Beyond a symptom count

Acevedo Mesa, Maria Angélica

DOI:
10.33612/diss.195932294

Publication date:
2021

05

Does Functional Somatic Symptoms measurement differ across Sex and Age? Differential Item Functioning in Somatic Symptoms measured with the CIDI

Angélica Acevedo-Mesa, Rei Monden, Sebastian Castro-Alvarez, Judith G. M. Rosmalen, Annelieke M. Roest, and Jorge N. Tendeiro

Assessment
2021

Citation for published version (APA):
Acevedo-Mesa, A., Monden, R., Castro-Alvarez, S., Rosmalen, J. G., Roest, A. M., & Tendeiro, J. N. (2021). Does Functional Somatic Symptoms Measurement Differ Across Sex and Age? Differential Item Functioning in Somatic Symptoms Measured With the CIDI. Assessment, 10731911211017228.

Abstract

Functional Somatic Symptoms (FSS) are physical symptoms that cannot be attributed to underlying pathology. Their severity is often measured with sum-scores on questionnaires; however, this may not adequately reflect FSS severity in subgroups of patients. We aimed to identify the items of the somatization section of the Composite International Diagnostic Interview that best discriminate FSS severity levels, and to assess their functioning in sex and age subgroups. We applied the Two Parameter Logistic Model to 19 items in a population-representative cohort of 962 participants. Subsequently, we examined Differential Item Functioning (DIF). "Localized (muscle) weakness" was the most discriminative item of FSS severity. "Abdominal pain" consistently showed DIF by sex, with males reporting it at higher FSS severity. There was no consistent DIF by age, however, "joint pain" showed poor discrimination of FSS severity in older adults. These findings could be helpful for the development of better assessment instruments for FSS, which can improve both future research and clinical care.

Keywords: Functional Somatic Symptoms, Item response Theory, Differential Item Functioning, Two Parameter Logistic Model, Composite International Diagnostic Interview, sex, age.

Introduction

Functional Somatic Symptoms (FSS) are physical symptoms that cannot be attributed to underlying pathology. Although they are recognized as prevalent in primary care and debilitating when becoming persistent (van Eck van der Sluijs, Jonna, et al., 2018), research is often hindered due to the difficulties of measuring FSS and its severity. More than 40 questionnaires have been developed to assess FSS (van Driel et al., 2017; Zijlema et al., 2013), but no consensus has been reached concerning the crucial aspects of FSS measurement, including which and how many symptoms should be assessed, or how to measure the severity of FSS. As a result, very different questionnaires are used in FSS research, which makes it difficult to compare the results between studies (Zijlema et al., 2013).

The traditional approach to measure FSS severity is based on a total sum-score of the severity of individual items, or a count of the presence of symptoms, however, this is known to be problematic. For example, even when two persons have the same sum-score, they may be suffering from different symptoms with varying severity. In the clinical setting, this hinders clinicians from grasping the severity of FSS in patients. Another issue in the measurement of FSS severity is that it is still unclear which symptoms should be included in FSS questionnaires, as highlighted in previous reviews (Barsky, Peekna, & Borus, 2001; Zijlema et al., 2013). The symptoms that have been included more frequently in FSS questionnaires are "headache", "nausea", "shortness of breath", and "(low) back pain" (Zijlema et al., 2013). However, it is not certain if these symptoms are the most useful to discriminate between different levels of FSS severity. To address these issues, in a previous study Item Response Theory (IRT) was used to determine which items of the somatization scale of the Symptom Checklist-90 (SCL-90) were the most discriminative for FSS severity (Acevedo-Mesa, Tendeiro, Roest, Rosmalen, & Monden, 2020). IRT is a very useful framework to select the most appropriate items to include in questionnaires. It explores the properties of each item by calculating its ability to discriminate between varying levels of severity of FSS (parameter α), and its severity level in a severity continuum of FSS (parameter β; Embretson & Reise, 2013). This previous study showed that "headache" was the least discriminative item, and that "heavy feelings in arms or legs" and "feeling weak in parts of your body" were the most discriminative and more informative items to measure the severity of FSS (Acevedo-Mesa et al., 2020). This finding is relevant since these two items are not frequently included in FSS scales.

However, the somatization scale of the SCL-90 assesses somatic symptoms irrespective of their cause. The fact that this and many similar instruments include both symptoms caused by a chronic disease and symptoms that are medically unexplained might influence the items' parameters and functioning. To further improve the assessment of FSS, we need insight into the item functioning of instruments focusing on medically unexplained somatic symptoms (Zijlema et al., 2013). In addition, previous IRT studies included instruments with low numbers of items and identified only very few discriminative items. FSS assessment could thus profit from studying instruments with a wider diversity of symptoms included. Therefore, the first aim of the current study is to apply IRT models to an FSS instrument that assesses a wide variety of medically unexplained symptoms, namely the somatization section of the Composite International Diagnostic Interview (CIDI). The CIDI is a particularly relevant instrument for this aim since it is also capable of distinguishing FSS from symptoms caused by chronic somatic conditions.

An additional issue about using sum-scores and symptom counts for measuring FSS is that these may not have the same meaning for different groups of persons, which may lead to bias in interpretation. More specifically, some factors can affect the probability of a person to report a specific symptom. For instance, in a construct such as anxiety, items related to fatigue seem to be reported mainly by older people, which may be due to aging processes rather than anxiety (Correa & Brown, 2019). When the probability of reporting an item is different for persons belonging to different groups, even when they possess the same level of the construct measured, *differential item functioning* (DIF) is considered to be present. Items with DIF systematically bias the results of one group compared to another. In the assessment of FSS, it seems likely that the sex and age of the individuals influence the probability to report a specific symptom. Therefore, the second aim of this study is to explore if there are indications of DIF regarding sex and age.

Overall, males and females have been found to report different symptoms, with females generally reporting more numerous, intense, burdensome, and frequent FSS, even when gynecological symptoms are excluded (Barsky et al., 2001). Previously, it has been found that headache, fatigue, and gastrointestinal symptoms, among others were more commonly related to the female sex (Kroenke & Price, 1993) . More recently, a study found that both biological sex and psychosocial gender were independently associated with symptoms of the SCL-90 somatization scale, with symptoms such as headache and feeling hot/ cold being more strongly associated with female sex than symptoms such as weakness in body parts and chest pain (Ballering, Bonvanie, Hartman, Monden, & Rosmalen, 2020). It has been proposed that differences in symptom reporting could be due to sex differences in symptom perception and pain sensitivity, a higher prevalence of depressive and anxiety disorders in females, which are also accompanied by somatic symptoms, and different types of life stressors linked with gender roles assigned to each sex (Barsky et al., 2001; Michael et al., 2005). The resulting differences in symptom reporting for males and females may imply that there is DIF in the assessment of FSS. Current FSS questionnaires do not address this issue, which may lead to bias in the measurement of FSS by sex. DIF analysis for age and sex, could enrich the current approaches for measuring FSS.

Regarding age, according to a systematic review, most studies suggest that the prevalence rates of FSS decrease in older populations (Hilderink, Collard, Rosmalen, & Voshaar, 2013). However, this might be explained by the fact that most FSS questionnaires are not validated for use in the elderly. Older patients often attribute bodily symptoms to somatic diseases, as they tend to have more somatic co-morbidity than other age groups (Creed, Henningsen, & Fink, 2011). In addition, symptoms can also represent the aging process and/or common geriatric syndromes, such as fatigue, tiredness, loss of energy and feeling weak, loss of appetite and weight loss, and insomnia, and are thus not considered as FSS (van Driel et al., 2017). Consequently, the measurement of FSS in older adults may be biased and DIF may be present in questionnaire items in different age groups. Indeed, DIF has been found in adolescents and younger adults, compared to adults, in the somatization scale of the Four-Dimensional Symptom Questionnaire (4DSQ; Terluin, van der Wouden, J C, & de Vet, H C W, 2019). Small to moderate DIF was found for the items “neck pain” and “blurred vision”. To our knowledge, DIF has not been studied for other age groups in FSS questionnaires.

In this study, we aim to analyze which items are best at discriminating between different levels of FSS severity and to assess their functioning in relation to sex and age. Firstly, we aim to identify the most discriminative symptoms of FSS severity by applying the 2-parameter logistic model (2PLM) from IRT to the somatization section of the CIDI in a population cohort. Secondly, we aim to examine DIF between males and females, as well as between adults and older adults. The results of this study can provide information about which symptoms to include in questionnaires in order to improve the measurement of FSS severity, and about which items perform differentially depending on the participant’s subgroup.

Method

Participants

The participants of this study are part of a sub-cohort of the Prevention of Renal and Vascular End Stage Disease (PREVEND), which is a population cohort study with participants from the city of Groningen, the Netherlands. The PREVEND study aims to investigate micro-albuminuria as a risk factor for renal and cardiovascular disease (Pinto-Sietsma et al., 2000). The recruitment process and the PREVEND cohort development have been described elsewhere (Pinto-Sietsma et al., 2000; Rosmalen, Tak, & de Jonge, 2011). From the 85,421 subjects contacted, 40,856 responded by sending a morning urine sample and filling out a questionnaire about demographics and cardiovascular history. From these, 7,768 participants with an elevated urinary albumin concentration (≥10 mg/l), and a randomly selected control group of 3,395 participants with a urinary albumin concentration of ≤10 mg/1, were contacted for further investigations. After these investigations, 8,592 participants formed the cohort for the current study.

The sub-cohort data employed for the present study consists of a group of albuminuria-negative participants combined with a random sample of albuminuria-positive participants, emulating the population-representative ratio of albuminuria-positive cases. Initially, 2554 participants from the PREVEND study were contacted to form the sub-cohort, for which they were requested to provide additional psychiatric and psychosocial data. Finally, 1094 (42.8%) participants completed the additional information, forming the sub-cohort used for this study (Rosmalen et al., 2011). Of the 1094 participants of the sub-cohort, 132 did not complete the somatization section of the Composite International Diagnostic Interview (CIDI) and were therefore excluded from the analysis. Thus, data from 962 participants of the PREVEND sub-cohort was used in this study.

Measures

Functional Somatic Symptoms. FSS were assessed with the somatization section of the CIDI. The CIDI is a structured instrument that was developed for the assessment of mental disorders according to the diagnostic criteria from the ICD-10 and the DSM-IV, suitable for epidemiological studies, and with validation in multiple languages and cultures (Andrews & Peters, 1998). For this study, a self-administered computerized version of the CIDI 2.1 12-month was applied. Inter-rater reliability for the self-administered CIDI and the CIDI delivered by an interviewer was good to excellent (Andrews & Peters, 1998). Interviewers were present during the application to solve questions and support with the use of the computer.

The somatization section of the CIDI assesses the presence of 43 symptoms during the previous 12 months. Although this section was originally developed to assess the outdated DSM-IV diagnosis of somatoform disorders, in this study, it was used because the interview structure enabled us to differentiate between functional symptoms and symptoms caused by chronic disease. Symptoms are classified as "present" when they are clinically relevant (i.e., require a healthcare visit); symptoms that did not lead to healthcare seeking were considered absent. When a symptom is considered present, the questionnaire assesses whether it is due to a physical illness or injury, based on patient self-report of evaluation by a health care professional, or whether it is caused by using a medication, drugs, or alcohol. When a symptom is present but is not explained by the two previous reasons, it is considered an FSS. Self-reported medical diagnoses from participants who indicated FSS were checked by the researchers to ensure that the symptoms were medically explained. When a symptom was due to a medical diagnosis of a functional syndrome (i.e., Irritable Bowel Syndrome, Chronic Fatigue Syndrome, and fibromyalgia), symptoms were coded as FSS. Thus, items are coded as "Absent/Not FSS (0)", and "Present FSS (1)".

Age: For the analysis, age was dichotomized into a group of adults (<60 years old), and Older adults (≥60 years old).

Statistical Analysis

All analyses were performed in R version 3.6.3 (R Core Team, 2020). For confidentiality reasons, the data used for this study is not available. The analysis code, results, and appendices are available on the Open Science Framework (https://osf.io/3n6fh//).

Item selection. For the analysis, we used 19 items of the somatization section of the CIDI. We first selected 23 of the 43 original symptoms by excluding items related to menstrual and reproductive health and symptoms that were reported by less than 10 persons, to avoid bias and inaccuracy of our estimates. We also based these selection criteria on a previous study in which latent class analysis was performed with these 23 items of the CIDI (Rosmalen et al., 2011). After fitting the 2PLM and assessing model fit, we decided to exclude four additional items, given that they did not meet the assumptions of the 2PLM, and affected model fit (see Appendices B and C). Thus, we performed the analysis with the 19 remaining items. We calculated internal consistency with the Kuder-Richardson formula, which is suitable for dichotomous items. Kuder-Richardson indicated an internal consistency of 0.70. More information about the excluded items can be found in Appendix A.

Assumption Check. We tested four assumptions of the 2PLM:
1) Unidimensionality: a single latent trait variable should account for a large proportion of the common variance among item responses (Embretson & Reise, 2013). Given that the items were dichotomous, we first calculated a polychoric correlations matrix and then performed an exploratory factor analysis (EFA) with the “factanal” function. We extracted one factor, using the minimal residuals method. As a rule-of-thumb, a factor should account for at least 20% of the variance for the questionnaire to minimally meet the assumption and to obtain stable parameter estimates in the IRT model (Reckase, 1979).

2) Local independence: Items must be independent of each other, conditional on the severity level, so that the probability of reporting a symptom in the questionnaire is strictly determined by the participant’s FSS severity level (Embretson & Reise, 2013). To check this assumption, we used the “residuals” function (Chalmers, 2012). We calculated the Cramer’s V effect sizes for each item. Cramer’s V calculates goodness of fit to indicate if data are independent of each other. A small (≤ .05) to medium (≤ .15) Cramer’s V effect size is interpreted as weak evidence against the local independence assumption (Cohen, 2013).

3) Item fit: Significant differences between the observed and expected item-scores indicate poor item fit to the model. We tested item fit with the Kang and

Chen's signed chi-squared test (S-x^2), which is suitable for polytomous models (Kang & Chen, 2008). Given that p-values are very sensitive to big sample sizes, we observed item fit with the Root Mean Square Error of Approximation (RMSEA), which gives a more precise estimation of item fit. The RMSEA is a measure of discrepancy between the covariance matrices of the observed and expected parameter values. According to prior criteria, an RMSEA < 0.06 shows goodness of fit (Cook, Kallen, & Amtmann, 2009; Kang & Chen, 2008). We checked item fit with the function "itemfit" (Chalmers, 2012).

4) Monotonicity: The relationship between the latent trait (severity of FSS) and the participant's responses should respond to an increasing function. To check this we used the function "check.monotonicity" (Van der Ark, L Andries, 2007).

The Two Parameter Logistic model (2PLM) fitting. The 2PL model is an IRT model for dichotomous data. The aim of IRT models is to model the association between a latent construct, such as FSS, and each item of a questionnaire (e.g., a symptom of FSS) through the response patterns of persons to a set of items (Reise & Waller, 2009). To model this relationship, the 2PL model describes the items by two parameters: 1) The discrimination parameter (α), and 2) the severity or location parameter (β). The discrimination parameter (α) is a slope parameter that reflects the ability of an item to distinguish between different levels of FSS severity. The severity or location parameter (β) reflects the location of the item on the FSS latent trait scale, that is, how severe a symptom is. The 2PL also calculates a person's location (θ) for each person based on their patterns of responses and the item parameters, which indicates the level of severity of FSS for each person. The person and item location parameters coexist on the same FSS severity continuum, allowing for direct comparisons between persons and symptoms. Location parameters are typically standardized scores with a mean of 0 and a standard deviation of 1 (Embretson & Reise, 2013).

We calculated the item parameters of the 19 selected items, with the "ltm" function. We plotted the item parameters and their Standard Error (SE). It is necessary to check the SEs of item parameters to ensure that the IRT model is well estimated (Tay, Meade, & Cao, 2015). Additionally, we plotted the Item Characteristic Curves (ICCs). Each curve represents the probability to report an item as a function of the level of FSS severity (θ). The item severity is the θ at which an individual has a .50 probability to report an item. Steeper curves reflect higher discrimination (Coulacoglou & Saklofske, 2017). Furthermore, we plotted the Test Information Function (TIF) which shows the amount of psychometric information, or preciseness, that a test provides, in relationship to the level of FSS severity (θ; Reise & Waller, 2009).

Complementary analysis: Zero-Inflated IRT. To explore the robustness of our item parameter estimations, we also fitted the zero-inflated IRT (ZI IRT) model to the 19 items of the CIDI. Given that psychiatric constructs are often skewed, an important part of general population samples may not report any symptom (i.e., zero inflation), leading to floor effects. As the estimation of IRT model parameters almost always operates under the assumption that the underlying trait is normally distributed (Wall, Park, & Moustaki, 2015) it is necessary to use IRT models that can handle these kinds of data in order to obtain unbiased estimates. The ZI IRT model addresses this problem, by identifying a group with a unique response pattern (i.e., zero responses) based on latent class analysis and allowing the latent trait to follow a mixture of normals, instead of a normal distribution. The model detects a pathological and a non-pathological group in the sample, and then estimates, and scales the discrimination and severity parameters only for the pathological group (Wall et al., 2015). Since approximately 95% of the symptoms were reported as absent/non-FSS (see results), and more than 50% of the sample did not report any symptoms (see results), we decided to explore a ZI IRT to avoid bias in our estimations. We performed the ZI IRT analysis using Mplus, with the example code provided by Wall, Park, and Moustaki (2015).

Differential Item Functioning. Firstly, we fitted the 2PLM for each sex and age subgroup with the function “itemParEst”. Secondly, we plotted the ICCs for each of these subgroups to have a visual exploration of DIF. Thirdly, we obtained the DIF statistics with three different methods, given that it is relevant to rely on at least two indicators when testing for DIF (Tay et al., 2015). This allowed us to compare the results and accrue more robust evidence of DIF. The methods used were: 1) Lord’s chi-square: this is a DIF detection model that requires a pre-estimation of an IRT model. It tests the null hypothesis of equal item parameters in both groups of subjects. It is based on a test statistic with a chi-square distribution under the null hypothesis. 2) Raju’s method: it computes the area between the ICCs of both groups. The null hypothesis is that the area is zero. 3) Mantel Haenszel (MH): this is a non-IRT method for detecting DIF. It tests if there is an association between group membership and item response, conditional on the total sum-score (Tay et al., 2015). When the difference of the probabilities for both groups to report an item are independent of the common severity, we talk about uniform DIF. This is shown graphically in ICCs that do not intersect. When the difference between the probabilities to report an item are not constant across the severity levels but depend on them, we talk about non-uniform DIF. This is reflected in intersecting ICCs curves (Magis, Béland, Tuerlinckx, & De Boeck, 2010). Lord and Raju’s methods allow for the identification of both uniform and non-uniform DIF, whereas the Mantel-Haenszel method only allows for the identification of uniform DIF.

Given that we did not pre-specify anchor items, we used item purification for the analyses, on each of the used methods. Item purification refers to an iterative elimination of the items with DIF, while the model rescales the item parameters of both groups to a common metric. This is important since the presence of a DIF for one item could influence the results of tests for DIF in other items. To have an indication of the effect size of DIF, we calculated the Delta scale (Δ_MH), based on the Mantel Haenszel odds-ratio method, which has been used as the standard for DIF effect size assessment (Cervantes, 2017; Suh, 2016). The scale usually ranges from 3 and – 3, and classifies the effects as A = negligible (-1<Δ_MH<1), B = moderate (1 ≤Δ_MH<1.5 or-1 ≥Δ_MH>-1.5), and C = large (Δ_MH≥1.5 or Δ_MH≤-1.5) The Δ_MH is negative when the focal group odds of reporting an item are less than the reference group odds (Zwick, 2012).

Sex: After fitting the 2PLM for both sex groups, we found that three items showed large SEs in their item parameters (see results), which resulted in an inaccurate reflection of DIF in their ICC plots. We excluded those items from the DIF analysis since they could influence the DIF estimation in other items. We used female sex as the reference group for sex analyses.

Age: After fitting the 2PLM, we found that one item had a very large SE in the severity parameter (see results), reflecting an unusual ICC. We decided to exclude this item from the DIF analysis to avoid its influence on the DIF estimation of other items. We used the adult age group as a reference group for age analyses.

Results

Sample descriptive statistics

Of the 962 participants, 502 (52%) were females and 460 (48%) were males. The mean age was 55.8 (Standard Deviation = 11.1), with a minimum age of 35.9 and a maximum age of 82.3. Regarding the age groups, 644 (67%) participants were grouped as "adults", while 313 (33%) participants were grouped as "older adults" (60+).

When calculating the number of symptoms reported, almost 50% of the participants reported having at least one symptom unexplained by physical pathology, out of the 19 symptoms included. The maximum number of symptoms was 13, reported by one participant. Figure 1 shows the number of unexplained symptoms reported by the participants.

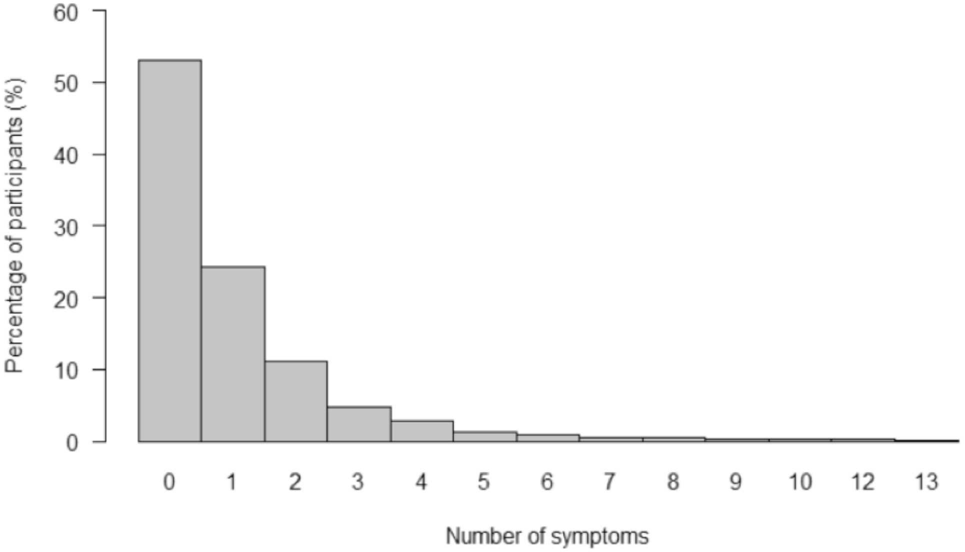
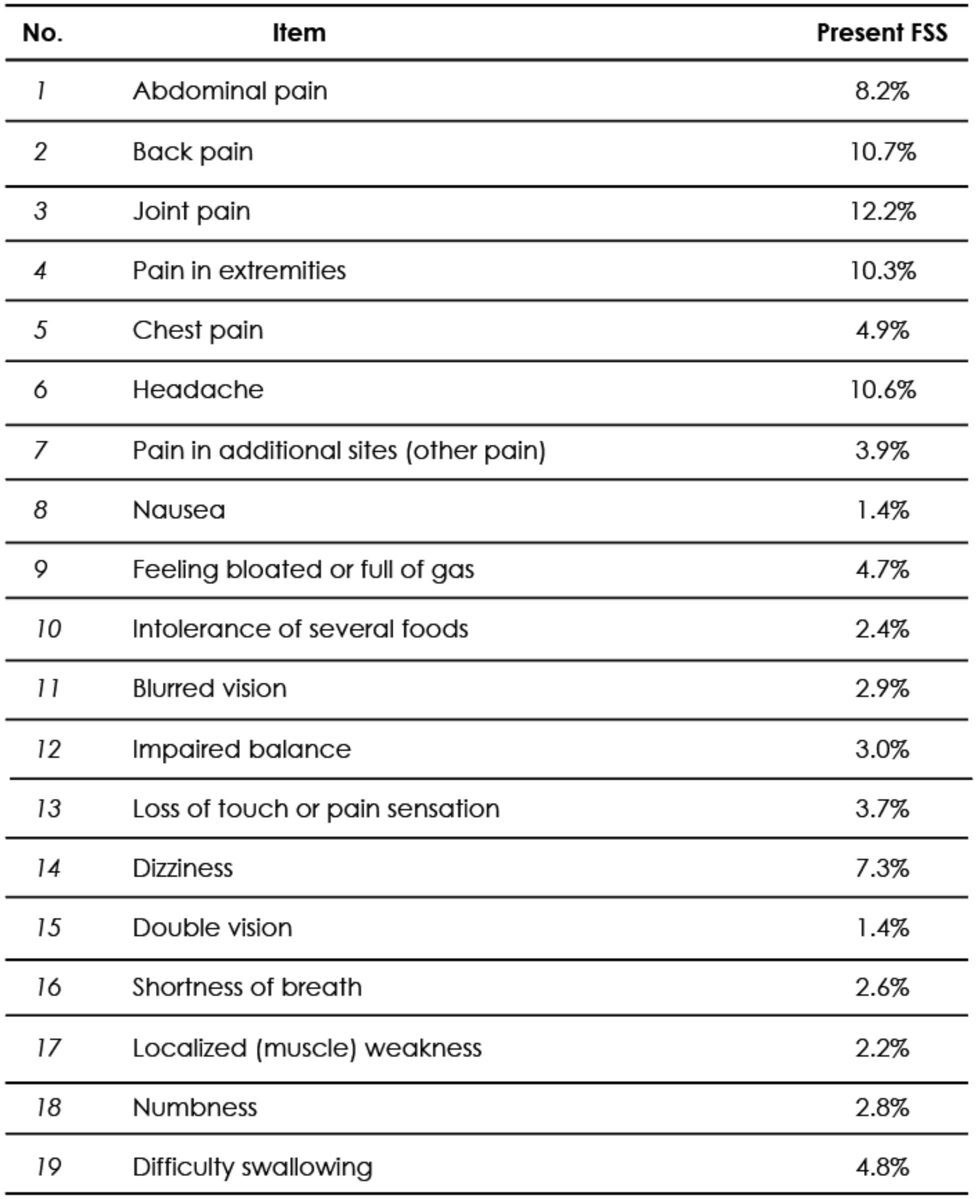


Figure 1. Percentage of unexplained symptoms reported by the participants from the 19 selected items of the somatization section of the Composite International Diagnostic Interview.

Distribution of response choices

Table 1 shows the percentage of participants who reported each item as an FSS present in the past 12 months. As it is expected with symptom-based constructs, our data was extremely skewed. On average, 94.8% of the responses in all FSS items signaled an absent/not FSS symptom, whereas 5.2% of the answers on average reflected the presence of an FSS symptom. Given that the application of the CIDI was computerized, there was no missing data.

Table 1. Distribution of response choices per item of the CIDI somatization section



No.	Item	Present FSS
1	Abdominal pain	8.2%
2	Back pain	10.7%
3	Joint pain	12.2%
4	Pain in extremities	10.3%
5	Chest pain	4.9%
6	Headache	10.6%
7	Pain in additional sites (other pain)	3.9%
8	Nausea	1.4%
9	Feeling bloated or full of gas	4.7%
10	Intolerance of several foods	2.4%
11	Blurred vision	2.9%
12	Impaired balance	3.0%
13	Loss of touch or pain sensation	3.7%
14	Dizziness	7.3%
15	Double vision	1.4%
16	Shortness of breath	2.6%
17	Localized (muscle) weakness	2.2%
18	Numbness	2.8%
19	Difficulty swallowing	4.8%

Note. CIDI = Composite International Diagnostic Interview; FSS = Functional Somatic Symptoms.

Model assumption check

Regarding the unidimensionality check, the EFA with a one-factor solution explained 34% of the total variance and all items had medium to high factor loadings (between 0.47 and 0.80). Regarding local independence, small to medium Cramer's V effect sizes were observed in all items (between -0.07 and 0.14), which indicates that there is weak evidence for the violation of the local independence assumption. All the RMSEA values of the S - x^2 test were lower than 0.06, indicating that the items showed a good fit to the model. There were no strong indications of violations of the monotonicity assumption. The results from these analyses can be found in Appendix B.

The 2PLM

Item parameters. Table 2 shows the estimated item parameters from the 2PL in the 19 selected items. These are listed from the least to the most severe item according to their β value. The severity parameter (β) ranged from 2.03 to 3.72. If we take into account that the mean θ score is 0, this means that these items are best at measuring levels of severity of FSS that deviate from 2.03 standard deviations from the mean, that is, these are best when measuring people with relatively high levels of severity of FSS. The item with the highest discrimination is 17 "Localized (muscle) weakness". This means that this item is the ablest to distinguish between people with different levels of FSS, in comparison to the rest of the items. This is followed by item 8 "Nausea". The least discriminative item was 7 "Pain in additional sites", followed by 11 "Blurred vision".

Table 2. 2PL IRT parameters of the 19 items

No.	Item	Discrimination (*a*)	SE (*a*)	Severity (ß)	SE (ß)
3	Joint pain	1.21	0.18	2.03	0.22
4	Pain in extremities	1.39	0.20	2.03	0.20
1	Abdominal pain	1.56	0.23	2.10	0.20
2	Back pain	1.20	0.18	2.19	0.24
6	Headache	1.17	0.18	2.23	0.25
17	Localized (muscle) weakness	**2.36**	0.45	2.57	0.23
14	Dizziness	1.17	0.20	2.63	0.33
13	Loss of touch or pain sensation	1.64	0.29	2.66	0.29
9	Feeling bloated or full of gas	1.42	0.24	2.70	0.31
12	Impaired balance	1.73	0.32	2.74	0.30
10	Intolerance of several foods	1.72	0.33	2.91	0.34
19	Difficulty swallowing or lump in the throat	1.17	0.22	3.05	0.43
5	Chest pain	1.13	0.22	3.10	0.45
16	Shortness of breath	1.48	0.30	3.11	0.42
18	Numbness/tingling	1.42	0.29	3.12	0.43
8	Nausea	**1.87**	0.43	3.15	0.42
7	Pain additional sites (Other pain)	1.02	0.23	3.60	0.64
15	Double vision	1.45	0.37	3.65	0.65
11	Blurred vision	1.09	0.26	3.72	0.70

Note. 2PL = two-parameter logistic, SE = Standard Error. Numbers in bold represent the highest discrimination parameters.

Person parameters. Regarding person location (θ), the mean score was 0.12 (Standard Deviation = 0.68), the minimum score was -0.41 and the maximum score was 3.02. The median was -0.41. From the 962 participants, 53% (510) had a score below 0, 34.8% (335) had a score between 0 and 1, 10.5% (101) had a score between 1 and 2, 1.6% (15) had a score between 2 and 3, and only 0.1% (1) had a score of 3 or more. This indicates that most of the participants had a low severity of FSS.

Test Information Function (TIF): The TIF peaks at around a θ score of 3 and provides information in a θ range between 0 and 6 (see Appendix D). This means that this set of items as a whole can provide some information for participants with lower to large FSS levels, but that the preciseness is higher for participants who show a θ score of approximately 3, that is the participants with the highest FSS severity levels in this sample (around 1% of the participants).

Complementary analysis: Zero-Inflated IRT. The parameters from the ZI IRT and the 2PLM had a correlation of 0.99, meaning that both models showed almost the same results (see Appendix E).

Differential Item Functioning

2PLM model fitting by groups. Regarding sex, the severity parameters (panel A, Figure 2), indicate that items 7 "Pain in additional sites (other pain)" and 12 "Impaired balance" show large SEs for the males' group. This results in ICC curves that inaccurately reflect DIF for these items between males and females. This is also the case for item 17 "Localized (muscle) weakness", which shows a very large SE for males' in the discrimination parameter (panel C). For this reason, these three items were excluded from the DIF analysis by sex, given that they inaccurately show DIF in the ICCs, and influence the estimation of DIF in the rest of the items.

Regarding age, panels B and D of Figure 2 show the parameters of the 19 items by age. As it is shown, item 11 "Blurred vision", has a very large SE (4.37) in the severity parameter of the older adults (Panel B). This also results in a very inaccurate ICC, for which we decided to remove this item from the DIF analyses.

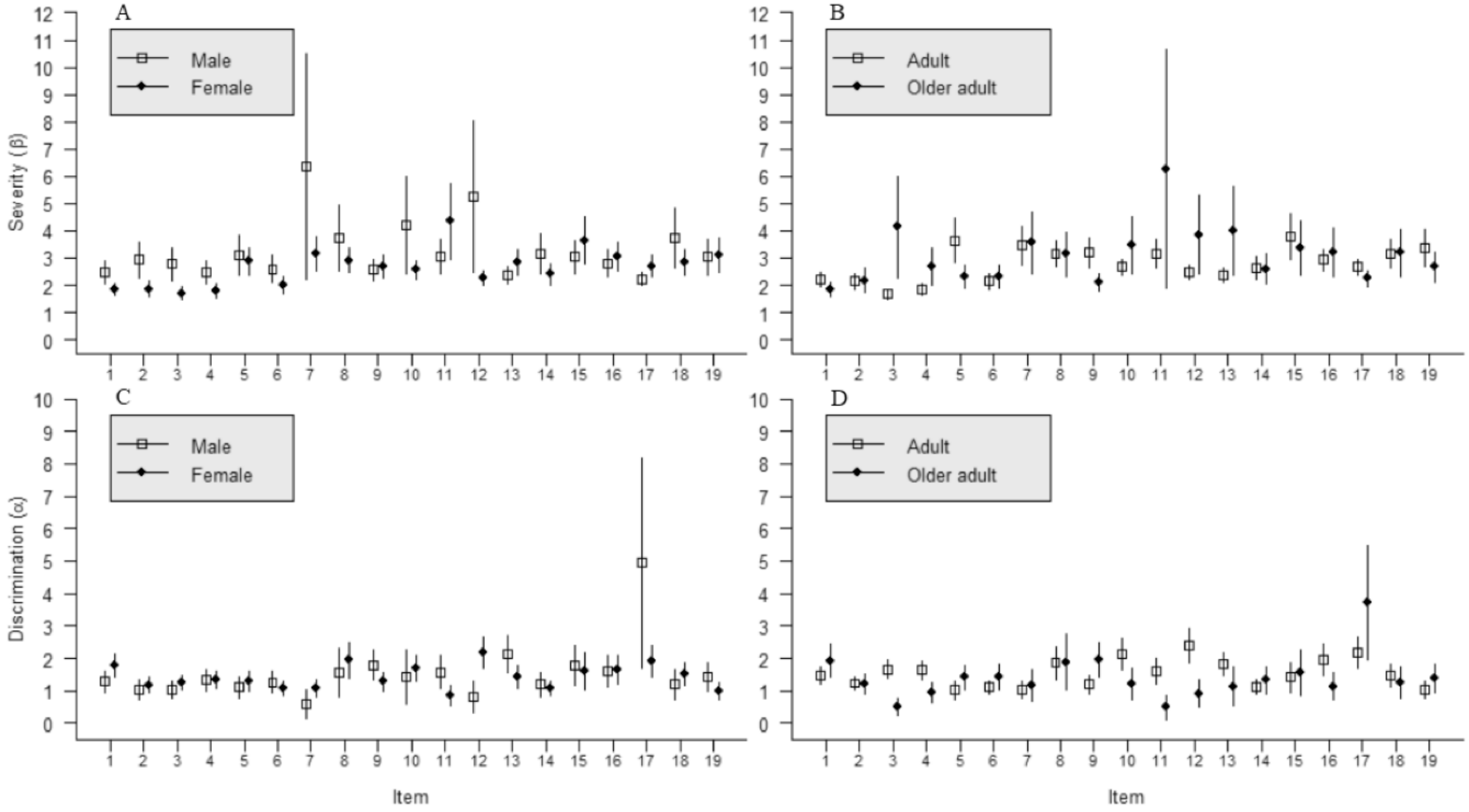


Figure 2. Two-parameter logistic model parameters by sex and age.

Note: Panel A shows the severity parameters by sex. Panel B shows the severity parameters by age. Panel C shows the discrimination parameters by sex. Panel D shows the discrimination parameters by age.

DIF by sex. The left section of Table 3 shows the statistics for the DIF tests by sex with the three selected methods. The three methods provided different results; however, item 1 “Abdominal pain” was consistently highlighted as a DIF item in all of them. A slight difference in the ICCs of males and females in this item can be seen in Figure 3 (Panel A) with males reporting this item at higher levels of FSS severity than females, and with this item being slightly more discriminative for females than for males. Four other items were signaled as DIF items by only two methods. According to Lord’s and Mantel Haenszel’s tests, items 5 “chest pain”, 15 “double vision”, and 16 “shortness of breath” showed an indication of DIF. When inspecting the ICCs of these items, it can be seen that item 5 (Figure 3, Panel B) shows that males tend to report chest pain when having higher FSS severity levels than females, although this item does not seem to be very discriminative for any sex. Item 15 (Figure 3, Panel D) shows an indication of uniform DIF, with the item showing almost the same discrimination for both males and females, but with females reporting this item at higher FSS severity levels. Item 16 (Figure 3, Panel E) on the other hand, does not seem to show important differences between males and females according to the ICCs, although it shows a large DIF effect size difference according to the Δ_MH . Item 13 was flagged as a DIF item by Lord’s and Raju’s method, showing a non-uniform DIF according to the ICCs (Figure 3, Panel C), where the likelihood to report this item is slightly higher for females at lower severity of FSS, and it becomes lower when females have higher severity levels of FSS. This item shows higher discrimination for males than for females. According to the Δ_MH, all the items signaled with DIF by more than two methods showed a large positive DIF effect size, meaning that males have a higher chance to report these items as FSS at lower severity levels than females.

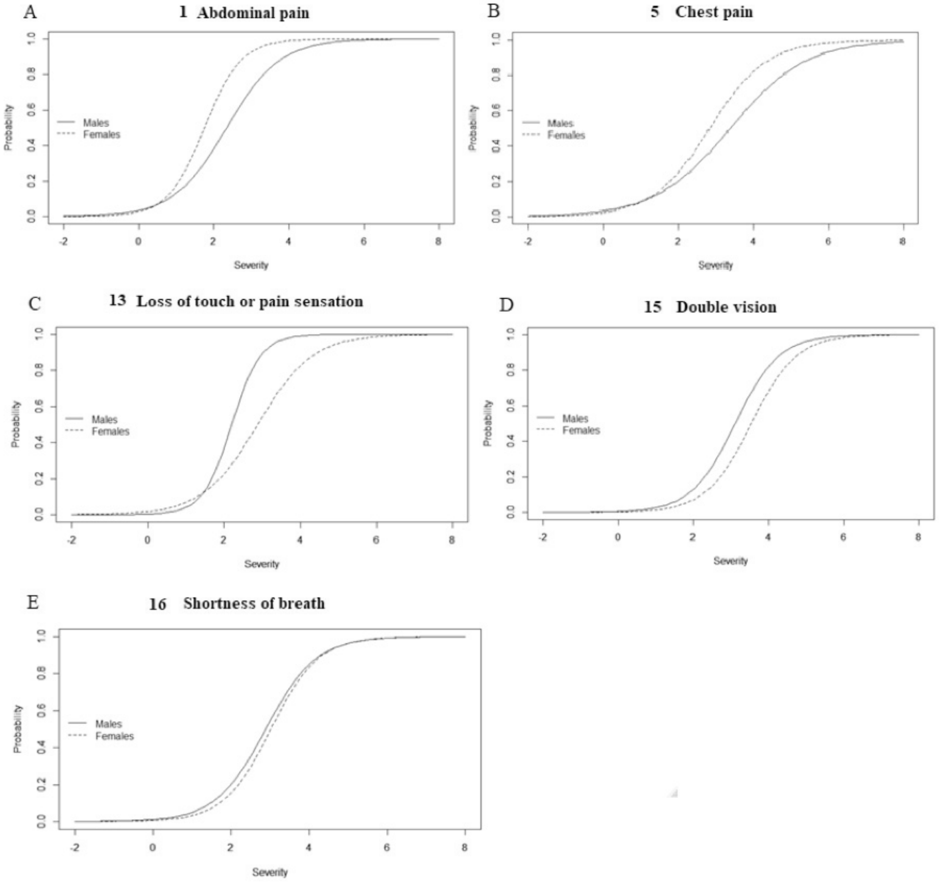


Figure 3. Item characteristic curves of the items detected as DIF items for sex, by more than one method.

DIF by age. The right section of Table 3 shows the statistics of the DIF analysis by age with the three methods. There is no consistent evidence of DIF for age with these methods. Regarding Lord's chi-square test, adults and older adults show a significant difference in the parameters of item 3 "Joint pain". When exploring the ICCs of this item, a difference between the curves of adults and older adults can be seen, reflecting a non-uniform DIF (Figure 4, panel A), with older adults reporting this item more frequently than adults when having lower FSS severity levels, and less frequently than adults when having higher FSS severity levels. Moreover, this item does not show good discrimination for the older adults' group. However, the Δ_MH shows that this item has a negligible DIF effect. No item was detected as a DIF item with Raju's method. Regarding the Mantel Haenszel's method, items 5 "Chest pain", 6 "Headache", and 16 "Shortness of breath" show DIF. All of these items show a large negative Δ_MH DIF effect size, meaning that older adults have a lower chance to report these symptoms as FSS when adjusting for the FSS severity levels of both groups. When exploring the ICCs (Figure 4), item 5 seems to have a large difference in the ICC curves, showing a mostly uniform DIF, with older adults reporting this item when having less FSS severity levels than adults, and with this item being more discriminative for older adults. Items 6 and 16 seem to show smaller differences in their curves and a non-uniform DIF.

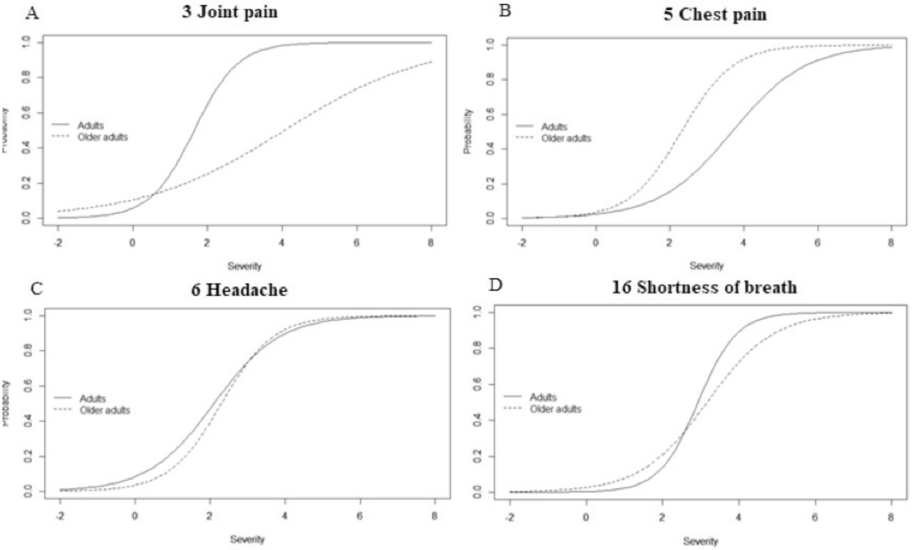


Figure 4. Item characteristic curves of the items selected as DIF for age, by at least one method.

Table 3. Statistics of DIF analysis for sex and age with three different methods

		DIF by sex							DIF by age						
		Lord's		Raju's		MH			Lord's		Raju's		MH		
#	**Item**	**Stat.**	**P-value**	**Stat.**	**P-value**	**Stat.**	**ΔMH**	**P-value**	**Stat.**	**P-value**	**Stat.**	**P-value**	**Stat.**	**ΔMH**	**P-value**
1	*Abdominal pain*	14.24	<0.001 ***	2.92	<0.01 **	5.08	1.61^{C}	0.02 *	0.80	0.67	-1-15	0.25	0.38	-0.52^{A}	0.54
2	*Back pain*	4.97	0.08	0.93	0.35	0.14	-0.31^{A}	0.71	0.03	0.98	0.14	0.89	0.00	0.06^{A}	0.98
3	*Joint pain*	6.33	0.04 *	1.08	0.28	0.00	0.09^{A}	0.99	11.72	<0.01 **	1.28	0.20	0.05	0.20^{A}	0.82
4	*Pain in extremities*	7.81	0.02 *	1.36	0.17	0.01	-0.02^{A}	0.91	3.49	0.17	1.17	0.24	0.12	0.30^{A}	0.73
5	*Chest pain*	6.98	0.03 *	1.04	0.30	6.47	2.09^{C}	0.01 *	2.72	0.26	-1.64	0.10	4.89	-1.82^{C}	0.03 *
6	*Headache*	3.50	0.17	0.52	0.60	0.61	-0.55^{A}	0.44	5.34	0.07	-1.08	0.28	5.65	1.67^{C}	0.02 *
7	*Pain in additional sites (other pain)*	*Excluded*							0.91	0.64	-0.36	0.72	1.54	1.46^{B}	0.21
8	*Nausea (without vomiting)*	2.60	0.27	0.53	0.60	0.00	0.43^{A}	0.94	0.01	1.00	-0.24	0.81	0.06	-0.16^{A}	0.80
9	*Feeling bloated or full of gas*	10.00	<0.01 **	0.96	0.34	0.50	0.76^{A}	0.48	2.89	0.24	-1.87	0.06	1.10	-1.05^{B}	0.29
10	*Intolerance of several foods*	2.02	0.36	0.65	0.52	2.05	-2.48^{C}	0.15	2.19	0.33	0.65	0.51	0.60	-1.10^{B}	0.44
11	*Blurred vision*	2.05	0.36	-1.12	0.26	0.40	0.91^{A}	0.53	*Excluded*						
12	*Impaired balance*	*Excluded*							5.92	0.05	1.01	0.31	2.16	-1.84^{C}	0.13
13	*Loss of touch or pain sensation*	18.25	<0.001 ***	-2.48	<0.01 *	3.46	1.93^{C}	0.06	0.97	0.62	0.89	0.37	3.70	2.50^{C}	0.05
14	*Dizziness*	1.39	0.50	0.25	0.80	0.85	-0.76^{A}	0.36	1.04	0.59	-0.43	0.66	0.79	0.77^{A}	0.37

15	*Double vision*	12.68	<0.01 **	0.81	0.42	7.68	5.12^{C}	<0.01 **	0.30	0.86	-0.53	0.60	0.00	-0.40^{A}	0.97
16	*Shortness of breath*	14.75	<0.001 ***	1.28	0.20	5.66	2.82^{C}	0.02 *	4.24	0.12	0.85	0.40	58.27	-2.81^{C}	0.02 *
17	*Localized (muscle) weakness*	*Excluded*							1.34	0.51	-1.47	0.14	0.00	-0.30^{A}	1.00
18	*Numbness*	3.92	0.14	0.58	0.56	0.66	1.10^{B}	0.42	0.99	0.61	0.27	0.79	0.70	-1.09^{B}	0.40
19	*Difficulty swallowing*	2.96	0.23	0.63	0.53	0.03	0.00^{A}	0.86	0.78	0.68	-1.04	0.30	0.02	-0.27^{A}	0.88
	Detection threshold	5.99		-1.96	-1.96		3.84		5.99		-1.96	-1.96		3.84	

Note: *p-value ≤ 0.5; **p-value ≤ 0.01, ***p-value ≤ 0.001. MH = Mantel Haenszel. ΔMH = Delta Mantel Haenszel for effect sizes. A = Negligible effect, B = Moderate effect, C = Large effect.

Discussion

This study aimed to identify the most discriminative symptoms of FSS from a set of 19 items of the somatization section of the CIDI in a general population. In addition, we examined DIF related to the sex or the age of the respondents. To this end, the 2PLM was fit to the data, followed by an inspection of the ICCs and three DIF indicators (Lord's Chi-square, Raju's, and Mantel Haenszel's methods).

On average, the answer option "absent/not FSS symptom" received 94.8% of the answers on each symptom, which resulted in a large fraction of the sample (53%) having a very low level of severity of FSS ($\theta < 0$), and only 1.7% presenting a high level of severity of FSS ($\theta \geq 2$). However, this is expected since the data was collected from a non-clinical sample, and since it was dichotomous data. The severity (β) indicated that the tested items are best suited to measure participants with higher levels of FSS severity ($\theta \geq 2.03$). The item with the highest discrimination parameter was item 17 "Localized (muscle) weakness", which is consistent with previous studies where the item "feeling weak physically" was found to be the most discriminative item from the somatization scale of the SCL-90 in a clinical psychiatric population (Paap et al., 2011), neuromuscular diagnosed patients (Hart, Werneke, George, & Deutscher, 2012) and a general population cohort without somatic conditions (Acevedo-Mesa et al., 2020). These results suggest that, even in different questionnaires (i.e., CIDI and SCL-90), items related to "weakness" are especially discriminating in the measurement of the severity of FSS, both in clinical and non-clinical populations.

According to experts' opinion, the most relevant symptoms to measure FSS severity include headache, nausea, shortness of breath, dizziness, and low back pain (Kroenke & Mangelsdorff, 1989; Zijlema et al., 2013). However, the current study showed that "headache", "dizziness" and "back pain" had rather low discrimination parameters (a =1.17 ~ 1.20), which is consistent with previous results (Acevedo-Mesa et al., 2020). In contrast, the item "nausea" (a = 1.87) had the second-highest discrimination parameter, which shows agreement with experts' opinion, but not with previous results (a = 1.16; Acevedo-Mesa et al., 2020). In our previous work, it was not possible to identify if symptoms were unexplained by pathology, and since nausea is a common symptom with many causes, it did not appear to be discriminative. However, the present study focused on unexplained symptoms, based on patient self-report of evaluation by a health care professional. This could suggest that specifically unexplained nausea is a discriminative symptom of FSS. In addition to the different assessment instruments, population characteristics could explain discrepancies between the previous (Acevedo-Mesa et al., 2020) and the current study. The mean age of the population in the previous study was considerably lower (42 years versus

55 years in the current study), which might also have influenced the IRT results. The inconsistencies between experts' opinions and empirical data highlight the relevance of studying the properties of the individual items to improve the construction of FSS questionnaires. These discrepancies may be due to the fact that according to experts' experience, the above-mentioned symptoms are observed in many FSS patients, but those items are also frequently reported in the general population and therefore not informative when it comes to measuring the severity of FSS.

Regarding the study of bias in symptom reporting using DIF for sex subgroups, item 1 "abdominal pain" showed consistent evidence of DIF across the three methods, with females being more likely to report abdominal pain than males at lower levels of FSS severity. Previously, it has been shown that females tend to report gastrointestinal symptoms more often than males (Kroenke & Price, 1993), and that chronic functional abdominal pain tends to be more common in girls (Korterink, Diederen, Benninga, & Tabbers, 2015; Rajindrajith, Zeevenhooven, Devanarayana, Perera, & Benninga, 2018). According to our results, it could be the case that males reporting this item have higher FSS severity levels than females reporting this item, implying that it is important for clinicians to pay differential attention to this symptom. Item 13 "loss of touch or pain sensation" showed an indication of non-uniform DIF according to the ICCs, with females reporting this symptom more frequently at lower levels of severity of FSS, and less frequently than males at higher levels of severity of FSS. This item was detected as a DIF item by Lord's and Raju's methods, but not by Mantel Haenszel's method, however, the Δ_MH showed a large DIF effect. This discrepancy could be explained because of the lack of ability of the Mantel Haenszel method to detect non-uniform DIF. In clinical practice, this means that this item could be a relevant symptom to discriminate FSS severity in males, but not in females, and that if females report it, they will generally have higher FSS severity levels than males who report it. It has been reported before that estrogen may have a role in pain sensitivity, with females being more sensitive to pain in periods of low estrogen such as menopause, and that the experience of pain differs by sex (Nikolov & Petkova, 2010; Sun et al., 2019), which may explain why the perception of loss of touch or pain sensation could not be a discriminative symptom of FSS in females.

Items 5 "chest pain", 15 "double vision", and 16 "shortness of breath" were flagged as DIF items by Lord's and Mantel Haenszel methods. The ICCs show that males report chest pain at slightly higher levels of FSS severity, however, this item is not very discriminative for any sex. This item has also been previously reported as one of the few symptoms that is more frequent in males than in females (Ballering et al., 2020). In clinical practice, this could mean that males presenting this symptom have higher FSS severity levels than females presenting this symptom. There are very small differences in the ICCs of males and females in item 15 “double vision”, and almost no visible differences in the ICCs of males and females in item 16 “shortness of breath”, although the Δ_MH indicate large DIF effects. One of the reasons for the discrepancy between the different DIF methods and ICCs could be that Lord’s DIF method is generally more sensitive than other methods (Lee & Suh, 2018; Özdemir, 2015). It has been reported that this method may detect DIF even with small differences in the area between two ICCs (Lee & Suh, 2018; Millsap & Everson, 1993). On the other hand, the Mantel Haenszel method has been reported to indicate DIF inaccurately when complex IRT models are used (Lee & Suh, 2018), especially in short questionnaires (e.g., less than 20 items). This means that due to the different assumptions of each method, there could be important discrepancies in the detection of DIF, which reinforces the need to use more than one method when studying DIF (Tay et al., 2015), and to also perform a visual exploration of the ICCs to contrast the statistics with the plots. Given these discrepancies, the evidence of DIF for these items is rather inconclusive, however, given that effect sizes indicated large effects, apparently small differences could have clinical relevance.

Regarding DIF for age, there was no consistent indication of DIF for any item. Four items were flagged as DIF items, namely, item 3 “joint pain”, 5 “chest pain”, 6 “headache, and 16 “shortness of breath”. However, all of them were detected by only one of the three tested methods. Nonetheless, when inspecting the ICCs, the item 3 “joint pain” showed that older adults (60+) have a higher probability to report this item at lower severity levels of FSS, but a low probability of reporting it at higher levels of severity of FSS. This may be because it is more likely that joint pain is explained by a pathology (e.g., osteoarthritis) in older adults, hence, this age group may attribute their joint pain to a physical cause and thus report it less frequently than adults as an FSS. This can also explain why this item has very low discrimination for older adults, as shown in the ICCs. In clinical practice, this means that this symptom would not provide much information about an older adult’s FSS severity level, given that it is a common symptom in the elderly. Overall, although there is no consistency in the results of DIF for age, most of the items signaled with DIF such as musculoskeletal symptoms (e.g., joint pain), chest pain, and shortness of breath are often reported by older adults (Michael et al., 2005). Even though these symptoms may not be typical for older populations (van Driel et al., 2017), our results show that the reporting of these symptoms may differ between adults and older adults. This means that these symptoms may not provide enough information for the assessment of FSS in older patients in the clinical setting as well as in research.

This study has several strengths. Firstly, we employed data from a reasonably large sample, from which symptoms were ensured to be medically unexplained. This is an important advantage compared to other studies since it is often unclear whether the symptoms reported are explained by a physical pathology. Secondly, it is the first time, to our knowledge, that sophisticated psychometric methods such as IRT are employed to analyze the somatization section of the CIDI. The psychometric properties of the CIDI have been studied before, but the analysis on the item level of this instrument could provide more insightful information. It is important to highlight that although the CIDI was constructed based on the criteria of the outdated somatoform disorder diagnosis, its strength for the present study is that it can distinguish between symptoms that are explained by a chronic condition, and symptoms that despite medical examination cannot be attributed to an underlying somatic condition. Furthermore, to ensure that the symptoms were adequately classified as FSS, medical diagnoses were checked by the researchers. Thirdly, we not only used the 2PL model to explore item parameters, but we also explored the robustness of such parameters by using the ZI IRT model. Given that a large majority of the answers indicated the absence of FSS symptoms, it was necessary to use a model that could take this into account to make sure that the estimates were accurate. The fact that the results of the 2PLM and ZI IRT models are consistent, provides robustness to our results. Finally, the use of three different methods to explore DIF, in addition to the visual exploration of the ICCs, also provides more robust results. Additionally, our dataset had no missing data given that the CIDI was completed digitally.

It is important to consider several limitations of this study when interpreting the results. Firstly, given that the items were dichotomous, information about the severity of the symptoms was lost. This may have also influenced the DIF results because there may have been items that were not reported as “present FSS” highly enough for the models to be able to detect DIF. Secondly, we only used 19 items from the original 43 items of the somatization section of the CIDI. Although we were careful to select the most appropriate items theoretically and statistically, our selection may not resemble the original factor structure intended when constructing this scale. However, our EFA shows evidence of unidimensionality. Thirdly, in the Lord’s test for DIF analysis by sex, 50% of the items were detected as DIF items. This is problematic since it might be the case that there are non-DIF items appearing as DIF because true DIF items could have been inadvertently used as anchor items between groups, given that we did not choose anchor items a priori (Tay et al., 2015). However, in all our DIF analyses we used item purification to avoid problems with linking, given that we did not know which items to use as anchor beforehand. With item purification, we hope we have avoided the influence of DIF items in the results of tests of other DIF items. Finally, although the CIDI interview uses an extensive algorithm to classify a symptom as an FSS, it should be emphasized that this decision was based on patients’ self-report on evaluation of a health care professional, and thus subject to various sources of bias on the side of the clinician (e.g., diagnostic bias) and the patient (e.g., recall bias). Thus, some symptoms reported as FSS could have been explained by an unknown underlying pathology. It has been previously found that 19.1% of the participants in this study reported one or more chronic somatic diseases (Ockenburg et al., 2015) and that 11.8% fulfilled the criteria for a depression or anxiety disorder (Rosmalen, et al., 2011). Although we cannot rule out that these or other comorbidities have influenced our results, we showed in a previous study that there were no differences in IRT results when excluding participants with somatic conditions (Acevedo-Mesa et al., 2020).

The symptoms identified by this study as most discriminative and reflective of severity could be included in future FSS questionnaires given their precision to measure FSS compared to other items. Future studies could examine the clinical implications of using person location scores generated by IRT models as measures of FSS. Further analyzing DIF among FSS patients with clinical samples may reveal systematic report biases. The results of such a study could contribute to the assessment of FSS severity in clinical patients. Future studies must consider that dichotomous FSS data tends to be highly skewed (i.e., zero-inflated), which could impact the estimates and conclusions of FSS studies. For this reason, the use and development of methods that consider the nature of FSS data, such as ZI IRT, are important to this field of study.

In conclusion, we found that the item "Localized (muscle) weakness" is the best at discriminating between FSS severity levels. The items "Abdominal pain" and "Loss of touch or pain sensation" show evidence of DIF by sex, for which clinicians should pay differential attention to these symptoms when reported. Regarding age, the item "joint pain" was not good at discriminating FSS severity in older adults, meaning that this symptom may not provide useful information for the assessment of FSS in this age group. The characteristic of these items could be considered when constructing FSS questionnaires in order to obtain improved assessments of FSS severity in research and clinical practice.

References

Acevedo-Mesa, A., Tendeiro, J. N., Roest, A., Rosmalen, J. G., & Monden, R. (2020). Improving the measurement of functional somatic symptoms with item response theory. *Assessment*, 1073191120947153.

Andrews, G., & Peters, L. (1998). The psychometric properties of the composite international diagnostic interview. *Social Psychiatry and Psychiatric Epidemiology, 33*(2), 80-88.

Ballering, A. V., Bonvanie, I. J., Hartman, T. C. O., Monden, R., & Rosmalen, J. G. (2020). Gender and sex independently associate with common somatic symptoms and lifetime prevalence of chronic disease. *Social Science & Medicine*, 112968.

Barsky, A. J., Peekna, H. M., & Borus, J. F. (2001). Somatic symptom reporting in women and men. *Journal of General Internal Medicine, 16*(4), 266-275.

Cervantes, V. H. (2017). DFIT: An R package for raju's differential functioning of items and tests framework. *Journal of Statistical Software, 76*(5), 1-24.

Chalmers, R. P. (2012). Mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software, 48*(6), 1-29.

Cohen, J. (2013). *Statistical power analysis for the behavioral sciences* Routledge.
Cook, K. F., Kallen, M. A., & Amtmann, D. (2009). Having a fit: Impact of number of items and distribution of data on traditional criteria for assessing IRT's unidimensionality assumption. *Quality of Life Research, 18*(4), 447-460.

Correa, J. K., & Brown, T. A. (2019). Expression of generalized anxiety disorder across the lifespan. Journal of Psychopathology and Behavioral *Assessment, 41*(1), 53-59.

Coulacoglou, C., & Saklofske, D. H. (2017). *Psychometrics and psychological assessment: Principles and applications* Academic Press.

Creed, F., Henningsen, P., & Fink, P. (2011). *Medically unexplained symptoms, somatisation and bodily distress: Developing better clinical services* Cambridge University Press.

Embretson, S. E., & Reise, S. P. (2013). *Item response theory* Psychology Press.

Hart, D. L., Werneke, M. W., George, S. Z., & Deutscher, D. (2012). Single-item screens identified patients with elevated levels of depressive and somatization symptoms in outpatient physical therapy. *Quality of Life Research, 21*(2), 257-268.

Hilderink, P., Collard, R., Rosmalen, J., & Voshaar, R. O. (2013). Prevalence of somatoform disorders and medically unexplained symptoms in old age populations in comparison with younger age groups: A systematic review. *Ageing Research Reviews, 12*(1), 151-156.

Kang, T., & Chen, T. T. (2008). Performance of the generalized S- X2 item fit index for polytomous IRT models. *Journal of Educational Measurement, 45*(4), 391-406.

Korterink, J. J., Diederen, K., Benninga, M. A., & Tabbers, M. M. (2015). Epidemiology of pediatric functional abdominal pain disorders: A meta-analysis. *PloS One, 10*(5), e0126982.

Kroenke, K., & Mangelsdorff, A. D. (1989). Common symptoms in ambulatory care: Incidence, evaluation, therapy, and outcome. *The American Journal of Medicine, 86*(3), 262-266.

Kroenke, K., & Price, R. K. (1993). Symptoms in the community: Prevalence, classification, and psychiatric comorbidity. *Archives of Internal Medicine, 153*(21), 2474-2480.

Lee, S., & Suh, Y. (2018). Lord's wald test for detecting DIF in multidimensional IRT models: A comparison of two estimation approaches. *Journal of Educational Measurement, 55*(2), 328-353.

Magis, D., Béland, S., Tuerlinckx, F., & De Boeck, P. (2010). A general framework and an R package for the detection of dichotomous differential item functioning. *Behavior Research Methods, 42*(3), 847-862.

Michael, C. S., Callahan, C. M., Counsell, S. R., Westmoreland, G. R., Stump, T. E., & Kroenke, K. (2005). Physical symptoms as a predictor of health care use and mortality among older adults. *The American Journal of Medicine, 118*(3), 301-306.

Millsap, R. E., & Everson, H. T. (1993). Methodology review: Statistical approaches for assessing measurement bias. *Applied Psychological Measurement, 17*(4), 297-334.

Nikolov, V., & Petkova, M. (2010). Pain sensitivity among women with low estrogen levels. *Procedia-Social and Behavioral Sciences, 5*, 289-293.

Ockenburg, S. L. van, Bos, E. H., Jonge, P. de, Harst, P. van der, Gans, R. O. B., & Rosmalen, J. G. M. (2015). Stressful life events and leukocyte telomere attrition in adulthood: A prospective population-based cohort study. Psychological Medicine, 45(14), 2975–2984. https://doi.org/10.1017/S0033291715000914

Özdemir, B. (2015). A comparison of IRT-based methods for examining differential item functioning in TIMSS 2011 mathematics subtest. *Procedia-Social and Behavioral Sciences, 174,* 2075-2083.

Paap, M. C., Meijer, R. R., Van Bebber, J., Pedersen, G., Karterud, S., Hellem, F. M., et al. (2011). A study of the dimensionality and measurement precision of the SCL- 90- R using item response theory. *International Journal of Methods in Psychiatric Research, 20*(3), e39-e55.

Pinto-Sietsma, S., Mulder, J., Janssen, W. M., Hillege, H. L., de Zeeuw, D., & de Jong, P. E. (2000). Smoking is related to albuminuria and abnormal renal function in nondiabetic persons. *Annals of Internal Medicine, 133*(8), 585-591.

Rajindrajith, S., Zeevenhooven, J., Devanarayana, N. M., Perera, B. J. C., & Benninga, M. A. (2018). Functional abdominal pain disorders in children. *Expert Review of Gastroenterology & Hepatology, 12*(4), 369-390.

Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational Statistics, 4*(3), 207-230.

Reise, S. P., & Waller, N. G. (2009). Item response theory and clinical measurement. *Annual Review of Clinical Psychology, 5*, 27-48.

Rosmalen, J. G., Tak, L. M., & de Jonge, P. (2011). Empirical foundations for the diagnosis of somatization: Implications for DSM-5. *Psychological Medicine, 41*(6), 1133-1142.

Suh, Y. (2016). Effect size measures for differential item functioning in a multidimensional IRT model. *Journal of Educational Measurement, 53*(4), 403-430.

Sun, L., Zhang, W., Xu, Q., Wu, H., Jiao, C., & Chen, X. (2019). Estrogen modulation of visceral pain. *Journal of Zhejiang University-SCIENCE B, 20*(8), 628-636.

Tay, L., Meade, A. W., & Cao, M. (2015). An overview and practical guide to IRT measurement equivalence analysis. *Organizational Research Methods, 18*(1), 3-46.

Terluin, B., van der Wouden, J C, & de Vet, H C W. (2019). Measurement equivalence of the four-dimensional symptom questionnaire (4DSQ) in adolescents and emerging adults. *PloS One, 14*(8), e0221904.

Van der Ark, L Andries. (2007). Mokken scale analysis in R. *Journal of Statistical Software, 20*(11), 1-19.

van Driel, T., Hilderink, P., Hanssen, D., de Boer, P., Rosmalen, J., & Oude Voshaar, R. (2017). Assessment of somatization and medically unexplained symptoms in later life. *Assessment,* 1073191117721740.

van Eck van der Sluijs, Jonna, Ten Have, M., De Graaf, R., Rijnders, C. A. T., Van Marwijk, H. W., & van der Feltz-Cornelis, Christina M. (2018). Predictors of persistent medically unexplained physical symptoms: Findings from a general population study. *Frontiers in Psychiatry,* 9, 613.

Wall, M. M., Park, J. Y., & Moustaki, I. (2015). IRT modeling in the presence of zero-inflation with application to psychiatric disorder severity. *Applied Psychological Measurement, 39*(8), 583-597.

Zijlema, W. L., Stolk, R. P., Löwe, B., Rief, W., White, P. D., & Rosmalen, J. G. (2013). How to assess common somatic symptoms in large-scale studies: A systematic review of questionnaires. *Journal of Psychosomatic Research, 74*(6), 459-468.

Zwick, R. (2012). A review of ETS differential item functioning assessment procedures: Flagging rules, minimum sample size requirements, and criterion refinement. *ETS Research Report Series, 2012*(1), i-30.

Appendix A

Table S1. Items from the somatization section of the CIDI that were excluded from the statistical analyses.

Item	Reason for exclusion
Menstrual pain	Menstrual and reproductive health
Vomiting through pregnancy	Menstrual and reproductive health
Irregular menstrual period	Menstrual and reproductive health
Excessive menstrual bleeding	Menstrual and reproductive health
Sexual indifference	Menstrual and reproductive health
Pain during sexual intercourse	Menstrual and reproductive health
Unpleasant sexual intercourse	Menstrual and reproductive health
Other sexual problems	Menstrual and reproductive health
Pain during urination	Reported by less than ten participants
Urinary retention	Reported by less than ten participants
Burning sensation in genitals	Reported by less than ten participants
Vomiting (other than during pregnancy)	Reported by less than ten participants
Blindness	Reported by less than ten participants
Deafness	Reported by less than ten participants
Impaired coordination	Reported by less than ten participants
Paralysis	Reported by less than ten participants
Seizures	Reported by less than ten participants
Loss of consciousness other than fainting	Reported by less than ten participants
Dissociative symptoms (such as amnesia)	Reported by less than ten participants
Bad taste in mouth	Reported by less than ten participants
Diarrhea	Low factor loading and high SE in severity parameter
Aphonia	Low factor loading and high SE in severity parameter
Skin blotches or discoloration	Low factor loading and high SE in severity parameter
Frequent urination	Low factor loading and high SE in severity parameter

Appendix B

Table S2. Factor analysis and item fit of the 23 items of the CIDI somatization section

Name	***Factor loading***	**RMSEA S_X2**	**P-value S_X2**
Abdominal pain	0.67	0.02	0.21
Back pain	0.48	0.02	0.26
Joint pain	0.45	0.00	0.58
Pain in extremities	0.62	0.02	0.21
Chest pain	0.47	0.00	0.72
Headache	0.51	0.00	0.53
Pain additional sites (Other pain)	0.42	0.00	0.56
Nausea	0.65	0.00	0.67
Diarrhea	0.30	0.02	0.19
Feeling bloated or full of gas	0.51	0.00	0.46
Intolerance of several foods	0.62	0.04	0.01
Blurred vision	0.54	0.03	0.14
Impaired balance	0.69	0.00	0.47
Loss of touch or pain sensation	0.65	0.00	0.42
Aphonia	0.32	0.00	0.55
Dizziness	0.60	0.00	0.52
Double vision	0.54	0.03	0.11
Shortness of breath (without exertion)	0.61	0.00	0.44
Localized (muscle) weakness	0.78	0.00	0.89
Skin blotches or discoloration (pimples or spots)	0.27	0.04	0.05
Frequent urination	0.33	0.00	0.95
Numbness/tingling (feeling deaf or tingling)	0.56	0.03	0.10
Difficulty swallowing or lump in the throat	0.47	0.03	0.04
Proportion Var:	0.29		

Table S3. Factor analysis and item fit of the 19 items of the CIDI somatization section

Name	*Factor loading*	RMSEA S_X2	P-value S_X2
Abdominal pain	0.66	0.01	0.35
Back pain	0.50	0.00	0.63
Joint pain	0.49	0.00	0.55
Pain in extremities	0.63	0.00	0.58
Chest pain	0.48	0.00	0.59
Headache	0.51	0.00	0.61
Pain additional sites (Other pain)	0.41	0.00	0.99
Nausea (without vomiting)	0.66	0.03	0.07
Feeling bloated or full of gas	0.51	0.02	0.21
Intolerance of several foods	0.65	0.04	0.01
Blurred vision	0.52	0.03	0.11
Impaired balance	0.71	0.00	0.55
Loss of touch or pain sensation (numb arm or leg or other body part)	0.65	0.02	0.25
Dizziness (limp or dizzy)	0.59	0.00	0.51
Double vision	0.55	0.02	0.25
Shortness of breath (without exertion)	0.59	0.00	0.92
Localized weakness (muscle weakness)	0.80	0.00	0.82
Numbness	0.54	0.03	0.05
Difficulty swallowing or lump in the throat	0.47	0.01	0.29
Proportion Var:	0.34		

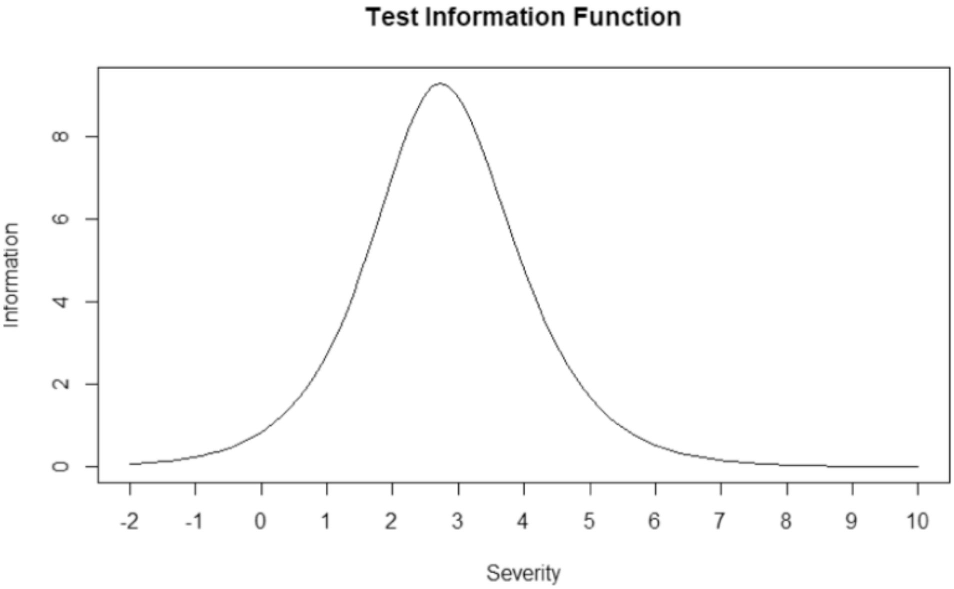
Appendix C

Table S4. The estimated item parameters from the 2PL in the 23 items selected from the CIDI somatization section.

No.	Item	Discrimination (α)	Severity (ß)	SE (α)	SE (ß)
19	*Localized (muscle) weakness*	2.39	2.55	0.46	0.23
8	*Nausea*	1.89	3.11	0.44	0.41
11	*Intolerance of several foods*	1.79	2.84	0.34	0.32
13	*Impaired balance*	1.71	2.75	0.31	0.30
14	*Loss of touch or pain sensation (numb arm or leg or other body part)*	1.64	2.66	0.29	0.29
1	*Abdominal pain*	1.59	2.07	0.24	0.19
18	*Shortness of breath (without exertion)*	1.57	3.00	0.31	0.38
22	*Numbness/tingling*	1.46	3.07	0.29	0.41
17	*Double vision*	1.45	3.64	0.38	0.66
10	*Feeling bloated or full of gas*	1.44	2.68	0.25	0.31
4	*Pain in extremities*	1.37	2.05	0.20	0.21
16	*Dizziness (limp or dizzy)*	1.21	2.57	0.20	0.31
23	*Difficulty swallowing or lump in the throat*	1.19	3.01	0.22	0.42
2	*Back pain*	1.19	2.20	0.18	0.25
6	*Headache*	1.18	2.21	0.18	0.25
5	*Chest pain*	1.15	3.06	0.22	0.44
3	*Joint pain*	1.15	2.10	0.17	0.24
12	*Blurred vision*	1.13	3.62	0.26	0.65
7	*Pain additional sites (Other pain)*	1.04	3.56	0.23	0.63
9	*Diarrhea*	0.85	4.99	0.30	1.49
20	*Skin blotches or discoloration*	0.82	5.10	0.29	1.56
21	*Frequent urination*	0.66	5.69	0.26	2.00
15	*Aphonia*	0.62	6.22	0.27	2.49

Appendix D

Figure S1. Test Information Function (TIF) of the 19 selected items from the somatization scale of the CIDI



Appendix E

Table S5. Zero Inflated item parameters

No.	Item	*a*	SE (*a*)	ß	SE (ß)
1	*Abdominal pain*	1.540	0.236	2.114	0.206
2	*Back pain*	1.197	0.189	2.191	0.256
3	*Joint pain*	1.225	0.196	2.025	0.235
4	*Pain in extremities*	1.400	0.217	2.025	0.210
5	*Chest pain*	1.114	0.209	3.139	0.455
6	*Headache*	1.162	0.182	2.253	0.263
7	*Pain additional sites (Other pain)*	1.026	0.230	3.584	0.640
8	*Nausea*	1.859	0.416	3.150	0.416
9	*Feeling bloated or full of gas*	1.419	0.236	2.710	0.313
10	*Intolerance of several foods*	1.697	0.370	2.931	0.380
11	*Blurred vision*	1.080	0.314	3.747	0.838
12	*Impaired balance*	1.701	0.345	2.759	0.330
13	*Loss of touch or pain sensation*	1.625	0.280	2.673	0.290
14	*Dizziness*	1.189	0.219	2.617	0.346
15	*Double vision*	1.427	0.392	3.686	0.704
16	*Shortness of breath (without exertion)*	1.463	0.327	3.128	0.464
17	*Localized (muscle) weakness*	2.319	0.438	2.585	0.238
18	*Numbness/tingling*	1.415	0.293	3.133	0.444
19	*Difficulty swallowing or lump in the throat*	1.169	0.218	3.049	0.427

Note: The ZI model classified the whole sample in the pathological group. Table S5 shows the estimated item parameters from the 2PL ZI model for the 19 selected items in the pathological group.