Original Article

Impact of Rating Scale Categories on Reliability and Fit Statistics of the Malay Spiritual Well-Being Scale using Rasch Analysis

Aqil Mohammad Daher[1], Syed Hassan Ahmad[2], Than Winn[3], Mohd Ikhsan Selamat[4]

Submitted: 30 Oct 2014
Accepted: 14 Mar 2015

[1] *Department of Community Medicine, Faculty of Medicine and Defence Health. National Defence University of Malaysia, Sungai Besi Camp, 57000 Kuala Lumpur, Malaysia*

[2] *Department of Psychiatry, faculty of Medicine. University Teknologi MARA, Taman Prima Selayang, 68100 Batu Caves Selangor, Malaysia*

[3] *Department of Community Medicine, Faculty of Medicine, MAHSA University, 50490 Kuala Lumpur, Malaysia*

[4] *Population Health and Preventive Medicine, Faculty of Medicine, University Teknologi MARA, Jalan Hospital, 47000 Sungai Buloh, Selangor, Malaysia*

Abstract

***Background*: Few studies have employed the item response theory in examining reliability. We conducted this study to examine the effect of Rating Scale Categories (RSCs) on the reliability and fit statistics of the Malay Spiritual Well-Being Scale, employing the Rasch model.**

***Methods:* The Malay Spiritual Well-Being Scale (SWBS) with the original six; three and four newly structured RSCs was distributed randomly among three different samples of 50 participants each.**

***Results:* The mean age of respondents in the three samples ranged between 36 and 39 years old. The majority was female in all samples, and Islam was the most prevalent religion among the respondents. The predominating race was Malay, followed by Chinese and Indian. The original six RSCs indicated better targeting of 0.99 and smallest model error of 0.24. The Infit Mnsq (mean square) and Zstd (Z standard) of the six RSCs were "1.1"and "-0.1"respectively. The six RSCs achieved the highest person and item reliabilities of 0.86 and 0.85 respectively. These reliabilities yielded the highest person (2.46) and item (2.38) separation indices compared to other the RSCs.**

***Conclusion:* The person and item reliability and, to a lesser extent, the fit statistics, were better with the six RSCs compared to the four and three RSCs.**

Keywords: *reliability, rating scale, spirituality, analysis*

Introduction

In order to ensure the validity and the reliability of an instrument, researchers have formulated various strategies that optimise the psychometric properties of the instrument. The reliability of a questionnaire represents the reproducibility of scores across different test situations. Reliability quantifies the measurement error and can be defined as the percentage of explained variance in relation to the total possible variance of scores (1). According to Classical Test Theory (CTT), reliability is expressed in terms of stability, equivalence, and consistency. The methods of assessing each aspect include test-retest reliability as a measure of stability, alternate or parallel form as a measure of equivalence, and the internal consistency is usually quantified using Cronbach's alpha and/or Kuder–Richardson-20 (KR-20) (2).

With the advent of the modern Item Response Theory (IRT), a new paradigm has emerged in assessing reliability. The Rasch model is a statistical model which is closely related to IRT: it examines the observed data against a standard model and provides insight as to the reliability, validity and the fitness of that instrument relative to the target population. Rasch quantifies person

ability and allocates it along a continuum of item difficulty to provide insight about performance of the test. The more the information about person ability and item difficulty obtained, the less the measurement error, and the better the reliability.

The main difference in the concept of reliability with regards to CTT and IRT is the Standard Error (SE) used in the calculation of the reliability coefficient. In the CTT, the item difficulty estimate is sample rated, computed as a sample response to that item, and assumes that the ordinal raw score is linear when it is not. The SE of measurements is then derived from the average of sample ability, so the estimation is based on average person ability and not individual person ability. Thus the error variance will be overestimated especially with inclusion of extreme persons who got less error variance (3).

In contrast to the conventional CTT, Rasch provides two indicators of reliability: person and item reliability related to quantified person ability, and item difficulty. The Rasch model offers a 'direct estimate of the modeled error variance for each estimate of a person's ability and an item's difficulty' measured on interval scale scores (4). Individual SEs provide more information than a sample or test average, and 'extreme scores are usually excluded because their SEs are infinitely large', and those measures therefore carry little information about those person's or item's location at the extreme of the ability continuum (5,6).

Furthermore, person and item reliability are then transformed into a separation indices to overcome the restriction of range of reliability value of being, 0–1. This separation index is used to classify a person's abilities and item difficulties into distinct groups. The number of categories of the rating scale and its effect on reliability has been debated over the last decades. An earlier attempt to identify the effect of category number on reliability revealed that there is no effect of the number of categories on the reliability (7). Studies showed unchanged test-retest reliability using rating categories with three, five, seven or nine response categories (8,9). Further research confirmed those findings (10–13).

Contradictory to the above, authors have concluded that increasing the number of response categories would improve the reliability of the study; inter-rater reliability (14) and test-retest reliability (15–17). Findings of other published work confirmed that reliability increases with increasing response categories, but not beyond six categories (18–20). Regarding the optimum number of categories, authors argued that reliability is better with seven responses (14,21), while others recommended only five instead (22,23). Using Rasch analysis, few studies have confirmed that the smaller number of rating categories yields better reliability and fit statistics including separation indices (24–26). The argument in these studies is the underutilisation of some rating categories based on post-hoc analysis.

The original spiritual Well-Being Scale (the English version) under investigation is a widely used self-administered scale that assesses spiritual well-being. To the best of our knowledge, no published study has so far examined the suitability of the rating scale of SWBS whether in the original setting and language or when translated, particularly into the Malay Language. The objective of this paper is to empirically investigate the potential effects of the number of response categories on the reliability and fit statistics of the Malay SWBS using Rasch analysis.

Materials and Method

The current work is a part of the validation and reliability testing study of the Malay SWBS, in which pre-testing was a continuous process to ensure a valid and reliable instrument. The respondents in this study were participants in ongoing community screening programs that cover mostly the Selangor state in Malaysia. As the aim of this study was to determine the impact of the number of rating categories on reliability and fit statistics, the choice of the newly modified rating categories was based on an analysis of the first sample collected at the first community screening program. The analysis of responses from the original version with six rating categories suggested that the new rating scale with four and three categories might yield better reliability.

Sample

The final linguistically-checked version of the Malay SWBS with the original six response categories and the newly modified (introduced) versions with four and three rating categories were administered to three samples of 50 respondents each, who were participating in the ongoing community screening program. All samples were selected with simple random sampling from a list of participants in the community program. It is worth noting that all participants fulfilled the inclusion criteria that the participants be Malaysian and be able to read and write in the Malay language as the requirements to participate in this study. The main exclusions were as follows:

non-Malaysian, and an inability to read and write in the Malay language.Ethical approval was obtained from the relevant institutional ethics committee, reference number 600-RMI (5/1/6).

Instrument

Palotzian and C Ellison are credited with developing the original SWB (27). The adaptation of the Malay SWBS from the original American English version was done according to commonly adopted guidelines of translating and adapting health questionnaires (28,29). The adaptation was performed by two independent native Malay speakers who carried out the forward translation, and whose quality was checked by other independent translators. The backward translation into English was carried out by another two independent translators. Discrepancies resulting from this process were resolved and consensus was reached regarding the harmonised Malay version of the SWBS. Debriefing sessions were carried out with 15 participants. Comments and ambiguity were recorded and discussed with translators in order to develop the finalised version of the Malay SWBS.

Statistical Analyses

Data were entered and analysed using the Winstep Rasch analysis software (30). Rasch analysis provides a customary summary statistics of fit indices like Infit Mnsq (mean square), Infit Zstd (standardised mean square), and reliability indices including person reliability, item reliability, and person and item separation indices achieved by the instrument and the target sample. 'Fit statistics indicate how accurately or predictably data fit the model' (31). The Rasch model aims at transferring the ordinal data into a continuous scale which provides a ruler against which to gauge person ability against item difficulty. The Rasch model assumes that the higher the person ability, the higher the possibility to endorse (answer) difficult item(s). Thus, fit statistics reflect how the data under study fit the stipulated model. On the other hand, reliability refers to the replication of the results in a different sample, i.e. the same item difficulty and person ability measures are expected to be replicated in a different sample. A separation index is calculated as the number of SE of spread among the items/persons, and the ability to define distinct groups of item difficulty and persons' abilities (32). Different iterations were obtained to identify which categorisation would perform better, and the better rating scale combination was decided by item and person Infit Mnsq and Zstd.

Results

Table 1 shows the characteristics of the study samples. The response rate was 98%, 90%, and 96% for first, second and third samples respectively. The mean age was almost comparable in all three samples. In terms of gender, females predominated the three samples with a higher proportion reported in the third sample (70.8%). Islam was the majority religion in the first and third samples, while Christianity predominated the second sample. Regarding educational level, the third sample reported a higher proportion of primary and secondary education (27.7% and 44.7% respectively) compared to the other samples, while the second and third samples reported higher proportions of university degrees, 75.6% and 46.9% respectively. Malay participants were the majority in the first and third samples, followed by the Chinese, while the Chinese were majority in the second sample, followed by the Malay. Employed participants were highly represented in all samples compared to other working status groups. More than half of the participants in all the samples reported as being married, and less than half (43.8%) and around a third (28.9%) reported single status in the first and second samples respectively.

The restructuring of the original six rating scale categories was based on the results of the first sample through post-hoc iteration. Data iteration is the default algorithm of Rasch analysis, in which the data is processed to provide estimates of Rasch measurement. The method is repeated (iteration) until the best fit measures of the data are obtained. Different post-hoc iterations (combination of categories) were obtained (Table 2). We found that the four and the three categories shown in Table 2 were the best among other iterations. The six rating scale categories of 'strongly agree', 'moderately agree', 'agree', 'disagree agree', 'moderately disagree', and 'strongly disagree' were collapsed into four categories ('strongly agree', 'agree', 'disagree', and 'strongly disagree') in one version, and into three categories in another version ('agree', 'neither agree nor disagree', and 'disagree'). It was observed that the four-category targeting was very good compared to the original six categories, and yielded the highest person reliability and separation. It was noticeable that the mean person Infit Mnsq of the four categories achieved the ideal value of 1, and that the person mean Infit Zstd was unchanged between different categorisations. The smallest model error was achieved by the 4 categories. The mean item Infit Mnsq was 1 for the four and three

Table 1: Descriptive statistics of the study samples

		6 categories	4 categories	3 categories
Age Mean (SD)		36 (11)	37 (9)	39(12)
Gender n (%)	Male	19 (38.8)	18 (40.0)	14 (29.2)
	Female	30 (61.2)	27 (60.0)	34 (70.8)
Religion n (%)	Muslim	21 (42.9)	11 (24.4)	31 (64.6)
	Buddhist	13 (26.5)	8 (17.8)	10 (20.8)
	Hindu	12 (24.5)	5 (11.1)	4 (8.3)
	Christian	3 (6.1)	15 (33.3)	3 (6.3)
	Others	0 (0)	6 (13.3)	0 (0)
Educational Level n (%)	No formal education	0 (0)	0 (0)	3 (6.4)
	Primary	1 (2)	0 (0)	13 (27.7)
	Secondary	15 (30.6)	5 (11.1)	21 (44.7)
	Diploma	10 (20.4)	6 (13.3)	8 (17.0)
	University Degree	23 (46.9)	34 (75.6)	2 (4.3)
Race n (%)	Malay	21 (42.0)	11 (24.4)	31 (64.6)
	Chinese	16 (32.0)	16 (35.6)	10 (20.8)
	Indian	13 (26.0)	10 (22.2)	7 (14.6)
	Others	0 (0)	8 (17.8)	0 (0)
Occupation n (%)	Unemployed	1 (2.1)	1 (2.2)	8 (17.0)
	Employed	45 (93.8)	34 (75.6)	18 (38.3)
	Pensioner	1 (2.1)	0 (0)	3 (6.4)
	Student	0 (0)	7 (15.6)	3 (6.4)
	Housewife	1 (2.1)	3 (6.7)	15 (31.9)
Marital Status n (%)	Single	21 (43.8)	13 (28.9)	7 (14.6)
	Married	25 (52.1)	31 (68.9)	39 (81.3)
	Separated/Divorced	2 (4.2)	1 (2.2)	0 (0.0)
	Widow	0 (0)	0 (0)	2 (4.2)

RSCs achieving the ideal value, and the mean item Infit Zstd was 0 for the six and three RSCs. Item and person reliability and separation indices were not dramatically changed, and all were in a good range.

Table 3 depicts the fits statistics and reliability indices for the three forms of categorisations of the rating scale. We observed that the original six categories achieved good targeting of less than one error, and achieved the smallest model error compared to the newly introduced four and three categories. Although the difference was not significant, the three categories achieved the smallest mean person Infit Mnsq of 1.03 with the smallest SD of 0.38.

The three categories registered the ideal mean person Infit Zstd of 0, and the others of −0.1. The SD of person Infit Zstd decreased slightly toward the smallest number of categories. The three categories registered the lowest mean and SD of the item Infit Mnsq compared to other categories. On the other hand, six categories registered the lowest mean item Infit Zstd of 0. The SD of the item measure increased slightly with four categories, and sizably with three categories to reflect more variation in the estimate of item difficulty.

In terms of reliability and separation index, it is observable that the six categories yielded the highest person and item reliability compared to the others. Similarly, the separation indices were higher for the six categories.

Table 2: Summary statistics of different post-hoc iterations

Categories	Original		123456		Post-hoc		111234		Post-hoc		111223	
	Persons		Items		Persons		Items		Persons		Items	
Statistics	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Measure	0.99	0.78	0.00	0.4	0.37	1.00	0.00	0.55	0.70	1.50	0.00	0.86
Model Error	0.24	0.07	0.15	0.01	0.29	0.07	0.18	0.01	0.45	0.11	0.28	0.01
INFIT MNSQ	1.1	0.73	1.01	0.33	1.00	0.44	1.00	0.32	1.03	0.50	1.00	0.31
INFIT ZSTD	−0.1	2.2	0.00	1.6	−0.1	1.5	-0.1	1.6	−0.1	1.7	0.0	1.6
OUTFIT MNSQ	1.06	0.89	1.06	0.45	1.02	0.72	1.02	0.42	1.00	0.59	1.01	0.35
OUTFIT ZSTD	−0.2	2.2	0.2	1.7	−0.1	1.5	0.0	1.5	−0.2	1.7	0.1	1.5
RELIABILITY	0.86		0.85		0.89		0.88		0.88		0.88	
SEPARATION	2.46		2.38		2.86		2.70		2.76		2.77	
SE	0.11		0.9		0.14		0.13		0.22		0.20	

Abbreviations: INFIT MNSQ = Infit Mean square; OUTFIT MNSQ = Outfit Mean square; INFIT ZSTD = Infit standardised mean square; OUTFIT ZSTD = Outfit standardised mean square; SE = Standard error.

Table 3: Summary statistics of three different forms of RSCs

Categories	Original		123456		112344		112233					
	Persons		Items		Persons		Items		Persons		Items	
Statistics	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Measure	0.99	0.78	0	0.4	1.55	1.2	0	0.6	1.75	1.18	0	1.19
Model Error	0.24	0.07	0.15	0.01	0.44	0.21	0.26	0.01	0.51	0.2	0.37	0.18
INFIT MNSQ	1.1	0.73	1.01	0.33	1.07	0.88	1.04	0.3	1.03	0.38	0.93	0.24
INFIT ZSTD	−0.1	2.2	0	1.6	−0.1	2.1	0.1	1.4	0	0.8	−0.2	0.9
OUTFIT MNSQ	1.06	0.89	1.06	0.45	1.04	0.88	1.06	0.4	1.21	1.44	1.21	0.99
OUTFIT ZSTD	−0.2	2.2	0.2	1.7	−0.2	2.1	0.2	1.4	0.1	0.9	0.2	1.2
RELIABILITY	0.86		0.85		0.8		0.77		0.75		0.83	
SEPARATION	2.46		2.38		1.98		1.84		1.75		2.23	
SE	0.11		0.9		0.18		0.13		0.19		0.27	

Abbreviations: INFIT MNSQ = Infit Mean square; OUTFIT MNSQ = Outfit Mean square; INFIT ZSTD = Infit standardised mean square; OUTFIT ZSTD = Outfit standardised mean square; SE = Standard error.

Discussion

Admittedly, reliability is a complex concept as there is no single parameter that quantifies this psychometric property. To the best of our knowledge, this is the first study which examines the effect of the number of response categories on the reliability of the Malay SWBS using the Rasch model. One of the advantages of Rasch model measurement is that it allows the estimation of the source of error arising from persons and items, whereas other methods of testing reliability report statistics about the performance of the respondents only.

The observed difference of the Infit Mnsq of the persons was not sizable between the three groups and it shows a good fit as it approaches "1" (31,33,34). The minimal changes of person Infit Zstd did not show a significantly better fit of the four and three categories compared to six, and did not cause a considerable 'noise' to the fit statistics. The variation in the estimate of person ability is comparable between the six and four categories, and lower still than the three categories. Knowing the Mnsq represent the average of the residual, we may conclude that fit statistics are minimally affected by the number of RSCs; rather it is affected by the structure of item and the sample size.

The highest possibility that people will behave in a similar way when they are subjected to items of similar difficulties was conferred by the six categories (0.86). We expected that six categories would perform better to separate the respondents into more than three strata which is considered good. Similarly, item reliability decreased with four categories when compared to six categories, and we can say that the items give less reliable information than the persons in this sample. In addition, the item separation was better for the six categories (2.38) and for the three categories (2.23) compared to the four categories (1.84), i.e. item endorsement (difficulty) by respondents were separated into 'three' levels compared to 'two'.

The findings may be explained by the fact that reducing the number of categories means reducing the endorsement spectrum, which in turn yields scores with little variance, affecting the separation index that is required to be as high as possible (35–37). The rating categories which yielded high reliability and a separation index were found to be those with many response categories. Our findings are consistent with those of previous studies which showed that increasing the number of categories would increase reliability (14,17–21). However, Garner (38) opined that 20 categories are necessary for better reliability.

Our results are not in tandem with the findings that reliability is not affected by the number of response categories (11–13). Furthermore, our findings are inconsistent with the findings of Zhu et al. (25) that a smaller number of categories improves the reliability in which the authors' judgment was based solely on post-hoc analysis. The post-hoc analysis in our case is in favor of three categories, but when the newly proposed three categories were tested, it showed that they were not better in terms of reliability and separation. Despite its not being investigated in our paper, it is pertinent to mention that increasing the number of RSCs adds advantages to better the discriminant ability of the test and the amount of information transmitted (38).

Although Rasch has been perceived to handle small sample size, the study might be limited by the small sample size which would increase the SE and hence reduce reliability. However, the reliability indices were high enough even with such a small sample. Moreover, due to the exploratory nature of this study, a small sample size is warranted as previously reported (39), especially with the Likert scale rating (40). It is our intention to present real life data rather than rely on the simulation of a large database. Using different groups of respondents to answer the three different forms of RSCs might affect the comparison. It would seem more useful if in the future, the same sample is subjected to the three different forms of the questionnaire and the order of entry randomised to avoid a carryover effect.

Conclusions

This study has showed that the person and item reliability and, to a lesser extent, the fit statistics, seemed better with six categories compared to four and three categories. No recommendations are made regarding reducing number of rating scale categories of the Malay SWBS. This study adds to the scope of investigating the optimal number of RSCs, and it affirms previous findings that a wider rating scale yields better reliability.

Acknowledgment

The authors wish to express gratitude to the Research Management Institute of the Universiti Teknologi MARA for providing a research grant (600-RMI/ST/DANA 5/3/ Dst (82/2011) that enables researcher to accomplish this humble work as a part of PhD project at UiTM. The authors thank Dr Azrilah AbdulAziz for her great help in analyzing the data. The authors also thank Professor Trevor Bond from James Cook University, Australia, and Mohd Zali Mohd Nor from Universiti Putra Malaysia whom comments and review helped finalizing this manuscript.

Conflict of Interest

None.

Funds

Research grant (600-RMI/ST/DANA 5/3/ Dst (82/2011).

Authors' Contributions

Conception and design: AMD, SHA, TW
Analysis and interpretation of the data, obtaining of funding: AMD
Drafting of the article, final approval of the article: AMD, SHA, TW, MIS
Critical revision of the article for the important intellectual content: SHA, TW
Collection and assembly of data: AMD, MIS

Correspondence

Dr Aqil Mohammad Daher
MBChB (Baghdad University), MPH (University Malaya), PhD (UiTM)
Community Medicine
Faculty of Medicine and Defence Health
National Defence University of Malaysia
Sungai Besi Camp
57000 Kuala Lumpur
Malaysia
Tel: +603 9051 3041
Fax: +603 9051 3042
Email: Aqil702001@yahoo.com

References

1. Carlson NR, Buskist W, Enzle ME, Heth CD. *Psychology: The Science of Behaviour (Canadian Edition)*. Allyn and Bacon Canada (CN): Scarborough, Ontario; 2000.

2. Progar Š, Sočan G. An empirical comparison of item response theory and classical test theory. *Horizons Psychol*. 2008;**17(3)**:5–24.

3. Schumacker RE, Smith EV. A Rasch Perspective. *Educ Psychol Measurement*. 2007;**67(3)**:394–409. doi: 10.1177/0013164406294776.

4. Wright BD, Mok MMC. *An overview of the family of Rasch measurement models*. In: Smith Jr EV, Smith RM, editors. Introduction to Rasch measurement. Maple Grove (MN): JAM Press; 2004. p. 1–24.

5. Linacre J. True-score reliability or Rasch statistical validity. *Rasch Measurement Transactions*. 1996;**9(4)**:455.

6. Clauser B, Linacre J. Relating Cronbach and Rasch reliabilities. *Rasch Measurement Transactions*. 1999;**13(2)**:696.

7. Schutz HG, Rucker MH. A Comparison of Variable Configurations Across Scale Lengths: An Empirical Study'. *Educ Psychol Measurement*. 1975;**35(2)**:319–324. doi: 10.1177/001316447503500210.

8. Bendig AW. The reliability of self-ratings as a function of the amount of verbal anchoring and of the number of categories on the scale. *J Appl Psychol*. 1953;**37(1)**:38–41. doi: 10.1037/h0057911.

9. Bendig AW. Reliability and the number of rating-scale categories. *J Appl Psychol*. 1954;**38(1)**:38–40. doi: 10.1037/h0055647.

10. Komorita SS. Attitude content, intensity, and the neutral point on a Likert scale. *J Soc Psychol*. 1963;**61(2)**;327–334. doi: 10.1080/00224545.1963.9919489.

11. Matell MS, Jacoby J. Is there an optimal number of alternatives for Likert scale items? Study I: Reliability and validity. *Educ Psychol Measurement*. 1971;**31(3)**:657–674. doi: 10.1177/001316447103100307.

12. Remington M, Tyrer P, Newson-Smith J, Cicchetti D. Comparative reliability of categorical and analogue rating scales in the assessment of psychiatric symptomatology. *Psychol Med*. 1979;**9(4)**:765–770. doi: 10.1017/S0033291700034097.

13. Brown G, Widing RE, Coulter RL. Customer evaluation of retail salespeople utilizing the SOCO scale: a replication, extension, and application. *J Academy Marketing Sci*. 1991;**19(4)**:347–351. doi: 10.1007/BF02726510.

14. Cicchetti DV, Shoinralter D, Tyrer PJ. The effect of number of rating scale categories on levels of interrater reliability: A Monte Carlo investigation. *Appl Psychol Measurement*. 1985;**9(1)**:31. doi: 10.1177/014662168500900103.

15. Oaster T. Number of alternatives per choice point and stability of Likert-type scales. *Perceptual Motor Skills*. 1989;**68**:549–550. doi: 10.2466/pms.1989.68.2.549.

16. Finn HR. Effects of some variations in rating scale characteristics on the means and reliabilities of ratings. *Educ Psychol Measurement*. 1972;**32(2)**:255–265. doi: 10.1177/001316447203200203.

17. Ramsay J. The effect of number of categories in rating scales on precision of estimation of scale values. *Psychometrika*. 1973;**38(4)**:513–532. doi: 10.1007/BF02291492.

18. Aiken LR. Number of response categories and statistics on a teacher rating scale. *Educ Psychol Measurement*. 1983;**43(2)**:397–401. doi: 10.1177/001316448304300209.

19. Boote AS. Reliability testing of psychographic scales:Five-point or seven-point? Anchored or labeled? *J Advert Res*. 1981;**21**:53–60.

20. McCallum DM, Keith BR, Wiebe DJ. Comparison of response formats for Multidimensional Health Locus of Control Scales: Six levels versus two levels. *J Personal Assessment*. 1988;**52(4)**:732–736. doi: 10.1207/s15327752jpa5204_12

21. McKelvie SJ. Graphic rating scales—How many categories? *British J Psychol*. 1978;**69(2)**:185–202. doi: 10.1111/j.2044-8295.1978.tb01647.x.

22. Jenkins GD, Taber TD. A Monte Carlo study of factors affecting three indices of composite scale reliability. *J Appl Psychol*. 1977;**62(4)**:392–398. doi: 10.1037/0021-9010.62.4.392.

23. Lissitz RW, Green SB. Effect of the number of scale points on reliability: A Monte Carlo approach. *J Appl Psychol*. 1975;**60(1)**:10–13. doi: 10.1037/h0076268.

24. Zhu W, Timm G, Ainsworth B. Rasch calibration and optimal categorization of an instrument measuring women's exercise perseverance and barriers. *Res Q Exerc Sport*. 2001;**72(2)**:104–116. doi: 10.1080/02701367.2001.10608940.

25. Zhu W, Updyke WF, Lewandowski C. Post-hoc Rasch analysis of optimal categorization of an ordered-response scale. *J Outcome Measurement*. 1997;**1(4)**:286–304.

26. Pesudovs K, Noble BA. Improving subjective scaling of pain using Rasch analysis. *J Pain*. 2005;**6(9)**:630–636. doi: 10.1016/j.jpain.2005.04.001.

27. Ellison CW. Spiritual Well-Being: Conceptualization and Measurement. *J Psycholand Theol*. 1983;**11**: 330–340.

28. World Health Organization. *Process of translation and adaptation of instruments*. Geneva (CH); World Health Organization: 2007.

29. Bullinger M, Alonso J, Apolone G, Leplège A, Sullivan M, Wood-Dauphinee S, et al. Translating health status questionnaires and evaluating their quality: the IQOLA Project approach. International Quality of Life Assessment. J Clin Epidemiol. 1998;**51(11)**: 913–923. doi: 10.1016/S0895-4356(98)00082-1.

30. Linacre J. *WINSTEPS Rasch measurement (Version 3.60. 1)*. Chicago (USA): WINSTEPS. com. Computer program; 2006.

31. Linacre JM. What do infit and outfit, mean-square and standardized mean. *Rasch Measurement Transactions*. 2002;**16(2)**:878.

32. Green KE, Frantom C, editors. *Survey development and validation with the Rasch model*. International Conference on Questionnaire Development, Evaluation, and Testing. Charleston (SC): (publication unknown); 2002.

33. Fisher WPJ. Rating Scale Instrument Quality Criteria. *Rasch Measurement Transactions*. 2007;**21(1)**:1095.

34. BD W, JM L. Reasonable mean-square fit values. *Rasch Measurement Transactions*. 1994;**8(3)**:370.

35. Chang L. A psychometric evaluation of four-point and six-point Likert-type scales in relation to reliability and validity. *Appl Psychol Measurement*. 1994;**18(3)**:205–215. doi: 10.1177/014662169401800302.

36. Nunnally Jr JC. *Introduction to psychological measurement*. New York (NY); McGraw-Hill: 1970.

37. Martin WS. Effects of scaling on the correlation coefficient: Additional considerations. *J Marketing Res*. 1978:**15(2)**;304–308. doi: 10.2307/3151268.

38. Garner WR. Rating scales, discriminability, and information transmission. *Psychol rev*. 1960;**67(6)**:343–352. doi: 10.1037/h0043047.

39. Chen W-H, Lenderking W, Jin Y, Wyrwich KW, Gelhorn H, Revicki DA. Is Rasch model analysis applicable in small sample size pilot studies for assessing item characteristics? An example using PROMIS pain behavior item bank data. *Qual Life Res*. 2014;**23(2)**:485–493. doi: 10.1007/s11136-013-0487-5.

40. Smith AB, Rush R, Fallowfield LJ, Velikova G, Sharpe M. Rasch fit statistics and sample size considerations for polytomous data. *BMC Med Res Methodology*. 2008;**8(1)**:33. doi: 10.1186/1471-2288-8-33.