ORIGINAL RESEARCH

Construct validity of the Depression and Somatic Symptoms Scale: evaluation by Mokken scale analysis

Ya-Hsin Chou[1]
Chin-Pang Lee[1,2]
Chia-Yih Liu[1,2]
Ching-I Hung[1,2]

[1]Department of Psychiatry, Chang-Gung Memorial Hospital at Linkou,
[2]School of Medicine, Chang Gung University College of Medicine, Taoyuan, Taiwan

Objective: Previous studies of the Depression and Somatic Symptoms Scale (DSSS), a free scale, have been based on the classical test theory, and the construct validity and dimensionality of the DSSS are as yet uncertain. The aim of this study was to use Mokken scale analysis (MSA) to assess the dimensionality of the DSSS.

Methods: A sample of 214 psychiatric outpatients with mood and anxiety disorders were enrolled at a medical center in Taiwan (age: mean [SD] =38.3 [10.5] years; 63.1% female) and asked to complete the DSSS. MSA was used to assess the dimensionality of the DSSS.

Results: All 22 items of the DSSS formed a moderate unidimensional scale (H_s=0.403), supporting its construct validity. The DSSS was divided into 4 subscales (H_s ranged from 0.35 to 0.67), including a general somatic scale (GSS), melancholic scale (MS), muscular pain scale (MPS), and chest symptom scale (CSS). The GSS is a weak reliable Mokken scale; the other 3 scales are strong reliable Mokken scales.

Conclusion: The DSSS is a psychometrically sound measure of depression and somatic symptoms in adult psychiatric outpatients with depression or anxiety. The summed score of the DSSS and its 4 subscales are valid statistics. Further research is required for replication of the 4 subscales of the DSSS.

Keywords: depression, somatization, Mokken scale analysis, item response theory, construct validity

Introduction

Somatic symptoms among patients with depression are important;[1,2] they may confound the diagnosis of depression,[3] are often residual symptoms of depression, and might increase the risk of relapse.[4,5] Somatic symptoms in patients with major depressive disorder are also associated with a negative treatment outcome and a poor quality of life.[6,7] Although somatic symptoms have a negative impact on depression, most conventional scales for depression do not include appropriate items related to somatic symptoms, which may hinder the monitoring, evaluation, and quantification of somatic symptoms.[8] For these reasons, the Depression and Somatic Symptoms Scale (DSSS) was developed.[8,9] The DSSS is a free scale that can be used to evaluate depression and somatic severity simultaneously. The DSSS and its subscales are significantly correlated with the Hamilton Depression Rating Scale (HAMD) as well as the mental and physical subscale scores of a health-related quality of life scale.[9] It is also sensitive to pharmacotherapy.[9] Moreover, the predictive ability of the DSSS for the prognosis of depression is not inferior to that of the HAMD.[10] Furthermore, previous studies

Correspondence: Ching-I Hung
Department of Psychiatry, Chang-Gung Memorial Hospital–Linkou, 5, Fuxing Street, Taoyuan, Guishan District 333, Taiwan
Tel +886 3 328 1200
Fax +886 3 328 0267
Email a36250@cgmh.org.tw

205

have demonstrated that the DSSS has not only a good reliability, but also acceptable convergent, factorial, and discriminative validities.[9–11]

Summed scores assume that all items are equally correlated with the measured underlying construct; in addition, the point intervals are equal on the scale. However, these assumptions are not always true.[12] For example, the items on the DSSS are not linear and continuous measurements, which means that calculating item scores might be meaningless.

There are 2 different approaches for evaluation of the psychometric properties of rating scales: the classical test theory (CTT) and the item response theory (IRT).[12] The limitations of CTT include the summed score problems and sample-dependent statistics, which may result in different psychometric properties when based on different samples.[13,14]

IRT provides item-level statistics that are not influenced by differences between samples.[12–14] IRT assumes that scale items can be ordered along levels of a latent trait, with item "difficulty" demonstrating whether items are difficult (rare) or less difficult (common).[15]

Mokken scale analysis (MSA) is a nonparametric form of IRT derived from Guttman scaling.[16–18] On a Guttman scale, a single response can be used to predict responses to all items on the scale.[16] In the field of health construct measurements, MSA can be used to scrutinize the appropriateness and performance of the measurements.[19] Being a nonparametric analytical method, MSA is robust according to the underlying distribution of the data, avoiding the methodological limitations of previous studies.

To the best of our knowledge, no study has used MSA to examine the psychometric properties of the DSSS. Therefore, the purposes of this study were as follows: 1) to examine the construct validity of the DSSS; 2) to assess the dimensionality of the DSSS; and 3) to examine the item hierarchy of the DSSS and determine whether the DSSS has an invariant item ordering (IIO) property, which means that the items of a scale have the same difficulty ordering.

Methods

Participants

This was a secondary analysis of data obtained for a cross-sectional study conducted at the Chang-Gung Memorial Hospital, Linkou, which is a tertiary medical center in Northern Taiwan. Participants were enrolled in the psychiatric outpatient clinic between September 2007 and August 2009. Three inclusion criteria were established: (1) aged between 20 and 60 years; (2) consecutive outpatients with depression or anxiety; (3) patients who had not taken antidepressants

within the previous 4 weeks. Four exclusion criteria were as follows: 1) a history of substance dependence or abuse without full remission in the index month; 2) psychotic disorders, such as schizophrenia, delusional disorder, and other psychotic disorders; 3) dementia, delirium, mental retardation, and mental disorders due to general medical conditions; and 4) patients with psychotic symptoms, catatonic features, severe psychomotor retardation, or a current manic episode in the previous month, which may cause difficulty in completing self-administered scales or cooperating with the study process. In total, we recruited 214 participants (age: mean [SD] =38.3 [10.5] years; 63.1% female) in this study. Table 1 shows the demographic characteristics of the sample.

Procedures

This study was approved by the Institutional Review Board of Chang-Gung Memorial Hospital. Participants provided explicit written informed consent to participate in the study. Participants were interviewed by a senior board-certified psychiatrist (C-I Hung), and psychiatric diagnoses were made according to the Structured Clinical Interview for *Diagnostic and Statistical Manual of Mental Disorders* (DSM)-IV–text revision (TR) Axis I Disorders (SCID-I).[20] Patients also completed the DSSS during their intake visits. All information was protected by delinking identifying information from main data sets and sources, and the data were available only to the investigators.

The DSSS

The DSSS is a self-administered scale. It is composed of 12 items in the Depression Subscale (DS) and 10 items in the Somatic Subscale (SS), the latter including 5 pain and 5 nonpain somatic symptoms.[9] Each question is rated on a 4-point Likert scale from 0 to 3, and the total score ranges from 0 to 66.

Statistical analysis

All analyses were conducted in R version 3.3.0 (R Foundation for Statistical Computing, Vienna, Austria). The *P*-values

Table 1 Demographic and clinical characteristics of the sample (N=214)

Characteristics	Statistics	Range
Age, mean (SD), years	38.3 (10.5)	20–58
Female, n (%)	135 (63.1)	
Education, mean (SD), years	12.5 (2.9)	0–18
Paid employment, n (%)	142 (66.4)	
Married, n (%)	121 (56.5)	
With current major depressive episode, n (%)	95 (44.4)	

were 2 tailed, and the significance level was set at 0.05. This method was also applied in a previous publication.[21]

MSA

MSA consists of 2 parts: 1) an automated item selection algorithm that partitions a set of ordinal items into Mokken scales, possibly leaving some items unselected; and 2) methods to investigate assumptions of nonparametric IRT models. The underlying assumptions of Mokken models are unidimensionality, local independence, and latent monotonicity.[16] Unidimensionality means that the scale under consideration measures a single latent trait θ. Local independence means that one's response to an item is not affected by one's responses to the other items in the same test.[22] Latent monotonicity assumes that for each item, the probability of a particular response level is a monotonically nondecreasing function of the latent trait θ.[19]

The 2 main Mokken models are as follows: the monotone homogeneity model (MHM) and the double monotonicity model (DMM).[16] The MHM assumes unidimensionality, monotonicity, and local independence of the items within a scale. If these assumptions are met, respondents can be ordered according to the simple sum score of items.[16,23,24] In addition to the features of the MHM, the DMM has the property of IIO.[22]

IIO refers to items that have the same "difficulty ordering" irrespective of the value of the latent trait.[23–25] The ordering of items is based on item difficulty and shows whether items are difficult (rare) or less difficult (common). IIO is therefore important in order to establish a scale hierarchy that is replicable across samples.[19] The IIO property can be expected to hold in any subgroup from the same population and thus is considered to be in some sense "person free".

MSA was conducted using the *mokken* package.[23,24] First, the assumption of unidimensionality was checked using Loevinger's scalability coefficients H.[16,23,26,27] Loevinger's scalability coefficients comprise 3 indexes: item-pair (H_{ij}), item (H_i), and scale (H_s) scalability coefficients. If $H_s=1$, then the scale is a perfect Guttman scale. The MHM implies that these 3 indexes should be between 0 and 1; higher H values hint a higher item discrimination power. The rules of thumb for Loevinger's scalability coefficients are described as follows:

A scale is considered weak if $0.3 \leq H < 0.4$, moderate if $0.4 \leq H < 0.5$, and strong if $H \geq 0.5$.[16,28]

First, an automated item selection procedure with a genetic algorithm was used to automatically search for a set of items with established cutoffs of H_i (lower bound c),[19,29] which started from 0.30, increased subsequently

in steps of 0.05, and stopped at 0.60.[16,17,23,27] By executing multiple analyses in this sequence, the strategy can provide important insights into the relationships among items. A scale that contains fewer than 4 items was considered unfavorable.[19] Second, to identify positive and negative dependence in the MHM, the assumption of local independence was checked using a conditional association procedure.[29–31] Locally dependent items were removed. Third, the assumption of monotonicity was checked using item-rest regression.[16,17,23,27,32] If there were violations of monotonicity, then their seriousness could be assessed via consideration of the *crit* statistic. Items with a *crit* statistic ≤ 40 can be safely included in any Mokken scale.[27] Fourth, the method *restscore* was used to check the assumption of nonintersection.[15,23,27] If there were violations of nonintersection, then their seriousness could be assessed via consideration of the *crit* statistic: *crit* values < 40 indicate no serious violation; *crit* values between 40 and 80 indicate minor violations; and *crit* values > 80 indicate serious violations.[27] Finally, the method *manifest IIO* was used to check IIO.[24] We used backward selection to remove items violating IIO. If there were an equal number of violations for ≥ 2 items, the item with the lowest scalability was removed. Subsequently, these selected items were checked for accuracy in IIO using the statistic H^T.[33] If manifest IIO holds, then $0.3 < H^T \leq 0.4$ was interpreted as a weak ordering, $0.4 < H^T \leq 0.5$ as a moderate ordering, and $H^T > 0.5$ as a strong ordering. Molenaar and Sijtsma's ρ was calculated to measure the reliability of the Mokken scales.[27] A scale with $\rho > 0.7$ was considered highly reliable.[34]

Results

The demographic and clinical characteristics of the sample are shown in Table 1. The age range of the subjects was between 20 and 58 years; the range of years of education was between 0 and 18 years.

Results of MSA

All of the corresponding item scalability coefficients (H_i) of the 22 items were larger than 0.3, and the scale scalability H_s of the DSSS was 0.40, representing a moderate Guttman scale. The reliability of the DSSS was excellent ($\rho=0.92$). There was no violation of local independence for any of the 22 items. The items also demonstrated good monotonicity without violation of monotonicity for any of the DSSS items. Among the 22 items of the DSSS, only Item 8 had serious violations of the assumption of nonintersection (*crit* =83); Item 8 was therefore removed. The backward item selection

Table 2 Number of items in each subscale during the iterative automatic item selection procedures

Lower bound, c^a	0.30	0.35	0.40	0.45	0.50	0.55	0.60
Subscale 1	22	3	2	3	6	7	9
Subscale 2	0	19	14	13	8	7	5
Subscale 3	0	0	2	7	4	4	2
Subscale 4	0	0	2	2	4	4	4
Subscale 5	0	0	2	0	0	0	2

Notes: [a]Cutoff values of H_i. H_i item scalability coefficient.

procedure for the other 21 items revealed that Items 6 and 7 should be removed. The remaining 19 items formed a reliable moderate Mokken scale without IIO (H_s=0.41, H^T=0.17, ρ=0.92), which fit MHM but still did not fit DMM.

Thus, we retained all 22 items to further explore the dimensions of the DSSS, and iterative automated item selection procedures were executed with lower bound c, which started from 0.30 and increased to 0.60 in increments of 0.05 (Table 2). Except for c=0.50 or 0.55, all other solutions included subscales that contained only 2 items. The solutions of c=0.50 and 0.55 were similar. So, the final solution to the Mokken scaling was set at c=0.55. Regarding the dimensionality of the DSSS, 4 reliable Mokken scales were identified (Table 3).

Subscales

The first subscale, the general somatic scale (GSS), consisted of 7 items mainly concerning vegetative symptoms, headache, dizziness, and anxious/irritable mood. These symptoms are nonspecific to depression and are present in various physical and mental disorders. The GSS was a weak reliable Mokken scale (H_s=0.35, ρ=0.76). There was no violation of local independence for any of the 7 items; the items also all demonstrated good monotonicity without violation. All 7 items had either no serious violations or minor violations of the assumption of nonintersection. All 7 items were retained in the backward item selection procedure. The GSS did not demonstrate the IIO property (H^T=0.19).

The second subscale, the melancholic scale (MS), consisted of 7 items and corresponded to the core symptoms of depression, as listed in the DSM-5 or International Statistical Classification of Diseases and Related Health Problems (ICD)-10 criteria of major depressive disorder. The MS was a strong Mokken scale (H_s=0.60, ρ=0.90). There was no violation of local independence for any of the 7 items, which also all demonstrated good monotonicity without any violation. All 7 items had no serious violations of the assumption of nonintersection. Item 10 was removed in the

Table 3 Four Mokken scales of the Depression and Somatic Symptoms Scale

Subscale	Item number	Label	Mean	H_i	Critical violation[a]
General somatic scale (GSS): H_s=0.35; ρ=0.76; H^T=0.19	4	Insomnia	1.97	0.32	0
	16	Anxious or nervous	1.85	0.39	0
	6	Irritable mood	1.73	0.36	0
	9	Dizziness	1.36	0.41	0
	14	Loss of interest in sex	1.33	0.29	0
	1	Headache	1.15	0.35	0
	22	Decreased appetite or loss of appetite	1.12	0.32	0
Melancholic scale (MS): H_s=0.60; ρ=0.90; H^T=0.27[b]	8	Unable to feel happy or decreased ability to feel happy	1.86	0.66	0
	10[c]	Depressed mood or tearful	1.76	0.63	201
	18	Unable to concentrate	1.73	0.57	122
	21	Fatigue or loss of energy	1.72	0.55	115
	2	Loss of interest in daily or leisure activities	1.55	0.63	0
	12	Feelings of self-reproach or guilt	1.32	0.58	0
	20	Thoughts of death or suicidal ideas	0.99	0.59	0
Muscular pain scale (MPS): H_s=0.67; ρ=0.87; H^T=0.41[b]	13	Neck or shoulder pain (or soreness)	1.69	0.66	0
	5	Muscle tension	1.58	0.66	0
	17	Soreness in more than half of the body's muscles	1.09	0.72	140
	7[c]	Back pain	1.09	0.65	144
Chest symptom scale (CSS): H_s=0.57; ρ=0.84; H^T=0.35	3	Tightness in the chest	1.64	0.61	0
	19	Palpitations or increased heart rate	1.38	0.60	0
	15	Shortness of breath or difficulty breathing	1.11	0.64	0
	11	Chest pain	0.87	0.57	0

Notes: [a]The number of significant violations of invariant item ordering (IIO) was calculated using the method of manifest IIO (MIIO). If 2 or more items had the maximum number of violations, then the item producing the lowest value of Loevinger's H was removed. If there were an equal number of violations for 2 or more items, then the item with the lowest scalability (H_i) was removed. [b]Coefficient H^T for the remaining items. [c]The items were removed by the backward item selection procedure. H_i item scalability coefficient; H_s scale scalability coefficient. $0.3<H^T≤0.4$ was interpreted as a weak IIO, $0.4<H^T≤0.5$ as a moderate IIO; a scale with $\rho>0.7$ was considered highly reliable scale.

backward item selection procedure. The remaining 6 items formed a strong Mokken scale and did not demonstrate the IIO property (H_s=0.59, H^T=0.27, ρ=0.88).

The third subscale, the muscular pain scale (MPS), consisted of 4 items related to painful symptoms. The MPS was a reliable strong Mokken scale (H_s=0.67, ρ=0.87). There was no violation of local independence for any of the 4 items, and all items also demonstrated good monotonicity without any violation. All 4 items had no serious violations of the assumption of nonintersection. Item 7 was removed in the backward item selection procedure. The remaining 3 items formed a strong Mokken scale with a moderate IIO property (H_s=0.70, H^T=0.41, ρ=0.85).

The final subscale, the chest symptom scale (CSS), consisted of 4 items mainly concerning cardiorespiratory symptoms. This was a reliable strong Mokken scale (H_s=0.57, ρ=0.84). There was no violation of local independence for any of the 4 items, and all items also demonstrated good monotonicity without any violation. All 4 items had no serious violations of the assumption of nonintersection. All 4 items were retained in the backward item selection procedure. The CSS had a weak IIO property (H^T=0.35).

Discussion

Our study demonstrated that the 22 items of the DSSS formed a reliable moderate unidimensional scale, which met the criteria for a MHM. The simple sum score of these 22 items within the scale can be used for ordinal personal measurement of depression in psychiatric adult outpatients with depression or anxiety. This finding supports the construct validity of the DSSS. The finding of unidimensionality was different from that of the validation study of the DSSS, which, using exploratory factor analysis (EFA), found a 2-factor solution.[9] As an exploratory study in nature, our study provides evidence to show that the sum scores of the DSSS and its 4 subscales are valid statistics. This study does not support the original bidimensional scoring. It is noteworthy that, being a parametric approach, the EFA is not a robust statistical method when its underlying assumption is not fulfilled. For example, the EFA requires the assumption of a normal distribution, which is frequently unrealistic with Likert-type scale data.[35] Further research using confirmatory factor analysis is required to verify whether the DSSS is uni- or bidimensional.

Using Mokken analysis, we identified 4 reliable Mokken scales among the 22 items of the DSSS. All 4 subscales of the DSSS met the MHM criteria, and 2 (the MPS and the CSS) further met the DHM criteria, indicating that their sum scores were sufficiently useful statistics. However, the GSS had a

H_s of 0.35, whereas the other 3 subscales obtained H_s values of >0.50. Regarding item reduction, these 3 subscales (the MS, the CSS, and the MPS) could form a shorter scale with better scalability than the original DSSS.

Our findings suggested that the MS might be particularly clinically relevant. The items in the MS were similar to those in the 6-item HAMD (HAMD-6), consisting of depressed mood, guilt feelings, work and interests, psychomotor retardation, psychic anxiety, and tiredness/pain. Recent studies have shown that the HAMD-6 is a strong unidimensional scale and is more suitable as an outcome measure than the traditional 17-item HAMD (HAMD-17).[35–37] In this regard, the MS might be used along with the HAMD-6 when evaluating the effect of antidepressant therapy.[35,37]

The GSS seemed less useful in measuring the severity of depression. The 7 items of the GSS cover to some degree the neurovegetative symptoms of depression such as appetite, sleep, sex, and anxiety symptoms. These items are traditionally considered to be atypical depressive features.[38] Regarding measurement of atypical depressive features, they are covered much more accurately by the Inventory of Depressive Symptomatology.[39]

Our study demonstrated IIO in the CSS and the MPS. Such item hierarchy will assist clinicians to more efficiently assess somatic symptoms. For example, Items 3 (chest tightness) and 13 (neck or shoulder pain) could be used as screening questions related to somatic symptoms. Patients who achieved high scores on Items 11 (chest pain) and 17 (soreness in more than half of the body's muscles) probably suffered significant chest discomfort and muscular pain.

Our study had several strengths. We used a validated questionnaire to measure depression and had a full range of data available for all participants. We applied MSA to assess the construct validity and dimensionality of the DSSS.

Our study also had several limitations. First, this study was a cross-sectional analysis of a single-site sample. Future research should attempt to replicate these findings to determine whether the 4 subscales of the DSSS are reasonable and useful. Second, our sample only contained adult psychiatric outpatients with depression or anxiety. Further research is needed to clarify the utility of the DSSS for other psychiatric disorders. Third, this study used data that were not obtained specifically for MSA. Therefore, the study would merit replication with a larger sample size.

Conclusion

The DSSS is a psychometrically sound measure of depression in adult psychiatric outpatients with depression or anxiety. The DSSS may further be divided into 4 Mokken scales,

2 of which had IIO properties. Future research should be performed to attempt to replicate the dimensionality and to determine whether similar items demonstrate IIO.

Acknowledgments

Disclosure

The authors report no conflicts of interest in this work.

References

1. Kapfhammer HP. Somatic symptoms in depression. *Dialogues Clin Neurosci*. 2006;8(2):227–239.
2. Bagayogo IP, Interian A, Escobar JI. Transcultural aspects of somatic symptoms in the context of depressive disorders. *Adv Psychosom Med*. 2013;33:64–74.
3. Croicu C, Chwastiak L, Katon W. Approach to the patient with multiple somatic symptoms. *Med Clin North Am*. 2014;98(5):1079–1095.
4. Fava M. Depression with physical symptoms: treating to remission. *J Clin Psychiatry*. 2003;64(7):24–28.
5. Greden JF. Physical symptoms of depression: unmet needs. *J Clin Psychiatry*. 2003;64(7):5–11.
6. Bair MJ, Robinson RL, Eckert GJ, Stang PE, Croghan TW, Kroenke K. Impact of pain on depression treatment response in primary care. *Psychosom Med*. 2004;66(1):17–22.
7. Raison CL, Hale MW, Williams LE, Wager TD, Lowry CA. Somatic influences on subjective well-being and affective disorders: the convergence of thermosensory and central serotonergic systems. *Front Psychol*. 2015;5:1580.
8. Hung CI, Weng LJ, Su YJ, Liu CY. Preliminary study of a scale measuring depression and somatic symptoms. *Psychol Rep*. 2006;99(2):379–389.
9. Hung CI, Weng LJ, Su YJ, Liu CY. Depression and somatic symptoms scale: a new scale with both depression and somatic symptoms emphasized. *Psychiatry Clin Neurosci*. 2006;60(6):700–708.
10. Hung CI, Liu CY, Wang SJ, Juang YY, Yang CH. Somatic symptoms: an important index in predicting the outcome of depression at six-month and two-year follow-up points among outpatients with major depressive disorder. *J Affect Disord*. 2010;125(1–3):134–140.
11. Hung CI, Liu CY, Wang SJ, Yao YC, Yang CH. The cut-off points of the Depression and Somatic Symptoms Scale and the Hospital Anxiety and Depression Scale in detecting non-full remission and a current major depressive episode. *Int J Psychiatry Clin Pract*. 2012;16(1):33–40.
12. Streiner D, Norman G. *Health Measurement Scales: A Practical Guide to Their Development and Use*. 4th ed. New York, NY: Oxford University Press; 2008.
13. Amin L, Rosenbaum P, Barr R, et al. Rasch analysis of the PedsQL: an increased understanding of the properties of a rating scale. *J Clin Epidemiol*. 2012;65(10):1117–1123.
14. Chang C-C, Su J-A, Tsai C-S, Yen C-F, Liu J-H, Lin C-Y. Rasch analysis suggested three unidimensional domains for Affiliate Stigma Scale: additional psychometric evaluation. *J Clin Epidemiol*. 2015; 68(6):674–683.
15. Embretson S, Reise S. *Item Response Theory for Psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates; 2000.
16. Mokken R. *A Theory and Procedure of Scale Analysis*. Berlin: De Gruyter; 1971.
17. Sijtsma K, Molenaar I. *Introduction to Nonparametric Item Response Theory*. Thousand Oaks, CA: Sage; 2002.
18. Guttman L. *The Basis for Scalogram Analysis*. Princeton: Princeton University Press; 1950.
19. Stochl J, Jones PB, Croudace TJ. Mokken scale analysis of mental health and well-being questionnaire item responses: a non-parametric IRT method in empirical research for applied health researchers. *BMC Med Res Methodol*. 2012;12(1):1–16.
20. First MB, Spitzer RL, Gibbon M, Williams JBW. *Structured Clinical Interview for DSM-IV-TR Axis I Disorders, Research Version, Patient Edition (SCID-I/P)*. New York, NY: Biometrics Research, New York State Psychiatric Institute; 2002.
21. Lee CP, Chu CL, Chen Y, Jiang KH, Chen JL, Chen CY. The Chinese Version of the Gotland Male Depression Scale (GMDS): Mokken scaling. *J Affect Disord*. 2015;186:48–52.
22. Sijtsma K, Junker BW. A survey of theory and methods of invariant item ordering. *Br J Math Stat Psychol*. 1996;49(pt 1):79–105.
23. van der Ark LA. Mokken Scale Analysis in R. *J Stat Softw*. 2007; 20(11):1–19.
24. van der Ark LA. New developments in Mokken Scale Analysis in R. *J Stat Softw*. 2012;48(5):1–27.
25. Doyle F, Watson R, Morgan K, McBride O. A hierarchy of distress and invariant item ordering in the General Health Questionnaire-12. *J Affect Disord*. 2012;139(1):85–88.
26. Loevinger J. The technic of homogeneous tests compared with some aspects of scale analysis and factor analysis. *Psychol Bull*. 1948;45(6): 507–529.
27. Molenaar I, Sijtsma K. *User's Manual MSP5 for Windows. Software Manual*. Groningen: IEC ProGAMMA; 2000.
28. Galindo-Garre F, Hidalgo MD, Guilera G, Pino O, Rojo JE, Gomez-Benito J. Modeling the World Health Organization Disability Assessment Schedule II using non-parametric item response models. *Int J Methods Psychiatr Res*. 2015;24(1):1–10.
29. Straat JH, Van der Ark LA, Sijtsma K. Comparing optimization algorithms for item selection in Mokken scale analysis. *J Classif*. 2013; 30(1):75–99.
30. Holland PW, Rosenbaum PR. Conditional association and unidimensionality in monotone latent variable models. *Ann Stat*. 1986;14: 1523–1543.
31. Straat JH. Using Scalability Coefficients and Conditional Association to Assess Monotone Homogeneity. Ridderkerk: Ridderprint; 2012.
32. Junker BW, Sijtsma K. Latent and manifest monotonicity in item response models. *Appl Psychol Meas*. 2000;24(1):65–81. Available from https://pure.uvt.nl/portal/files/1459839/Straat_using_23-11-2012.pdf. Accessed on June 20, 2016.
33. Ligtvoet R, van der Ark LA, te Marvelde JM, Sijtsma K. Investigating an invariant item ordering for polytomously scored items. *Educ Psychol Meas*. 2010;70(4):578–595.
34. Sijtsma K, Molenaar IW. Reliability of test scores in nonparametric item response theory. *Psychometrika*. 1987;52(1):79–97.
35. Bech P, Fava M, Trivedi MH, Wisniewski SR, Rush AJ. Factor structure and dimensionality of the two depression scales in STAR*D using level 1 datasets. *J Affect Disord*. 2011;132(3):396–400.
36. Licht RW, Qvitzau S, Allerup P, Bech P. Validation of the Bech-Rafaelsen Melancholia Scale and the Hamilton Depression Scale in patients with major depression; is the total score a valid measure of illness severity? *Acta Psychiatr Scand*. 2005;111(2):144–149.
37. Østergaard SD, Bech P, Miskowiak KW. Fewer study participants needed to demonstrate superior antidepressant efficacy when using the Hamilton melancholia subscale (HAM-D6) as outcome measure. *J Affect Disord*. 2016;190:842–845.
38. Angst J, Gamma A, Benazzi F, et al. Atypical depressive syndromes in varying definitions. *Eur Arch Psychiatry Clin Neurosci*. 2006;256(1): 44–54.
39. Rush AJ, Giles DE, Schlesser MA, Fulton CL, Weissenburger J, Burns C. The inventory for depressive symptomatology (IDS): preliminary findings. *Psychiatry Res*. 1986;18(1):65–87.