WOMEN study)

Pedersen, Susanne S; Mathiasen, Kim; Christensen, Karl Bang; Makransky, Guido

Published in:
Journal of Psychosomatic Research

DOI:
10.1016/j.jpsychores.2016.09.010

Publication date:
2016

Document version:
Accepted manuscript

Psychometric Analysis of the Patient Health Questionnaire in Danish Patients with an Implantable Cardioverter Defibrillator

Data from the DEFIB-WOMEN study

Susanne S Pedersen (PhD)[1,2,3], Kim Mathiasen (MSc)[1,4], Karl Bang Christensen (PhD)[5], Guido Makransky (PhD)[1]

[1]Department of Psychology, University of Southern Denmark, Odense, Denmark
[2]Department of Cardiology, Odense University Hospital, Odense, Denmark
[3]Department of Cardiology, Erasmus Medical Center, Rotterdam, the Netherlands
[4]Mental Health Services, Centre for Telepsychiatry, Odense, Denmark
[5]Section of Biostatistics, University of Copenhagen, Copenhagen, Denmark

Word count (text only): 3622

Short running head: PHQ-9 in Danish ICD patients

Total number of tables and figures: 6

Disclosures: None of the authors have any conflicts of interest to declare related to this manuscript.

***Corresponding author:** Susanne S Pedersen (PhD), Department of Psychology, University of Southern Denmark, Campusvej 55, DK-5230 Odense M, Denmark. *Phone:* +45 65 50 79 92; *Fax (none); E-mail:* sspedersen@health.sdu.dk

Abstract

Objective: To assess the psychometric properties of the Patient Health Questionnaire (PHQ-9), a measure of depressive symptoms, in a large Danish national cohort of patients with heart disease, implanted with an implantable cardioverter defibrillator (ICD), using item response theory.

Methods: A prospective cohort of patients implanted with an ICD (n=1531; 80.4% men) completed the PHQ-9 at the time of implant. Data were analyzed using two item response theory models, the partial credit model and the generalized partial credit model.

Results: The analysis showed disordered response thresholds in eight of nine items for the partial credit model and five of nine items for the generalized partial credit model, indicating that respondents have difficulty discriminating between response options. When collapsing response options 2 and 3, the rescored PHQ-9 had a better fit to both models. The unidimensionality and the precision of the rescored PHQ-9 were confirmed. Items did not have any differential functioning (DIF) across educational level, age, indication for ICD implantation, and severity of heart failure that influence depression outcomes in patients with an ICD. One item exhibited DIF by gender. Three items did not fit the partial credit model, but the generalized partial credit model could be fitted to the full item set.

Conclusion: The unidimensionality and reliability of the Danish version of the PHQ-9 were confirmed. However, the associated consequences of the number of response options (3-point versus 4-point Likert scale) need to be further examined for the PHQ-9 both as a screening tool and outcome measure.

Keywords: Depressive symptoms; implantable cardioverter defibrillator; patient health questionnaire (PHQ-9); item response theory (IRT).

Abstract (word count): 248

Introduction

Ischemic heart disease and depression are listed as first and second among the projected top 10 diseases in 2020 contributing to the largest disease burden worldwide [1]. However, the two diseases also interact to exacerbate risk of mortality, and depression is a common co-morbid disorder in patients with heart disease, with prevalence rates of around 20% [2]. Depression also increases the risk for mortality by 2-fold, even when adjusting for traditional biomedical risk factors, irrespective of whether patients have ischemic heart disease [3], heart failure [4], atrial fibrillation [5], peripheral arterial disease [6], or ventricular arrhythmias that are treated with implantable cardioverter defibrillator (ICD) therapy [7].

Given the substantial gains to be made from identifying and treating patients with co-morbid heart disease and depression, the last decade has witnessed increased focus on depression in the cardiovascular community. In 2012, the European Society of Cardiology included psychosocial risk factors, such as depression, in the guidelines on cardiovascular disease prevention in clinical practice [8]. In 2014, the American Heart Association published a scientific statement advocating that depression be raised to risk factor status [9].

To facilitate the identification of patients at risk of depression and thus patients at high-risk for poorer health outcomes, it is important to have adequate measures available. Hence, the Patient Health Questionnaire, a measure of depressive symptoms, has been the subject of considerable investigation. Generally, the PHQ-9 has been shown to have good psychometric properties in both general and medical populations [10-12], including patients with heart disease [13]. However, these evaluations have predominantly used Classical Test Theory (CTT) rather than Item Response Theory (IRT). IRT, which has been denominated by many researchers as the “measurement paradigm of the 21st century” (e.g. [14, 15]) has become increasingly popular, as it offers several advantages over CTT. These include a richer description of the performance of each item, greater detail on a measure’s precision, and when the assumptions are met, scores are independent of the items and invariant across different samples [16-20].

To our knowledge, 13 studies have previously assessed the psychometric properties of the PHQ-9 using IRT. None have targeted a Danish population and none specifically ICD patients. In somatic settings, studies focused on visually impaired populations in Australia [21] and Southern India [22, 23], pregnant women in Peru [24], in Côte d'Ivoire and Ghana [25], and patients with spinal cord injury [26], and traumatic brain injury in the USA [27]. Only one study was found in patients with heart failure [28], and it only assessed the possibility of using the PHQ-9 as a computer-adaptive test. One other study assessed the psychometric properties of the PHQ-9 on patients waiting for coronary bypass graft surgery [29]. The remaining studies were conducted in different populations of either specific demographic groups (university students in Japan [30], elderly in Germany [31] or in psychiatric or neurological settings (outpatient clinic in Sweden [32] and neurological setting in the USA [33]). The findings generally indicate that the PHQ-9 fits well, is unidimensional, has no local dependence, and satisfactory person separation reliability. Some studies have, however, found problems with the response categories suggesting that the two middle categories ('several days' and 'more than half the days') should be collapsed into one category (e.g. [21, 26]) while others did not (e.g. [22, 25]).

Also, one study found differential item functioning (DIF) for one item across levels of the variable duration of visual impairment [22], while other studies did not detect any DIF across a number of demographic variables including age and gender. The study in patients with heart failure did not address the potential problem with the middle response categories or DIF but focused primarily on diagnostic accuracy of different screening methods for depression [28]. No substantial DIF was found in the study on coronary artery bypass patients [29].

Overall, there is a need to further investigate the psychometric properties of the PHQ-9 in populations with heart disease to assess if the category problems identified in other populations are also present in these patient groups, and whether measurement invariance in the form of DIF exists across different groups of respondents stratified by gender, age,

education, as well as indication for ICD (primary versus secondary) and severity of heart failure. Hence, using a large national sample of patients with heart disease implanted with an ICD, we examined the psychometric properties of the Danish version of the PHQ-9 using IRT models.

In the current study, we used two IRT models; the partial credit model (PCM) [17-19] and the generalized partial credit model (GPCM) [34] to investigate the psychometric properties of the PHQ-9. Both models describe the relationship between an underlying trait and the probability of a specific item response. This relationship places the individual's level of the underlying trait and the item location on the same metric. Observed data are tested against the assumptions of the model, and if met, the score of a scale can be said to reflect the severity of the underlying trait. In the PCM, specifically the sum of item responses is sufficient for the latent trait and the model is an extension of the Rasch model [18] to items with more than two response options (polytomous items). The GPCM allows discrimination parameters to vary across items and is a more general model where sufficiency does not hold. That is, the sum score is not necessarily appropriate within the GPCM. It is thus important to test if the data fit the PCM in applied settings, where the use of raw total scores is common as support for diagnostic assignment and clinical decision-making.

Methods

Study design and population

DEFIB-WOMEN is a national, multi-center, prospective, observational study that included patients implanted consecutively with a first-time ICD or cardiac resynchronization therapy device with defibrillator (CRT-D). A total of 1531 (80.4% men) patients were included in this study. The data was collected between June 2010 and April 2013 from all implanting centers in Denmark (Odense University Hospital, Aarhus University Hospital, Aalborg University Hospital, Copenhagen University Hospital (Rigshospitalet), and Gentofte University Hospital), of which there were 5 at the time. Patients were eligible for study inclusion, if they were implanted with a first-time ICD or CRT-D, were above 18 years of age, spoke and

understood Danish, provided written informed consent, and completed all of the PHQ-9 items. Patients with a history of severe psychiatric illness (e.g. schizophrenia), on the waiting list for heart transplantation, with a left ventricular assist device, or with insufficient knowledge of the Danish language were excluded.

Study procedure

The study protocol was submitted to the Regional Committees on Health Research Ethics for Southern Denmark, who indicated that no written consent was required by Danish law. The protocol was also submitted and approved by the Danish Data Protection Agency, in order to be able to access information from the Danish registers. The study was conducted according to the Helsinki Declaration and every patient received both oral and written information about the study. An ICD nurse at the participating hospitals approached patients for study participation one day post implant and prior to discharge. Patients signed an informed consent form in hospital and were asked to complete and return a package of standardized and validated questionnaires that included the PHQ-9 within 7 days post discharge. An overview of the package of the questionnaires that patients completed is provided elsewhere [35]. If patients did not respond within the designated timeframe, they received a written reminder together with a new questionnaire and a self-addressed, stamped envelope.

Measures

Information on patients' demographic and clinical characteristics was either captured from purpose-designed questions in the questionnaire package or from the patients' medical records.

Patients completed the Patient Health Questionnaire-9 (PHQ-9) at baseline, which is the self-administered version of the PRIME-MD diagnostic instrument for common mental disorders and is based on the diagnostic criteria of the DSM IV [11]. All PHQ-9 items are rated on a four-point Likert scale: 0 (not at all), 1 (several days), 2 (more than half the days), and 3 (nearly every day).

Statistical analysis

Descriptive statistics were calculated using IBM SPSS software (version 23.0.0; SPSS Inc., Chicago, IL).

The psychometric properties of the PHQ-9 were assessed by investigating the fit of the scale to the PCM and to the GPCM [17, 36-38]. Analyses were run using RUMM2030 [36], Mplus [37, 38], R [39], and SAS [40, 41]. Among the evaluation criteria applied were ordered response thresholds, unidimensionality, item fit, local independence, measurement invariance (DIF), and the standard error of measurement as evaluated by test information functions. These methods have been described in detail elsewhere [19, 42, 43] and are only described briefly below.

The ordered response thresholds refer to the degree to which the responses provided by examinees are functioning as intended by the item developer [42, 44, 45]. This is assessed by means of examining whether the category structure of the 4-point Likert scale is suitable in this sample. An ordered set of response thresholds for each item is expected when responses to the items are consistent with the metric estimate of the underlying construct [19]. Disordered thresholds occur when respondents have difficulty discriminating between the response options. This means that there is interchangeability of categories and a category that is expected to be “harder” than an adjacent category is actually “easier”.

Unidimensionality is the assumption that the items in the scale measure only one underlying trait, and is examined with confirmatory factor analysis (CFA) using polychoric correlations [46] in Mplus. Reported goodness-of-fit indices include the comparative fit index (CFI), the Tucker-Lewis index (TLI), and the root mean square of approximation (RMSEA) [46]. The estimation method used is Muthen’s three-step procedure [47].

Item fit is the assessment of whether items fit the IRT model. A commonly used method for assessing item fit in the PCM is the Chi-square statistic that compares the difference between observed values and expected values for groups representing different severity levels across the latent trait (depression). Residuals in the range of ± 2.5 indicate a

good fit. For the GPCM, fit is evaluated graphically. The estimated item parameters of the PCM and the GPCM are also reported as an additional method to judge the quality of the individual PHQ-9 items.

Local dependence (LD) between items occurs when items are redundant or dependent, such that the response on one item may influence the probability of the response to another.

The residual correlation matrix can be used to examine LD. Consistent with other literature, residuals of more than 0.2 are labeled as being locally dependent in this study (e.g., [48]).

Item invariance requires that item estimation is independent of the subgroups of individuals completing the measure [49]. Items not demonstrating invariance are commonly referred to as exhibiting DIF [48]. For example, DIF occurs when different subgroups within the sample (e.g., men versus women) have different scores on a specific item, despite equal levels of the latent trait (depression). Items with significant Chi-square statistics at the .05 level (two-sided and with a Bonferroni correction applied separately within each DIF-variable) are reported as exhibiting DIF across the variables of gender (males/females), educational level (unskilled, skilled, higher education $\leq$ 4 years, higher >4 years, other), age (50 and below, 51 to 60, 61 to 70, 71 to 80, above 80), indication for ICD implantation (primary versus secondary), and heart failure severity (non-symptomatic versus symptomatic).

Results

Description of the cohort

A description of the participants' demographic and clinical characteristics is provided in **Table 1**.

Psychometric properties of the PHQ-9

The initial results indicated that the PHQ-9 had suboptimal fit to both IRT models (PCM and GPCM) in its current format for the given sample. An investigation into the primary cause of the problem indicated that that the four category response options did not function well, as disordered response thresholds occurred in eight of nine items for the PCM and five of nine items for the GPCM (results not shown). Disordered thresholds are an indication that the responses provided by the participants are not functioning as intended. They can occur when there are too many response options, or when the labeling of the options is too similar or ambiguous [49, 50]. The problem for all of these items was that respondents seemed to have difficulty discriminating between the two middle categories 1 *"several days"* and 2 *"more than half the days"*. The left panel of **Figure 1** provides an illustration of this problem for item 1 **("Little interest or pleasure in doing things"**), indicating that respondents did not consistently choose these categories. That is, the difficulty of a higher threshold was lower than that of its adjacent lower threshold, showing that category 2 ("*several days"*) was never the most likely response and was thus redundant. Similar results were found for the other items with the exception of item 5 which did not have reversed thresholds, but also had highly overlapping response functions for response options 1 and 2. Therefore, we used a three-category response format, combining the two middle categories, for all nine items in the remaining analyses. This is referred to as the rescored PHQ-9 in the remainder of the paper. The bottom panel of **Figure 1** provides an illustration of the category probability curves for item 1 as an example in the rescored PHQ-9.

Table 2 reports the item parameters for the rescored PHQ-9 items in order of difficulty with the items that are easiest to endorse appearing first. The discrimination parameter describes the relationship between the item and the latent trait. For items with higher discrimination, the probability of responding in a higher category increases more for small changes in the latent trait of depression.

Construct validity of the rescored PHQ-9

The rescored PHQ-9 showed sub-optimal fit to the PCM in the given sample. A CFA for the one factor model with all 9 items in the rescored PHQ-9 yielded acceptable fit indices (CFI = 0.973; TLI = 0.964; RMSEA = 0.075), thus supporting combining the nine items to create a unidimensional scale. Furthermore, there was no evidence of local dependence among items. The test of item fit indicated that none of the items had positive fit residuals above 2.5; additionally, items 2, 6, and 9 had negative fit residuals indicating that the items discriminated better than predicted by the PCM. Plots of observed and expected item mean scores indicated acceptable fit to the GPCM. This is illustrated in **Figure 2** for the item with the worst fit to the PCM (item 6).

Figure 3 shows the test information functions, with the PCM in the left panel, and the GPCM in the right panel. It is clear from these graphs that the rescored PHQ-9 has the highest test information, and therefore the highest measurement precision for people with a high level of depression.

Item invariance in the form of DIF was assessed for gender (men/women), education (5 categories), age (5 categories), indication for ICD implantation (primary/secondary prevention), and severity of heart failure (symptomatic versus non-symptomatic) for all items. Significant uniform DIF was only found for item 2 **("Feeling down, depressed, or hopeless"**) across gender, while no DIF was found across the other four variables. **Figure 4** illustrates the DIF for item 2, where it is clear that women are more likely to endorse this item compared to men even though they have the same level of the latent trait of depression. When DIF is identified it can be accounted for by splitting the item into two items with gender specific item parameters (e.g. [50, 51]).

Discussion

To our knowledge, the psychometric properties of the PHQ-9 have not yet been assessed in a Danish setting, and not all studies have assessed the psychometric properties of the PHQ-9 using IRT. In the current study, we investigated the fit of the Danish version of the PHQ-9

in a large Danish cohort of patients with heart disease that were implanted with an ICD. The results of the study indicated that there were disordered thresholds in eight of the nine items for the PCM and in five items for the GPCM, suggesting that the four-point response format did not function appropriately for this sample. Disordered thresholds can occur when there are too many response options, or when the labeling of the options is too similar or ambiguous. In this cohort, the respondents seemed to have difficulty discriminating between response categories 1 *"several days"* and 2 *"more than half the days".* This finding is consistent with the results of previous studies in non-cardiac populations investigating the fit of the PHQ-9 with the PCM [21, 24, 26].

This study adds to the evidence that the four-category response format does not function appropriately for the PHQ-9, favoring a rescored 3-category response format. The practical implications of ignoring this in the PHQ-9 can be pronounced, in particular when the PHQ-9 is used as a measure of change. The results of the current study and previous studies [21, 26, 52] show that the validity of the PHQ-9 is compromised when the total sum score based on the four response options is used, which could limit the validity of conclusions drawn from studies that evaluate the clinical efficacy of a given intervention.

The rescored PHQ-9 had a better fit to the IRT models after combining the two middle response categories (i.e., categories 1 and 2). The CFA analysis showed that all items measured a unidimensional trait. This finding supports previous research, showing that the PHQ-9 measures a single unidimensional trait [22, 52] and contradicts previous studies that have not been able to fully support such unidimensionality [31]. Our results also showed that there was no local dependence (LD) across the items in the scale.

The test of item fit indicated that none of the items had positive fit residuals. However, items 2 ("**Feeling down, depressed, or hopeless"**), 6 **("Feeling bad about yourself - or that you are a failure or have let yourself or your family down"**), and 9 **("Thoughts that you would be better off dead or of hurting yourself in some way"**) had negative residuals, indicating that the items discriminated better than expected by the PCM. This was also evident in the graphical illustration of item fit for the PCM (**Figure 2**, left panel). This is

consistent with the results from the GPCM that showed these items to have the highest discrimination parameters, indicating that these items discriminated highest between respondents with low and high levels of the latent trait of depression. Notably, in the graphical illustration of item fit the GPCM showed better agreement between the observed and expected item mean scores (**Figure 2**, right panel).

Given that variables, such as educational level, age, indication for ICD implantation (primary versus secondary) and severity of heart failure, have been shown to influence depression outcomes in patients with an ICD [50, 53], we examined whether the different items of the PHQ-9 could be biased by these variables and thus potentially the total PHQ-9 score and prevalence of depressive symptoms. However, we found no support for such a bias in this cohort across educational level, age, indication for ICD implantation, and severity of heart failure. Hence, given this measurement invariance, the rescored PHQ-9 can confidently be used to assess depression in patients with an ICD without such risk of bias. However, the findings that some items did not fit the PCM do not support the usage of the sum score. Furthermore, we did find a significant uniform DIF for item 2 (**"Feeling down, depressed, or hopeless"**) across gender. The practical implications of this DIF were investigated by correcting for the DIF in this item by splitting the item into two items with gender specific item parameters as described by others [50, 51]. The results showed that the difference in person location estimates were practically non-existent, indicating that the impact of DIF in this item is not substantial. Nevertheless, future research should investigate whether this item has DIF in other populations, as well as whether the DIF can influence the depression score when comparisons are made across gender groups with the PHQ-9.

The results of our study also showed that the targeting of the rescored PHQ-9 was not optimal for all respondents due to the fact that the test information was largest for values of the latent variable corresponding to high levels of depression. Items in the scale assess depression for the most depressed individuals but the targeting is worse for respondents who do not show a lot of symptoms. This is not a problem in most settings, as the PHQ-9 is used to screen for depression, and not as a tool to discriminate between individuals who do not

show a lot of depressive symptoms. However, this finding suggests that the rescored PHQ-9 may not be able to reliably identify sub-threshold depression, and consequently may not be useful in studies that focus on sub-threshold depression.

The results of this study should be interpreted with the following limitations in mind. We focused on a cohort of patients with heart disease implanted with an ICD. Thus, we do not know whether our results are generalizable to other populations with chronic disease nor to the healthy population. This study also has several strengths. To our knowledge, it is the first study to examine the psychometric properties of the Danish version of the PHQ-9. Second, our patient group was derived from a national cohort, recruited from all implanting ICD centers in Denmark. Third, we used IRT, which offers several advantages over CTT.

In conclusion, in a large Danish national cohort of patients implanted with an ICD, we found disordered thresholds in eight of the nine items, suggesting that the scale might contain too many response options or use labeling of the options that are too similar. Our results confirmed the unidimensionality of the PHQ-9 using IRT and CFA. We found no support for items being biased by educational level, age, indication for ICD implantation, and severity of heart failure that are known to influence depression outcomes in patients with an ICD.

Acknowledgement

The DEFIB-WOMEN study was supported with grant no. 09-10-R75-A2713-22565 from the Danish Heart Foundation.

References

[1] C.M. Michaud, C.J. Murray, B.R. Bloom, Burden of disease--implications for future research, JAMA 285(5) (2001) 535-9.
[2] G. Magyar-Russell, B.D. Thombs, J.X. Cai, T. Baveja, E.A. Kuhl, P.P. Singh, M. Montenegro Braga Barroso, E. Arthurs, M. Roseman, N. Amin, J.E. Marine, R.C. Ziegelstein, The prevalence of anxiety and depression in adults with implantable cardioverter defibrillators: a systematic review, J Psychosom Res 71(4) (2011) 223-31.
[3] M. Zuidersma, H.J. Conradi, J.P. van Melle, J. Ormel, P. de Jonge, Self-reported depressive symptoms, diagnosed clinical depression and cardiac morbidity and mortality after myocardial infarction, Int J Cardiol 167(6) (2013) 2775-80.
[4] G.C. Reeves, A.S. Alhurani, S.K. Frazier, J.F. Watkins, T.A. Lennie, D.K. Moser, The association of comorbid diabetes mellitus and symptoms of depression with all-cause mortality and cardiac rehospitalization in patients with heart failure, BMJ Open Diabetes Res Care 3(1) (2015) e000077.
[5] N. Frasure-Smith, F. Lesperance, M. Habra, M. Talajic, P. Khairy, P. Dorian, D. Roy, F. Atrial, I. Congestive Heart Failure, Elevated depression symptoms predict long-term cardiovascular mortality in patients with atrial fibrillation and heart failure, Circulation 120(2) (2009) 134-40, 3p following 140.
[6] G. Cherr, P. Zimmerman, J. Wang, H. Dosluoglu, Patients with Depression are at Increased Risk for Secondary Cardiovascular Events after Lower Extremity Revascularization, Journal of General Internal Medicine 23(5) (2008) 629.
[7] M.H. Mastenbroek, H. Versteeg, L. Jordaens, D.A. Theuns, S.S. Pedersen, Ventricular tachyarrhythmias and mortality in patients with an implantable cardioverter defibrillator: impact of depression in the MIDAS cohort, Psychosom Med 76(1) (2014) 58-65.
[8] J. Perk, G. De Backer, H. Gohlke, I. Graham, Z. Reiner, M. Verschuren, C. Albus, P. Benlian, G. Boysen, R. Cifkova, C. Deaton, S. Ebrahim, M. Fisher, G. Germano, R. Hobbs, A. Hoes, S. Karadeniz, A. Mezzani, E. Prescott, L. Ryden, M. Scherer, M. Syvanne, W.J. Scholte op Reimer, C. Vrints, D. Wood, J.L. Zamorano, F. Zannad, P. European Association for Cardiovascular, Rehabilitation, E.S.C.C.f.P. Guidelines, European Guidelines on cardiovascular disease prevention in clinical practice (version 2012). The Fifth Joint Task Force of the European Society of Cardiology and Other Societies on Cardiovascular Disease Prevention in Clinical Practice (constituted by representatives of nine societies and by invited experts), Eur Heart J 33(13) (2012) 1635-701.
[9] J.H. Lichtman, E.S. Froelicher, J.A. Blumenthal, R.M. Carney, L.V. Doering, N. Frasure-Smith, K.E. Freedland, A.S. Jaffe, E.C. Leifheit-Limson, D.S. Sheps, V. Vaccarino, L. Wulsin, E. American Heart Association Statistics Committee of the Council on, Prevention, C. the Council on, N. Stroke, Depression as a risk factor for poor prognosis among patients with acute coronary syndrome: systematic review and recommendations: a scientific statement from the American Heart Association, Circulation 129(12) (2014) 1350-69.
[10] A. Martin, W. Rief, A. Klaiberg, E. Braehler, Validity of the Brief Patient Health Questionnaire Mood Scale (PHQ-9) in the general population, General hospital psychiatry 28(1) (2006) 71-7.
[11] K. Kroenke, R.L. Spitzer, J.B.W. Williams, The PHQ-9: Validity of a brief depression severity measure, Journal of General Internal Medicine 16(9) (2001) 606-613.
[12] S. Gilbody, D. Richards, S. Brealey, C. Hewitt, Screening for depression in medical settings with the Patient Health Questionnaire (PHQ): a diagnostic meta-analysis, J Gen Intern Med 22(11) (2007) 1596-602.
[13] M.H. Hammash, L.A. Hall, T.A. Lennie, S. Heo, M.L. Chung, K.S. Lee, D.K. Moser, Psychometrics of the PHQ-9 as a measure of depressive symptoms in patients with heart failure, Eur J Cardiovasc Nurs 12(5) (2013) 446-53.
[14] R.D. Hays, L.S. Morales, S.P. Reise, Item response theory and health outcomes measurement in the 21st century, Med Care 38(9 Suppl) (2000) II28-42.
[15] J.E. Ware, Jr., Conceptualization and measurement of health-related quality of life: comments on an evolving field, Arch Phys Med Rehabil 84(4 Suppl 2) (2003) S43-51.

[16] T.H. Nguyen, H.-R. Han, M.T. Kim, K.S. Chan, An introduction to item response theory for patient-reported outcome measurement, The Patient: Patient-Centered Outcomes Research 7(1) (2014) 23-35.
[17] G.N. Masters, A Rasch model for partial credit scoring, Psychometrica 47 (1982) 149-174.
[18] G. Rasch, Probabilistic models for some intelligence and attainment tests, Copenhagen, 1960.
[19] A. Tennant, P.G. Conaghan, The Rasch measurement model in rheumatology: what is it and why use it? When should it be applied, and what should one look for in a Rasch paper?, Arthritis Rheum 57(8) (2007) 1358-62.
[20] A. O'Neil, B. Taylor, D.L. Hare, K. Sanderson, S. Cyril, K. Venugopal, B. Chan, J.J. Atherton, A. Hawkes, D.L. Walters, B. Oldenburg, T. MoodCare Investigator, Long-term efficacy of a tele-health intervention for acute coronary syndrome patients with depression: 12-month results of the MoodCare randomized controlled trial, Eur J Prev Cardiol 22(9) (2015) 1111-20.
[21] E.L. Lamoureux, H.W. Tee, K. Pesudovs, J.F. Pallant, J.E. Keeffe, G. Rees, Can clinicians use the PHQ-9 to assess depression in people with vision loss?, Optometry and vision science : official publication of the American Academy of Optometry 86(2) (2009) 139-45.
[22] V.K. Gothwal, D.K. Bagga, R. Sumalini, Rasch validation of the PHQ-9 in people with visual impairment in South India, Journal of Affective Disorders 167 (2014) 171-177.
[23] V.K. Gothwal, D.K. Bagga, S. Bharani, R. Sumalini, S.P. Reddy, The patient health questionnaire-9: validation among patients with glaucoma, PLoS One 9(7) (2014) e101295.
[24] Q.Y. Zhong, B. Gelaye, M.B. Rondon, S.E. Sanchez, G.E. Simon, D.C. Henderson, Y.V. Barrios, P.M. Sanchez, M.A. Williams, Using the Patient Health Questionnaire (PHQ-9) and the Edinburgh Postnatal Depression Scale (EPDS) to assess suicidal ideation among pregnant women in Lima, Peru, Arch Womens Ment Health 18(6) (2015) 783-92.
[25] D. Barthel, C. Barkmann, S. Ehrhardt, S. Schoppen, C. Bindt, C.D.S.S.G. International, Screening for depression in pregnant women from Cote dIvoire and Ghana: Psychometric properties of the Patient Health Questionnaire-9, Journal of affective disorders 187 (2015) 232-40.
[26] R.T. Williams, A.W. Heinemann, R.K. Bode, C.S. Wilson, J.R. Fann, D.G. Tate, Improving measurement properties of the Patient Health Questionnaire–9 with rating scale analysis, Rehabilitation Psychology 54(2) (2009) 198-203.
[27] J.R. Dyer, R. Williams, C.H. Bombardier, S. Vannoy, J.R. Fann, Evaluating the Psychometric Properties of 3 Depression Measures in a Sample of Persons With Traumatic Brain Injury and Major Depressive Disorder, J Head Trauma Rehabil 31(3) (2016) 225-32.
[28] H.F. Fischer, C. Klug, K. Roeper, E. Blozik, F. Edelmann, M. Eisele, S. Störk, R. Wachter, M. Scherer, M. Rose, C. Herrmann-Lingen, Screening for mental disorders in heart failure patients using computer-adaptive tests, Quality of Life Research: An International Journal of Quality of Life Aspects of Treatment, Care & Rehabilitation 23(5) (2014) 1609-1618.
[29] F. Kendel, M. Wirtz, A. Dunkel, E. Lehmkuhl, R. Hetzer, V. Regitz-Zagrosek, Screening for depression: Rasch analysis of the dimensional structure of the PHQ-9 and the HADS-D, Journal of affective disorders 122(3) (2010) 241-6.
[30] Y. Umegaki, N. Todo, Psychometric Properties of the Japanese CES-D, SDS, and PHQ-9 Depression Scales in University Students, Psychol Assess (2016).
[31] T. Forkmann, S. Gauggel, L. Spangenberg, E. Brahler, H. Glaesmer, Dimensional assessment of depressive severity in the elderly general population: psychometric evaluation of the PHQ-9 using Rasch Analysis, Journal of affective disorders 148(2-3) (2013) 323-30.
[32] M. Adler, J. Hetta, G. Isacsson, U. Brodin, An item response theory evaluation of three depression assessment instruments in a clinical sample, BMC Med Res Methodol 12 (2012) 84.
[33] E.R. Walker, G.J. Engelhard, N.J. Thompson, Rasch measurement theory to assess three depression scales among adults with epilepsy, Seizure 21(6) (2012) 437-443.

[34] E.A. Muraki, A Generalized Partial Credit Model: Application of an EM Algorithm Applied Psychological Measures 16(2) (1992) 159176.
[35] S.S. Pedersen, J.C. Nielsen, S. Riahi, J. Haarbo, R. Videbaek, M.L. Larsen, O. Skov, C. Knudsen, J.B. Johansen, Study Design and Cohort Description of DEFIB-WOMEN - a National Danish Study in Patients with an ICD, (2016).
[36] D. Andrich, B. Sheridan, G. Luo, Rasch models for measurement: RUMM2030., Perth, Australia: Rumm Laboratory, 2010.
[37] L.K. Muthen, B.O. Muthen, Mplus (Version 7), Los Angeles, California, 2012.
[38] MIRT: Multidimensional Item Response Theory, in: C.A.W. Glas (Ed.) University of Twente, the Netherlands, 2010.
[39] D. Rizopoulos, ltm: An R package for latent variable modeling and item response theory analyses, J Stat Softw 17(5) (2006).
[40] K.B. Christensen, M. Olsbjerg, Marginal maximum likelihood estimation in polytomous Rasch models using SAS, . Pub. Inst. Stat. Univ. 57 (2013) 69-84.
[41] M. Olsbjerg, K.B. Christensen, LIRT : SAS Macros for Longitudinal IRT Models Department of Biostatistics University of Copenhagen 2014.
[42] J.F. Pallant, A. Tennant, An introduction to the Rasch measurement model: An example using the Hospital Anxiety and Depression Scale (HADS), British Journal of Clinical Psychology 46 (2007) 1-18.
[43] G. Makransky, C.W. Schnohr, T. Torsheim, C. Currie, Equating the HBSC Family Affluence Scale across survey years: a method to account for item parameter drift using the Rasch model, Qual Life Res 23(10) (2014) 2899-907.
[44] S.K. Nam, E. Yang, S.M. Lee, S.H. Lee, H. Seol, A psychometric evaluation of the career decision self-efficacy scale with Korean students: A Rasch model approach, Journal of Career Development 38 (2011) 147-166.
[45] S. Messick, Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning
, American Psychologist 50 (1995) 741-749.
[46] L. Hu, P.M. Bentler, Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives., Structural Equation Modeling: A Multidisciplinary Journal 6(1) (1999) 1-55.
[47] B. Muthen, A General Structural Equation Model with Dichotomous, Ordered Categorical, and Continuous Latent Variable Indicators, Psychometrika 49(1) (1984) 115-132.
[48] G. Makransky, N. Bilenberg, Psychometric properties of the parent and teacher ADHD Rating Scale (ADHD-RS): measurement invariance across gender, age, and informant, Assessment 21(6) (2014) 694-705.
[49] T.G. Bond, C.M. Fox, Applying the Rasch model: Fundamental measurement in the human sciences. , Mahwah, NJ: Erlbaum, 2001.
[50] C. Hagquist, M. Bruce, J.P. Gustavsson, Using the Rasch model in nursing research: an introduction and illustrative example, International Journal of Nursing Studies 46(3) (2009) 380-393.
[51] G. Makransky, C.A.W. Glas, Modeling differential item functioning with group-specific item parameters: A computerized adaptive testing application., Measurement 46 (2013) 3228-3237.
[52] Q. Zhong, B. Gelaye, J.R. Fann, S.E. Sanchez, M.A. Williams, Cross-cultural validity of the Spanish version of PHQ-9 among pregnant Peruvian women: A Rasch item response theory analysis, Journal of Affective Disorders 158 (2014) 148-153.
[53] J.B. Johansen, S.S. Pedersen, H. Spindler, K. Andersen, J.C. Nielsen, P.T. Mortensen, Symptomatic heart failure is the most important clinical correlate of impaired quality of life, anxiety, and depression in implantable cardioverter-defibrillator patients: a single-centre, cross-sectional study in 610 patients, Europace 10(5) (2008) 545-51.

Table 1. Participants' characteristics (N = 1531)*

Characteristics		
Men *(n=1531)*	1231	(80.4)
Age (years) *(n=1531)*		
≤ 50	135	(8.8)
51-60	246	(16.1)
61-70	473	(30.9)
71-80	546	(35.7)
>80	131	(8.6)
Educational level *(n=1531)*		
Unskilled	289	(18.9)
Skilled	554	(36.2)
Higher education ≤4 years	302	(19.7)
Higher >4 years	174	(11.4)
Other	203	(13.3)
Arrhythmia *(n=1483)*		
Primary prevention intervention	834	(54.5)
Secondary prevention intervention	649	(42.4)
Heart failure severity *(n=1349)*		
NYHA class I-II	1026	(67.0)
NYHA class III-IV (symptomatic heart failure)	323	(21.1)

** Listed as n (%) unless otherwise indicated*

NYHA = New York Heart Association functional class

Table 2. Item hierarchy and fit statistics for the PHQ-9 before and after combining response categories 2 and 3

	PCM				GPCM		
	Discr.	Threshold 1 (s.e.)	Threshold 2 (s.e.)	Fit resid.	Discr.	Thereshold 1 (s.e.)	Threshold 2 (s.e.)
4. Feeling tired or having little energy	1.000	-1.225 (0.070)	1.795 (0.081)	-0,349	1.885	-0.845 (0.053)	1.250 (0.061)
3. Trouble falling or staying asleep, or sleeping too much	1.000	0.040 (0.066)	2.029 (0.096)	0,765	1.507	0.032 (0.050)	1.550 (0.078)
1. Little interest or pleasure in doing things	1.000	0.846 (0.068)	2.608 (0.122)	1,554	1.517	0.625 (0.058)	1.963 (0.100)
5. Poor appetite or overeating	1.000	0.899 (0.068)	2.682 (0.126)	1,121	1.302	0.735 (0.067)	2.122 (0.117)
7. Trouble concentrating on things, such as reading the newspaper or watching television	1.000	1.512 (0.074)	2.850 (0.147)	-2,203	1.739	1.019 (0.066)	2.070 (0.105)
8. Moving or speaking so slowly that other people could have noticed. Or the opposite – being fidgety or restless and moving around a lot more than usual	1.000	1.579 (0.075)	3.044 (0.158)	-1,041	1.713	1.074 (0.068)	2.201 (0.114)
2. Feeling down, depressed, or hopeless	1.000	1.024 (0.068)	3.525 (0.167)	-4,885	2.686	0.599 (0.042)	2.155 (0.092)
6. Feeling bad about yourself - or that you are a failure or have let yourself or your family down	1.000	1.543 (0.074)	3.278 (0.169)	-5,241	2.849	0.861 (0.046)	2.077 (0.088)
9. Thoughts that you would be better off dead or of hurting yourself in some way	1.000	3.188 (0.113)	3.668 (0.299)	-2,572	2.225	1.861 (0.106)	2.571 (0.161)

Figure 1. Category probability curves for item 1 (**"Little interest or pleasure in doing things"**) with response categories 0 = "*not at all"*, 1 = *"several days"*, 2 = *"more than half the days"*, and 3 = *"nearly every day"* before (upper panel) and after (lower panel) collapsing categories 1 and 2. PCM shown in left panel, GPCM shown in right panel based on a standard normal distribution.

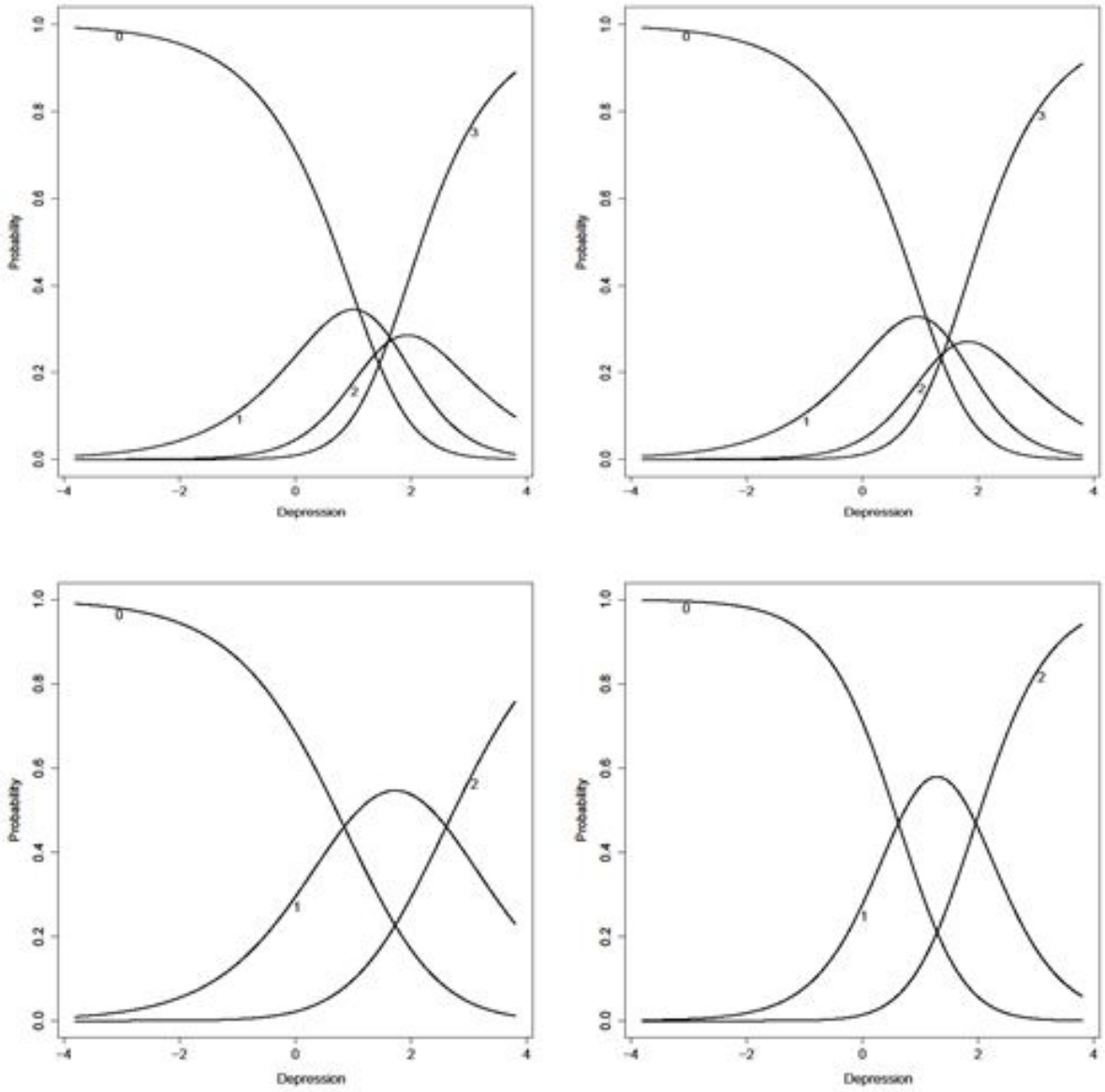

Figure 2. Observed item mean scores across score groups for item 6 ('feeling bad about yourself – or that you are a failure or have let yourself or your family down') plotted together with expected scores in the PCM (left panel) and the GPCM (right panel) based on a standard normal distribution

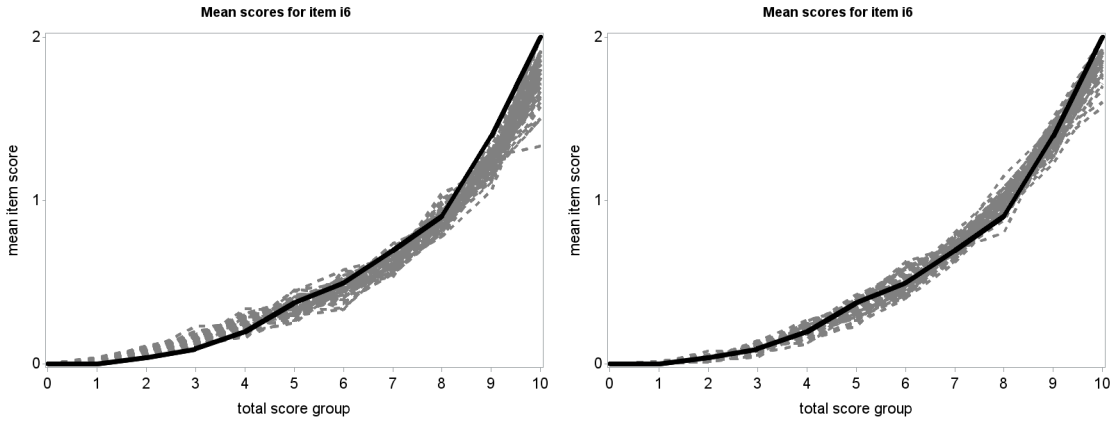

Figure 3. Test information functions for the rescored PHQ-9 in the PCM (left panel) and the GPCM (right panel) based on a standard normal distribution for the latent variable

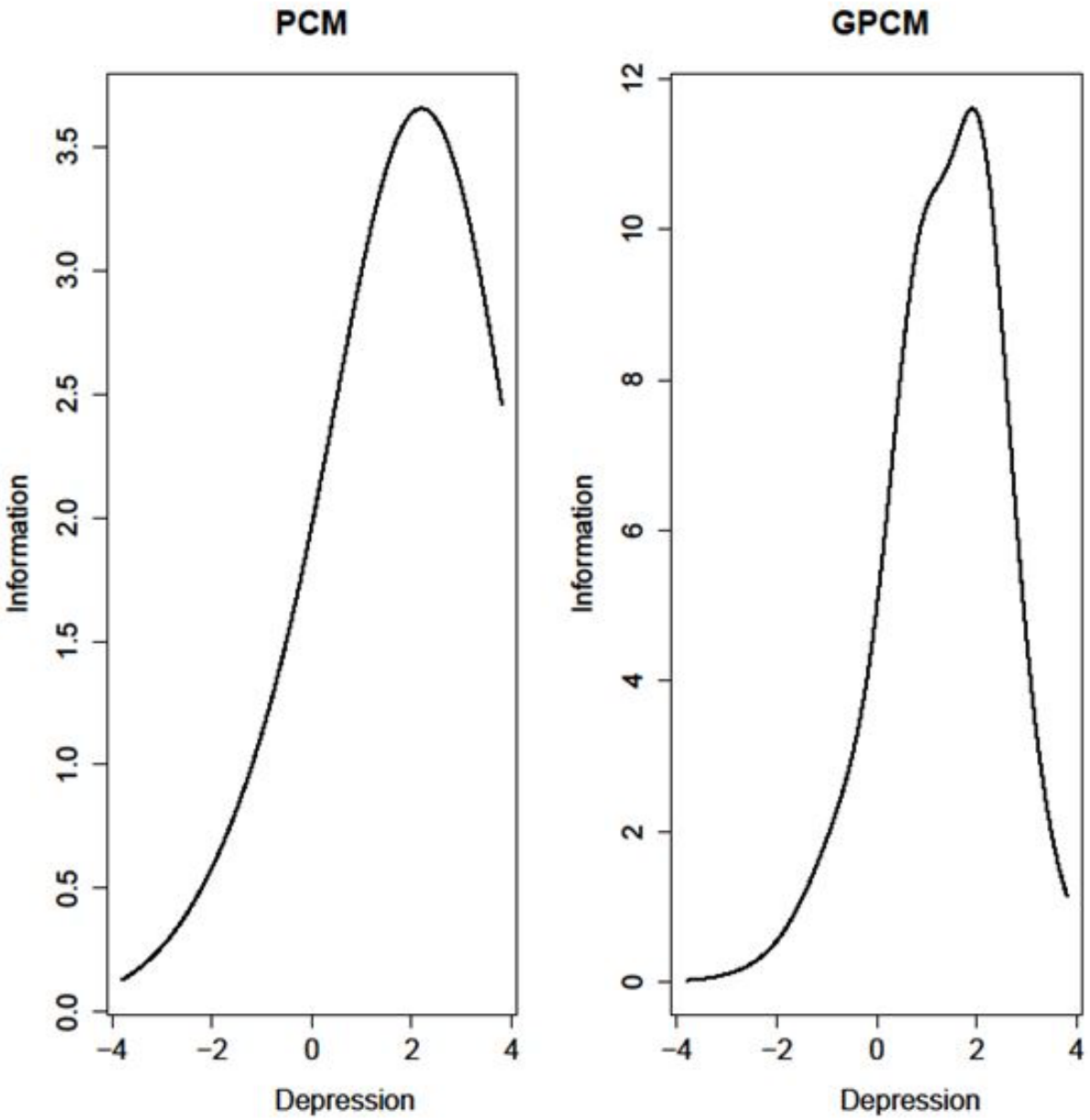

Figure 4. Item characteristic curves for item 2 **("Feeling down, depressed, or hopeless")** illustrating uniform DIF across gender based on conditional maximum likelihood (CML) estimation without an assumption of the distribution of the latent trait but with a linear restriction on the threshold parameters