Original Article

Evaluation and improvement of the psychometric properties of the Hamilton Rating Scale for Depression using Rasch analysis for applying in Belorusian Population

Marat Assanovich[1*]

1. Ph.D., M.D. Head of the Department of Clinical Psychology and Psychotherapy, Grodno State Medical University, Grodno, Belarus

** (Corresponding Author: Marat Assanovich mail: 70malas@gmail.com)*

(Received: 30 Jan 2017; Revised: 20 Feb 2017; Accepted: 15 Mar 2017)

Abstract

Introduction: The Hamilton Rating Scale for Depression (HAM-D) has not been tested for the psychometric properties in Belarusian population. In this regard conducting psychometric assessment of the HAM-D used in Belarusian population has special importance. The aim of this study was to investigate and improve the psychometric properties of the HAM-D applying in Belarusian population through Rasch analysis.

Methods: Data from 551 Belarusian patients with confirmed diagnosis of depression were used in this study. The Partial Credit Model was used in the psychometric analysis.

Results: The analysis of item-person map showed disagreement between distributions of person measures and item categories. Only 8 items had appropriate values of Rasch fit-indexes. I have improved the psychometric properties of the HAM-D. Five items with bad values of fit-indexes were recombined into two new items. New items were added to the remaining 8 items with acceptable fit-indexes values and then Rasch analysis was repeated. All items showed adequate construct validity. The person separation index (PSI) for original HAM-D was 2.32. The person reliability was 0.84. Reliability analysis of the final model included 10 items showed higher estimates: the PSI was 3.61, the number of strata rose to 5.14, the person reliability was 0.93.

Conclusion: This study revealed that the HAM-D applied in Belarusian population has several psychometric problems. After removing invalid items and changing score rules of bipolar items psychometric properties of the scale were improved.

Declaration of Interest: None.

Key words: Rasch analysis, Hamilton Rating Scale Depression.

Introduction

Clinical rating scales are designed to assess the severity of mental disorder after the diagnosis of the disorder has been identified (1). Rating scales can be applied at anytime throughout treatment to help define changes in the severity of the disorder. Information given by these scales can help to provide treatment planning particularly frequency of sessions or drug dosages (1, 2).

The Hamilton rating scale for depression (HAM-D) is one of the most popular objective scales to be developed for measure of depression severity (2–4). It is a clinician rated scale aimed to assess depression severity among patients with confirmed diagnosis of depression (2).

The original HAM-D includes 21 items, but the last four items are not counted toward the total score because these items do not measure depression severity (2). Therefore, the 17-item version of the HAM-D has become the standard for assessment depression severity and the most widely used clinical rating scale in clinical trials and in clinical practice (1, 2).

The HAM-D has not been tested for the psychometric properties in Belarusian population. A number of studies (3, 5) in recent years have revealed that the HAM-D has several psychometric problems relating to multidimensional structure, unsatisfied construct validity of individual items. The Hamilton rating scale for depression is widely used in a clinic in Belarus. Over the last 20 years, this scale is the only psychometric scale used in Belarus to assess the severity of depression. In this regard conducting psychometric assessment of the HAM-D used in Belorusian population has special importance.

Classical test theory has traditionally been used for psychometric analysis of clinical rating scales (6, 7). However, during the last decades, more efficient psychometric technologies has been increasingly adopted. The most suitable form of these technologies for the analysis of clinical rating scales is the Rasch model (8). Use of Rasch metric system involves a comprehensive and fundamental analysis of the data and provides important psychometric information that cannot be reached through the CTT approach (9, 10). The data are examined for fitting the Rasch model, allowing for a rigorous investigation of the construct validity of each item, unidimensionality, reliability and separation characteristics of the scale (8). The main question solving by Rasch model is whether the scale is consistent for principles of scientific measurement which is required to justify summation of raw scores and linearly transformation total scores from their original scale to an interval measuring scale (8, 11, 12). The aim of this study was to evaluate and improve the psychometric properties of the HAM-D using Rasch analysis for Belorussian population. This is the first Rasch analysis of the HAM-D in Belarus.

Methods

The HAM-D was administered to 551 Belorusian patients with confirmed diagnosis of depression according to ICD-10 criteria. Inclusion criteria were the consistency of the patient's disorder with the criteria of the current depressive episode, obtaining a methodologically complete protocol using HAM-D. The criteria for exclusion were the presence of comorbid disorders and somatic diseases. Patients were examined during outpatient treatment (with a mild depressive episode) and inpatient treatment (with a moderate and severe depressive episode) in the Grodno Regional Clinical Center "Psychiatry-Narcology" and in the Grodno City Clinical Hospital No. 3 (Republic of Belarus). There were 231 males and 320 females aged between 23 and 54 years. This age range was the result of the sampling of subjects in accordance with inclusion and exclusion criteria. The HAM-D data were transferred into the Rasch partial credit model using the Jmetrik (13). Rasch analysis was conducted for the HAM-D to examine targeting, model fit statistics, dimensionality, person separation statistics and reliability.

Targeting means matching the distribution of category difficulties to the distribution of persons (14). It is important, particularly in clinical psychometrics, that measuring instrument is well-targeted. Item maps visualize the distribution of persons and distribution of item category difficulties for an entire scale. They are useful for determining whether items match with the population and identifying unacceptable shifts such as “ceiling and floor” effects (15).

To examine how well the empirical data fit the Rasch model, special chi-square fit statistics were calculated for each item. These fit statistics are the weighted mean square (WMS) and the unweighted mean square (UMS) (12). The WMS is the information-weighted mean square residual difference between observed and expected responses. These statistics take into consideration unexpected responses near the person’s construct level (8). The UMS is the simple unweighted mean square residual and is more sensitive to outliers (12). The ideal infit or outfit mean square values are 1.0. In practice values should range between 0.6 and 1.4 for clinical interviews (16). High values reflect lack of predictability of an item. Low WMS and UMSt reflect overfit, which means over-predictability of an item (8).

The separation statistics of Rasch model indicate the internal consistency and discriminative characteristics of the scale. The Person separation index and number of strata index show the ability of the measure to differ statistically significant levels of construct severity amongst respondents. The reliability index is interpreted in a comparable way to Cronbach's alpha coefficient where 0.85 is considered a minimal value for individual or clinical use (8, 11).

According to the Rasch methodology, when the data fit the Rasch model, the Rasch scale has only one dimension in the data. Unidimensionality was assessed by running a Rasch principal components analysis (PCA) of the residual differences between observed and expected responses (10). By running a Rasch PCA of residuals, researcher is searching for evidence of a component that explains a large amount of variance in the residuals. This component is named as the first contrast. Generally, the presence of unidimensionality is defined when the first contrast have an eigenvalue less than 2 (17, 18).

Results

Table 1. presents difficulty measures for each item. Items are arranged in ascending order of difficulty from top to bottom.

Table 1 Item difficulties of the HAM-D

№ item	Name of item	Difficulty	Standard error
14	Genital symptoms	-0.51	0.07
1	Depressed mood	-0.21	0.05
13	Somatic symptoms	-0.39	0.08
7	Work, activities	0.4	0.05
12	Somatic symptoms gastrointestinals	0.51	0.08
16	Loss of weight	0.83	0.07
4	Insomnia early	0.87	0.07
6	Insomnia late	0.87	0.07
10	Anxiety psychic	0.94	0.05
2	Feeling of guilt	0.98	0.05
17	Insight	1.12	0.08
3	Suicide	1.24	0.05
15	Hypochondriasis	1.25	0.05
8	Retardation	1.35	0.05
9	Agitation	1.35	0.05
5	Insomnia middle	1.43	0.08
11	Anxiety somatic	1.47	0.05

Data in Table 1 shows that the item difficulties range from -0.51 to 1.47 logits. This range has moderate width shifted towards moderately high difficulty. There were no items with a significantly low and excessively high levels of difficulty. The

most difficult items didn't reach very high levels. The easiest item was item 14 "genital symptoms" (difficulty score was -0.51). The most difficult item was item 11 "anxiety somatic" (difficulty score was 1.47). In general, the distance between the minimum and maximum level of difficulty was not large and amounted to 1.98 logits.

Figure 1 presents the person–item map. The item categories on the right side (the dots om the vertical segments) are located against the person measures on the left side. As follows from the item-person map, the distribution of person measures was in range of -2.5 to 1.0 logits. The spread of range was 3.5 logits.

Visual analysis of distribution of the person measures shows shift towards low and moderately levels of depression severity within the scale. The right side of the map contains distribution of item categories. Category difficulties ranged from -3.5 to 2.6 logits. The spread of range was 6.1 logits, which significantly exceeds the range of person measures (3.5 logit). Visual analysis of the distribution of item categories shows that most categories are located in the range of high difficulty above the scale mark equal to 0.5 logits. At the same time most person measures are in the range below 0.5 logits.

Infit and outfit statistics for several items were not within the accepted range. Table 2 represents item fit statistics for each item during through six Rasch calibrations. According to table 2 only 8 items had WMS and UMS values within range 0.6 – 1.4. These items were depressed mood, feeling of guilt, suicide, work activites, somaic symptoms gastrointestinals, somatic symptoms, genital symptoms and loss of weight. Other 9 items were outside of accepted range and were removed from the scale.

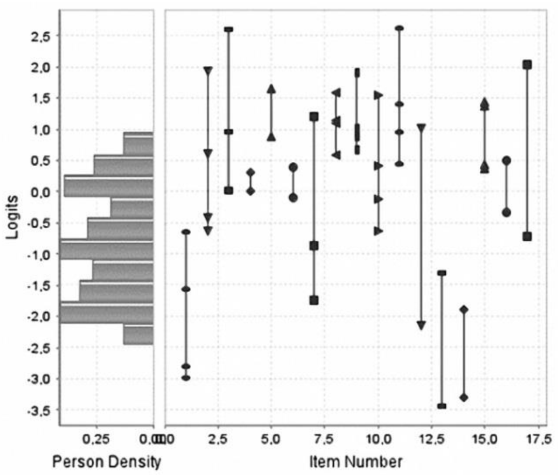

Fig 1. Item-person map of the original HAM-D. The item categories on the right side (the dots on the vertical segments) are located against the person measures on the left side.

PCA of residuals showed that the unexplained variance in the first factor was 21.0% (3.6 eigenvalue units). This bipolar factor consisted of retardation versus agitation, anxiety psychic and anxiety somatic. An interesting finding was that the fourth factor (1.4 eigenvalue units) consisted insomnia early versus insomnia late. Given the presence of these two bipolar factors consisting of the classic depressive symptoms I decided to recombine items in the following way. I combined item 8 “retardation” and item 9 “agitation” on the basis of “either-or” rule into one item called “motor disturbances”. “Either-or” principle means that the maximum score for any of the two items was the score of the new item “motor disturbances”. Then using the same principle I combined all three symptoms of insomnia in one item called “insomnia”. New items were added to the remaining 8 items with acceptable fit-indexes values and then Rasch analysis was performed. Table 3 represents item fit statistics for each item after this Rasch calibration. All 10 items had WMS and UMS values within range 0.6 – 1.4.

tems with the added 2 new items

№	Item	WMS	UMS
1	Depressed mood	0.84	0.88
2	Feeling of guilt	0.87	0.83
3	Suicide	1.24	1.15
4	Insomnia	0.99	1.06
5	Work, activities	0.96	0.87
6	Motor disturbances	0.93	0.91
7	Somatic symptoms gastrointestinals	1.11	1.17
8	Somatic symptoms	1.25	1.37
9	Genital symptoms	0.67	0.63
10	Loss of weight	1.22	1.03

PCA of residuals showed that the first contrast had only 2.0 eigenvalue units and the eigenvalues of the others factors were below 2.

Figure 2 presents the person–item map consisting of the distributions of the person measures on the left side and the distribution

of 10 item category difficulties on the right side. Visual analysis of the item-person map shows that the item categories are uniformly distributed over the entire range of the person measures. The distribution of the item category difficulties completely covers the spread of the person measures.
The PSI for original SIGH-D (HAM-D) exceeded the critical threshold of 2.0 but was not high (2.32). The number of strata was 3.43 indicating an ability to discriminate between patients with 3 or 4 different levels of depression. The person reliability was also not very high (0.84).
Reliability analysis of the final model consisting of 10 items showed significantly higher estimates. The PSI was 3.61. The number of strata rose to 5.14, indicating an ability to discriminate between patients with 5 statistical different levels of depression. The person reliability was 0.93.

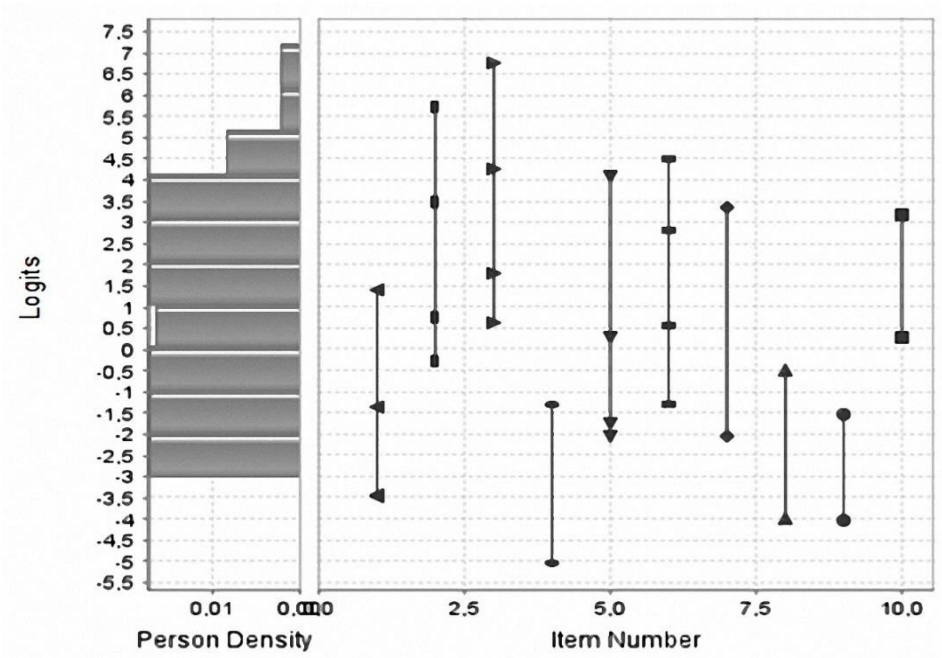

Fig 2. Item-person map of the revised HAM-D. The item categories on the right side (the dots on the vertical segments) are located against the person measures on the left side.

Conclusion

The present work was aimed at analyzing the psychometric properties of the Hamilton rating scale for depression in a sample of Belarusian patients applying the Rasch model. This was the first study to undertake a rigorous examination of the psychometric properties of the HAM-D using Rasch analysis in Belarus. The results showed that the HAM-D had unsatisfactory properties according to the Rasch model requirements. Several items had very high values of fit-indexes. This means that these items were characterized by poor construct validity. Such items as Agitation – Retardation and Insomnia-items had low construct validity due to conflicting behavior towards each other within the scale. The antagonistic relationship between these points was not taken into account in the original design of the scale.
Targeting was not good; most item categories were located in the range of high difficulty. The analysis of item-person map shows disagreement between distributions of person measures and item categories on the scale continuum of the HAM-D. Most categories are designed to measure much severe depression, which is not currently defined in patients. The explanation of this phenomenon may lie in the fact that HAM-D was developed in the 1950-1960's (4, 19), when clinical cases of considerably severe depression were fairly common. Today the frequency of severe

depression decreased significantly due to the clinical pathomorphism of mental disorders and more effective treatments. In general, Figure 1 shows that distribution range of item categories covers the distribution range of person measures. Nevertheless shifting of item category difficulties towards more severe depression may cause a decrease in the accuracy of assessment of mild to moderate depression severity. Poor targeting led to decrease the accuracy of assessing the severity of depression, especially in the cases of mild degrees of its severity.

The dimensionality of the original scale had more than one dimension. It included dimensions, not related to the construct of depression. Person separation statistics and reliability values were not high. These results are similar to those from Rasch analysis of the HAM-D (3, 5, 20), which have also shown that the HAM-D contains some misfitting items, has poor targeting of item categories and heterogeneous dimensionality.

After removing misfitting items and recombining 5 items with reciprocal behavior Rasch-analysis showed much better results. All items had appropriate construct validity; item categories showed good targeting on person measures. Modified scale had unidimensional structure. Person separation statistics and reliability indexes obtained high values. Improving the psychometric characteristics will increase the accuracy of assessing severity of depression using HAM-D. The next stage of the study should be aimed at developing new cut-off scores for assessing the severity of depression using a modified version of the scale.

In conclusion, this study revealed several valuable insights into the psychometric properties of the HAM-D. We confirmed that the Belarusian version of the Hamilton scale has several psychometric flaws. The main problem lies in the fact that several items of the scale has bad construct validity which was manifested in values of Rasch fit-indexes. Some items have bipolar relationships with each other which leads to a reduction of their validity. Separation reliability is also insufficient. After removing invalid items and changing score rules of bipolar items psychometric properties of the scale were improved.

References

1. Cusin C, Yang H, Yeung A, Fava M. Rating Scales for Depression. In: Baer L, Blais MA, editor. Handbook of Clinical Rating Scales and Assessment in Psychiatry and Mental Health. Totowa, NJ: Humana Press; 2009. p. 7–35.
2. McDowell I. Measuring Health: A Guide to Rating Scales and Questionnaires. 3rd ed. New York: Oxford University Press; 2006.
3. Bagby RM, Ryder AG, Schuller DR, Marshall MB. The Hamilton Depression Rating Scale: Has the Gold Standard Become a Lead Weight? American Journal of Psychiatry 2004; 161:2163-2177 .
4. Hamilton M. A rating scale for depression. J. Neurol. Neurosurg. Psychiatr. 1960:56–62.
5. Bech P, Allerup P, Larsen ER, Csillag C, Licht RW. The Hamilton Depression Scale (HAM-D) and the Montgomery–Åsberg Depression Scale (MADRS). A psychometric re-analysis of the European Genome-Based Therapeutic Drugs for Depression Study using Rasch analysis. Psychiatry Research 2014; 217(3):226–32.
6. Lord, F. M., Novick M.R. Statistical theories of mental test scores: Addison-Wesley Pub. Co.; 1968.
7. Crocker L, Algina J. Introduction to Classical and Modern Test Theory: Cengage Learning; 2008.
8. Bond TG, Fox CM. Applying the Rasch Model: Fundamental Measurement in the Human Sciences. 2nd: LEA; 2007.
9. Green K, Frantom C. Survey development and validation with the Rasch model: Paper presented at the International Conference on Questionnaire Development, Evaluation and Testing, Charleston, SC, 14-17 November.; 2002.
10. Jafari P, Bagheri Z, Ayatollahi S, Soltani Z. Using Rasch rating scale model to reassess the psychometric properties of the Persian version of the PedsQLTM 4.0 Generic Core Scales in school children. Health and Quality of Life Outcomes 2012; 10(1):27–38.
11. DeMars C. Item response theory: Oxford University Press; 2010.
12. Wright BD, Stone MH. Measurement essentials. 2nd: Wilmington, Delaware: WIDE RANGE, INC; 1999.
13. Meyer JP. Applied Measurement with jMetrik: Routledge; 2014.
14. Pallant J, Tennant A. An introduction to the Rasch measurement model: An example of suing the Hospital Anxiety ande Depsression Scale (HADS). B J Clin Psych 2007; 46:1–18.

15. Zucca A, Lambert SD, Boyes AW, Pallant JF. Rasch analysis of the Mini-Mental Adjustment to Cancer Scale (mini-MAC) among a heterogeneous sample of long-term cancer survivors: A cross-sectional study. Health and Quality of Life Outcomes 2012; 10(1):55.
16. Wright BD, Linacre J.M. Reasonable mean-square fit values. Rasch Measurement Transaction 1994; 8(3):370.
17. Linacre JM. Detecting multidimensionality: Which residual data-type works best? Journal of Outcome Measurement 1998; 2(3):266–83.
18. Linacre JM. Local independence and residual covariance: A study of olympic figure skating ratings. Journal of Applied Measurement 2009; 10(2):157–69.
19. Hamilton M. Development of a rating scale for primary depressive illness. The British journal of social and clinical psychology 1967; 6(4):278–96.
20. Bech P, Allerup P, Gram LF, Reisby N, Rosenberg R, Jacobsen O et al. The Hamilton depression scale. Evaluation of objectivity using logistic models. Acta psychiatrica Scandinavica 1981; 63(3):290–9.