Original Article

Rasch validation of the SF-36 for assessing the health status of Korean older adults

Sae-Hyung Kim, PhD[1]), Wi-Young So, PhD[2])*

[1)] *Department of Measurement and Evaluation of Physical Education, Chungbuk National University, Republic of Korea*
[2)] *Sports and Health Care Major, College of Humanities and Arts, Korea National University of Transportation: 50 Daehak-ro, Chungju-si, Chungbuk 380-702, Republic of Korea*

Abstract. [Purpose] To verify, using Rasch analysis, the applicability of the 36-Item Short Form Health Survey (SF-36) to elderly Koreans, as this instrument would be useful for determining elderly individuals' overall performance and providing them with health information. [Subjects and Methods] The SF-36 was administered to a sample of 510 individuals aged over 60 living in the Seoul and Gyeonggi areas of South Korea. When testing for goodness-of-fit, we considered items with infit and outfit indexes of over 1.30 or less than 0.70 to be incongruent. SF-36 factors that contained over three items, including physical functioning, role limitations (physical and emotional), mental health, vitality, and general health, were analyzed. Each factor was examined through step calibration of the response categories in the probability curve. [Results] The response categories were found to be appropriate because the adjustment values of each factor increased. We found five items in physical functioning, two items in role limitation–emotion, one item in mental health, and one item in general health to be incongruent; all items in the role limitation–physical and vitality factors were congruent. [Conclusion] We conclude that the SF-36 could be revised to more accurately measure the health status of elderly Koreans.
Key words: Elderly, Health survey

INTRODUCTION

By 2017, Korea will be an aging society, with over 14.0% of its population being over 65 years old; by 2026, it will be a super-aged society, with the elderly accounting for over 20.0% of the population[1]. For this reason, gerontology research is becoming increasingly important, especially that centered on the measurement and evaluation of elderly health status[2, 3].

In order to measure the health status of the elderly accurately, the concept of health should be clearly defined in terms of present societal demands. Health, which a century ago simply meant "survival" or being "disease free"[4], nowadays contains aspects related to quality of life in addition to the absence of disease[5]. Due to this change in the definition of health, the use of a scale developed according to the former definition would be obviously problematic in the present aging society. Unfortunately, numerous existing measures of health status of adults or the elderly adhere to the former definition. One of these, the 36-Item Short Form Health Survey (SF-36) has come to serve as a representative general health status and treatment efficacy assessment scale since its development[6–8]. The SF-36 has been used in health science, medicine, and physical education among other fields. Furthermore, it has been translated into several languages, including Korean, and it is widely used as a measure of the health status of Korean elderly adults. Its validity in the adult and elderly population of Korea has also been confirmed. For example, Nam and Lee[4], who studied adults, and Han and Lee[9], who studied the elderly, demonstrated the construct validity of the SF-36 through exploratory and confirmatory factor analyses. However, factor analysis by itself does not accurately confirm a scale's validity, especially when it is applied to a group other than the original[10]. Furthermore, it cannot verify the structure of the sub-components and the difficulty level of the items[11].

The above problems can be circumvented via the Rasch model[12]. This model is based on item response theory, and it evaluates the appropriateness of the number of items in a scale and calculates the fit of the items to the data and the difficulty of these items. The Rasch model is also able to evaluate the suitability of the response method (e.g., five- or four-point scales) for subjects[13]. Furthermore, it can detect overlapping and incongruent items, which can be confusing to subjects[14–16]. For these reasons, it is feasible and desirable to employ the Rasch methodology to confirm the validity of the SF-36 for elderly Korean adults.

SUBJECTS AND METHODS

Setting and samples: This study surveyed elderly people over the age of 60 by distributing the SF-36 questionnaire

*Corresponding author. Wi-Young So (E-mail: wowso@ut.ac.kr)

at senior welfare centers in the Seoul and Gyeonggi areas of South Korea. An investigator explained and completed each item for those elderly persons who were unable to complete the self-administered questionnaire. Five hundred and ten responses were collected after excluding one with inconsistent answers. The distribution of subjects according to gender and age is shown in Table 1.

This study was approved by the Public Institutional Review Board of the Korea Ministry of Health and Welfare (P01-201303-SB-07-00) and conforms to the principles of the Declaration of Helsinki. All study participants provided their informed consent before completing the questionnaire. Furthermore, we received permission to use the SF-36 from the company that owns it (OptumInsight Life Sciences, Inc., USA); the approval number is C009024.

Measurements/Instruments: The SF-36 scale was developed by Ware and Sherboune[17], and was adapted by Koh et al.[18] to measure the quality of life of elderly Koreans. It consists of 36 items categorized as follows: 10 items of physical functioning, four items of role limitation–physical, two items of bodily pain, five items of general health, four items of vitality, two items of social functioning, three items of role limitation–emotion, five items of mental health, and one item of health changes. The items concerning physical functioning are answered on a three-point scale, while those concerning bodily pain are answered on a five or six-point scale; the rest of the items are answered on a five-point scale. In general, higher scores indicate better health status; however, three items of general health, one item of social functioning, two items of vitality, two items of mental health, the two items of bodily pain, and the item of health changes are reverse scored. The sequence of the items in the present study was randomized to increase measurement validity. The SF-36 subscales and the number of items are shown in Table 2.

We used Rasch analysis to examine the validity of the SF-36. First, we conducted a principal component analysis to determine the unidimensionality of the items, as this is a fundamental condition of the Rasch model. This was done by calculating whether the items of each factor share equal variance[19] using SPSS 20.0 (SPSS, Chicago, IL, USA). In principal component analysis, the scale is considered unidimensional if the eigenvalue of the first component is greater than 1.00 or if the largest principal component accounts for over 20% of the total variance[20]. Goodness-of-fit was then examined using Winsteps 3.65.0 (Rasch Measurement Software, USA). An item was regarded as inappropriate if its goodness-of-fit (infit and outfit) indexes were more than 1.30 or less than 0.70[21]. Furthermore, the suitability of the rating scales of the items was evaluated through step calibration in the probability curve. Step calibration examines the appropriateness of the rating scales of the items by dividing subjects' numerical response scores by the number of responses in the scale (three-point, five-point, or six-point scale). The response scale appropriateness is expressed as thresholds, which are calculated from each factor's item characteristic curve[22]. Health status is considered to have an appropriate response scale if the step calibration of the scale applied to each factor (the adjustment value) increases gradually towards as high scale score[23, 24]. Social function-

ing, bodily pain, and health change, which consist of less than two items each, were excluded from this analysis.

RESULTS

Unidimensionality verification: The results of the principal component analysis used to test the unidimensionality are displayed in Table 3. All six factors together accounted for over 20% of the variance (effect size) of the first component eigenvalue. Furthermore, the variances of the first component eigenvalue in physical functioning, role limitation–physical, role limitation–emotion, mental health, vitality, and general health were 64.13%, 89.51%, 91.31%, 62.43%, 65.34%, and 61.46% respectively; thus, the criteria of unidimensionality were satisfied.
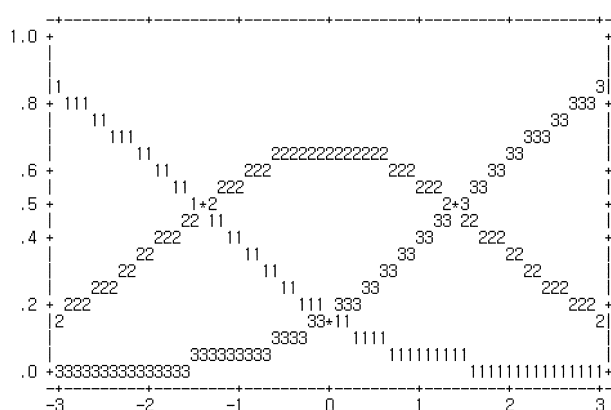
Table 1. Distribution of subjects according to age and gender

Age	Male	Female	Total
60–64	12 (2.4%)	16 (3.1%)	28 (5.5%)
65–69	53 (10.4%)	109 (21.4%)	162 (31.8%)
70–74	86 (16.9%)	103 (20.2%)	189 (37.1%)
75–79	31 (6.1%)	43 (8.4%)	74 (14.5%)
Over 80	22 (4.3%)	35 (6.9%)	57 (11.2%)
Total	204 (40.0%)	306 (60.0%)	510 (100%)

Table 2. SF-36 dimensions and the number and sequence of items

Dimension	Number of items	Sequence of items
Physical functioning	10	3, 4, 5, 6, 7, 8, 9, 10, 11, 12
Social functioning	2	20, 32
Role limitation–Physical	4	13, 14, 15, 16
Role limitation–Emotional	3	17, 18, 19
Mental health	5	24, 25, 26, 28, 30
Vitality	4	23, 27, 29, 31
Bodily pain	2	21, 22
General health	5	1, 33, 34, 35, 36
Health changes	1	2
Total	36	

SF-36, 36-item Short Form Survey

Table 3. Results of principal component analysis of SF-36

Factor	First component eigenvalue	Second component eigenvalue	Variance of first component eigenvalue (%)
Physical functioning	6.41	1.18	64.13
Role limitation–Physical	3.58	0.19	89.51
Role limitation–Emotion	2.73	0.18	91.31
Mental health	3.12	0.72	62.43
Vitality	2.61	0.73	65.34
General health	3.07	0.85	61.46

SF-36, 36-item Short Form Survey

Fig. 1. Response category probability curve of physical functioning

Fig. 2. Response category probability curve of role limitation–physical

Table 4. Step calibration of the response category of physical functioning

Category	Response rate (%)	Observed average value	Step calibration
1	21	−2.74	
2	30	0.32	−1.42
3	49	3.63	1.42

Table 5. Step calibration of the response category of role limitation–physical

Category	Response rate (%)	Observed average value	Step calibration
1	9	−5.26	
2	22	−2.74	−5.11
3	22	0.40	−1.02
4	31	3.05	1.38
5	16	4.94	4.75

It is possible to examine whether the number of item categories is appropriate by using the category probability curve (Fig. 1) and step calibration (Table 4), both of which are integral in Rasch analysis. The x-axis of the category probability curve presents the personal latent characteristic of health status and the logit difference of item difficulty, while the y-axis presents the category probability of a specific response category. Step calibration involves presenting the x-axis value of the point that each category curve crosses. A response category is considered appropriate if the step calibration increases gradually. The response rate of category 3 was 49%, which was the highest. From categories 1 to 3, the observed average value increased gradually from −2.74 to 3.63. Furthermore, for the step calibration, the numerical value of the two points that each of the three category probability curves crossed was shown to increase gradually—specifically, the value between categories 1 and 2 was −1.42 and that between 2 and 3 was 1.42.

Figure 2 and Table 5 present the category probability curve and step calibration of the four items of the role limitation–physical factor, which were answered using a five-point scale. As observed from the figure, the crossing points increased gradually, as follows: between categories 1 and 2, the crossing point was −5.11; between 2 and 3, it was −1.02; between 3 and 4, it was 1.38; and between 4 and 5, it was 4.75. Figure 3 and Table 6 show the category probability curve and step calibration of the three items of the role limitation–emotion factor, which were also answered using a five-point scale. The crossing points increased gradually, as follows: between categories 1 and 2, the crossing point was −6.15; between categories 2 and 3, it was −2.19; between categories 3 and 4, it was 2.16; and between categories 4
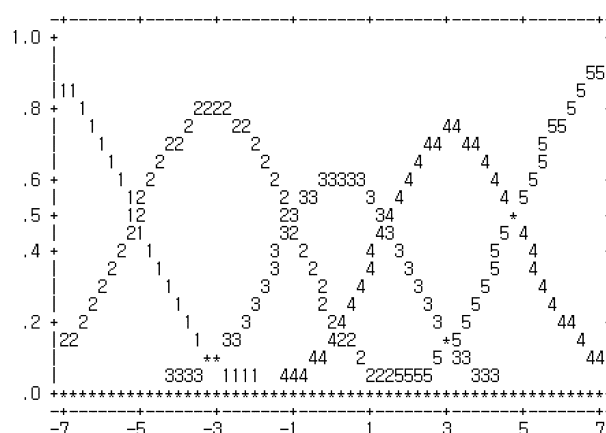
and 5, it was 6.17. Figure 4 and Table 7 show the category probability curve and step calibration of the five items of the mental health factor, which were answered on a five-point scale. As before, the crossing points increased gradually, as follows: between categories 1 and 2, the crossing point was −1.54; between categories 2 and 3, it was −0.65; between categories 3 and 4, it was 0.75; and between categories 4 and 5, it was 1.44. Figure 5 and Table 8 present the category probability curve and step calibration of the four items of the vitality factor, which were answered on a five-point scale. The crossing points increased gradually, as follows: between categories 1 and 2, the crossing point was −1.52; between categories 2 and 3, it was −0.68; between categories 3 and 4, it was 0.74; and between categories 4 and 5, it was 1.46. Figure 6 and Table 9 display the category probability curve and step calibration of the five items of the general health factor, which were answered on a five-point scale. The crossing point increased gradually, as follows: between categories 1 and 2, the crossing point was −2.17; between categories 2 and 3, it was −0.60; between categories 3 and 4, it was 0.59; and between categories 4 and 5, it was 2.18.

Test of goodness-of-fit: Next, we examined the item difficulty and the goodness-of-fit of the factors (Table 10). In Rasch analysis, the closer the fit index is to 1.0, the more appropriate the item. In physical functioning, the infit and outfit indexes of the first item (engaging in vigorous activities such as running, lifting heavy objects, or participating in strenuous sports) were respectively 0.93 and 1.96 (the latter being over the threshold); those of the fourth item (climbing several
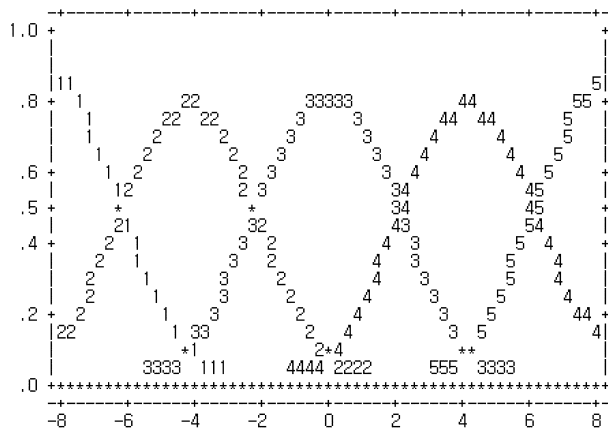
Fig. 3. Response category probability curve of role limitation–emotion

Fig. 4. Response category probability curve of mental health

Table 6. Step calibration of the response category of role limitation–emotion

Category	Response rate (%)	Observed average value	Step calibration
1	5	−5.99	
2	17	−3.33	−6.15
3	32	0.28	−2.19
4	34	3.97	2.16
5	12	5.92	6.17

Table 7. Step calibration of the response category of mental health

Category	Response rate (%)	Observed average value	Step calibration
1	11	−1.88	
2	14	−0.77	−1.54
3	21	0.21	−0.65
4	21	1.25	0.75
5	32	2.56	1.44

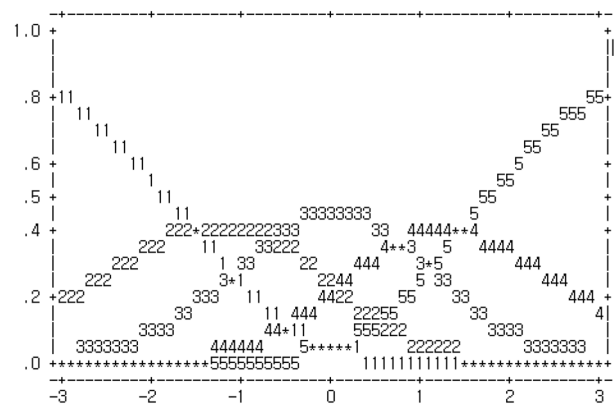**Fig. 5.** Response category probability curve of vitality
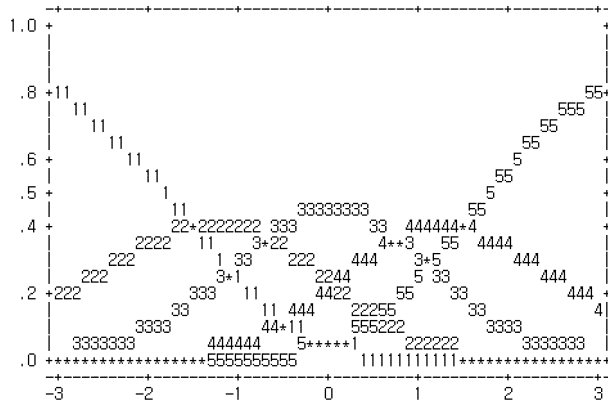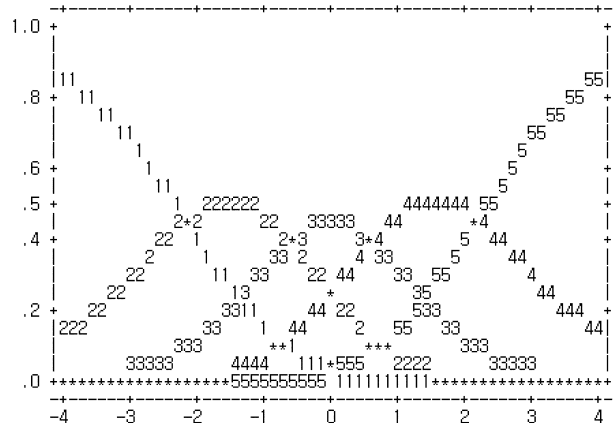

Fig. 6. Response category probability curve of general health

Table 8. Step calibration of the response category of vitality

Category	Response rate (%)	Observed average value	Step calibration
1	17	−1.99	
2	19	−0.95	−1.52
3	24	0.11	−0.68
4	19	1.01	0.74
5	21	2.05	1.46

Table 9. Step calibration of the response category of general health

Category	Response rate (%)	Observed average value	Step calibration
1	16	−2.27	
2	25	−1.18	−2.17
3	26	−0.11	−0.60
4	23	1.05	0.59
5	10	1.81	2.18

Table 10. Item difficulty and goodness-of-fit index

	No.	Items	Calibration logit	SE logit	Infit MnSq[a]	Outfit MnSq
Physical functioning	1	Engaging in vigorous activities such as running, lifting heavy objects, or participating in strenuous sports	3.68	0.11	0.93	1.96
	2	Engaging in moderate activities such as moving a table, pushing a vacuum cleaner, bowling, or playing golf	1.59	0.10	1.06	1.22
	3	Able to lift or carry groceries	−0.09	0.11	0.97	1.03
	4	Able to climb several flights of stairs	1.22	0.10	0.76	0.66
	5	Able to climb one flight of stairs	−0.96	0.11	1.02	0.98
	6	Able to bend, kneel, or stoop	0.88	0.10	1.43	1.67
	7	Able to walk more than a mile	−0.07	0.11	0.88	0.90
	8	Able to walk several blocks	−1.12	0.12	0.81	0.71
	9	Able to walk one block	−1.93	0.13	0.91	0.69
	10	Able to bathe or dress yourself	−3.19	0.16	1.20	0.64
Role limitation–physical	1	You have to cut down on the amount of time you spend on work or other activities	−0.60	0.10	1.13	1.14
	2	You have accomplished less than you would have liked	−0.09	0.10	1.01	0.98
	3	You are limited in the kind of work or other activities you can perform	0.53	0.10	0.92	0.90
	4	You have difficulty performing work or other activities (for example, it takes extra effort)	0.16	0.10	0.88	0.85
Role limitation–emotion	1	You have to cut down on the amount of time spent on work or other activities due to feeling depressed or anxious	−0.27	0.13	0.88	0.86
	2	You have accomplished less than you would have liked due to feeling depressed or anxious	0.04	0.13	0.69	0.65
	3	You did not do work or other activities as carefully as usual due to feeling depressed or anxious	0.23	0.13	1.36	1.32
Mental health	1	Are you a very nervous person?	−1.24	0.07	1.44	1.58
	2	Do you feel so down in the dumps that nothing could cheer you up?	−0.73	0.06	0.78	0.70
	3	Do you feel calm and peaceful?	1.28	0.06	1.05	1.00
	4	Do you feel downhearted and blue?	−0.55	0.06	0.88	0.85
	5	Are you a happy person?	1.24	0.06	0.97	0.92
Vitality	1	Do you feel full of pep?	0.63	0.06	0.84	0.84
	2	Do you have a lot of energy?	0.81	0.06	1.04	1.01
	3	Do you feel worn out?	−1.36	0.06	1.02	1.07
	4	Do you feel tired?	−0.08	0.06	1.10	1.14
General health	1	In general, would you say your health is	0.99	0.06	0.78	0.74
	2	I seem to get sick a little easier than other people do	−0.94	0.06	1.01	0.99
	3	I am as healthy as anybody I know	−0.15	0.06	0.95	0.90
	4	I expect my health to get worse	0.22	0.06	1.39	1.46
	5	My health is excellent	−0.12	0.06	0.87	0.82

Item names are descriptions of the content and do not reflect the exact Korean wording of the question. [a] MnSq: mean square residuals

flights of stairs) were respectively 0.76 and 0.66 (the latter being below the threshold); those of the sixth item (being able to bend, kneel, or stoop) were respectively 1.43 and 1.67 (both of which were over the threshold); those of the ninth item (being able to walk one block) were respectively 0.91 and 0.69 (the latter being below the threshold); and those of the 10th item (being able to bathe or dress) were respectively 1.20 and 0.64 (the latter being below the threshold). For role limitation–emotion, the infit and outfit indexes of the second item (accomplished less than would have liked due to feeling depressed or anxious) were respectively 0.69 and 0.65 (both of which were below the threshold), and those of the third

item (did not do work or other activities as carefully as usual due to feeling depressed or anxious) were respectively 1.36 and 1.32 (both of which were over the threshold). For mental health, the infit and outfit indexes of the first item ("Are you a very nervous person?") were respectively 1.44 and 1.58 (both of which were over the threshold). For general health, the infit and outfit indexes of the fourth item ("I expect my health to get worse") were respectively 1.39 and 1.46 (both of which were over the threshold). The other items, such as the four items of role limitation–physical and the four items of vitality, were appropriate, with goodness-of-fit (infit and outfit) indexes over 0.70 and under 1.30.

DISCUSSION

This study examined whether the SF-36 is appropriate for elderly Koreans using Rasch analysis. The six factors of the SF-36 were found to satisfy the criterion of unidimensionality. Furthermore, step calibration showed that the three-point scale used to respond to items of the physical functioning factor and the five-point scales for the role limitation–physical, role limitation–emotion, mental health, vitality, and general health factors were appropriate. This implies that the SF-36 response categories are appropriate for elderly Koreans, and that the elderly who participated in this study understood the reverse scored items of the SF-36.

In our study, fewer than three items from each factor—which tested content-based evidence such as bodily pain, social functioning, and health changes—were excluded from the analysis. This corresponds with the results of Bollen[25] and Gilbert et al.[26], the latter of whom conducted a psychometric scaling study, which indicated that at least three items were appropriate for exclusion. However, the results of the test of goodness-of-fit suggested that five of the ten items in the physical functioning factor are inappropriate. The infit indexes of items 1 (engaging in vigorous activities such as running, lifting heavy objects, participating in strenuous sports) and 6 (being able to bend, kneel, or stoop) were over 1.30. This implies that these items were incongruent with other items and confused subjects[27], making these items inappropriate for measuring the physical functioning level of elderly Koreans. In addition, the outfit indexes of items 4 (being able to climb several flights of stairs), 9 (being able to walk one block), and 10 (being able to bathe or dress) were less than 0.70. This suggests that these items overlap with each other, and excluding them would simplify and clarify the measurement[28].

Furthermore, items 2 (accomplished less than you would have liked due to feeling depressed or anxious) and 3 (did not do work or other activities as carefully as usual due to feeling depressed or anxious) of role limitation–emotion, which consists of three items, were found to be inappropriate. Item 2 overlapped with role limitation–physical and Item 3 had the highest logit difficulty, which likely means that it was confusing to subjects. Both the infit and outfit indexes of one item each of the role limitation–emotion and general health factors were over 1.30, making these items completely inappropriate for the participants. Item 1 in mental health ("Are you a very nervous person?") and item 4 in general health ("I expect my health to get worse") were incongruent with other items within the same factors; that is, subjects who scored highly on other items scored low for item 1 in mental health and item 4 in general health. This may be because the word "nervous" can be misinterpreted and the word "worse" could have been confusing because it is the only negatively phrased item in the general health factor.

This study investigated the validity of the SF-36 for measuring the health status of elderly Koreans by applying the Rasch model. We found that nine items on this scale are inappropriate for measuring the health status of elderly Koreans. Therefore, it is necessary to revise this scale to make it better able to measure the health status of elderly Koreans. Future studies should explore ways of improving the validity of the SF-36 for elderly Koreans.

REFERENCES

1) Statistics Korea: 2011 Statistics data for elderly. Statistics Korea. http://sgis.kostat.go.kr/publicsmodel/ (Accessed Aug. 8, 2014)
2) Morala D, Shiomi T: Assessing reliability and validity of physical performance test for the Japanese elderly. J Phys Ther Sci, 2004, 16: 15–20. [CrossRef]
3) Sangpring P, Vongsirinavarat M, Hiengkaew V, et al.: Development of a geriatric fear of falling questionnaire for assessing the fear of falling of Thai elders. J Phys Ther Sci, 2012, 24: 359–364. [CrossRef]
4) Nam BH, Lee SW: Testing the validity of the Korean SF-36 health survey. J Korean Soc Health Stat, 2003, 28: 3–24.
5) Kim YG: The research for health perception and quality of life in elderly. J Korean Soc Occup Ther, 2008, 16: 100–110.
6) Ware JE, Kosinski M: SF-36 physical & mental health summary scale: A menual for users of version 1, 2nd ed. RI: Quality Metric Incorporated, 2001.
7) Brazier J, Usherwood T, Harper R, et al.: Deriving a preference-based single index from the UK SF-36 Health Survey. J Clin Epidemiol, 1998, 51: 1115–1128. [Medline] [CrossRef]
8) Gandek B, Ware JE Jr, Aaronson NK, et al.: Tests of data quality, scaling assumptions, and reliability of the SF-36 in eleven countries: results from the IQOLA Project. International Quality of Life Assessment. J Clin Epidemiol, 1998, 51: 1149–1158. [Medline] [CrossRef]
9) Han CW, Lee EJ: Use of structural equation modeling to test construct validity of the SF-36 health survey among community-dwelling elderly in Korea. J Korea Acad Care Manage, 2009, 4: 45–59.
10) Hammond SM: An IRT investigation of the validity of non-patient anlogue research using the Beck Depression Inventory. Eur J Psychol Assess, 1995, 11: 14–20. [CrossRef]
11) Beck AT, Steer RA, Garbin MG: Psychometric properties of the Beck Depression Inventory: twenty-five years of evaluation. Clin Psychol Rev, 1988, 8: 77–100. [CrossRef]
12) Hong SH, Cho YR: An application of Rasch rating scale model: construction of a short version of the dysfunction beliefs test. Korean J Clin Psychol, 2006, 25: 865–880.
13) Kim SH, Kang SJ, Yang ES: Development and validation of obesity risk measurement scale through Rasch model. J Meas Eval Phys Educ Sport Sci, 2011, 13: 9–22.
14) Humphry S: Item set discrimination and the unit in the Rasch model. J Appl Meas, 2012, 13: 165–180. [Medline]
15) Massof RW: Is the partial credit model a Rasch model? J Appl Meas, 2012, 13: 114–131. [Medline]
16) Olson JR, Belohlav JA, Cook LS: A Rasch model analysis of technology usage in Minnesota hospitals. Int J Med Inform, 2012, 81: 527–538. [Medline] [CrossRef]
17) Ware JE Jr, Sherbourne CD: The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection. Med Care, 1992, 30: 473–483. [Medline] [CrossRef]
18) Koh SB, Chang SJ, Kang MG, et al.: Reliability and validity on measurement instrument for health status assessment in occupational workers. J Prev Med Pub Health, 1997, 30: 251–265.
19) Henning G: Dimensionality and construct validity of language test. Lang Test, 1992, 9: 1–11. [CrossRef]
20) Reckase MD: Unifactor latent trait models applied to multifactor tests: Results and implications. J Educ Stat, 1979, 4: 207–230. [CrossRef]
21) Smith RM, Schumacker RE, Bush JM: Using item mean squares to evaluate fit to the Rasch model. Paper presented at the annual meeting of the American educational research association. San Francisco, 1995.
22) Linacre JM: Generalizability theory and many-facet Rasch measurement. Paper presented at the American Educational Research Association, 1993.
23) Chi E: Comparing holistic and analytic scoring for performance assessment with many-facet Rasch model. J Appl Meas, 2001, 2: 379–388. [Medline]
24) McNamara TF: Measuring second language performance. London: Longman, 1996.
25) Bollen KA: Structural equations with latent variables. New York: Wiley, 1989.
26) Gilbert DT, Fiske ST, Lindzey G: The handbook of social psychology 4th ed. Boston: McGraw Hill, 1998.
27) Hong S, Kim B, Wolfe M: A psychometric reversion of the European American values scale for Asian Americans using the Rasch model. Meas Educ Couns Dev, 2005, 37: 194–207.
28) Hong S, Wong EC: RASCH rating scale modeling of the Korean version of the beck depression inventory. Educ Psychol Meas, 2005, 65: 124–139. [CrossRef]