Heliyon 5 (2019) e02300

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.heliyon.com

Evaluating hierarchical items of the geriatric depression scale through factor analysis and item response theory

Nahathai Wongpakaran *, Tinakon Wongpakaran, Pimolpun Kuntawong

Department of Psychiatry, Faculty of Medicine, Chiang Mai University, Thailand

ARTICLE INFO

Keywords:
Clinical psychology
Psychology
Psychiatry
Rasch analysis
Mokken analysis
Elderly
Multiple indicator multiple cause model
GDS
Confirmatory factor analysis
Short

ABSTRACT

Background: Geriatric depression scale (GDS) is a common screening tool for measuring depression among older adults. It employs a multi-factor structure and some differential item functioning (DIF) allowing different versions of GDS across cultures. The present study aimed to identify the short version of the hierarchical scale of GDS in which all items comprised the invariant item ordering, and items without DIF.

Methods: Participants and Measurement: A total of 803 participants, 70% female, with a mean age of 69.24 years (SD = 6.88) were enrolled from three geriatric units of tertiary care hospitals. All completed the 15-item GDS. Three methods of confirmatory factor analysis (CFA) with multiple indicators, multiple cause model, Mokken analysis and Rasch analysis were applied.

Results: Item 9 (prefer to stay at home) showed poor discriminatory power among all three methods. After removing DIF items due to sex and age, nine items remained suitable for the shortened version by CFA. When Mokken and Rasch analysis were applied, only six items remained for the hierarchical scale. Compared with other related shortened version of GDS, the new GDS-6 proved to have a comparable ability to detect depression as did the original 15-item GDS.

Limitation: The new GDS-6 needs to be investigated for test-retest reliability to ensure temporal stability of the scale. This cross-sectional analysis needs replication.

Conclusion: The GDS-6 derived from IRT had measurement properties and met criteria related to unidimensionality and ability to separate levels of depression. It was shown to be equal to or better in predicting performance compared with the original 15-item GDS.

1. Introduction

Geriatric depression is a common psychiatric disorder. Its prevalence ranges from 10 to 15% of older adults in the general population (Kok and Reynolds, 2017), and 27% in the outpatient setting (Wang et al., 2017). In Thailand, which has now become an aging society, we found 23% receiving long term care (Wongpakaran and Wongpakaran, 2012).

To screen for depression among the elderly in epidemiological studies, the Geriatric Depression Scale (GDS), developed by Yesavage et al. (1982), has become one of the common instruments for clinical use with older adults. The 15-item version of the GDS is widely used and has been translated to many languages (Kim et al., 2013; Mitchell et al., 2010a). In investigating its psychometric properties, researchers found the GDS to be a reliable and valid instrument (Chau et al., 2006; Friedman et al., 2005; Incalzi et al., 2003; Malakouti et al., 2006). Even though the GDS aims to screen for depression, in terms of facture structure, whether the GDS has a unidimensional or multidimensional construct remains unclear. If it has multidimensionality, then whether it has sufficient unidimensionality that individuals can be reliably measured using the sum and cut-off scores for screening remains unsure. Related studies have found clear evidence of language differences in the factor structure of the GDS as well as the possibility of diverse cultural groups (Kim et al., 2013). Consequently, false estimation of depressive symptoms may be produced. When total scores are not unidimensional, they are, as stated by Thurstone (1931), technically invalid because their meaning is uncertain because what the scores represent remains unclear. Also unclear is when two individuals with the same score can be considered comparable. The comparisons of mean GDS can be problematic when the scale has different dimensionalities. In this review, Kim et al. suggested making some adjustments, such as removing some culturally or linguistically biased items.

In addition, the GDS faces a problem of differential item function

* Corresponding author.
E-mail address: nahathai.wongpakaran@cmu.ac.th (N. Wongpakaran).

https://doi.org/10.1016/j.heliyon.2019.e02300
Received 5 May 2019; Received in revised form 22 May 2019; Accepted 9 August 2019

(DIF) which indicated that the items in the scale are responded to differentially by distinct groups of participants, leading to item multidimensionality and undermining the construct validity.

The DIF items of GDS found were due to sex, age, ethnic, language, cognitive impairment status or clinical setting, such as residing in a nursing home (Broekman et al., 2008; Chiang et al., 2009; Kim et al., 2013; Midden and Mast, 2017; Mitchell et al., 2010b; Wongpakaran et al., 2013). Even though DIF items have been explored among elderly subjects, some studies have revealed contradicting results regarding age, level of education and cognitive impairment for which the discordance may be contributing to other factors such as culture, sample size or method of analysis (Tang et al., 2005).

Attempts have been made to eliminate the DIF as well as to shorten the 15-tem GDS. In addition to an attempt to acquire robust items in the scale, many researchers are motivated to shorten the GDS to increase compliance among older adults by endeavoring to take out some problematic items, while maintaining accuracy for screening. Most investigators have used Classic Test Theory (CTT) or True Score Theory, e.g., factor analysis, by removing items with low or poor loading coefficients as well as DIF items to reduce the original version to a shorter one, e.g., the GDS-10, GDS-7, GDS-5, GDS-4, and even GDS-1 (Pocklington et al., 2016), hoping that all the items of the scale measure only one construct (unidimensionality), while retaining the benefit of a shortened scale.

Despite the numerous GDS revisions made to remedy the original flaws or to modify it to fit a particular population, the problem remains when the new version was applied in another sample with a different language and culture. The method of analysis, based on CTT, clearly yielded different results especially on DIF, even when conducted on the same sample. Notably, most investigations used factor analysis to develop (shorten) a new scale, which has some limitations, for example, estimates of item difficulty are group dependent (Zanon et al., 2016). In comparison between CTT and IRT in the same test, a study has shown the advantages of using IRT approaches over CTT on estimating item difficulties, internal consistencies and standard errors. In CTT, investigators may not only rely on previous reliability estimates but to estimate their own and report noted differences, while for the IRT both person and item reliability were evaluated providing more stable results (Magno, 2009).

In addition, CTT relies largely on the principle of correlation. As expected, items show high factor loadings and contribute to reliability through high item-intercorrelation. CTT not only encourages the recruitment of similar items but also eliminates items with lower correlation with other terms in the scale, e.g., difficult or easy to endorse items. In comparison, Rasch modeling, a kind of IRT, emphasizes investigating the entire breadth of the construct, not just high correlation items (Wright and Stone, 1979). The similarity of wording rather than the relationship of items with the construct can be misleading in that the scale is superior in quality, albeit with inflated internal consistency indices like Cronbach's alpha (Steinberg and Thissen, 1996). In comparison, the Rasch Model is based on a strict mathematical model of a theoretical relationship (Bond, 2015). Item and person fit in relation to the model are computed and taken to investigate unidimensionality. In addition, the outstanding advantage of the Rasch model is to not only provide both item and person fit indices but also render graphic displays to enable the researcher to decide whether the chosen items are spread sufficiently along a latent trait continuum, and where extra items might be included in the scale.

Investigations to shorten GDS using IRT are limited (Chachamovich et al., 2010), despite the fact that IRT or Rasch provides some advantage over CTT because IRT has two main models, nonparametric and parametric. Roughly speaking, both have the same principle, focusing on the relation between individual item responses and individual latent trait values, represented by an item response function (Sijtsma and Molenaar, 2002). The key feature of NIRT is the nonparametric definition of scalability based on homogeneity and the concept of nondecreasing item response function. A widely used nonparametric IRT is Mokken scaling analysis (Watson et al., 2012). The latter parametric developments of IRT were Rasch model, focusing on the scientific properties of measurement models and complementary to the nonparametric models. Therefore, investigators testing for NIRT before PIRT is reasonable.

To create a new scale, the fundamental rule of measurement concerning the ideal property of scales using nonparametric IRT is that items are scored in the same order by all respondents at all levels of the latent trait being measured – which means the 'hierarchical' structure (also called 'invariant item ordering-IIO') of the scale. IIO is considered by experts to be an exacting but important property of scales and a strong requirement in measurement practice (Ligtvoet et al., 2010; Sijtsma and Molenaar, 2002).

The advantage of the hierarchical scale is that when an instrument contains hierarchically ordered items, the items can be ordered from those indicating mild depressive symptoms to those indicating severe depressive symptoms. Thus, the total score is clearly interpreted. For example, a certain total score explains not only how many items but also which items were scored. The advantage of the IIO in addition to ranking individuals according to item difficulty, the IIO can detect DIF. An invariant item ordering connotes that different subgroups from the population of interest have the same item ordering. When item orderings differ, this may be an indication of differential item functioning. In a certain clinical practice, IIO can be advantageous for clinicians as it can differentiate types of dementia (McGrory et al., 2015). Whether IIO retains a set of items can be examined by IRT including parametric methods such as Rasch scaling and the nonparametric method of Mokken scaling analysis (Meijer et al., 1990) (Watson et al., 2012). The hierarchical scale; on the other hand, is not provided by factor analysis. While no shortened GDS version is universally valid due to problems of varying underlying construct and DIF across studies, a shortened one developed based on hierarchical ordered items has never been proposed.

In the present study, the authors aimed to find hierarchical items while shortening the GDS. The shorter version should maintain its validity and provide the same acceptable screening accuracy as the original version of GDS-15 in a tertiary care setting. To ensure these requirements; item hierarchy, unidimensionality and DIF-free were met for the new, shortened scale; we used three different methods from both CTT and IRT. First, confirmatory factor analysis (CFA) was used as well as exploring for DIF items using the Multiple-Indicators Multiple-Cause, due to its being a common method for most investigators. Second concerns the nonparametric IRT and finally, parametric IRT models. However, all methods were compared to identify the potential hierarchical scale for the shortened GDS, but the IRT was mainly analyzed. Finally, we tested the performance of this new robust and hierarchical scale in detecting depression by comparing overall performance using Areas under the Receiver Operating Characteristic curves.

2. Methods

2.1. Participants and design

These study participants were from the study of depressive disorders, anxiety disorders and suicide risk (DAS) among the elderly, with data collected between January 2012 and April 2013. Participants were recruited from the geriatric departments of four tertiary hospitals across Thailand. Enrolled participants were aged $\geq$60 years, with one of the following symptoms: dysphoric mood, feelings of boredom, sleep problems, eating problems, fatigue, memory problems or unexplained somatic symptoms. Patients with a severe physical illness that may have affected the interview or completion of questionnaires were excluded including those who had language barriers, inability to complete the questionnaires, a history of schizophrenia, bipolar disorder or schizoaffective disorder. The enrollment rate was high (93 %). The Eligible participants (57%) were screened for clinical disorders using the Mini-Neuropsychiatric Interview for DSM-IV-TR by trained research assistant nurses Along with interviews, 794 participants (99% response rate)

completed the self-report questionnaires including the 15-item Thai Geriatric Depression Scale. Written informed consent was obtained from all subjects before participating. This study was approved by the institutional ethical committee of the Central Research Ethics Committee of Thailand, the ethics committee of Faculty of Medicine, Chiang Mai University, the ethics committee of Prasat Neurological Institute, and the ethics committee of Songkhla Rajanagarindra Psychiatric Hospital.

2.2. Instruments

The GDS evaluates the extent to which an individual experiences depressive symptoms. The GDS is an easy to administer self-report questionnaire containing 15 items, with a "yes-no" response. The total scores range from 0 to 15 after 5-positivele-keyed items have been reversed. For the Thai version, a standard process using forward and backward translation was used for translation; details may be found elsewhere (Wongpakaran et al., 2013). A recent study showed, the GDS-15 provided a sensitivity of 0.87 and specificity of 0.83 (Dias et al., 2017). A 2 weeks test-retest reliability of the GDS-15 calculated by intraclass correlation was 0.83 (Nyunt et al., 2009). For this study, Cronbach's alpha was 0.81 for the whole sample.

2.3. Analyses

2.3.1. Confirmatory factor analysis

Because the sum score of GDS was nonnormally distributed CFA with categorical outcome was performed to evaluate the nature of and relations between latent constructs. Weighted least squares with correction to means and variances (WLSMV) considering suitable estimators for categorical data were used as estimation the method. A unidimensional construct of the GDS was analyzed; when the model did not fit the data, other dimensional models were compared. The fit indexes used for comparing the goodness of fit included: a Comparative Fit Index (CFI) of $\geq$0.95, a NonNormed Fit Index (NFI) or the Tucker-Lewis Index (TLI) $\geq$0.9, a root-mean-square error of approximation (RMSEA) $\leq$0.6 (0.08 considered acceptable fit (Hu and Bentler, 1998, 1999)) and a χ^2/df result $<$3 (Kline, 1998).

DIF or "Item bias is constituted by the fact that subgroups might perceive questionnaire items differently. In analyzing DIF for sex and age, the Multiple-Indicators Multiple-Cause (MIMIC) model approach was applied using Mplus 8 Software (1998–2017, Muthén & Muthén). The MIMIC model approach to DIF was considered an ordered logistics CFA model with covariates. With this model, the covariates such as sex and age could be tested simultaneously (Jones, 2006). Modification indices were used after initial analysis and error terms correlation was used when indicated (Byrne, 2010).

2.3.2. Item response theory (IRT)

The rationale of IRT models is based on the set of item-response functions (IRF) accounting for the relationship between item responses and the latent trait. The difference between parametric (e.g. Rasch model) and nonparametric IRT (e.g. Mokken model) concerns the specification of the probabilistic relationship between individuals and items of the latent variable. Using mathematical calculations makes the form of IRF more smoothly parametric than a nonparametric IRT. However, both models need to meet the four required assumptions for IRT. First, unidimensionality (UD) is the scale measuring only one latent trait. Second, the local independence (LI) response is only regulated by the abilities not by other factors, for which this assumption is established when unidimensionality is achieved. Third, monotonicity (M), constitutes the monotonely nondecreasing function of the latent trait - higher score and the higher latent trait values and finally, nonintersecting IRF (NI) comprises items can be ranked according to their frequency of endorsement (difficulty), providing hierarchical ordering of scale items (Molenaar et al., 2000). When the three assumptions of UD, LI, and M were met, the "monotonic homogeneity (MH) model" is achieved. NI added to MH is called "double monotonicity (DM) model". Because nonparametric IRT places fewer restrictions than the parametric IRT, it creates a first impression of the overall quality of the measurement. Therefore, nonparametric IRT should be conducted before parametric IRT. Mokken analysis was used for nonparametric IRT, whereas Rasch analysis was used for parametric IRT(Molenaar et al., 2000).

2.3.2.1. Mokken analysis. Mokken analysis demonstrates whether a scale is unidimensional, and provides an impression of the overall quality of the scale, indicating that further parametric IRT methods are justifiable. To test for unidimensionality, Mokken analysis uses the method of scalability, akin to factor analysis to see how many of the scales can be factored. Two types of scalability are employed. First, scalability coefficients *H* are the considered-capacity of the full scale to order persons according to their sum score on the dimension representing latent traits. Second, item scalability involves the coefficient-ability of an item to contribute to the total score for ordering persons on the dimension representing latent traits (Rob R. Meijer and Baneke, 2004; Sijtsma and Molenaar, 2002). In general, H coefficient should be at least 0.3 to be acceptable.

Mokken analysis tests monotonic homogeneity by calculating of Guttman error in which the response does not conform to the Guttman scaling as hypothesized.

Critical value (*Crit*) per item was used to determine monotonic homogeneity by combining evidence about the item's *H*-value, the frequency and the size of the violations and their significance. As recommended by Molenar and Sajitsma, under normal circumstances, e.g., 100–3000 sample, 4–40 items scale, with 2–5 response categories, *Crit* values $>$ 80 strongly indicate that an item violates the assumption. *Crit* values $<$ 40; on the other hand, are less perfect than *Crit* value $=$ 0 but may well be attributed to sampling variation rather than due to systematic model violations (Molenaar et al., 2000).

Finally, Mokken analysis calculates nonintersecting IRF by showing invariant item ordering (IIO) (Ligtvoet et al., 2010) using methods of Restscore, Restplit, and Pmatrix. The item exhibiting this assumption uses the statistic HTrans (HT) of at least 0.3. For H or HT coefficients, 0.3 to 0.4 should be interpreted as weak, 0.4 to 0.5 as moderate, and more than 0.5 as strong. H and HT are considered indicative of the strength and structure of a scale. In other words, the data allow invariant orderings of both individuals (H is high) and items (HT is high). As proposed by Sijtsma et al., if the DMM fits the data, and IIO can be demonstrated, it can be resolved that item ordering is invariant across populations and population subgroups (Sijtsma, 2011). In terms of reliability, Mokken analysis allows for a reliability of the scale statistic rho, which is comparable to Cronbach's alpha (DeJong and Molenaar, 1987).

In summary, to investigate IIO, we followed the method suggested by Sijtsma et al.(Sijtsma and Ark, 2017) (1) the scalability of the items was investigated using an automated item selection procedure, (2) monotonicity was investigated by inspecting item rest-score regressions, (3) IIO was investigated using inspecting methods, i.e. the rest-score method, and the Pmatrix method, and finally (4) the accuracy of the item ordering was investigated by HT coefficient. For dichotomous items of the GDS, analysis was performed using the program MSP5 for Windows (Groningen: ProGamma).

2.3.2.2. Rasch analysis. While Mokken analysis illustrates that the item can be useful for measurement, parametric models can demonstrate the performance of an item. The Rasch model estimates measures the positions of both items and individuals on the continuum latent trait. It can also be modified to maximize the scale performance. By parametric method, Rasch model uses a mathematical formula to specify the form of the relationship between the respondents and the items that defines a single trait. The model assumes the construct to be unidimensional.

These two parameters are estimated by nonlinear transformation of raw score into logit (individual logit and item logit), and put them in the

same scale. Therefore, this common scale comprises two locations of individual and item. The more severe an individual has (high trait), a more severely the item will be endorsed (more difficult item).

To assess unidimensionality, the Rasch model computed outlier-sensitive fit statistics (outfit) mean square (MnSq) and information-weighted fit statistics (infit) mean square (MnSq). Items with misfit were considered to violate the assumption of unidimensionality and should be removed to maintain a unidimensional instrument. An item is considered to achieve acceptable fit when an outfit or infit value is of 0.5–1.5 (Wright and Linacre, 1994).

As recommended by Linacre, a value 1.5 to 2.0 denotes that item is unproductive for measurement (but not degrading). Value >2 (underfit) denotes that an item degrades the measurement properties of the scale. Value <0.5 (overfit) denotes poor local independence, implying that ratings on the items are not independent of each other. Discrimination values of 0.5–2 indicates that the Rasch model is appropriate to analyze the data (Linacre, 2017).

Rasch reports two reliability concepts, person reliability and item reliability. Person reliability means capacity of the scale to separate individuals in two or three levels (comparable to Cronbach's alpha). Item reliability denotes the relevance of the item for measurement and whether the sample is large enough to locate the items precisely on the dimension.

Person separation has different applications and implications from reliability. Person separation is used to classify people. Value of separation <2 and person reliability <0.8 denotes that the measurement may not be sensitive enough to differentiate between high and low performers. In this case, more items may be needed. Item separation is used to affirm the item hierarchy. Separation value less than 3 and item reliability less than 0.9, implying that the sample is not large enough to support construct validity or a difficulty with the item hierarchy of the instrument (Linacre, 2017).

Although DIF are related to parametric models, it also has relevance in Mokken analysis, where large differences between groups are observed, in terms of item scalability, perhaps being seen as a DIF in a broader, more unspecified sense (Adler et al., 2012). DIF was identified when DIF has a contrast size above 0.5 logits.

Finally, hierarchical scale consists of the set of selected items tested to predict a gold standard test diagnosis (for major depressive disorder). The area under the receiver operating characteristic curve (ROC) is used to predict performance. The sets of items were compared in terms for area under the ROC. Winsteps was used for Rasch analysis (Winsteps® (Version 4.3.3) Beaverton, Oregon: Winsteps.com). IBM SPSS version 22 was used for descriptive and ROC analyses (IBM Corp., Armonk, N.Y., USA).

3. Results

Table 1 shows the sociodemographic data of the sample. The participants' age average was 69 years old, most were female (70%), with elementary level of education on average. Most lived with their partner (63%). More than one half had low income. Twenty-four percent received a diagnosis of depressive disorder, most of which (72.5%) were major depressive disorder.

The proportion for each item of the GDS is shown in Table 2. One-factor CFA had an acceptable fit to the data with Chi-square 316.29, df 90, CFI = 0.963; TLI = 0.956; RMSEA = 0.0561 (90% CI .049, .063), and WRMR = 1.372. It appeared that item 9, 'Do you prefer to stay at home, rather than go out and do new things?' had negative and low coefficient value, implying that it did not load on any factor and should be removed from further analysis.

To find item bias, one covariate of age and sex was added at one time separately to the structural model and then built up in the measurement model. We found evidence of DIF associated with sex for two GDS item, i.e., 13. 'Do you feel full of energy?' and 14 'Do you feel that your situation is hopeless?', and with age for three GDS items, namely, 3 'Do you feel that your life is empty?', 7.'Do you feel happy most of the time?', 11 'Do you think it is wonderful to be alive?' and 15 'Do you think that most people are better off than you are?'. In summary, from CFA analysis, 7 items were suggested to be removed; 1 for invalidness and 6 for DIF items. Therefore, only 9 items were left for the short version.

Table 1
Demographic data of respondent patients (n = 803).

Characteristics	N (%) or Mean ± SD
Age, mean (SD)	69.24 ± 6.88 (60–89)
Sex	
Male	239 (30)
Female	557 (70)
Years of education, mean (SD)	6.63 ± 4.9
Marital status	
Single	29 (3.61)
Live together	509 (63.39)
Divorced widow	36 (4.48)
Widowed spouse	228 (28.39)
Income (US dollars/month)*	
<143	469 (59.4)
143–285	110 (13.9)
286–572	114 (14.4)
>572	96 (12.2)
History of illness	
Alcoholism or abuse	9 (1.2)
Suicide attempt	11 (1.7)
Other psychiatric diseases	55 (8.7)
Family history	
Alcohol or other substance abuse	21 (2.6)
Bipolar disorder	0 (0)
Cognitive disorders	5 (0.6)
Depressive disorders	18 (2.2)
Other disorders (schizophrenia, autism, anxiety disorder)	23 (2.9)
DSM-IV clinical disorder	
All depressive disorder	190 (23.7)
Major depressive disorder	138 (17.19)
Dysthymia	40 (4.98)
Double Depression	12 (1.49)
Geriatric Depression Scale-15, mean (SD) of total score	52.05 (14.16)

Note.
* Converted 1 USD from 35 THB

3.1. Mokken scale analysis results

The results of the separate MSA are shown in Table 3. In general homogeneity coefficients exhibited moderate strength (Scale: H = 0.34). Six items (6, 9, 10, 11, 13, and 15 were unscalable (Hj < 0.3). Then 9 items were further checked for monotonicity, all except item 13, and item 4, which had Crit value less than 40. Next, items 13 and 4 were removed and re-analyzed. None had a Crit value over 40; therefore, 8 items were retained.

We then assessed noninteraction and invariant item ordering of the scale using the manifest invariant item ordering method. The results confirmed that the remaining 8 items made up a moderate Guttman scale with moderate support for invariant item ordering. The reliability of the GDS-8 scale was good: Cronbach's α 0.81, Molenaar Sijtsma's ρ 0.82, Guttman's λ 0.82. However, Item 4 had a minimal violation of Crit value (actual value 49); therefore, it was retained for further analysis.

3.2. Rasch analysis based on Mokken

All 9 items were included for analysis for Rasch analysis. Item 2 showed misfit when outfit mean square was higher than 2. In exploring DIF, no DIF value was observed for sex among 9 items whereas items 3, 5, and 7 showed value. Item 7 showed the most DIF impact as DIF contrast exceeded a logit of 0.5.

Item 2 was a misfit item while items 7 and 3 had potential DIF. Items 3 and 7 yielded a DIF contrast of 0.50 and 0.71, respectively; therefore, they were removed. Thus, only 6 items were left for the final set of

Table 2
Proportion of each item, factor loadings, and DIF items of GDS.

GDS item	Yes	No	EC	S.E.	EC/ S.E.	P-Value	DIF for sex		DIF for age	
							Est./ S.E.	P-Value	Est./ S.E.	P-Value
1. Are you basically satisfied with your life?	0.881	0.119	0.763	0.039	19.326	0.000	0.456	0.649	−0.934	0.351
2. Have you dropped many of your activities and interests?	0.431	0.569	0.546	0.042	13.075	0.000	−0.433	0.665	1.709	0.088
3. Do you feel that your life is empty?	0.687	0.313	0.757	0.029	26.197	0.000	−1.568	0.117	2.662	0.008
4. Do you often get bored?	0.664	0.336	0.841	0.025	33.707	0.000	−1.006	0.314	0.211	0.833
5. Are you in good spirits most of the time?	0.815	0.185	0.754	0.035	21.462	0.000	0.771	0.441	−1.697	0.090
6. Are you afraid that something bad is going to happen to you?	0.711	0.289	0.563	0.041	13.587	0.000	−0.979	0.328	−1.611	0.107
7. Do you feel happy most of the time?	0.770	0.230	0.786	0.030	25.794	0.000	0.135	0.892	−2.957	0.003
8. Do you often feel helpless?	0.809	0.191	0.855	0.025	33.868	0.000	0.889	0.374	0.447	0.655
9. Do you prefer to stay at home, rather than go out and do new things?	0.232	0.768	−0.039	0.055	−0.716	0.474	-	-	-	-
10. Do you feel you have more problems with memory than most?	0.597	0.403	0.369	0.047	7.859	0.000	−0.293	0.769	0.197	0.844
11. Do you think it is wonderful to be alive?	0.657	0.343	0.480	0.043	11.091	0.000	1.641	0.101	−2.045	0.041
12. Do you feel worthless the way you are now?	0.810	0.190	0.866	0.025	34.167	0.000	−1.124	0.261	1.504	0.133
13. Do you feel full of energy?	0.623	0.377	0.551	0.040	13.819	0.000	2.970	0.003	0.940	0.347
14. Do you feel that your situation is hopeless?	0.812	0.188	0.877	0.023	37.321	0.000	−2.031	0.042	0.651	0.515
15. Do you think that most people are better off than you are?	0.731	0.269	0.528	0.044	12.087	0.000	1.427	0.154	2.065	0.039

DIF = Differential Item Functioning, EC = estimated coefficient, SE = standard error.

Table 3
Mokken Scale Analysis results of GDS.

GDS item	ItemH	#ac	#vi	#vi/#ac	maxvi	sum	sum/#ac	zmax	#zsig	crit
1	0.46	16	0	0.00	0.00	0.00	0.0000	0.00	0	0
2	0.38	15	0	0.00	0.00	0.00	0.0000	0.00	0	0
3	0.36	21	0	0.00	0.00	0.00	0.0000	0.00	0	0
4	0.42	21	0	0.00	0.00	0.00	0.0000	0.00	0	0
5	0.42	21	0	0.00	0.00	0.00	0.0000	0.00	0	0
6	0.28	21	0	0.00	0.00	0.00	0.0000	0.00	0	0
7	0.42	15	0	0.00	0.00	0.00	0.0000	0.00	0	0
8	0.43	21	0	0.00	0.00	0.00	0.0000	0.00	0	0
9	−0.02	21	8	0.38	0.16	0.77	0.0366	2.63	4	170
10	0.20	21	0	0.00	0.00	0.00	0.0000	0.00	0	0
11	0.23	21	1	0.05	0.04	0.04	0.0017	0.42	0	18
12	0.44	21	0	0.00	0.00	0.00	0.0000	0.00	0	0
13	0.28	15	0	0.00	0.00	0.00	0.0000	0.00	0	0
14	0.45	21	0	0.00	0.00	0.00	0.0000	0.00	0	0
15	0.26	21	0	0.00	0.00	0.00	0.0000	0.00	0	0
Scale H	0.337 (0.016)									
MS	0.827									
alpha	0.816									
lamda	0.824									

Note.
GDS = Geriatric depression scale.
#ac = the number of active comparisons carried out to check the DM properties of the data.
#vi = the number of violations of DM found in these comparisons.
Crit = Critical value, calculated by summation of the coefficient values of ItemH, #ac, #vi, #vi/#ac, maxvi, sum, sum/#ac, zmax, and #zsig.
MS = Molenaar Sijtsma's ρ

Table 4
Rasch Analysis results of the GDS.

Item	Difficulties	Infit MnSq	Outfit MnSq	Sex		Age	
				DIF contrast	p-value	DIF contrast	p-value
2	−2.91	1.28	**3.23**	0.04	0.723	0.23	0.586
1	1.61	1.12	1.27	0.30	0.333	0.24	0.644
3	−0.65	1.05	1.03	0.21	0.442	**0.50**	**0.022**
5	0.72	1.05	0.89	0.38	0.170	0.45	0.027
7	0.19	0.96	0.84	0.27	0.267	**0.71**	**0.000**
12	0.61	0.89	0.82	0.26	0.329	0.38	0.097
14	0.64	0.86	0.73	0.00	1.000	0.18	0.355
4	−0.82	0.85	0.85	0.46	0.176	0.06	0.712
8	0.62	0.85	0.84	0.13	0.932	0.21	0.408

Note: Significant DIF contrasts are in bold.
GDS = Geriatric depression scale.
Infit = information-weighted fit statistic; Outfit = outlier-sensitive fit statistics; MnSq = mean square. a Loadings are derived from one-factor model.

hierarchical scale (shown in Table 4).

3.3. ROC analysis

In terms of accuracy of the hierarchical scale of GDS in predicting depression against the gold standard clinical interview diagnosis, we used the area under the ROC curve as the criterion to compare the following set of items: the 15-item (original version) and the 6-item hierarchical scale (Fig. 1).

Both scales provided AUCs of >0.8 denoting good accuracy performance (Somoza et al., 1989). No difference in area under the ROC curve between both scales ($p > .05$); however, GDS-6 yielded a sensitivity of 73.29 and specificity of 81.24%, whereas GDS-15 yielded a sensitivity of 66.20 and specificity of 84.84%.

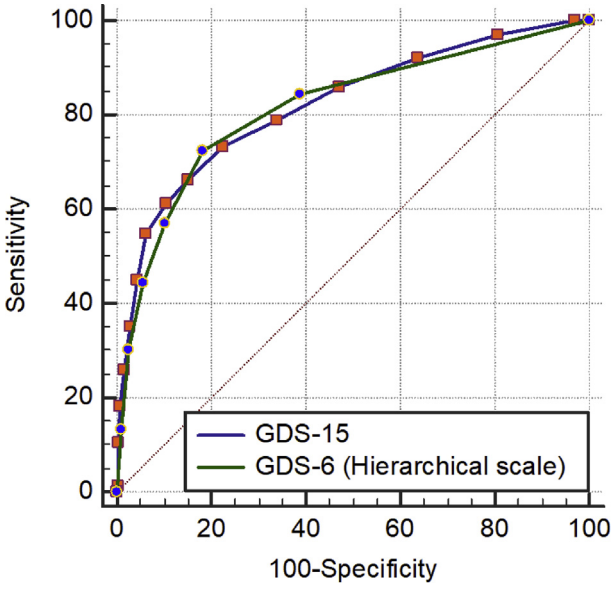
Table 5 compared the present findings and other short versions both using Classic test theory (factor analysis) and Rasch model. Item 1 was selected from all versions. Likewise, item 10 was deselected from all versions. Item 9 was removed from most. Internal consistency, illustrated by Cronbach's alpha, showed all short versions were acceptable.

4. Discussion

The aim of the present research was to identify the hierarchically ordered items and exclusion of the DIF items to form a shortened version from the existing GDS that have no clear multidimensional structure. This hierarchical scale comprised 6 items that showed sufficient internal consistency and validity. Factor analysis, Mokken analysis and Rasch analysis yielded quite similar, even though not exact results. The main difference was from the DIF items identified.

CFA yielded 9 items that corresponded to Mokken and Rasch except for item 14, 'Do you feel that your situation is hopeless?' This item was found to show DIF value for sex but not with IRT method. This discrepancy of the significant DIF was due to the different method of analysis and length of the scale as well (Linacre, 2017).

Scalability of an item set by Mokken analysis was comparable to factor analysis, but Mokken provided better results in yielding unidimensionality as we can see that 6 items were removed after using the aisp method. To put it another way, GDS-15 had insufficient unidimensionality to use the sum or total score as all items did not measure the same latent trait. This did not only affect the validity of the sum score but also the cut-off score for screening for depression. Mokken analysis results demonstrated that only 10 items should be used; however, when using DIF taking sex and age into account only 6 items met all criteria.



Note

ROC = Receiver Operating Curve
GDS = Geriatric depression scale

Fig. 1. ROC curves of the original and shortened version.

The advantage of Mokken and Rasch analysis over CFA is that they provide a better understanding of item reliability in that the (hierarchical) ordering of the items is the same for all patients. Some may argue the necessity of noncognitive measurements like GDS requiring that kind of IIO. To have IIO property implies no DIF is found for any subgroups. We can see from CFA that GDS (except item 9) met the acceptable criteria for unidimensionality provided that DIF was ignored. The source of DIF may be from cultural or racial differences. For example, Item 9 was considered to exhibit a prominent 'cultural' bias due to the differences involving the elderly way of life. Thai elderly are likely to stay, regardless of feeling depressed or not. Therefore, this item was invalid as could not assess depression.

Compared with related studies using the same method of Rasch analysis, there seems to be more item agreement than when using the Classic test theory in other related studies. Our present findings had 67.7% item agreement as with other Rasch analysis conducted by Chachamovich et al. (2010) and 83.3% of item agreement with Tang et al. (Tang et al., 2005).

In comparison with factor analysis, Mokken scaling and Rasch analysis presented some advantages. Mokken scaling and Rasch analysis can demonstrate the systematic ordered relationships between items, whereas factor analysis describes groups of highly correlating items. Creating ordered relationships improves construct validity (DeJong and Molenaar, 1987). In addition, the order of difficulty of the items often has an important theoretical interpretation that is not allowed for in factor analyses.

When testing for the usefulness of the hierarchical scale GDS, it demonstrated a comparable ability to the original 15-item GDS despite comprising only 6 items.

Our results were consistent with Tang et al. in that no difference was found for the area under the ROC curve between the shorter version and the original (Tang et al., 2005). Likewise, GDS-7 was developed by Broekman et al. using factor analysis (Broekman et al., 2011). They yielded a similar AUC as the original GDS-15. The slight difference between the present study and GDS-10 and GDS-7 was that ours yielded a higher positive predictive value than the original, whereas GDS-10 and GDS-7 yielded lower positive predictive values than the original.

The GDS-4 developed by Allgaier et al. (2013) was found to have lower AUC than the original GDS-15 (0.82 vs. 0.77) but with low Cronbach's alpha among all (0.60). Because they adopted the GDS-4 from D'ath's study, we assumed that DIF items occurred in their studied sample. In addition, the Cronbach's alpha of the original GDS-4 in D'ath was indeed low (0.55) (D'Ath et al., 1994).

Castello et al. used GDS-4 from Almeida and Almeida's study (different from GDS-4 of D'ath) with primary care patients and found Cronbach's α coefficient value for the GDS-4 was 0.74. However, the AUC of this GDS-4 was lower than the GDS-15 (Almeida Osvaldo and Almeida Shirley, 1999; Castelo et al., 2010). Similar to Allgaier et al., we thought that the DIF items could not be removed in the Castelo's sample. Compared with the hierarchical scale of GDS-6, the Cronbach's alpha was satisfied.

In sum, the hierarchical scale demonstrated a robust construct validity of the test which was supported by a related study involving testing hierarchical scale (Hidalgo et al., 2015; Li and Zumbo, 2009). Hierarchical scale is not only ordered but free from DIF. It can also detect aberrant item score patterns (Meijer et al., 2008) and to detect differential item functioning. Moreover, the IIO may be of help to test hypotheses concerning psychological constructs (Sijtsma and Molenaar, 2002). Therefore, the items included in the scale should be hierarchically ordered or else, the cutoffs for screening may be biased (DeJong and

Table 5
Comparing to other short GDS.

	present study		GDS10	GDS11	GDS-7	GDS-4	GDS4
Method of analysis	M + R	CTT	Rasch	Rasch	CTT	CTT	CTT
item							
1. Are you basically satisfied with your life?	6*	+	9	11	+	+	+
5. Are you in good spirits most of the time?	5	+		5	+		
14. Do you feel that your situation is hopeless?	4	+	4	2			
8. Do you often feel helpless?	3	+	7	9	+		
12. Do you feel pretty worthless the way you are now?	2	+	8	3			
4. Do you often get bored?	1	+	3		+		
10. Do you feel you have more problems with memory than most?		+					
11. Do you think it is wonderful to be alive?			10	10			
13. Do you feel full of energy?			2	1			
15. Do you think that most people are better off than you are?		+		6	+		
2. Have you dropped many of your activities and interests?			1	4		+	
3. Do you feel that your life is empty?			5		+		+
6. Are you afraid that something bad is going to happen to you?		+	6	8			+
7. Do you feel happy most of the time?				7	+	+	+
9. Do you prefer to stay at home, rather than go out and do new things?						+	
Cronbach's alpha	.80	.76	0.768	n/a	0.80	0.74	0.60
Criteria validity parameters							
AUC (original: short)	.82 (.82)	.82 (.82)	n/a	.86 (.86)	.98 (.99)	.91 (.85)	(.85)
% Sensitivity (original: short)	66 (73)	66 (72)	n/a	79 (65)	97 (93)	89 (84)	69 (88)
% Specificity (original: short)	85 (81)	85 (81)	n/a	74 (90)	95 (91)	70 (75)	88 (53)
Positive predictive value (original: short)	49.2 (46.9)	49.2 (45.4)	n/a	30 (47)	42 (27)	38 (41)	69 (43)
Negative predictive value (original: short)	91.9 (93.1)	91.9 (92.9)	n/a	96 (95)	99.8 (99.7)	97 (96)	88 (92)

Note.
The number reflects the order of relationship, the lowest number (1) indicates the least difficult item; the highest number indicates the most difficulty item. (provided only in Mokken or Rasch model).
CTT = Classic Test Theory.
M = Moken scale analysis, R = Rasch analysis, n/a = not applicable.

Molenaar, 1987).

However, even though IIO reflects that DIF items were removed, it may be impossible to have a scale that is free from any DIF in depression because DIF can occur due to many factors involved in depression, such as individual's perception of depression based on culture, personality trait, neuroticism, depression history, childhood stress, neuroticism and stressful life events (Fried et al., 2014; Wongpakaran, Wongpakaran et al., 2015). To manage DIF, investigators usually focus on the sizeable magnitude and impact of DIF(Teresi et al., 2008). However, it has been recommended that adjustments of scale scores were frequently recommended instead of attempting to remove all DIF items (Teresi et al., 2008).

In addition, the hierarchical scale provides a shorter but comparable ability of the test compared with the original longer version. Hierarchical scales can add to clinical interpretation, for instance, a patient responding 'no' to feelings of worthlessness' is unlikely to respond 'yes' to feelings of helplessness or hopelessness, thus, appealing to their ease of use and scoring (Doyle et al., 2012). This kind of hierarchy is, in fact, clearly advantageous in cognitive function tests as well as in tests with physical function such as movement function of any part of the body in rehabilitation or orthopedics. For a subjective measure like depression, the study of ordered relationships of depressive symptom has scarcely been reported. One study on the hierarchy of depressive symptoms was explored in melancholic features using the Hamilton Rating Scale for Depression. The results showed that the rank of severity of the melancholic features were as shown below. First, depressed mood followed by work and interests, somatic symptoms, psychic anxiety, feelings of guilt and finally, psychomotor retardation. The items "anxiety", "psychomotor retardation" and "guilt feelings" seemed more severe than the "somatic symptoms", they did not show the capacity to differentiate individuals in the person-item map (Primo de Carvalho Alves et al., 2017). For self-report measurement, a 21-item Beck depression inventory-II was investigated for hierarchical and dimensional scale using a bifactor model. Despite the fact that this study was not directly tapped for a hierarchical scale of BDI, irritability was the least difficulty, whereas suicidal thoughts was the most difficult item (Al-Turkait and Ohaeri, 2010). For this hierarchical scale of GDS-6, except for the positive-worded items, the hierarchy of the item made much clinical sense with the depressive continuum, starting from the least to the most severe, i.e. 'get bored', 'worthless', helpless and hopeless.

4.1. Limitation and future study

The new GDS-6 needs to be investigated for test-retest reliability to ensure temporal stability of the scale. Despite the fact that the shortened version conducted was based on a different approach, equivalent accuracy in predicting depression was evidenced. We strongly encourage replicating the study using a similar analysis method to compare the results with the present study. In addition, envisioning how this GDS-6 would play out in another sample would be difficult. Thus, it should be investigated using subjects with different cultural backgrounds to examine item performance as well as to see whether DIF still exists in another sample including the performance of accuracy for predicting depression using the area under the ROC curve between the GDS-6 and the original. In addition, to maximize its usefulness, sensitivity analysis when using the GDS-6 as an outcome measure should be examined, for example, during follow-up after treatment either by psychotropic medication or by psychosocial therapy.

5. Conclusion

Due to the flaws related to unstable structure and the DIF of the GDS the results could not automatically be generalized or used in a population with different language and cultural background. The unequivalent briefer scale makes it difficult to study comparability. We have created another short version of the GDS based on the notion of hierarchical items using the current common analysis methods. The hierarchical scale of GDS-6, derived from IRT, showed that it was equal to or better in predicting performance compared with the original 15-item GDS and made better sense in clinical interpretation as compared with the Classic

Test Theory method. Applying this GDS-6 is in other languages and cultures would be encouraging.

Declarations

Author contribution statement

N. Wongpakaran, T. Wongpakaran, P. Kuntawong: Conceived and designed the experiments; Performed the experiments; Analyzed and interpreted the data; Contributed reagents, materials, analysis tools or data; Wrote the paper.

Funding statement

This study was part of the "Depressive Disorders, Anxiety Disorders, and Suicide Risk and the Associated Factors among Elderly People (DAS) Program, and was funded by the National Research Council of Thailand.

Competing interest statement

The authors declare no conflict of interest.

Additional information

No additional information is available for this paper.

References

Adler, M., Hetta, J., Isacsson, G., Brodin, U., 2012. An item response theory evaluation of three depression assessment instruments in a clinical sample. BMC Med. Res. Methodol. 12, 84.

Al-Turkait, F.A., Ohaeri, J.U., 2010. Dimensional and hierarchical models of depression using the Beck Depression Inventory-II in an Arab college student sample. BMC Psychiatry 10, 60-60.

Allgaier, A.K., Kramer, D., Saravo, B., Mergl, R., Fejtkova, S., Hegerl, U., 2013. Beside the Geriatric Depression Scale: the WHO-Five Well-being Index as a valid screening tool for depression in nursing homes. Int. J. Geriatr. Psychiatry 28 (11), 1197–1204.

Almeida Osvaldo, P., Almeida Shirley, A., 1999. Short versions of the geriatric depression scale: a study of their validity for the diagnosis of a major depressive episode according to ICD-10 and DSM-IV. Int. J. Geriatr. Psychiatry 14 (10), 858–865.

Bond, T.G., 2015. Applying the Rasch model : fundamental measurement. In the Human Sciences/Authored by Trevor G. Bond and Christine M. Fox. Routledge, Taylor & Francis Group, New York.

Broekman, B.F., Niti, M., Nyunt, M.S., Ko, S.M., Kumar, R., Ng, T.P., 2011. Validation of a brief seven-item response bias-free geriatric depression scale. Am. J. Geriatr. Psychiatry 19 (6), 589–596.

Broekman, B.F., Nyunt, S.Z., Niti, M., Jin, A.Z., Ko, S.M., Kumar, R., et al., 2008. Differential item functioning of the geriatric depression scale in an Asian population. J. Affect. Disord. 108 (3), 285–290.

Byrne, B.M., 2010. Structural Equation Modeling with AMOS: Basic Concepts, Applications, and Programming, second ed. Taylor and Francis Group, LLC, New York, London.

Castelo, M.S., Coelho-Filho, J.M., Carvalho, A.F., Lima, J.W., Noleto, J.C., Ribeiro, K.G., Siqueira-Neto, J.I., 2010. Validity of the Brazilian version of the Geriatric Depression Scale (GDS) among primary care patients. Int. Psychogeriatr. 22 (1), 109–113.

Chachamovich, E., Fleck, M.P., Power, M., 2010. Is geriatric depression scale-15 a suitable instrument for measuring depression in Brazil? Results of a Rasch analysis. Psychol. Health Med. 15 (5), 596–606.

Chau, J., Martin, C.R., Thompson, D.R., Chang, A.M., Woo, J., 2006. Factor structure of the Chinese version of the geriatric depression scale. Psychol. Health Med. 11 (1), 48–59.

Chiang, K.S., Green, K.E., Cox, E.O., 2009. Rasch analysis of the geriatric depression scale-short form. Gerontol. 49 (2), 262–275.

D'Ath, P., Katona, P., Mullan, E., Evans, S., Katona, C., 1994. Screening, detection and management of depression in elderly primary care attenders. I: the acceptability and performance of the 15 item Geriatric Depression Scale (GDS15) and the development of short versions. Fam. Pract. 11 (3), 260–266.

DeJong, A., Molenaar, I.W., 1987. An application of Mokken's model for stochastic, cumulative scaling in psychiatric research. J. Psychiatr. Res. 21 (2), 137–149.

Dias, F.L.d.C., Teixeira, A.L., Guimarães, H.C., Barbosa, M.T., Resende, E.d.P.F., Beato, R.G., Caramelli, P., 2017. Accuracy of the 15-item geriatric depression scale (GDS-15) in a community-dwelling oldest-old sample: the Pietà study. Trends Psychiatr. Psychother. 39, 276–279.

Doyle, F., Watson, R., Morgan, K., McBride, O., 2012. A hierarchy of distress and invariant item ordering in the General Health Questionnaire-12. J. Affect. Disord. 139 (1), 85–88.

Fried, E.I., Nesse, R.M., Zivin, K., Guille, C., Sen, S., 2014. Depression is more than the sum-score of its parts: individual DSM symptoms have different risk factors. Psychol. Med. 44 (10), 2067–2076.

Friedman, B., Heisel, M.J., Delavan, R.L., 2005. Psychometric properties of the 15-item geriatric depression scale in functionally impaired, cognitively intact, community-dwelling elderly primary care patients. J. Am. Geriatr. Soc. 53 (9), 1570–1576.

Hidalgo, M.D., Galindo-Garre, F., Gómez-Benito, J., 2015. Differential Item Functioning and Cut-Off Scores: Implications for Test Score Interpretation. Retrieved from. http://revistes.ub.edu/index.php/Anuario-psicologia/article/download/12019/14791.

Hu, L., Bentler, P.M., 1998. Fit indices in covariance structure modeling: sensitivity to under parameterized model misspecification. Psychol. Methods 3, 424–453.

Hu, L., Bentler, P.M., 1999. Cut off criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. Struct. Equ. Model. 6 (1), 1–55.

Incalzi, R.A., Cesari, M., Pedone, C., Carbonin, P.U., 2003. Construct validity of the 15-item geriatric depression scale in older medical inpatients. J. Geriatr. Psychiatry Neurol. 16 (1), 23–28.

Jones, R.N., 2006. Identification of measurement differences between English and Spanish language versions of the Mini-Mental State Examination. Detecting differential item functioning using MIMIC modeling. Med. Care 44 (11 Suppl 3), S124–133.

Kim, G., Decoster, J., Huang, C.H., Bryant, A.N., 2013. A meta-analysis of the factor structure of the Geriatric Depression Scale (GDS): the effects of language. Int. Psychogeriatr. 71–81.

Kline, R.B., 1998. Principles and Practice of Structural Equation Modeling. Guilford, New York.

Kok, R.M., Reynolds, C.F., 2017. Management of depression in older adults: a review. J. Am. Med. Assoc. 317 (20), 2114–2122.

Li, Z., Zumbo, B.D., 2009. Impact of differential item functioning on subsequent statistical conclusions based on observed test score data. Psicologica 30 (2), 343–370.

Ligtvoet, R., van der Ark, L.A., te Marvelde, J.M., Sijtsma, K., 2010. Investigating an invariant item ordering for polytomously scored items. Educ. Psychol. Meas. 70 (4), 578–595.

Linacre, J.M., 2017. Winsteps® Rasch Measurement Computer Program User's Guide. Winsteps.com, Beaverton, Oregon.

Magno, C., 2009. Demonstrating the difference between classical test theory and item response theory using derived test data. Int. J. Edu. Psychol. Assess. 1 (1), 1–11.

Malakouti, S.K., Fatollahi, P., Mirabzadeh, A., Salavati, M., Zandi, T., 2006. Reliability, validity and factor structure of the GDS-15 in Iranian elderly. Int. J. Geriatr. Psychiatry 21 (6), 588–593.

McGrory, S., Starr, J.M., Shenkin, S.D., Austin, E.J., Hodges, J.R., 2015. Does the order of item difficulty of the Addenbrooke's cognitive examination add Anything to subdomain scores in the clinical assessment of dementia? Dement. Geriatr. Cognit. Disord. EXTRA 5 (1), 155–169.

Meijer, R.R., Baneke, J.J., 2004. Analyzing psychopathology items: a case for nonparametric item response theory modeling. Psychol. Methods 9 (3), 354–368.

Meijer, R.R., Egberink, I.J., Emons, W.H., Sijtsma, K., 2008. Detection and validation of unscalable item score patterns using item response theory: an illustration with Harter's Self-Perception Profile for Children. J. Personal. Assess. 90 (3), 227–238.

Meijer, R.R., Sijtsma, K., Smid, N.G., 1990. Theoretical and empirical comparison of the Mokken and the Rasch approach to IRT. Appl. Psychol. Meas. 14 (3), 283–298.

Midden, A.J., Mast, B.T., 2017. Differential item functioning analysis of items on the Geriatric Depression Scale-15 based on the presence or absence of cognitive impairment. Aging Ment. Health 1–7.

Mitchell, A.J., Bird, V., Rizzo, M., Meader, N., 2010a. Diagnostic validity and added value of the Geriatric Depression Scale for depression in primary care: a meta-analysis of GDS30 and GDS15. J. Affect. Disord. 125 (1-3), 10–17.

Mitchell, A.J., Bird, V., Rizzo, M., Meader, N., 2010b. Which version of the geriatric depression scale is most useful in medical settings and nursing homes? Diagnostic validity meta-analysis. Am. J. Geriatr. Psychiatry 18 (12), 1066–1077.

Molenaar, I.W., Sijtsma, K., Boer, P., 2000. User's Manual for MSP5 for Windows: A Program for Mokken Scale Analysis for Polytomous Items. Version 5.0. University of Groningen, Groningen, The Netherlands.

Nyunt, M.S., Fones, C., Niti, M., Ng, T.P., 2009. Criterion-based validity and reliability of the Geriatric Depression Screening Scale (GDS-15) in a large validation sample of community-living Asian older adults. Aging Ment. Health 13 (3), 376–382.

Pocklington, C., Gilbody, S., Manea, L., McMillan, D., 2016. The diagnostic accuracy of brief versions of the Geriatric Depression Scale: a systematic review and meta-analysis. Int. J. Geriatr. Psychiatry 31 (8), 837–857.

Primo de Carvalho Alves, L., Pio de Almeida Fleck, M., Boni, A., Sica da Rocha, N., 2017. The major depressive disorder hierarchy: Rasch analysis of 6 items of the Hamilton depression scale covering the continuum of depressive syndrome. PLoS One 12 (1), e0170000.

Sijtsma, K., 2011. Introduction to the measurement of psychological attributes. Measurement 44 (7), 1209–1219.

Sijtsma, K., Ark, L.A.v. d., 2017. A tutorial on how to do a Mokken scale analysis on your test and questionnaire data. Br. J. Math. Stat. Psychol. 70 (1), 137–158.

Sijtsma, K., Molenaar, I.W., 2002. Introduction to Nonparametric Item Response Theory, Vol. 5. Sage.

Somoza, E., Soutullo-Esperon, L., Mossman, D., 1989. Evaluation and optimization of diagnostic tests using receiver operating characteristic analysis and information theory. Int. J. Bio Med. Comput. 24 (3), 153–189.

Steinberg, L., Thissen, D., 1996. Uses of item response theory and the testlet concept in the measurement of psychopathology. Psychol. Methods 1 (1), 81–97.

Tang, W.K., Wong, E., Chiu, H.F., Lum, C.M., Ungvari, G.S., 2005. The Geriatric Depression Scale should be shortened: results of Rasch analysis. Int. J. Geriatr. Psychiatry 20 (8), 783–789.

Teresi, J.A., Ramirez, M., Lai, J.-S., Silver, S., 2008. Occurrences and sources of Differential Item Functioning (DIF) in patient-reported outcome measures: description of DIF methods, and review of measures of depression, quality of life and general health. Psychol. Sci. Q. 50 (4), 538-538.

Thurstone, L.L. (Ed.), 1931. The Measurement of Social Attitudes. American Psychological Association, pp. 249–269.

Wang, J., Wu, X., Lai, W., Long, E., Zhang, X., Li, W., et al., 2017. Prevalence of depression and depressive symptoms among outpatients: a systematic review and meta-analysis. BMJ Open 7 (8), e017173.

Watson, R., van der Ark, L.A., Lin, L.-C., Fieo, R., Deary, I.J., Meijer, R.R., 2012. Item response theory: how Mokken scaling can be used in clinical practice. J. Clin. Nurs. 21 (19–20), 2736–2746.

Wongpakaran, N., Wongpakaran, T., 2012. Prevalence of major depressive disorders and suicide in long-term care facilities: a report from northern Thailand. Psychogeriatrics 12 (1), 11–17.

Wongpakaran, N., Wongpakaran, T., Reekum, R.V., 2013. The use of GDS-15 in detecting MDD: a comparison between residents in a Thai long-term care home and geriatric outpatients. J. Clin. Med. Res. 5 (2), 101–111.

Wongpakaran, T., Wongpakaran, N., Boonyanaruthee, V., Pinyopornpanish, M., Intaprasert, S., 2015. The influence of comorbid personality disorders on recovery from depression. Neuropsychiatr. Dis. Treat. 11, 725–732.

Wright, B.D., Linacre, J.M., 1994. Reasonable mean-square fit values. Rasch Meas. Trans. 370–371.

Wright, B.D., Stone, M.H., 1979. Best Test Design: Rasch Measurement. Mesa Press, Chicago, IL.

Yesavage, J.A., Brink, T.L., Rose, T.L., Lum, O., Huang, V., Adey, M., Leirer, V.O., 1982. Development and validation of a geriatric depression screening scale: a preliminary report. J. Psychiatr. Res. 17 (1), 37–49.

Zanon, C., Hutz, C.S., Yoo, H., Hambleton, R.K., 2016. An application of item response theory to psychological test development. Psicol. Reflexão Crítica 29 (1), 18.