ORİJİNAL ARAŞTIRMA ORIGINAL RESEARCH

DOI: 10.5336/biostatic.2019-72009

The Effect of Different Strategies for Combining Disordered Thresholds on Rasch Model Fit

Sırasız Eşik Değerlerinin Birleştirilmesinde Farklı Stratejilerin Rasch Modeline Uyum Üzerindeki Etkisi

Ömer Faruk DADAŞ[a], Derya GÖKMEN[b], Timur KÖSE[a]

[a]Ege University Faculty of Medicine, Department of Biostatistics and Medical Informatics, İzmir, TURKEY
[b]Ankara University Faculty of Medicine, Department of Biostatistics, Ankara, TURKEY

This study was presented as oral presentation at XX. National and III. International Biostatistics Congress, 26-29 October 2018, Gaziantep, Turkey.

ABSTRACT Objective: The internal construct validity of the scales is examined by Rasch analysis. There are limited number of studies on how to combine disordered thresholds in Rasch analysis. In this study, new four different category-combining strategies are proposed in order to combine disordered thresholds. The effects of these strategies on the overall fit of the Rasch model are investigated. **Material and Methods:** The strategies are obtained by combining the categories with disordered thresholds, "to the left, to the right, to middle-left, and to middle-right". It is decided to use the Partial Credit Model as the appropriate Rasch model. The data is obtained using the measurement tool, which assesses the disability levels of patients with rheumatoid arthritis. The internal construct validity of this scale is evaluated using RUMM2030 software. **Results:** In the measurement tools obtained after the category combining, information loss occurred due to the combining of the categories and it has been seen that standard errors of the ability of individuals have risen. However, the model fit and internal consistency of measurement tools have increased after the combining strategies. **Conclusion:** The category combining strategies are carried out on only one sample data, and since the single strategy is applied for all items in the combining process, the size of the disordered thresholds is ignored. Therefore, it is necessary to carry out a simulation study in order to generalize the results obtained.

Keywords: Disordered thresholds; model fit; Rasch analysis; Rasch models

ÖZET Amaç: Ölçeklerin içsel yapı geçerliliği Rasch analizi ile incelenmektedir. Rasch analizinde sırasız eşik değerlerin nasıl birleştirilmesi gerektiğine dair sınırlı sayıda çalışma bulunmaktadır. Bu çalışmada sırasız eşik değerleri birleştirmek için kullanılabilecek yeni dört farklı kategori birleştirme stratejisi önerilmiştir. Bu birleştirme stratejilerinin Rasch modelin genel uyumu üzerindeki etkileri araştırılacaktır. **Gereç ve Yöntemler:** Stratejiler, sırasız eşik değerlere sahip kategorilerin "sola, sağa, orta-sola ve orta-sağa" birleştirilmesi ile elde edilmiştir. Uygun Rasch modeli olarak Kısmi Kredi Modelin kullanılmasına karar verilmiştir. Veriler, romatoid artritli hastaların özürlülük seviyelerini değerlendiren ölçme aracı kullanılarak elde edilmiştir. Bu ölçeğin içsel yapı geçerliliği RUMM2030 yazılımı kullanılarak değerlendirilmiştir. **Bulgular:** Kategori birleştirmeden sonra elde edilen ölçme araçlarında kategorilerin birleştirilmesine bağlı olarak bilgi kaybının meydana geldiği ve bireylerin incelenen özellik seviyelerine ait standart hatalarının arttığı görülmüştür. Ancak birleştirme stratejilerinden sonra ölçme araçlarının model uyumunun ve iç tutarlılığının arttığı görülmüştür. **Sonuç:** Kategori birleştirme stratejileri yalnızca bir örnek veri üzerinde gerçekleştirilmiştir ve birleştirme sürecinde tüm maddeler için tek bir strateji uygulandığından sırasız eşik değerlerin büyüklüğü göz ardı edilmiştir. Bu nedenle elde edilen sonuçları genelleştirmek için bir simülasyon çalışmasının yapılmasına ihtiyaç vardır.

Anahtar kelimeler: Sırasız eşik değerler; model uyumu; Rasch analizi; Rasch modeler

In the examination of the internal construct validity of the measurement tools, Rasch analysis is one of the most common used approaches in the context of item response theory (IRT). Rasch analysis is a method that

Correspondence: Timur KÖSE
Ege University Faculty of Medicine, Department of Biostatistics and Medical Informatics, İzmir, TURKEY/TÜRKİYE
E-mail: timur.kose@ege.edu.tr

Peer review under responsibility of Turkiye Klinikleri Journal of Biostatistics.

Received: 26 Oct 2019 ***Received in revised form:*** 28 Nov 2019 ***Accepted:*** 09 Dec 2019 ***Available online:*** 26 Dec 2019

can be applied in areas such as health, education, psychology, and social sciences,and it was first introduced in the Health Sciences Literature at the end of the 1970s. After the article written by Wright and Linacre in 1989, its use in this field has increased rapidly.[1,2] In the Rasch analysis, the responses obtained with the ordered scale are converted to the interval variable by examining whether the data set conforms to a mathematical measurement model (Rasch model).

The Rasch model was developed to analyze the internal construct validity of dichotomous measurement tools by Danish mathematician Georg Rasch in 1960. In the Rasch model, the probability of responding correctly to any item of an individual is determined by the logistic function of the difference between the individual's ability and the difficulty of the item. Approximately 20 years later, Andrich D. (1978) expanded the Rasch model family for the items more than two categories by developing the Rating Scale Model (RSM). A few years after this development, Masters G. (1982) incorporated the Partial Credit Model (PCM) into this family.[3] The RSM and PCM are the models used for polytomous items.[4,5]

The threshold concept comes up when there are polytomous items in the measurement tool. The thresholds between the response categories of items in the measurement tool are the values that indicate the measurements in which adjacent categories are answered equally or the transition points between categories.[6] A threshold indicates the ability at the point where the probability of choosing the category K in an item is equal to the probability of choosing the category K+1.[7]

The first stage of Rasch analysis is to examine whether the thresholds for these items are ordered when there are polytomous items. In cases where individuals have difficulty in consistently distinguishing between response categories, disordered thresholds occur. In fact, disordered thresholds mean that individuals cannot choose categories that are appropriate for ability levels. The reasons for this may be the situation in which there are categories with the possibility of confusing the measurement tool (sometimes, often, etc.) and the use of items with a large number of categories.[7]

For example, the following question was asked to a patient with rheumatoid arthritis (RA).[8]

Can you climb five steps?

0) I'm doing it without difficulty

1) I'm doing it a little hard

2) I'm doing it with much difficulty

3) I can't do it

A patient who cannot do this activity at all (3) is expected to have a higher level of disability than a patient who does it very hard (2). Those who do it without difficulty (0) should also have lower disability level than those who choose other categories. For thresholds to be sorted, it is necessary to check whether this condition is met. If the level of disability of those who choose 3 is lower than those who choose 2, the order of the threshold will be distorted.

The fact that the thresholds are not ordered does not coincide with the theory of measurement. Therefore, the categories of items with disordered thresholds need to be combined with one of the other categories.[6] In the Rasch model, disordered thresholds are determined by examining the category probability curves for each item. These curves show the possibility of selecting the relevant categories depending on the relationship between the individual's ability and item difficulty. For example, if the individual's ability is relatively lower than the item difficulty, the probability of Category 0 being chosen will be higher than the probability of Category 1 being chosen. Category probability curves showing ordered and disordered thresholds for hypothetical items are shown in Figure 1 and Figure 2, respectively. The item difficulty in Figure 1 is 0.536. The probability of selecting the categories (0) and (1) of an individual having a level of ability -2 is 0.72 and 0.25, respectively.

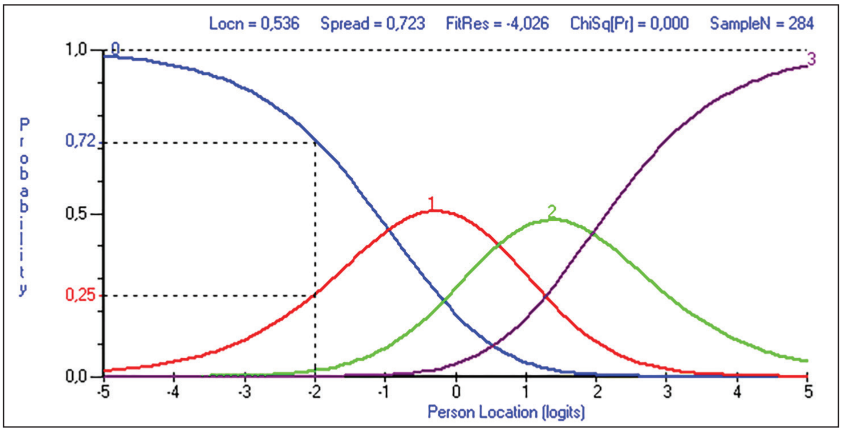

FIGURE 1: An example item with an ordered threshold.

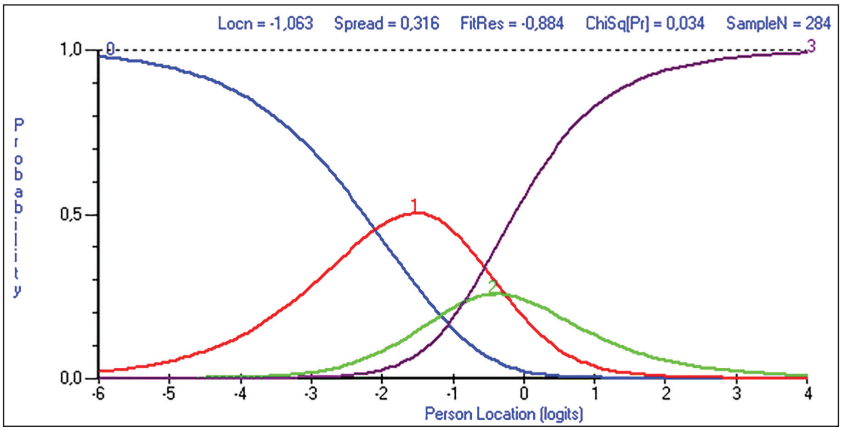

FIGURE 2: An example item with a disordered threshold.

The person who evaluates the measurement tool must combine the categories to provide the maximum amount of information. Most importantly, after the categories are combined, the new categories must be meaningful. For example, it is not correct to combine the category “agree” with the category “disagree”. On the other hand, the number of individuals who choose categories is important in combining the categories.[9] After the categories are combined, individuals need to be able to distinguish the categories in the best way and the categories that provide the most compatible data to the model must be formed.[10]

In general, it is stated that combining the categories with disordered thresholds improves the overall fit of the model, but the limited studies examining the effect of different combining strategies on the Rasch model fit have a motivating role for this study.[11] For this purpose, a measurement tool was used to evaluate the disability levels in the areas of “self-care, getting around, holding activities” of RA patients who were previously developed within the scope of The Scientific and Technological Research Council of Turkey (TUBITAK) project. Four different combining strategies have been developed in this study: “To the left, to the right, to the middle-left and to the middle-right”. These combining strategies were used for the items with disordered thresholds in this tool and differences/similarities in terms of fit to the Rasch model were evaluated.

MATERIAL AND METHODS

THE STUDY SAMPLE

Of the 300 patients with RA, 77 (26%) were female and 223 (74%) were male. The mean age of the patients (±standard deviation: SD) was 52.3±11.5 (minimum 18, maximum 82) and the mean duration of disease (±SD) was 11.3±8.0 years.

THE MEASUREMENT TOOL

As application data, a measurement tool, which was developed in a previous TUBITAK project (Evaluation of Disability with Computer Adaptive Testing Method in Rheumatoid Arthritis Patients. TUBITAK 1001 research projects, 109S342, 2010-2012) and evaluated disability levels in the areas of "self-care, getting around, holding activities" of RA patients, was used. The following scales were used in the development of the measurement tool used in the study:

- World Health Organization-Disability Assessment Schedule II – WHODAS-II
- Arthritis Impact Measurement Scales II – AIMS-II
- Nottingham Health Profile – NHP
- Health Assessment Questionnaire – HAQ

All three of these scales, except NHP, consist of polytomous items.

RASCH ANALYSIS

In the process of evaluating of the internal construct validity of the measurement tool with Rasch analysis, first the appropriate Rasch model is decided. In Rasch analysis, there are two models that can be used in case of polytomous items: Rating Scale Model (RSM) and Partial Credit Model (PCM). The main difference between RSM and PCM is: The distance between the thresholds in RSM is the same for all items, but not the same in PCM.[12] On the other hand, the likelihood ratio test can be used to decide which of these two models will be used.[13] According to the visual examination and likelihood ratio test performed for the application data, it was found appropriate to use PCM in the analysis of the data. The equation of PCM is:[7]

$$ln\left(\frac{P_{nij}}{1-P_{nij-1}}\right)=\theta_n-\beta_{ij}$$

where P is the probability of person n affirming item i; θ is the person ability, and β is the item difficulty.

After determining the model to be used to examine of the internal construct validity, the following protocol must be fulfilled in Rasch analysis.[13]

- Testing whether the thresholds for items are ordered
- Examination of whether the items in the measurement tool comply with the model
- Examination of local independence assumption
- Examination of unidimensionality assumption
- Examination of differential item functioning
- Testing internal consistency of the measurement tool (reliability)

THE ORDERING OF THE THRESHOLDS

Four different strategies were used to combine the categories with disordered threshold.

Strategy 1 (Combining to the left): It is the method of combining the disordered threshold category to the left category.

Strategy 2 (Combining to the right): It is the method of combining the disordered threshold category to the right category.

Strategy 3 (Combining to the middle-left): It is the method of combining the disordered threshold category to the middle category. When the middle category has the disordered threshold, it is the method of combining the middle category with the left category.

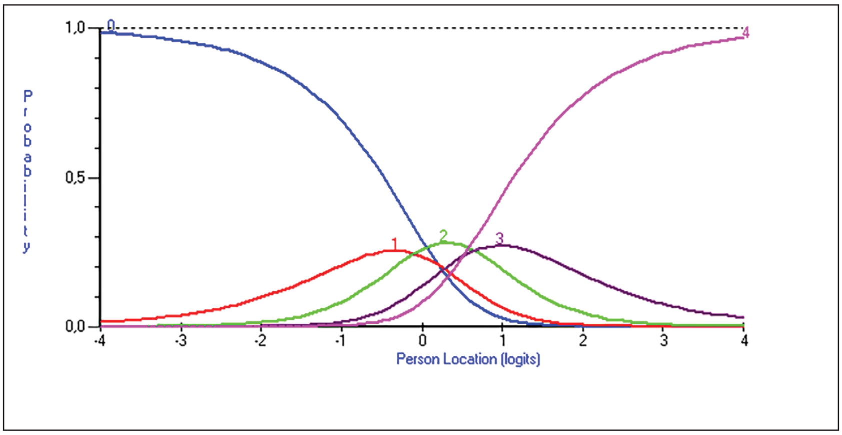

FIGURE 3: Graphical representation of categories 1 and 3 with disordered threshold.

Strategy 4 (Combining to the middle-right): It is the method of combining the disordered threshold category to the middle category. When the middle category has the disordered threshold, it is the method of combining the middle category with the right category.

An example of these different category-combining strategies is given below. The situation in which the measurement tool is 5 categories (0-1-2-3-4), and the categories 1 and 3 indicate the categories with disordered threshold are given in Figure 3.

0 1 2 3 4

Categories 1 and 3 can be combined as follows.

a) Combining to the left category (category 1 to category 0, category 3 to category 2)

0 0 1 1 2

b) Combining to the right category (category 1 to category 2, category 3 to category 4)

0 1 1 2 2

c) Combining to the middle category (The categories 1 and 3 are combined to category 2)

0 1 1 1 2

MODEL FIT

Model fit tests the extent to which individual responses to the measurement tool are compatible with the expected responses from the Rasch model. In evaluating the goodness of fit of measurement tools, the item interaction statistics, person interaction statistics and item-trait interaction statistics are used. After item and person interaction statistics are transformed into standardized z scores, if their mean is near 0 and standard deviation is close to 1, it is considered that the items and persons are compatible with the model. The item-trait interaction statistics show the property of invariance across the level of ability and are expressed by the chi-square value. When the p value associated with this chi-square value is greater than the Bonferroni-corrected p value, it is assumed that the hierarchical ordering of responses to the items of the measurement tool does not change throughout ability level. This means ensuring the property of invariance.[7] These statistical tests are discussed in the context of data analysis in this article.

In addition to the aforementioned model fit statistics, there are fit statistics calculated by chi-square statistics and residual values for items and persons. If the residual values of the items have values in the range of ±2.5 and the p-values of the chi-square statistics are greater than the Bonferroni-corrected p value, it is stated that the items in the measurement tool are compatible with the model.[13]

LOCAL INDEPENDENCE

The local independence assumption is tested by conducting principal component analysis (PCA) over the residuals of the items (Observed value - Expected value). If the residual correlation of any item pair is 0.30 or above, it can be decided to remove the item, having higher correlation with the other items, from the measurement tool.[14] In other words, item residuals should not be related to each other.

In this study, it was decided to remove the item from the model which is worse compatible with the model than the item pairs that show high correlation with each other.

UNIDIMENSIONALITY

The PCA should be applied on the residuals in order to determine the unidimensionality of the measurement tool. In order to ensure the local independence and unidimensionality assumption, there should be no meaningful pattern in the residuals. In order to achieve this assumption, the estimates obtained from the positive and negative loaded items on the first principal component must be compared with the paired t test.[7] In addition, the percentage of tests outside the range of -1,96 to 1,96 should not exceed 5%.[13] A confidence interval for a binomial test of proportions is calculated for the proportion of observed number of significant tests, and the lower bound should overlap the 5% expected value for the scale to be unidimensional.[2]

In this study, it was decided to remove the item from the model which has worse model fit in the items which may impair the unidimensionality assumption of the measurement tool (item with higher positive and negative loads than other items).

DIFFERENTIAL ITEM FUNCTIONING

The differential item functioning (DIF) is a condition that may affect model fit; it appears that different groups with the same ability level in the sample are more successful or unsuccessful in one item. There are two types of DIF which are uniform and non-uniform. For example, if women throughout the ability level have a significantly higher or lower score than men for any item in the measurement tool, it is accepted that the item shows a uniform DIF in terms of gender. On the other hand, non-uniform DIF is mentioned for an item in the measurement tool, when women have a significantly higher or lower score up to a certain value of ability level and beyond this value, men have a significantly higher or lower score. In the presence of uniform DIF, women and men can be grouped separately and the items can be calibrated. Conversely, in the presence of non-uniform DIF, it is generally recommended to remove the problematic item from the measurement tool.[7]

Whether the DIF is present in the measurement tool can be examined both statistically and graphically (with the curve of the item characteristic function). For each item in the measurement tool, DIF can be analyzed by performing two-way variance analysis of the different levels of trait (class intervals) and over each level of the variables of individuals. In cases where the main effect is significant in the variables of individuals, it is referred to as uniform DIF, whereas where the meaning of interaction is significant, it is considered to be non-uniform DIF.[7] In this study, age (≤43, 44-52, 53-60, ≥61), gender and disease duration (≤10 years, ≥11 years) were taken as variables of the individuals. For these three variables, it was examined whether the items had DIF.

RELIABILITY

While the internal consistency of the measurement tools is examined, the person separation index (PSI) and Cronbach's alpha coefficient are used. If there are missing responses in the data, the Cronbach's alpha coefficient cannot be calculated, but the PSI can be calculated. In order to divide individuals into two different groups in terms of ability, at least 0.70 of reliability is accepted, while at least 0.90 of reliability is acceptable to divide into four different groups.[15]

In this study, the student version of RUMM 2030 program was used to evaluate the internal construct validity of the measurement tool.[16]

RESULTS

As a result of the Rasch analysis performed using PCM, when the category probability curves were examined, 25 out of a total of 50 items had disordered thresholds. These categories with disordered threshold were combined according to four different strategies. After the four strategies were applied separately, the items that disrupt both the assumption of local independence and were incompatible with the model were removed from the model. At the next stage, both the items that are incompatible with the model and deteriorate the unidimensionality were removed from the model.

1. RASCH ANALYSIS RESULTS AFTER "TO THE LEFT COMBINING STRATEGY"

Whether the 50 items in the measurement tool provide model fit and local independence assumption were studied simultaneously. According to this, 21 items which are not compatible with the model and do not provide the assumption of local independence are excluded from the model. Then, within the remaining items, the items that are not compatible with the model and disrupt the unidimensionality were investigated and it was decided to remove 10 more items from the model. When the goodness of fit of the 19-item measurement tool was analyzed, the mean of the item interaction statistics (SD) was 0.000 (1.382) and the mean of the person interaction statistics (SD) was -2.318 (1.923). Since these values are close to the standardized z score (mean 0, standard deviation 1), it can be said that the remaining 19 items and the patients are compatible with the model. For the item-trait interaction statistics, the chi-square value was 119.494 (df=95; p=0.045). The p value here is greater than the p value with Bonferroni-corrected (0.05/19=0.0026) and is not statistically significant. As a result, since the hierarchical ordering of responses to the items of the measurement tool, which was obtained as a result of the "to the left combining strategy", did not change along the ability level, it was determined that the property of invariance of the measurement tool was provided.

In addition to the fit statistics given above, the residual values of each item in the "self-care, getting around, holding activities" measurement tool and the results of fit statistics calculated according to Chi-square statistics are given in Table 1. As the residual values according to Table 1 have values in the range of ±2.5 and the p-values of the chi-square statistics are larger than the p-values with Bonferroni-corrected, all items in the measurement tool are compatible with the model (Although the values for only two items were out of the limits, it was decided to keep them in the measurement tool because they did not create a problem with respect to the chi-square value).

The PCA was used to determine whether the 19-item measurement tool provided the unidimensionality assumption, and it was observed that there was no structure that disrupts this assumption. On the other hand, it has been determined that items do not have DIF in terms of age, sex and duration of illness.

Because of the missing observations in the measurement tool, PSI value was used to evaluate the internal consistency and this value was determined as 0.922. Therefore, it was concluded that the measurement tool was reliable.

TABLE 1: Item fit statistics after "to the left combining strategy".

Code	Items	*β*	SE	Residual	Chi-Square	df	p
wd23	In the past 30 days, how much difficulty did you have moving around inside your home?	0.482	0.119	1.310	5.509	5	0.357
wd31	In the past 30 days, how much difficulty did you have washing your whole body?	0.428	0.138	-1.611	3.000	5	0.700
wd33	In the past 30 days, how much difficulty did you have eating?	2.268	0.247	-0.828	3.286	5	0.656
wd54	In the past 30 days, how much difficulty did you have getting your household work done as quickly as needed?	-1.799	0.115	1.469	9.550	5	0.089
wd62	In the past 30 days, how much of a problem did you have because of barriers or hindrances in the world around you?	0.232	0.132	-0.197	2.179	5	0.824
wd68	In the past 30 days, how much of a problem did you have in doing things by yourself for relaxation or pleasure?	-0.053	0.181	-0.450	3.576	5	0.612
a3	Could you easily turn a key in a lock?	0.225	0.134	1.896	10.045	5	0.074
n10	I can walk about only indoors.	-0.038	0.180	-0.731	2.667	5	0.751
n11	I find it hard to bend.	-2.012	0.154	0.217	8.526	5	0.130
n14	I'm unable to walk at all.	2.089	0.313	0.561	13.408	5	0.020
n17	I have trouble getting up and down stairs and steps.	-3.173	0.161	2.127	5.201	5	0.392
h4	Get in and out of bed?	1.368	0.133	-1.862	5.690	5	0.338
h6	Lift a full cup or glass to your mouth?	0.805	0.121	-0.339	3.292	5	0.655
h10	Wash and dry your entire body?	-0.200	0.102	-2.622	6.842	5	0.233
h12	Get on and off the toilet?	0.480	0.116	-0.462	3.041	5	0.694
h14	Bend down to pick up clothing from the floor?	0.124	0.111	-3.625	15.269	5	0.009
h17	Turn taps on and off?	0.371	0.113	-0.010	3.132	5	0.680
h18	Run errands and shop?	-1.974	0.113	-0.426	8.163	5	0.147
h19	Get in and out of a car?	0.376	0.113	-2.082	7.117	5	0.212

SE: standard error, df: degrees of freedom.

2. RASCH ANALYSIS RESULTS AFTER "TO THE RIGHT COMBINING STRATEGY"

Whether the 50 items in the measurement tool provide model fit and local independence assumption were studied simultaneously. According to this, 21 items which are not compatible with the model and do not provide the assumption of local independence are excluded from the model. Then, within the remaining items, the items that are not compatible with the model and disrupts the unidimensionality were investigated and it was decided to remove 9 more items from the model. When the goodness of fit of the 20-item measurement tool was analyzed, the mean of the item interaction statistics (SD) was 0.000 (1.845) and the mean of the person interaction statistics (SD) was -1.223 (1.925). Since these values are close to the standardized z score (mean 0, standard deviation 1), it can be said that the remaining 20 items and the patients are compatible with the model. For the item-trait interaction statistics, the chi-square value was 138.700 (df=100; p=0.006). The p value here is greater than the p value with Bonferroni-corrected (0.05/20=0.0025) and is not statistically significant. As a result, since the hierarchical ordering of responses to the items of the measurement tool, which was obtained as a result of the "to the right combining strategy", did not change along the ability level, it was determined that the property of invariance of the measurement tool was provided.

In addition to the fit statistics given above, the residual values of each item in the "self-care, getting around, holding activities" measurement tool and the results of fit statistics calculated according to chi-square statistics are given in Table 2. As the residual values according to Table 2 have values in the range of ± 2.5 and the p-values of the chi-square statistics are larger than the p-values with Bonferroni-corrected, all items in the measurement tool are compatible with the model.

The PCA was used to determine whether the 20-item measurement tool provided the unidimensionality assumption, and it was observed that there was no structure that disrupts this assumption. On the other hand, it has been determined that items do not have DIF in terms of age, sex and duration of illness.

Because of the missing observations in the measurement tool, PSI value was used to evaluate the internal consistency and this value was determined as 0,933. Therefore, it was concluded that the measurement tool was reliable.

TABLE 2: Item fit statistics after "to the right combining strategy".

Code	Items	β	SE	Residual	Chi-Square	df	p
wd22	In the past 30 days, how much difficulty did you have standing up from sitting down?	0.826	0.100	1.736	9.389	5	0.095
wd25	In the past 30 days, how much difficulty did you have walking a long distance such as a kilometer [or equivalent]?	-2.675	0.163	-0.685	6.321	5	0.276
wd31	In the past 30 days, how much difficulty did you have washing your whole body?	0.683	0.119	-0.751	4.970	5	0.420
wd54	In the past 30 days, how much difficulty did you have getting your household work done as quickly as needed?	-1.339	0.099	0.942	9.549	5	0.089
wd61	In the past 30 days, how much of a problem did you have in joining in community activities (for example, festivities, religious or other activities) in the same way as anyone else can?	-0.344	0.104	1.453	4.197	5	0.521
wd62	In the past 30 days, how much of a problem did you have because of barriers or hindrances in the world around you?	0.422	0.122	1.787	2.793	5	0.732
wd68	In the past 30 days, how much of a problem did you have in doing things by yourself for relaxation or pleasure?	-0.952	0.147	-1.132	11.514	5	0.042
n10	I can walk about only indoors.	1.072	0.180	-1.105	2.522	5	0.773
n11	I find it hard to bend.	-0.826	0.147	-0.595	4.793	5	0.442
n14	I'm unable to walk at all.	3.145	0.317	0.970	16.357	5	0.006
n17	I have trouble getting up and down stairs and steps.	-1.964	0.152	0.351	9.337	5	0.096
n18	I find it hard to reach for things.	-2.088	0.153	-1.697	16.959	5	0.005
n27	I find it hard to stand for long.	-3.147	0.176	0.183	2.091	5	0.836
h4	Get in and out of bed?	2.373	0.132	-1.907	8.006	5	0.156
h6	Lift a full cup or glass to your mouth?	1.782	0.120	-0.326	2.556	5	0.768
h8	Walk outdoors on flat ground?	2.667	0.143	0.333	12.239	5	0.032
h12	Get on and off the toilet?	1.553	0.115	-0.674	5.711	5	0.335
h13	Reach and get down an object?	-1.321	0.103	-1.432	5.238	5	0.388
h17	Turn taps on and off?	1.380	0.111	0.193	2.705	5	0.745
h18	Run errands and shop?	-1.247	0.101	-0.454	1.453	5	0.918

SE: standard error, df: degrees of freedom.

3. RASCH ANALYSIS RESULTS AFTER "TO THE MIDDLE-LEFT COMBINING STRATEGY"

Whether the 50 items in the measurement tool provide model fit and local independence assumption were studied simultaneously. According to this, 26 items which are not compatible with the model and do not provide the assumption of local independence are excluded from the model. Then, within the remaining items, the items that are not compatible with the model and disrupt the unidimensionality were investigated and it was decided to remove 4 more items from the model. When the goodness of fit of the 20-item measurement tool was analyzed, the mean of the item interaction statistics (SD) was 0.000 (1.439) and the mean of the person interaction statistics (SD) was -1.656 (1.780). Since these values are close to the standardized z score (mean 0, standard deviation 1), it can be said that the remaining 20 items and the patients are compatible with the model. For the item-trait interaction statistics, the chi-square value was 136.901 (df=100; p=0.008). The p value here is greater than the p value with Bonferroni-corrected (0.05/20=0.0025) and is not statistically significant. As a result, since the hierarchical ordering of responses to the items of the measurement tool, which was obtained as a result of the "to the middle-left combining strategy", did not change along the ability level, it was determined that the property of invariance of the measurement tool was provided.

In addition to the fit statistics given above, the residual values of each item in the "self-care, getting around, holding activities" measurement tool and the results of fit statistics calculated according to chi-square statistics are given in Table 3. As the residual values according to Table 3 have values in the range of ± 2.5 and the p-values of the chi-square statistics are larger than the p-values with Bonferroni-corrected, all items in the measurement tool are compatible with the model.

The PCA was used to determine whether the 20-item measurement tool provided the unidimensionality assumption, and it was observed that there was no structure that disrupts this assumption. On the other hand, it has been determined that items do not have DIF in terms of sex and duration of illness. However, in the wd68 item, it was determined that there was uniform DIF in terms of age. Item characteristic function curves for age groups of this item are shown in Figure 4.

When Figure 4 is evaluated, it is concluded that item wd68 is a more difficult activity for patients over 61 years of age. However, since this item is important for the validity of the scope, it is decided to keep it in the measurement tool.

Because of the missing observations in the measurement tool, PSI value was used to evaluate the internal consistency and this value was determined as 0.925. Therefore, it was concluded that the measurement tool was reliable.

4. RASCH ANALYSIS RESULTS AFTER "TO THE MIDDLE-RIGHT COMBINING STRATEGY"

Whether the 50 items in the measurement tool provide model fit and local independence assumption were studied simultaneously. According to this, 23 items which are not compatible with the model and do not provide the assumption of local independence are excluded from the model. Then, within the remaining items, the items that are not compatible with the model and disrupt the unidimensionality were investigated and it was decided to remove 6 more items from the model. When the goodness of fit of the 21-item measurement tool was analyzed, the mean of the item interaction statistics (SD) was 0.000 (1,400) and the mean of the person interaction statistics (SD) was -1.846 (1.931). Since these values are close to the standardized z score (mean 0, standard deviation 1), it can be said that the remaining 21 items and the patients are compatible with the model. For the item-trait interaction statistics, the chi-square value was 135.001 (df=105; p=0.026). The p value here is greater than the p value with Bonferroni-corrected (0.05/21=0.0024) and is not statistically significant. As a result, since the hierarchical ordering of responses to the items of the measurement tool, which was obtained

TABLE 3: Item fit statistics after "to the middle-left combining strategy".

Code	Items	β	SE	Residual	Chi-Square	df	p
wd21	In the past 30 days, how much difficulty did you have standing for long periods such as 30 minutes?	-1.144	0.086	-0.235	6.504	5	0.260
wd22	In the past 30 days, how much difficulty did you have standing up from sitting down?	0.459	0.107	-0.282	7.909	5	0.161
wd23	In the past 30 days, how much difficulty did you have moving around inside your home?	1.022	0.115	-1.612	5.752	5	0.331
wd25	In the past 30 days, how much difficulty did you have walking a long distance such as a kilometer [or equivalent]?	-1.646	0.080	-0.088	5.328	5	0.377
wd32	In the past 30 days, how much difficulty did you have getting dressed?	0.576	0.104	-0.157	5.593	5	0.348
wd45	In the past 30 days, how much difficulty did you have sexual activities?	0.098	0.164	1.710	6.378	5	0.271
wd55	In the past 30 days, how much difficulty did you have your day-to-day work/school?	-1.607	0.109	0.728	1.829	5	0.872
wd61	In the past 30 days, how much of a problem did you have in joining in community activities (for example, festivities, religious or other activities) in the same way as anyone else can?	-0.542	0.107	1.159	9.448	5	0.092
wd62	In the past 30 days, how much of a problem did you have because of barriers or hindrances in the world around you?	-0.081	0.120	0.387	6.194	5	0.288
wd68	In the past 30 days, how much of a problem did you have in doing things by yourself for relaxation or pleasure?	0.502	0.176	-1.089	5.767	5	0.330
a3	Could you easily turn a key in a lock?	0.532	0.125	0.642	12.146	5	0.033
n10	I can walk about only indoors.	0.511	0.175	-1.299	4.983	5	0.418
n11	I find it hard to bend.	-1.338	0.143	-1.191	12.648	5	0.027
n14	I'm unable to walk at all.	2.583	0.318	0.639	4.989	5	0.417
n27	I find it hard to stand for long.	-3.553	0.173	-0.312	1.912	5	0.861
h6	Lift a full cup or glass to your mouth?	1.171	0.116	1.021	3.986	5	0.551
h8	Walk outdoors on flat ground?	2.060	0.138	-0.147	9.733	5	0.083
h12	Get on and off the toilet?	0.927	0.111	-1.230	9.206	5	0.101
h13	Reach and get down an object?	-1.350	0.106	-0.587	9.579	5	0.088
h19	Get in and out of a car?	0.820	0.109	-1.470	7.016	5	0.219

SE: standard error, df: degrees of freedom.

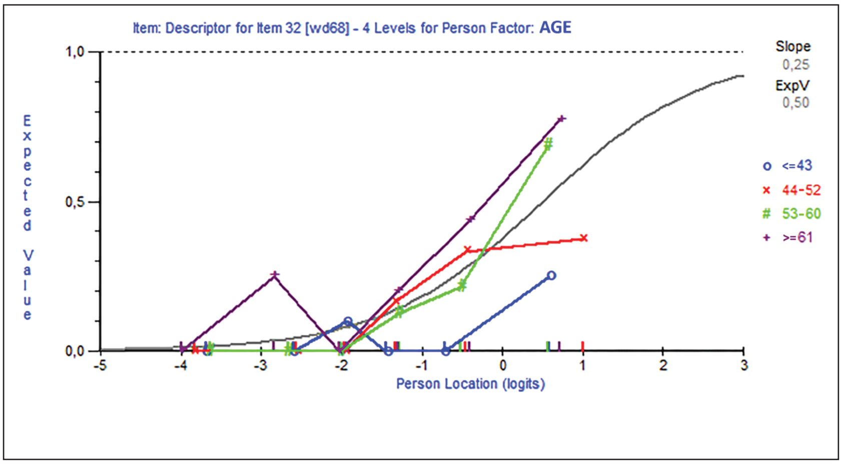

FIGURE 4: Item characteristic function curves for age groups of item wd68.

as a result of the "to the middle-right combining strategy", did not change along the ability level, it was determined that the property of invariance of the measurement tool was provided.

In addition to the fit statistics given above, the residual values of each item in the "self-care, getting around, holding activities" measurement tool and the results of fit statistics calculated according to chi-square statistics are given in Table 4. As the residual values according to Table 4 have values in the range of ±2.5 and the p-values of the chi-square statistics are larger than the p-values with Bonferroni-corrected, all items in the measurement tool are compatible with the model (Although the values for only one item were out of the limits, it was decided to keep it in the measurement tool because it did not create a problem with respect to the chi-square value).

The PCA was used to determine whether the 21-item measurement tool provided the unidimensionality assumption, and it was observed that there was no structure that disrupts this assumption. On the other hand, it has been determined that items do not have DIF in terms of age, sex and duration of illness. Because of the missing observations in the measurement tool, PSI value was used to evaluate the internal consistency and this value was determined as 0.935. Therefore, it was concluded that the measurement tool was reliable.

DISCUSSION

Rasch analysis is among the most common methods in evaluating the internal construct validity of the measurement tools. The first stage of this analysis is to examine whether items with disordered threshold are present in the measurement tool and to combine the categories appropriately if they exist. In this study, a measurement tool was used to evaluate the levels of disability in the areas of "self-care, getting around, holding activities" of patients with RA in the health field. As a result, the effects of combining the categories of items with disordered thresholds on model fit were investigated by using four different combining strategies **(to the left, to the right, to the middle-left and to the middle-right).**

There is a recommendation in the literature that the number of individuals in each category should not be less than 10 before the implementation of the category-combining strategies. For this purpose, the number of individuals who choose the item categories and categories with disordered threshold before applying the different category-combining strategies mentioned above are given in Table 5. In the table, colored cells show categories with disordered threshold, and the numbers marked with red represent categories with fewer than 10 individuals. According to this, the number of individuals in all categories with disordered thresholds is more than 10 (except for the category 3 of wd45

TABLE 4: Item fit statistics after "to the middle-right combining strategy".

Code	Items	β	SE	Residual	Chi-Square	df	p
wd21	In the past 30 days, how much difficulty did you have standing for long periods such as 30 minutes?	-1.300	0.123	0.895	6.338	5	0.275
wd23	In the past 30 days, how much difficulty did you have moving around inside your home?	0.834	0.109	1.641	6.866	5	0.231
wd32	In the past 30 days, how much difficulty did you have getting dressed?	0.524	0.106	-0.535	2.291	5	0.808
wd54	In the past 30 days, how much difficulty did you have getting your household work done as quickly as needed?	-1.292	0.112	1.182	4.472	5	0.484
wd61	In the past 30 days, how much of a problem did you have in joining in community activities (for example, festivities, religious or other activities) in the same way as anyone else can?	-0.647	0.111	1.646	8.952	5	0.111
wd62	In the past 30 days, how much of a problem did you have because of barriers or hindrances in the world around you?	-0.179	0.124	1.440	8.335	5	0.139
wd68	In the past 30 days, how much of a problem did you have in doing things by yourself for relaxation or pleasure?	-1.618	0.148	-1.616	8.313	5	0.140
a3	Could you easily turn a key in a lock?	0.463	0.129	1.286	4.961	5	0.421
n10	I can walk about only indoors.	0.420	0.179	-0.944	2.928	5	0.711
n11	I find it hard to bend.	-1.489	0.148	-0.282	2.796	5	0.731
n14	I’m unable to walk at all.	2.558	0.320	0.581	3.806	5	0.578
n17	I have trouble getting up and down stairs and steps.	-2.619	0.154	1.587	3.912	5	0.562
h4	Get in and out of bed?	1.810	0.132	-2.082	8.410	5	0.135
h6	Lift a full cup or glass to your mouth?	1.177	0.119	-0.068	2.750	5	0.739
h8	Walk outdoors on flat ground?	2.092	0.142	0.152	10.380	5	0.065
h11	Take a bath	-0.294	0.114	-3.037	11.210	5	0.047
h12	Get on and off the toilet?	0.912	0.114	-0.160	8.120	5	0.150
h13	Reach and get down an object?	-1.542	0.111	-1.438	11.179	5	0.048
h17	Turn taps on and off?	0.807	0.112	-0.127	3.924	5	0.560
h18	Run errands and shop?	-1.448	0.110	-1.139	6.890	5	0.229
h19	Get in and out of a car?	0.831	0.113	-2.107	8.168	5	0.147

SE: standard error, df: degrees of freedom.

TABLE 5: The number of individuals who choose the item categories and categories with disordered threshold in the measurement tool before applying the category-combining strategies.

Code	0	1	2	3	4	Total
wd21	72	59	38	71	44	284
wd22	109	87	43	42	3	284
wd23	170	72	20	17	4	283
wd25	74	55	38	38	79	284
wd31	189	36	28	14	17	284
wd32	188	37	29	18	12	284
wd33	222	30	16	14	2	284
wd34	194	29	17	24	20	284
wd45	181	17	10	7	65	280
wd52	124	52	32	23	53	284
wd53	123	54	27	25	55	284
wd54	99	62	30	33	60	284
wd55	67	58	36	47	74	282
wd61	144	37	36	20	47	284
wd62	132	51	37	42	22	284
wd68	152	26	27	23	55	283
a1	177	34	34	20	18	283
a2	191	33	31	15	14	284
a3	189	20	39	20	16	284
a4	188	25	36	17	16	282
a5	105	31	43	27	78	284
h11	170	45	24	43		282
h13	93	84	36	70		283
h18	97	75	41	70		283
h20	85	75	39	81		280

item). There are fewer than 10 individuals in Category 4 of only three items (wd22, wd23 and wd33). For this low number of categories, no combining has been performed outside of the combining strategies applied in the study.

There are items from each scale in the new measurement tools obtained after four different combining strategies. In terms of the sub-sections of the scales,

- All sub-sections of the WHODAS-II scale, except the “getting along with people” sub-section, are represented in the measurement tools.

- The “hand and finger function” part of the AIMS-II scale was represented in other strategies while it was not represented in the measurement tool obtained after the “to the right combining strategy”.

- The “physical abilities” part of the NHP scale is represented in the measurement tools that are formed by four strategies.

- In the HAQ scale, all sub-sections, except the “dressing and grooming” section, have been represented in new measurement tools. This may be due to the fact that the items (h1 and h2) in the section “dressing and grooming” relate to the “self-care” items on the scale of WHODAS-II. These items were excluded from the measurement tool because they disrupted the assumption of local independence. Therefore, they are not represented in the measurement tools obtained after four strategies.

As a result, in the measurement tools obtained after four different strategies, both the items from each scale and the representation of subsection of each scale are important for evidence of the validity of the scope of these measurement tools.

When categories with disordered thresholds are combined according to four different strategies, it is possible to obtain items of similar category order. Accordingly, when similarity rates for four different combining strategies are evaluated, the similarity rate of "to the left" and "to the middle-left" combining strategies was 52% and the similarity rate of "to the right" and "to the middle-right" combining strategies was 36%. In addition, the similarity rate of "to the middle-left" and "to the middle-right" combining strategies was 76%. In 19 of the 25 items with disordered threshold in the measurement tool, these two strategies were combined in the same way, while different combinations were made in 6 items. This is because both strategies actually combine into the middle category. However, when the middle category has the disordered threshold, the combination of the middle category to the left and right separates these two strategies. The similarity rates of other combining strategies are below 32%.

For these four different strategies, the results of the Rasch analysis before and after category-combining are given in Table 6.

Table 6 summarizes the results of the Rasch analysis obtained before and after the combining using the same items. When the findings in these tables are compared, after the combining strategies, there is an absolute increase in the mean and standard deviation values of the item interaction statistics and the person interaction statistics. This situation is thought to be due to the loss of information in the measurement tool based on the combining of the categories and the increase in the standard error in estimates of the persons' ability levels. On the other hand, after the combining strategies, according to the results of the item-trait interaction statistics, it was observed that the property of invariance of measurement tools were provided and model fit was increased. In addition, the internal consistency of measurement tools is higher according to PSI statistics obtained after the combining strategies. Specifically, in terms of PSI statistics, "to the middle-right" combining strategy seems more appropriate, while "to the left" combining strategy seems more appropriate in terms of model fit.

The agreement between the person ability estimates obtained after four combining strategies was calculated as 0.977 by intraclass correlation coefficient (ICC). As this agreement is high, it is understood that there is not much change in the ability of persons according to four strategies. Therefore, it is considered to be difficult to decide which of the four strategies to choose.

TABLE 6: Rasch analysis results before and after combining for four different strategies.

	The Combining Strategies			
Statistics	**Left**	**Right**	**Middle-Left**	**Middle-Right**
Mean of the item interaction statistics (SD)	0.000 (1.162) *0.000 (1.382)*	0.000 (1.534) *0.000 (1.845)*	0.000 (1.272) *0.000 (1.439)*	0.000 (1.219) *0.000 (1.400)*
Mean of the person interaction statistics (SD)	-1.784 (1.621) *-2.318 (1.923)*	-1.254 (1.631) *-1.223 (1.925)*	-1.252 (1.408) *-1.656 (1.780)*	-1.549 (1.558) *-1.846 (1.931)*
Chi-Square (df); p	207 (95); <0.001* *119 (95); 0.045*	177 (100); <0.001* *138 (100); 0.006*	295 (100); <0.001* *136 (100); 0.008*	227 (105); <0.001* *135 (105); 0.026*
PSI	0.851 *0.922*	0.892 *0.933*	0.882 *0.925*	0.878 *0.935*
n(Bonferroni Corrected α) +	19 (0.0026) *19 (0.0026)*	20 (0.0025) *20 (0.0025)*	20 (0.0025) *20 (0.0025)*	21 (0.0024) *21 (0.0024)*

The values in the first row in the cells correspond to the values "before", the values in the second row in italics correspond to "after" combining of the categories.
df: degrees of freedom; PSI: Person separation index; +uncorrected $\alpha=0,05$; *p<corrected α and the property of invariance is not provided.

In the literature, there are a limited number of studies examining the impact of combining categories of items which have disordered thresholds with different strategies on Rasch model fit. In a study conducted by Grondin and Blais in 2010, incompatible individuals were excluded from the measurement tool, which consists of 24 items of 6 categories, in order to combine disordered thresholds categories. Subsequently, the intermediate categories (including categories with ordered threshold), and then the remaining categories were combined for all items. Finally, each item category has been combined with at least 10 individuals. As a result of these combining strategies, it was observed that in some of the model fit worsened, some improved. Consequently, instead of all items, it is stated that performing category-combining for each item is positive in terms of model fit and that the categories resulting from category-combining should be meaningful.[3]

Adams et al. reported in their study in 2012 that ordered or disordered thresholds depend only on the number of individuals responding to the categories, and that the fact that items have disordered thresholds does not require the disordered categories. Therefore, it is stated that even if the items have disordered thresholds, the individuals with a high ability level can give them appropriate responses, and accordingly, it is concluded that the model fit of the items can be good.[17] Wetzel and Carstensen supported the aforementioned study and reported that more careful consideration should be given to avoid losing valuable information in combining disordered threshold categories.[18]

CONCLUSION

In this study, when combining categories, there were not at least 10 individuals in each item category; that is, the number of individuals responding to the categories has not been evaluated. Before these strategies are applied, it is thought that it is more appropriate to combine the categories with less than 10 frequencies and then combine them with the strategies used in the present study in terms of model fit. In addition, category-combining strategies were performed on only one data, and the size of disordered thresholds was ignored because a single strategy was applied for all items in the combining process. It is not appropriate to generalize the findings obtained due to these conditions. Therefore, the mentioned conditions can be considered as the limitations of the study.

It is thought that this study will benefit the researchers working on similar subjects since in the literature, there are a limited number of studies examining the effects of combining the categories of disordered thresholds items with different strategies on the Rasch model fit.

Acknowledgement

We would like to thank Prof. Atilla Halil Elhan for allowing the use of the data set and for his useful suggestions and comments.

Source of Finance

During this study, no financial or spiritual support was received neither from any pharmaceutical company that has a direct connection with the research subject, nor from a company that provides or produces medical instruments and materials which may negatively affect the evaluation process of this study.

Conflict of Interest

No conflicts of interest between the authors and / or family members of the scientific and medical committee members or members of the potential conflicts of interest, counseling, expertise, working conditions, share holding and similar situations in any firm.

Authorship Contributions

***Idea/Concept:** Timur Köse; **Design:** Timur Köse, Derya Gökmen; **Control/Supervision:** Timur Köse, Derya Gökmen; **Data Collection and/or Processing:** Derya Gökmen, Ömer Faruk Dadaş; **Analysis and/or Interpretation:** Derya Gökmen, Ömer Faruk Dadaş; **Literature Review:** Ömer Faruk Dadaş; **Writing The Article:** Ömer Faruk Dadaş, Derya Gökmen, Timur Köse; **Critical Review:** Derya Gökmen, Timur Köse.*

REFERENCES

1. Bond TG, Fox CM. Applying the Rasch Model: Fundamental Measurement in the Human Sciences. 2nd ed. New York, London: Psychology Press; 2007. p.340.
2. Elhan AH, Küçükdeveci AA, Tennant A. The rasch measurement model. In: Franchignoni F, ed. Advances in Rehabilitation. Vol. 19. Pavia: Maugeri Foundation Books; 2010. p.89-102.
3. Grondin J, Blais JG. A Rasch analysis on collapsing categories in item's response scales of survey questionnaire: maybe it's not one size fits all. Online Submission. 2010;29.
4. Andrich D. A rating formulation for ordered response categories. Psychometrika. 1978;43(4):561-73. [Crossref]
5. Masters GN. A rasch model for partial credit scoring. Psychometrika. 1982;47(2):149-74. [Crossref]
6. Andrich D. Category ordering and their utility. Rasch Meas Trans. 1996;9(4):464.
7. Pallant JF, Tennant A. An introduction to the Rasch measurement model: an example using the Hospital Anxiety and Depression Scale (HADS). Br J Clin Psychol. 2007;46(Pt 1):1-18. PMID: 17472198. [Crossref] [PubMed]
8. Elhan AH, Atakurt Y. [Why is it necessary to use Rasch analysis when evaluating measures?]. Ankara Üniversitesi Tıp Fakültesi Mecmuası. 2005;58:47-50.
9. Wright B, Linacre J. Combining (collapsing) and splitting categories. Rasch Meas Trans. 1992;6(3):233-5.
10. Lopez WA. Communication validity and rating scales. Rasch Meas Trans. 1996;10(1):482.
11. Pallant JF, Miller RL, Tennant A. Evaluation of the Edinburgh post natal depression scale using Rasch analysis. BMC Psychiatry. 2006;6(1):28. PMID: 16768803. [Crossref] [PubMed] [PMC]
12. Chang CH, Reeve BB. Item response theory and its applications to patient-reported outcomes measurement. Eval Health Prof. 2005;28(3):264-82. PMID: 16123257. [Crossref] [PubMed]
13. 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(8):1358-62. PMID: 18050173. [Crossref] [PubMed]
14. 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. PMID: 19545445. [Crossref] [PubMed] [PMC]
15. Fisher W. Reliability, separation, strata statistics. Rasch Meas Trans. 1992;6(3):238. [Crossref]
16. Andrich D, Sheridan B, Luo G. Rumm 2030. Perth, Australia: Rumm Laboratories; 2012.
17. Adams RJ, Wu ML, Wilson M. The rasch rating model and the disordered threshold controversy. Educ Psychol Meas. 2012;72(4):547-73. [Crossref]
18. Wetzel E, Carstensen CH. Reversed thresholds in partial credit models: a season for collapsing categories? Assessment. 2014;21(6):765-74. PMID: 24789857. [Crossref] [PubMed]