Original Article

Primary Principles in Developing Scale with Rasch Analysis: Portfolio Anxiety Assessment

L Tomak, O Midik[1]

Departments of Biostatistics and Medical Informatics, [1]Medical Education, Medical Faculty, Ondokuz Mayis University, Samsun, Turkey

ABSTRACT

Background: Rasch model is a useful method for developing a new scale. This study aims to determine the fitting between data obtained from answers for a portfolio anxiety scale and Rasch model and describes how the scale can be modified to increase the fitting through different steps. **Materials and Methods:** A portfolio scale was applied to 171 students of the Faculty of Medicine, Ondokuz Mayis University. The partial credit model was used, and fit statistics were assessed to determine the fitting of the data to Rasch model. Person separation index (PSI) was used for reliability. **Results:** For a satisfaction subscale, the average item fit residual value was 0.47 and the average person fit residual value was −0.29. For the item–trait χ^2 interaction, $P = 0.655$ and PSI = 0.81. For a writing anxiety subscale, the average item fit residual value was 0.08 and the average person fit residual value was −0.24. For the item–trait χ^2 interaction, $P = 0.698$ and PSI = 0.73. For a reflection anxiety subscale, the average item fit residual value was 0.64 and the average item fit residual value was 0.64. For the item–trait χ^2 interaction, $P = 0.195$ and PSI = 0.73. **Conclusion:** The validity and reliability of Rasch analysis portfolio scale were analyzed, and items that worked well were included in the study. The results show that Rasch model provides a more accurate analysis for developing and adapting scales. Both the fit statistics and fit graphs help improve the analyses.

KEYWORDS: *Partial credit model, portfolio, Rasch model, scale development*

INTRODUCTION

*P*sychometric analyses are important for developing scales to assess attitudes and behaviors about factors such as life quality and anxiety and for improving an existing scale.[1-4] A scale is developed to create a reliable, valid, and flexible assessment instrument that contains appropriate items. All items in a scale are used for assessment, and the ideal is to perform the most effective assessment using the least possible number of items.[5]

Classical test theory (CTT) is the most extensively and frequently used assessment method. In classical analysis, any test score consists of the sum of correct value and random error. In classical analysis hypothesis, error is normally distributed and the average of error equals to zero.[1,6] An individual's raw score from the assessment instrument is an indicator of the level he has or she the

characteristic that is being measured. An individual's score as a result of repeated measurements changes even if the individual's ability is stable and this variability is the basis of CTT. The most important advantage of CTT is its relatively weak theoretical assumptions and its being easily applicable for most tests.[7,8]

Rasch model, developed as alternative to CTT, aims to express and model the association between individuals' behaviors and the characteristics that influence these behaviors; the potential of these characteristics to exist is assumed through probability-based functions.[9] It explains an individual's degree of having a characteristic

Address for correspondence: Dr. L Tomak, Department of Biostatistics and Medical Informatics, Medical Faculty, Ondokuz Mayis University, Samsun, Turkey. E-mail: lemant@omu.edu.tr

Access this article online	
Quick Response Code:	**Website**: www.njcponline.com
	DOI: 10.4103/njcp.njcp_275_17

as a mathematical model based on the association between the individual's answer to an item and the parameters that describe this item. Individuals with higher potential variable (θ) degree are most probably those who give the correct answer.[10] In addition to calibrating items, this model uses answers to items and assesses the characteristics of items. Because the test statistics do not depend on the test structure, this method is easier to use than the classical method.[11] It enables correlated errors to be dealt with easily. Its main purpose is to measure the primary ability of the test, which shows its performance. This works independent of the sample.[12] There is an assumption of unidimensionality and according to this assumption, a test measures only one latent characteristic and an individual with a high score has a high probability of answering an item correctly. This model is mathematically correct and if the assumptions are met, it can easily solve most complicated problems.[13,14] Rasch model is frequently preferred for items with two choices; partial credit model and rating scale model are frequently preferred for items with multiple choices.[15]

Portfolios are files that record the progress of students or individuals in line with specific goals within a specific period of time.[16,17] They show progress in many different areas. In medical education, portfolios are increasingly being used for assessing education, mainly because they can express the development of occupational competence perception, contribution to assessment and evaluation (especially for areas such as personal development, self-oriented learning, reflective skills, professionalism, and reasoning that are difficult to assess using other assessment methods), contribution to personal attitude, encouragement of interaction between student and teacher, and increase in use of reflective strategies. In addition to achievement attainments, portfolios can also reflect some negative features such as students' anxieties about the processes for the creation and assessment of portfolios. These anxieties have been expressed qualitatively in previous studies; however, no study has developed a quantitative scale for this purpose.[18,19]

This study aims to develop a portfolio anxiety scale to determine the fitting between the data obtained from answers and Rasch model and to explain how the scale can be modified to improve this fitting through different steps.

MATERIALS AND METHODS

Participants

In this study, a scale developed to determine students' portfolio anxiety was applied to second-year students from the Faculty of Medicine, Ondokuz Mayis University, in 2015 and local ethics committee approved the study protocol (B.30.2.ODM.0.20.08/765). Students who were introduced to portfolios for the first time in elective medicine blocks in their second year were asked to reflect on their attitudes and behaviors. The questionnaire was administered to 200 students, and 171 valid responses were obtained.

Scale structure

The items developed were assessed in terms of content validity to determine whether they were quantitatively and qualitatively adequate for assessing students' behaviors.[20-22] Expert views were considered for assessing content validity; 6 items were removed, and 31 items were assessed.[20,21]

The items were categorized as strongly disagree, disagree, partly disagree, partly agree, agree, and strongly agree, with the corresponding scoring being 1, 2, 3, 4, 5, and 6; for inverse items 1, 2, and 5, the scoring was 6, 5, 4, 3, 2, and 1, respectively. A higher score indicates a higher fear of the portfolio.

To assess the validity and reliability of the items, all answers were analyzed using Rasch Unidimensional Measurement Model 2030 (RUMM 2030).[23] This program compares the observed and expected values for each class interval (CI) by sorting person locations and dividing them in equal numbers into CIs. Also other analyses were performed using SPSS 18.[24]

Rasch model

Rasch model is the first item response theory (IRT) model developed with one parameter. It includes only a difficulty parameter. The probability of answering an item correctly is defined as a function of the rate of a person's level of ability to an item's difficulty.[11,25] The probability of item "*i*" being answered correctly can be expressed as follows[1]

$$P_i(\theta) = P(x_i = 1/\theta) = \frac{e^{(\theta - b_i)}}{1 + e^{(\theta - b_i)}} = \frac{\exp(\theta - b_i)}{1 + \exp(\theta - b_i)} \quad (1)$$

where b_i is the item difficulty parameter for item i. The probability changes from 0 for $\theta = -\infty$ to 1 for $\theta = \infty$.[25]

Well-known Rasch models for answers with more than two categories include partial credit model and rating scale model. Partial credit model is a simple adaptation of Rasch model with two choices. It has no limitations for threshold values being the same in all items. The assessment is made with one parameter instead of two parameters as the difficulty of choices in an item and difficulty of items in the test.[26,27] In the partial credit model, categories change from 0 to m;

the probabilities for categories other than 0 are given as follows:[5]

$$P_{ik}(\theta) = \frac{\exp\left[\sum_{j=1}^{k}(\theta - \delta_{ij})\right]}{1 + \sum_{h=1}^{m}\exp\left[\sum_{j=1}^{h}(\theta - \delta_{ij})\right]} \quad (2)$$

Where $\ddot{a}_{ij}$ is the difficulty step probability, and the difficulty step parameter is given as

$P_{ik}/(P_{ik} + P_{i, k-1})$.[27]

Rating scale model is specifically used for Likert-type scales.[26] This model also includes the difficulty parameters of the thresholds. The subject's probability of answering an item correctly is obtained by assessing the item's difficulty level and the choice's threshold difficulty level together. The probability of item i to choose category k for the m + 1 score category can be calculated as follows:[26,28]

$$P_{ik}(\theta) = \frac{\exp\left[k(\theta - b_i) - \sum_{j=0}^{k}\tau_{j(m+1)}\right]}{\sum_{h=0}^{m}\exp\left[h(\theta - b_i) - \sum_{j=0}^{h}\tau_{j(m+1)}\right]} \quad (3)$$

$\tau_{j(m+1)}$ is calculated for all items with m + 1 category, and it expresses the cutoff point for each of these items; $\tau_{0(m+1)} = 0$.

In two-parameter logistic model from IRT family, item discrimination gets in the model in addition to item difficult. The relative significance of an individual's level of ability and item cutoff point is found by the discrimination power of the item. Three-parameter logistic model was developed to assess the effect of the parameter of chance on multiple choice tests which are used in the assessment of education. The third parameter is called pseudo-chance-level parameter.[1,11,29] The required sample size differs between one-, two-, and three-parameter IRT. One-parameter Rasch model needs less sample size when compared with two- or three-parameter IRT models, and this is one of the advantages of the model.[29]

RESULTS

When assessed according to the likelihood (LH) ratio test, the partial credit model was found suitable. The average item fit residual was 0.88, and the standard deviation was 1.32; the average person fit residual was −0.19, and the standard deviation was 1.89. For item–trait interaction, $\chi^2 = 92.23$ and $P = 0.007$. The person separation index (PSI) value was 0.92. More than one dimension was found, and the assessment was made in terms of subscales. In terms of principal component analysis (PCA), the scale was limited to

Table 1: Items in the satisfaction subscale

Item Number	Dimension 1: Satisfaction
1	I think my portfolio is important in terms of my own development.
2	I think my portfolio is useful for my learning process.
3	I have negative feelings about my portfolio.
4	I get bored while writing my portfolio.
5	I take pleasure in writing my portfolio.
6	I think that my portfolio should be canceled when the advantage/disadvantage rate is compared.
14	I think that my portfolio grade will not show my actual competence.
27	I think that my portfolio is time-consuming.
28	I think that the time I allocate for my portfolio is unnecessary.
29	I do not have enough time to write my portfolio.

Table 2: Threshold values for satisfaction subscale

Item	CenThr1	CenThr2	CenThr3	CenThr4	CenThr5
1	−2.59	−0.08	0.18	0.25	2.24
2	−2.59	−0.45	0.19	0.64	2.21
3	−1.79	−0.35	−0.12	0.24	2.02
4	−0.12	−0.40	−0.45	−0.05	1.01
5	−1.43	0.10	−0.02	−0.13	1.48
6	−0.96	−0.69	−0.09	0.60	1.13
14	−0.97	−0.47	−0.34	0.13	1.65
27	−0.22	−0.64	−0.31	0.32	0.84
28	−1.61	−0.32	0.22	0.54	1.16
29	−1.72	−0.25	−0.08	0.17	1.88

Table 3: Item fit statistics for satisfaction subscale

Item	Location	FitResid	χ^2	P
1	0.01	−0.46	5.65	0.059
2	0.01	−1.37	5.11	0.078
3	−0.08	0.96	0.81	0.667
4	0.75	4.13	2.05	0.358
5	0.48	0.46	3.74	0.154
6	0.09	2.28	4.37	0.112
14	−0.21	4.17	12.83	0.002
27	−0.08	−0.36	4.72	0.094
28	−0.55	−2.27	8.31	0.016
29	−0.40	−0.26	0.25	0.881

FitResid=Fit residual

three dimensions that explained 52% of the variance. Three dimensions were detected for portfolio scale.

Satisfaction subscale

Table 1 lists the items in the satisfaction subscale. While the thresholds for items 1, 2, 3, 6, 14, 28, and 29 were correctly ordered, the placement of thresholds for items 4, 5, and 27 was inconsistent with the logical order in the model [Table 2]. Figure 1 shows the assessment of the category probability curve (CPC) for item 2;

Table 4: Summary of Rasch analysis statistics for subdimensions

Satisfaction subscale	Item FitResid Mean (SD)	Person FitResid Mean (SD)	Item-trait χ^2 interaction		PSI
			χ^2	Prop	
Initial	0.77 (2.01)	−0.28 (1.38)	47.93	<0.001	0.87
Rescore	0.73 (2.18)	−0.21 (1.23)	47.86	<0.001	0.87
Eliminated	0.47 (0.82)	−0.29 (1.15)	9.55	0.655	0.81
Writing anxiety subscale					
Initial	0.91 (1.37)	−0.27 (1.41)	28.69	0.154	0.86
Rescore	0.52 (1.56)	−0.26 (1.44)	46.39	0.002	0.87
Eliminated	0.08 (0.88)	−0.24 (0.92)	7.28	0.698	0.73
Reflection anxiety subscale					
Initial	0.76 (1.59)	−0.41 (1.66)	46.30	<0.001	0.84
Rescore	0.22 (1.43)	−0.36 (1.41)	62.42	<0.001	0.79
Eliminated	0.64 (0.98)	−0.37 (1.05)	11.12	0.195	0.73

FitResid=Fit residual; SD=Standard deviation; Prob=Probability; PSI=Person separation index

Table 5: Item fit statistics and scoring structure (after rescoring) for subdimensions

	Item	Location	SE	FitResid	χ^2	Prob
Satisfaction subscale	2	0.03	0.09	−0.19	0.90	0.636
	3	−0.10	0.09	0.49	0.95	0.624
	5	0.48	0.14	1.49	0.99	0.610
	6	0.08	0.08	1.41	0.95	0.624
	27	−0.09	0.10	−0.39	4.80	0.090
	29	−0.40	0.09	0.00	0.97	0.616
Writing anxiety subscale	7	−0.29	0.08	−0.55	0.92	0.632
	11	−0.42	0.13	−0.05	2.95	0.229
	17	−0.50	0.08	−0.88	2.50	0.287
	19	1.42	0.14	0.53	0.50	0.781
	30	−0.21	0.13	1.33	0.42	0.812
Reflection anxiety subscale	23	0.12	0.07	−0.15	5.11	0.077
	24	−0.02	0.08	0.63	0.62	0.735
	25	0.36	0.08	0.08	3.45	0.177
	26	−0.46	0.13	2.02	1.95	0.378

SE=Standard error of mean; FitResid=Fit residual; Prob=Probability

Table 6: Items in the writing anxiety subscale

Dimension 2: Writing anxiety

7	I think that I will not be able to understand/fulfill the questions/tasks well enough while writing my portfolio.
8	I worry that I will use wrong expressions in my portfolio.
9	I worry that I will not be able to answer correctly when asked about the content of my portfolio.
10	I am undecided while writing because the content of my portfolio is flexible.
11	I worry about the grade I will receive for my portfolio.
17	I am scared about making mistakes while writing my portfolio.
18	I worry about how to write my portfolio best.
19	I cannot express my feelings in writing.
20	I think that I will have difficulty in finding the correct words or sentences and bringing them together.
30	I panic as the deadline approaches.
31	I am scared that I will not have the chance to develop what I do and that I will not submit the best while keeping a portfolio.

the thresholds of the items are in hierarchical order. Figure 2 shows the CPC for item 5; the thresholds are disordered. The second threshold is placed after the third and fourth thresholds along the logit scale. Before the fit statistics were found for items with irregularly placed thresholds, the answer categories were organized. To prevent irregularities, the items were rescored. Item 4 was rescored as 000012, 5 as 000112, and 27 as 000123. Figure 3 shows the CPC of rescored item 5. Figure 4 shows the threshold maps of the rescored items.

After ordering the items, fit analyses were assessed [Table 3]. The fit residual was found as 4.13 for item 4 and as 4.17 for 14, exceeding ± 2.5 limits. For item 14, $\chi^2 = 12.83$ and P was <0.005. No fit was found for these items. In a visual assessment of the fit to Rasch model, the item characteristic curve (ICC) was obtained for each item. Item 29 showed good fit; however, item 14 was not consistent with the
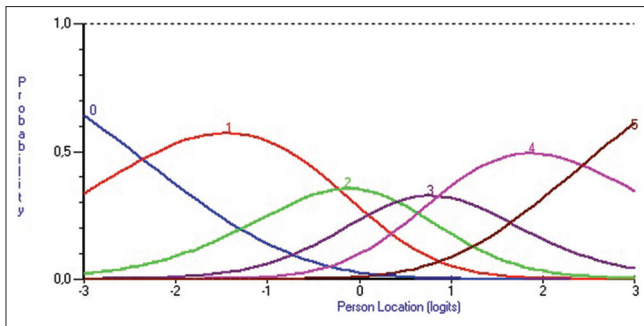
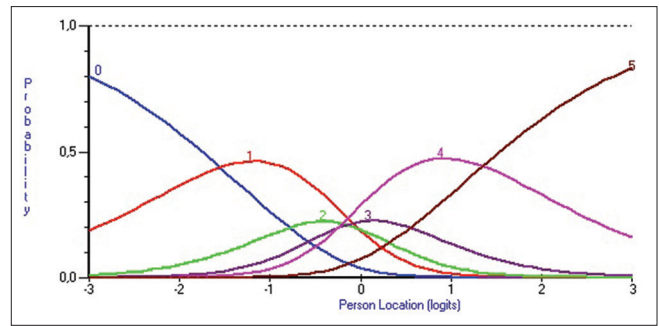**Figure 1:** Category probability curve for item 2 with ordered thresholds

Figure 2: Category probability curve for item 5 with disordered thresholds
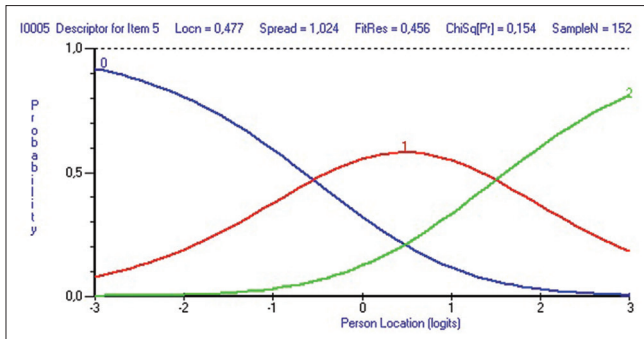

Figure 3: Category probability curve for item 5 with ordered thresholds
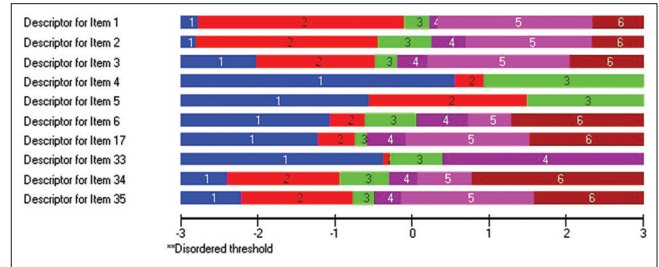

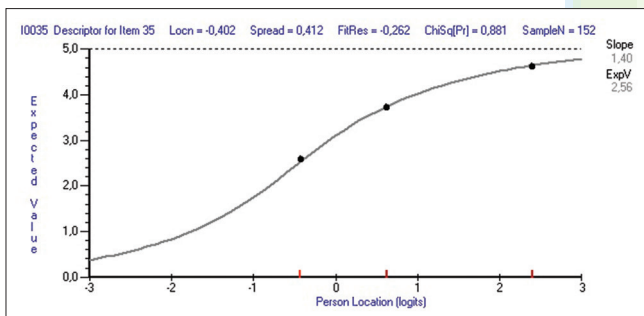**Figure 4:** Threshold map for satisfaction subscale

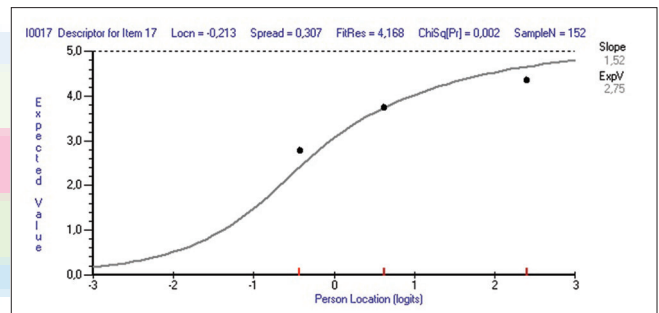**Figure 5:** Item characteristic curve for item 29

Figure 6: Item characteristic curve for item 14

model [Figures 5 and 6]. When the fit residuals for a person were examined, 12 people were found outside the ± 2.5 threshold. The residual correlation between items was assessed for local dependency; it was 0.59 for items 1 and 2, 0.34 for items 4 and 5, and 0.33 for items 27 and 28. The scale was revised by rescoring the items and excluding some. Table 4 shows the pre-revision and post-revision statistics. The item fit residual and item–trait χ^2 interaction values were better after revision. PSI values were over 0.80 initially and after revision. Unidimensionality was examined by PCA of the residuals followed by independent t-test. The satisfaction subscale met criteria for unidimensionality. When the scale was examined for differential item functioning (DIF) in terms of gender, no difference was found. The items and persons were assessed together using a person–item location distribution map. Figure 7

shows the person–item distribution diagram. Table 5 shows the items' locations on the logit scale after revisions. It was found that item 29 was the easiest and item 5 was the most difficult.

Writing anxiety subscale

Table 6 lists the items in the writing anxiety subscale. Items with disordered cutoff points were scored again. The scores for items 8, 9, 11, 18, 19, 30, and 31 were rescored as 000112, whereas those for the others remained the same. The fit residual was 3.15 for item 10 and 3.13 for item 20. P value for χ^2 for items 9 and 18 was <0.0045. The fit statistics of persons were assessed and kept. In terms of local dependency, the residual correlations were −0.32 between items 7 and 31, 0.31 between items 10 and 17, 0.34 between items 17 and 18, and 0.36 between items 19 and 20. The scale was revised by rescoring and excluding items. Table 4 lists the summary statistics before and after revision. With revisions, the item and person fit residuals and item–trait
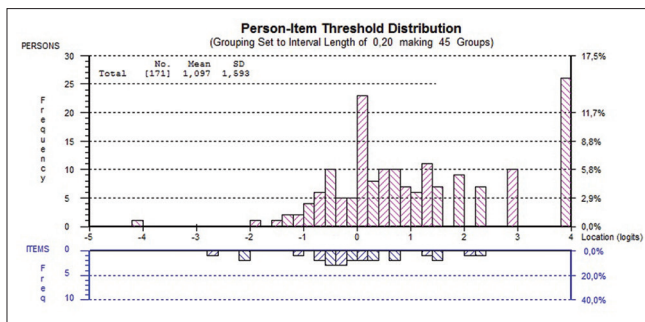
Figure 7: Person–item threshold map

Table 7: Items in the reflection anxiety subscale

	Dimension 3: Reflection anxiety
12	I do not think that my portfolio is graded fairly.
13	I do not want to write what I think owing to grade anxiety.
15	I worry that I will get lower marks because of my negative feelings.
16	I think that the evaluators will give wrong decisions.
21	While keeping a portfolio, I do not want to share my weaknesses.
22	I do not want to write things about myself in the portfolio.
23	I do not want to write and share my fears and anxieties while keeping a portfolio.
24	I do not know how to reflect.
25	I find it difficult to make a self-assessment.
26	What I write in the portfolio does not show my actual thoughts.

interactions were found to improve. The PSI value was 0.73. This subscale met criteria for unidimensionality. No DIF was found. Table 5 shows the locations of items after revisions on the logit scale. Item 17 was the easiest and item 19 was the most difficult.

Reflection anxiety subscale

Table 7 lists the items in the reflection anxiety subscale. In this scale, items with disordered threshold were rescored as 12, 13, 15, 16, 21, 22, and 26, and their scores were rescored as 011222. For all items, the fit residuals were within ±2.5. P value of χ^2 for items 21 and 22 was <0.005. The fit statistics of persons were assessed and kept. Multiple relations were found for many items. The residual correlation was over 0.3 between items 12 and 16–21–25; 13 and 15–23; 15 and 16–25; 16 and 21–25; 21 and 22; and 22 and 23. Items except those between 23 and 26 were excluded, and the subscale was finalized. Table 4 lists the summary statistics before and after revision. With revisions, the item and person fit residuals and item–trait interactions improved. The PSI value of the latest version was 0.73. The reflection anxiety subscale met criteria for unidimensionality. No DIF was found. Table 5 lists the

locations of items on the logit scale after revisions. Item 26 was the easiest and item 25 was the most difficult.

DISCUSSION

At the beginning of the study, the partial credit model was chosen according to the LH ratio test. The assumptions of this model are better than those of the rating scale model.[26,30] More than one dimension was found, and the analysis continued through these dimensions. The scale with three dimensions explained 52% of the variance.

In Rasch model, the answer is shown with CPC for each item section and category. For the items, the answer curves are shown with decision increasing from left to right. Misorder may be seen in the cutoff points because the answer category is too much, categories are overlapped, or there are too many dimensions.[31] High and low scores indicate high and low anxiety levels, respectively. It is very important to assess the logit locations of threshold points. The cutoff points for items 4, 5, and 27 in the satisfaction scale were not logically ordered. The disordered thresholds are very important from the viewpoint of validity and reliability. In case of such a problem, first, the problem should be solved, and then the fit statistics should be assessed.[32,33] Mislocated items were rescored first.[30]

Various methods are used to assess the fit of items using Rasch model.[33] One of these is the analysis of individual item fit values; standardized residuals are expected to be within ±2.5.[33,34] In the first dimension, the residuals of items 4 and 14 were over +2.5. High positive residuals indicate deviation, whereas high negative residuals indicate local dependency.[33]

With the χ^2 test for consistency, the difference between the observed and the expected values for a specific ICC is examined. The average responses of persons in each CI are indicated graphically by one point, whereas the expected values are indicated by a curve. When P value of the χ^2 test is less than 0.05, the difference between the values and the fitting with the model is weak.[35,36] The χ^2 test for item 14 and ICC showed that the model fit was not good. If the residual correlation values between items are over 0.3, there is a local dependency.[37,38] Thus, local dependency was found between items 1 and 2, 4 and 5, and 27 and 28, and items other than 2, 3, 5, 6, 27, and 29 were excluded.

In Rasch model, the average item fit residuals and item–trait interaction statistics assess the general fitting of items. The average value of all items should be close to 0, and the standard deviation should be close to 1. In the latest satisfaction subscale, the average item

fit residual was close to 0, and the standard deviation was close to 1. These values and item–trait interactions are indicators of fitting; they show that the items work together, have internal consistency, and measure only one characteristic.[32,33]

The reliability of the scale is assessed using PSI, which indicates the power of the scale for differentiating between persons. PSI was 0.81 for the latest version of the satisfaction scale, indicating that the reliability is good. For scale reliability, the required minimum cutoff point is 0.70.[39]

Unidimensionality was evaluated through PCA of the residuals. Items were divided into two subsets based on positive or negative loading on the first residual, and for each respondent, person estimates were derived from each subset and compared by t-test. Unidimensionality is regarded as supported if <5% of the t-tests are significant outside the ±1.96 range and the 95% binomial confidence intervals include 5%.[39-41] When positive and negative residuals were compared using a t-test, the scale was found to be unidimensional.

The DIF assesses whether subgroups give different answers to items systematically.[42] In an F-test to determine the DIF, if the variances of two subgroups are equal, they come from the same population and there is no DIF.[31] No difference was found in terms of gender.

The person and model fitting were assessed using the person fit residual. Fit residual value less than −2.5 shows the presence of a mental situation or thought, and a value higher than +2.5 shows carelessness or low motivation.[32] In a general assessment of persons' fit to the model, an approximate normal distribution with average of 0 and deviation of 1 is expected. These values were attained with the latest version, showing that person consistency is within acceptable limits.

The items and persons are assessed using a person–item location distribution map. The person and item locations are shown together on the same axis. This graph shows that the items have a good distribution; however, it also shows that it does not have sufficient measurement for assessing some people.[32]

The disordered items in the reflection anxiety subscale were rescored and assessed for fit statistics. Items with a fit residual beyond ± 2.5 for χ^2 value—10 and 20 and 9 and 18—were assessed as being inconsistent. Items 7–31, 10–17, 17 and 18, and 19 and 20 were assessed as being locally dependent. The scale was revised and finalized. After inconsistent items and items with local dependency were excluded, items 7, 11, 17, 19, and 30 were left. The summary fit residuals for person and item and

item–trait interaction show that the model is consistent. The PSI value of 0.73 indicates reliability. The scale is unidimensional, and it assesses only one trait.

Disordered items in the writing anxiety subscale were rescored and assessed for fit statistics. No items showed a fit residual less than ±2.5; items 21 and 22 for χ^2 value were assessed as being inconsistent. The correlation between many items (12, 13, 15, 16, 21, 23, 24, 25) was greater than 0.3, and these were assessed as being locally dependent. The scale was revised and finalized. After inconsistent items and items with local dependency were excluded, items 23, 24, 25, and 26 were left. The summary fit residuals for person and item and item–trait interaction show that the model is consistent. The PSI value of 0.73 indicates reliability. There is no DIF. The scale is unidimensional, and it assesses only one trait.

CONCLUSION

In this study, Rasch model was used for assessing a scale developed for portfolio anxiety. Rasch model is a valuable model with item–trait χ^2 interaction statistics, item and person detailed and summary fit statistics, CPC, ICC, and person-item map. A scale analyzed within three dimensions was finalized using this method. However, some items had to be excluded because they showed multiple correlations. These items can be reevaluated in future studies. In addition to these, this study has some limitations. These can be expressed as the lack of test–retest analysis to assess the reliability of the scale, the generalizability, and strong assumptions of Rasch analysis. This study, which is the first one on this subject, can be improved and extended through future studies.

Financial support and sponsorship
None.

Conflicts of interest
There are no conflicts of interest.

REFERENCES

1. De Grutijter DN, Van der Kamp LJ. Statistical Test Theory for the Behavioral Sciences. London: Chapman & Hall; 2008.
2. Chang CC, Su JA, Chang KC, Lin CY, Koschorke M, Rüsch N, Thornicroft G. Development of the Family Stigma Stress Scale (FSSS) for detecting stigma stress in caregivers of people with mental illness. Eval Health Prof 2017; DOI: 10.1177/0163278717745658.
3. Lin CY, Griffiths MD, Pakpour AH. Psychometric evaluation of Persian Nomophobia Questionnaire (NMP-Q): Differential item functioning and measurement invariance across gender. J Behav Addict 2018;7:100-8.
4. Lin CY, Broström A, Nilsen P, Griffiths MD, Pakpour AH. Psychometric validation of the Bergen Social Media Addiction Scale using classic test theory and Rasch models. J Behav Addict 2017;6:620-9.

5. Allen DD. Validity and reliability of the movement ability measure: A self-report instrument proposed for assessing movement across diagnoses and ability levels. Phys Ther 2011;87:899-916.

6. Chang CC, Su JA, Tsai CS, Yen CF, Liu JH, Lin CY. Rasch analysis suggested three unidimensional domains for Affiliate Stigma Scale: Additional psychometric evaluation. J Clin Epidemiol 2015;68:674-83.

7. Lin CY, Yang SC, Lai WW, Su WC, Wang JD*. Rasch models suggested the satisfactory psychometric properties of the World Health Organization Quality of Life—Brief among lung cancer patients. J Health Psychol 2017;22:397-408.

8. DeVellis RF. Classical test theory. Med Care 2006;44:50-9.

9. Crocker L, Algina J. Introduction to Classical and Modern Test Theory. Mason, OH: Cengage Learning; 2008.

10. Rasch G. Probabilistic Models for Some Intelligence and Attainment Tests. Chicago, IL: University of Chicago; 1960.

11. Recklase MD. Multidimensional Item Response Theory. New York: Springer; 2009.

12. Demars C. Item Response Theory. New York: Oxford University; 2010.

13. Chang KC, Wang JD, Tang HP, Cheng CM, Lin CY. Psychometric evaluation using Rasch analysis of the WHOQOL-BREF in heroin-dependent people undergoing methadone maintenance treatment: Further item validation. Health Qual Life Outcomes 2014;12:148.

14. Strong C, Lin YC, Tsai MC, Lin CY. Factor structure of Sizing Me Up, a self-reported weight-related quality of life instrument, in community children across weight status. Child Obesity 2017;13:111-9.

15. Thorpe GL, Favia A. Data analysis using item response theory methodology: An introduction to selected programs and applications. Psychol Faculty Scholarship 2012;20:1-33.

16. Driessen E, Van Tartwijk J. Portfolios in personal and professional development. In: Swanwick T, editor. Understanding Medical Education: Evidence, Theory and Practice. 2nd ed. New York: Wiley; 2013. p. 193-201.

17. Buckley S, Coleman J, Davison I, Khan KS, Zamora J, Malick S, et al. The educational effects of portfolios on undergraduate student learning: A best evidence medical education (BEME) systematic review. BEME Guide No. 11. Med Teach 2009;31:282-98.

18. Driessen E, Van Tartwijk J, Van Der Vleuten C, Wass V. Portfolios in medical education: Why do they meet with mixed success? A systematic review. Med Educ 2007;41:1224-33.

19. Davis MH, Ponnamperuma GG, Ker JS. Student perceptions of a portfolio assesment process. Med Educ 2009;43:89-98.

20. Davis MH, Friedman BD, Harden RM, Howie P, Ker J, McGhee C, et al. Portfolio assessment in medical students' final examinations. Med Teach 2001;23:357-66.

21. Shen M, Hu M, Sun Z. Development and validation of brief scales to measure emotional and behavioural problems among Chinese adolescents. BMJ Open 2017;7:e012961.

22. Kozlov E, Carpenter BD, Rodebaugh TL. Development and validation of the Palliative Care Knowledge Scale (PaCKS). Palliat Support Care 2016;1-11.

23. Andrich D, Sheridan B, Luo G. RUMM 2030 Version 5.4 for Windows. RUMM Laboratory Pty Ltd.; 2012.

24. IBM Corp. IBM SPSS Statistics for Windows, Version 21.0. Armonk, NY: IBM Corp; 2012.

25. Brodin U, Fors U, Laksov KB. The application of item response theory on a teaching strategy profile questionnaire. BMC Med Educ 2010;10:14.

26. Zheng X, Rabe-Hesketh S. Estimating parameters of dichotomous and ordinal item response models with gllamm. Stata J 2007;7:313-33.

27. Olsen R, Garrat A, Iversen H, Bjertnaes O. Rasch analysis of the psychiatric out-patient experiences questionnaire (POPEQ). BMC Health Serv Res 2010;10:282.

28. Jafari P, Bagheri Z, Ayatollahi SM, Soltani Z. Using Rasch rating scale model to reassess the psychometric properties of the Persian version of the PedsQL™ 4.0 Generic Core Scales in school children. Health Qual Life Outcomes 2012;10:27.

29. Chang CC, Lin CY, Gronholm PC, Wu TH. Cross-validation of two commonly used self-stigma measures, Taiwan versions of the Internalized Stigma Mental Illness scale and Self-Stigma Scale-Short, for people with mental illness. Assessment 2016 DOI: 10.1177/1073191116658547

30. Das Nair R, Moreton BJ, Lincoln NB. Rasch analysis of the Nottingham extended activities of daily living scale. J Rehabil Med 2011;43:944-50.

31. Andrich D. Controversy and the Rasch model: A characteristic of incompatible paradigms? Med Care 2004;42:I7-16.

32. Gibbons CJ, Mills RJ, Thornton EW. Rasch analysis of the hospital anxiety and depression scale (HADS) for use in motor neurone disease. Health Qual Life Outcomes 2011;9:82.

33. Hendriks J, Fyfe S, Styles I, Skinner SR, Merriman G. Scale construction utilizing the Rasch unidimensional measurement model: A measurement of adolescent attitudes towards abortion. Australas Med J 2012;5:251-61.

34. Neu D, Mairesse O, Hoffmann G, Valsamis JB, Verbanck P, Linkowski P, et al. Do 'sleepy' and 'tired' go together? Rasch analysis of the relationships between sleepiness, fatigue and nonrestorative sleep complaints in a nonclinical population sample. Neuroepidemiology 2010;35:1-11.

35. Adedoyin OO. Using IRT Approach to detect gender biased items in public examination: A case study of the Botswana junior certificate examination in mathematics. Educ Res Rev 2010;5:385-99.

36. Adedoyin OO, Adedoyin JA. Assessing the comparability between classical test theory (CTT) and item response theory (IRT) models in estimating test item parameters. Herald J Educ General Stud 2013;2:107-14.

37. Ramp M, Khan F, Misajon RA, Pallant JF. Rasch analysis of the Multiple Sclerosis Impact Scale MSIS-29. Health Qual Life Outcomes 2009;7:58.

38. Pallant JF, Keenan AM, Misajon R, Conaghan PG, Tennant A. Measuring the impact and distress of osteoarthritis from the patients' perspective. Health Qual Life Outcomes 2009;7:37.

39. Tennant A, Conaghan PG. The Rasch measurement model in rheumatology: What is it and why use it? When should it be applied, and what should one look for in a Rasch paper? Arthritis Rheum 2007;57:1358-62.

40. Tennant A, Pallant J. Unidimensionality matters. Rasch Meas Trans 2006;20:1048-51.

41. Smith EV Jr. Detecting and evaluating the impact of multidimensionality using item fit statistics and principal component analysis of residuals. J Appl Meas 2002;3:205-31.

42. Borsboom D. When does measurement invariance matter? Med Care 2006;44:176-81.