한국전문물리치료학회지 2014년 21권 4호 40-48 ISSN (Print) 1225-8962, ISSN (Online) 2287-982X
Phys Ther Korea 2014;21(4):40-48 http://dx.doi.org/10.12674/ptk.2014.21.4.040

Differential Item Functioning of the Oswestry Low Back Pain Questionnaire Between Participants With and Without Low Back Pain

Bong-sam Choi, PhD, MPH, PT
Dept. of Physical Therapy, College of Health and Welfare, Woosong University

Abstract

Differential item functioning (DIF) based on Rasch model can be used to examine whether the items function similarly across different groups and identify items that appear to be too easy or difficult after controlling for the ability levels of the compared groups. The Oswestry low back pain disability (Oswestry) has traditionally been proved as an effective instrument measuring disability resulting from low back pain (LBP). In this study, DIF method was used to explore whether items on the Oswestry perform similarly across two different groups (participants with LBP and no LBP). A series of Rasch analyses on the 10 items of the Oswestry were performed using Winsteps® software. Forty-two participants with back pain were recruited from 3 rehabilitation hospitals in Gainesville, Florida. Another 42 participants with no LBP were recruited from several public places in the rehabilitation hospitals. Based on the DIF analysis across the two groups, several items were found to have an uniform DIF. Participants with no LBP had more difficulty on lifting and personal care items and participants with LBP had more difficulty on sleeping and social life items. For non-LBP group, a high ceiling effects (83% of participants with non-LBP) was detected, which was not be able to be effectively measured with the Oswestry items. Although 4 items of the Oswestry function differently across the two groups, all items of the Oswestry were well targeted the LBP group.

Key Words: Assessment; Item response theory; Low back pain; Measurement; Rasch analysis; Rehabilitation.

Introduction

The Oswestry low back pain disability questionnaire (Oswestry), or versions of it, has been used for decades to measure disability resulting from low back pain (LBP). The Oswestry was created by John O'Brien in 1976 and originally called the Oswestry disability index. Fairbanks et al (1980) had first published the Oswestry in physiotherapy in 1980. A latest version from several updates through replacing an item with employment item was created (Fritz and Irrgang, 2001). Despite the many revisions, the Oswestry only provides an indication of how LBP impacts on an individual's ability to manage daily life tasks. It often fails to provide detail information on test items other than the summated score. A search performed by using MEDLINE found 161 citations in which disabilities resulting from LBP were evaluated by using the Oswestry (Fairbank and Pynsent, 2000). Today, the Oswestry, or versions of it, is the most widely accepted condition-specific outcome measure for LBP.

Like other classical test theory (CTT)-based instruments measuring function domain of a model of disablement, the Oswestry produce a summated score to represent the functional limitation pertaining to LBP. The score provides only an overall sense of disability resulting from LBP as a whole and no item level psychometrics such as how individual patients respond on the individual items (Davidson, 2008; Lu

Corresponding author: Bong-sam Choi bchoi@wsu.ac.kr

ISSN (Print) 1225-8962, ISSN (Online) 2287-982X
http://dx.doi.org/10.12674/ptk.2014.21.4.040

et al, 2013). Such information about individual test items is essential to precisely assess the domain of function within an instrument. Of the item level psychometric properties, an information on whether the items respond similarly across different groups and identify items would be valuable to determine items that appear to be too easy or difficult for the groups (i.e., differential item functioning). This would allow investigators suggest that the particular items can be removed to avoid potential biases that might result against a particular subgroup. Although the Oswestry is widely used to assess outcomes and classify patients with low back problems, no previous differential item functioning (DIF) investigations have performed on the Oswestry.

Compared to conventional CTT-based models, the Rasch model (1-parameter item response theory model) focus on item level statistics rather than the test as a whole. The Rasch model estimates person ability measure based on the probability of choosing a response category of item by conjoining item difficulty. The estimated item difficulty and person measure always represent invariant and consistent measures on the latent traits over time. It never changes like a ruler by using a unit of measurement called a logit (i.e., log-odds unit). The invariant logit scale is based on the rationale that patients with low disability will have higher probability of getting low response category on difficult item while patient with high disability will have lower probability of getting higher response category on easy items. The invariant property of item difficulty would allow one to estimate person ability with respect to the items used with an instrument and item statistics under different groups remain unchanged (Taherbhai and Young, 2004). In turn, these item characteristics should be consistent across different patient groups. This leads to the investigation in the consistency of item performance across groups such as DIF procedure.

DIF is a statistical procedure identifying items that appear to be having difficulty levels that are dependent on membership to a particular group after controlling for the ability levels of groups being compared (Finch and Hernandez Finch, 2014; Huang, 2014; Teresi, 2001). The DIF has become central to the investigation in the health-related measurement field to compare response patterns across gender, ethnicity, educational level, age, countries, severity groups and varied diagnostic groups of different item functioning across two groups (Fleishman and Lawrence, 2003; Haley et al, 2004). The Rasch model has a strong assumption that item discrimination parameters are equal across all items. This assumption makes Rasch model allows to detect uniform DIF where there is a relative advantage for one group over the other group through the entire ability range and non-uniform DIF where one group has a relative advantage over the other group at particular person ability range but has a relative disadvantage at other person ability range.

The purposes of this study are: 1) to demonstrate how the Rasch measurement model can be used to determine the dimensionality of the Oswestry items, 2) to inspect the hierarchical order of the Oswestry items, 3) to display how the Oswestry items function differently across the groups with and without LBP.

Methods

Participants

Eighty-four participants with and without LBP (42 participants for each group) participated in this study. Each group was recruited separately. The criteria for participants with LBP included if anyone else is: 1) currently having LBP, 2) having previous treatments for LBP, 3) having ability to read and understand English, and 4) being age 18 years or older. The participants with LBP were recruited from 3 out-patient rehabilitation clinics in Gainesville, Florida; Shands Hospital at University of Florida, Shands Orthopaedics and Sports Medicine Institute and Shands Rehabilitation Hospital. All selected participants presenting to the recruiting sites between

ISSN (Print) 1225-8962, ISSN (Online) 2287-982X
http://dx.doi.org/10.12674/ptk.2014.21.4.040

November 3, 2009 and June 30, 2010 were recruited and scheduled for the data collection. The participants were asked to complete the Oswestry following a screening for the inclusion criteria. During the same data collection period, forty-two participants without LBP were recruited from multiple public sites in Gainesville, Florida. The criteria for participants without LBP included if anyone else is: 1) currently having no LBP, 2) able to read and understand English, and 3) 18 years of age or older. This study was approved by the Institutional Review Board at the University of Florida (Approved by IRB #17-2009).

The mean age was 53.7 years ranging from 18 to 74 years for LBP group and 48.7 years ranging from 19 to 78 years for non-LBP group. The participants included 29 females (69%)/13 males (31%) for back pain group and 27 females (64.3%)/15 males (35.7%) for non-LBP group. The participants with more than a year of back-related problems were nearly 60%.

Instruments and measurements

The Oswestry is one of the most widely accepted and currently used self-report instruments measuring the impact on patients' ability to manage daily life tasks (Fritz and Irrgang, 2001). The Oswestry generally provides an indication of perceived disability resulting from back pain. The latest version of the Oswestry used in the present study contains ten items of pain intensity, personal care, lifting, walking, sitting, standing, sleeping, employment/home-making, and traveling. Participants were responded on a 6 point (5 to 0) ordinal scale according to how much difficulty they experience in daily life. For each test item, the possible score would be 5 (more disabled) to 0 (least disabled) and 50 to 0 for summated total score. The total score (i.e., sum of all item responses) is converted to a percentage score ranging from 0 (no disability) to 100 (most severe disability).

Goodness of fit statistics (fit statistics) were obtained using Winsteps® software program ver. 3.57.2 (Linacre, Chicago, IL, USA) to determine the dimensionality of the Oswestry. The dimensionality can be determined by the fit statistics from the Rasch measurement model (Bond and Fox, 2001). For each item, the fit statistics was produced by scrutinizing mean square standardized residuals (MnSq), which represents observed variance divided by expected variance. The original logit measures was converted to 0-100 scale by using the UMEAN and USCALE commands in Winsteps® software program which lead to direct comparisons between logit measures and the original scores of the Oswestry. The optimal value of the residuals for an item is 1.0. Wright and Linacre (1994) suggested that acceptable criterion of the residuals would be between .6 and 1.4 for general survey study. The criterion for the optimal range is determined by the intended purpose of the measure and the degree of rigor desired (Wright and Linacre, 1994). An item with low or high residual values suggests that the item may be redundant or not be belonging to the trait being measured.

The person measure and item difficulty for both groups were determined by applying a series of Rasch analyses. Rasch analysis linearly transformed the raw scores of the Oswestry to the logit estimates and places test items and persons on the same linear continuum along with the item difficulty and person measures (i.e., person-item map). That is, plotting measure of items and persons on the linear continuum can reveal how well the Oswestry items capture the disability levels expressed the two known groups. By using the person-item map, ceiling or floor effects will visually be inspected.

In addition, the DIF method used in this study was based on the differences between two parameters calibrated on the same item from two groups (Wright and Stone, 1979). Given the pairs of item calibrations and associated estimates of the standard error of estimate from the Rasch model, a t-statistic can be constructed for each item using the formula: where d_{i1} and d_{i2} are the item difficulty of item i in

$$t = \frac{d_{i1} - d_{i2}}{(s_{i1}^2 + s_{i2}^2)^{1/2}}$$

the calibration based on group 1 and 2, s_{i1} is the standard error of estimate for d_{i1}, and s_{i2} is the standard error of estimate for d_{i2}. A graphical representation method equivalent to the t-statistic method was also proposed in the method.

Results

Dimensionality

In order to investigate an evidence confirming the extent to which item represent a unidimensional construct, the fit statistics from the Rasch analysis were inspected. Table 1 and Table 2 presents the Oswestry items in order of item calibrations as well as the fit statistics for participants with and without LBP. The item calibrations for all the items more varied in the non-LBP group than LBP group. The Table 1 represents that all the items fits to the Rasch model except for the employment item. However the fit statistic of employment item was a nearly acceptable range for survey data.

The Table 2 represents that three items (pain, employment, and sleeping item) show unacceptable fit statistics for non-LBP group. Therefore, these 3 items were considered as either measuring other constructs than disability resulting from LBP or not fitting to the Rasch model.

Table 1. Fit statistics of the Oswestry for low back pain group

Items	Difficulty (logits)	Infit MnSq[a]	ZSTD[b]	Outfit MnSq	ZSTD
Lifting	56.78	.76	-1.1	.73	-1.2
Standing	55.39	1.11	.6	1.08	.4
Employment	51.67	.60	-2.3	.58	-2.3
Pain	51.55	1.25	1.2	1.18	.8
Social life	49.45	1.05	.3	1.04	.3
Sitting	48.24	1.31	1.4	1.33	1.4
Travel	46.73	.80	-.9	.87	-.5
Walking	44.41	1.28	1.2	1.17	.7
Sleeping	43.01	1.04	.3	1.01	.1
Personal care	41.66	.90	-.4	.83	-.7

[a]mean square standardized residuals, [b]Z-score standardized.

Table 2. Fit statistics of the Oswestry for non-low back pain group

Items	Difficulty (logits)	Infit MnSq[a]	ZSTD[b]	Outfit MnSq	ZSTD
Lifting	68.79	1.24	.8	1.32	1.0
Pain	59.29	.49	-1.8	.51	-1.6
Standing	59.23	.95	.0	.92	-.1
Personal care	55.85	.86	-.3	.73	-.7
Walking	53.36	.98	.1	.91	-.1
Employment	50.28	1.75	1.8	1.15	.5
Sitting	43.85	1.05	.3	.70	-.4
Travel	43.85	1.39	1.0	1.08	.3
Social life	29.25	1.33	.8	.92	.3
Sleeping	25.16	.47	-1.1	.24	-.5

[a]mean square standardized residuals, [b]Z-score standardized.

ISSN (Print) 1225-8962, ISSN (Online) 2287-982X
http://dx.doi.org/10.12674/ptk.2014.21.4.040

Hierarchical order of the Oswestry items

Figure 1 presents the hierarchical order of item difficulty of the Oswestry for the both groups. The items are displayed on the right side in order of the most difficult at the top and the easiest at the bottom, while the person ability are listed on the left side with the same fashion from the lowest at the bottom to the highest at the top. The lifting item was the most difficult for both groups, while per-

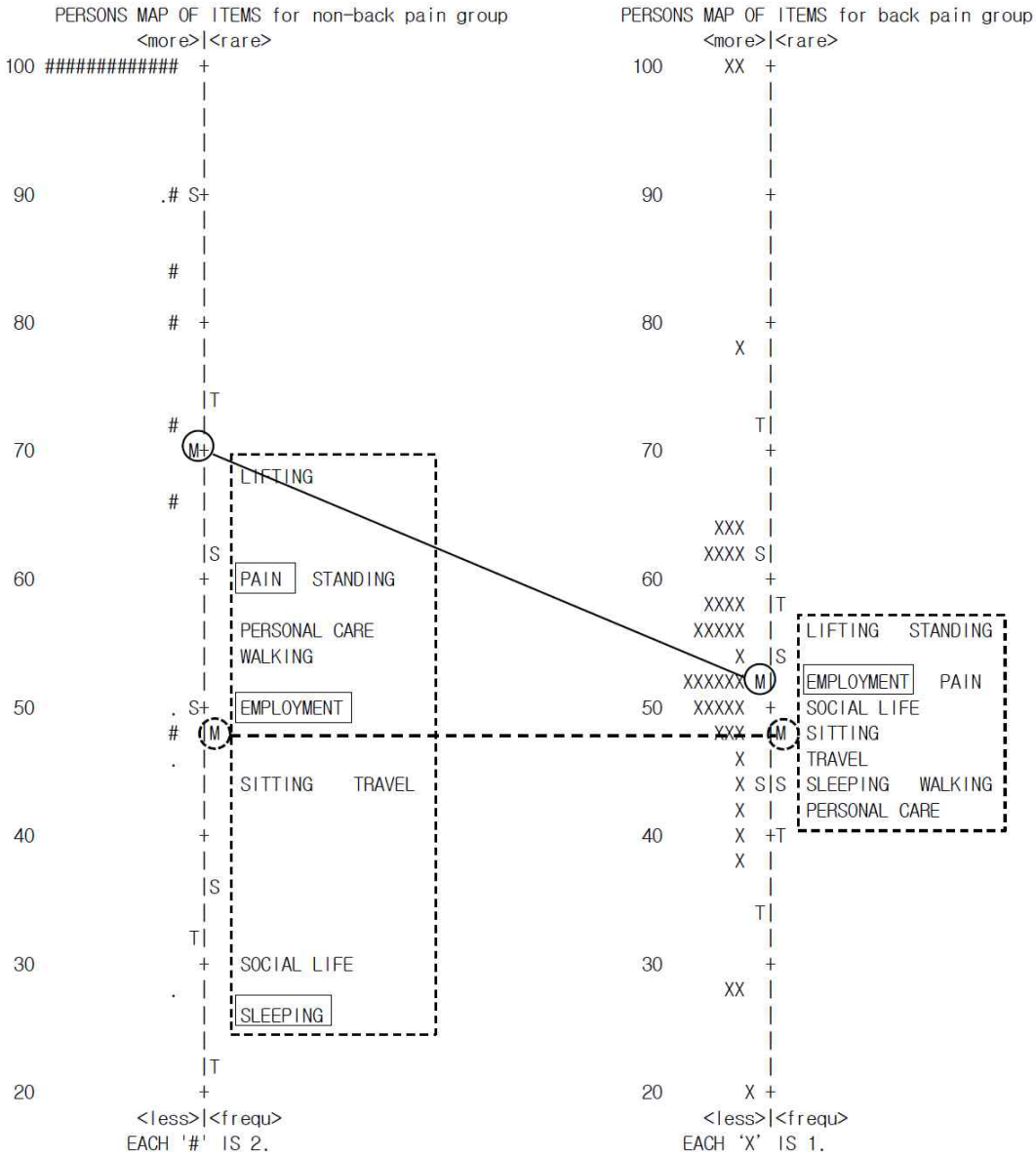

Figure 1. Item-person map of the Oswestry for non-back pain group (left) and back pain group (right) (The graph represents the person ability measures on the left side and item difficulty measures on the right side of each map in the 0-100 converted score from logits with the Rasch analysis. Each analysis is anchored the average person ability measure to compare item difficulties for both groups. The average item calibrations and person measures for both groups were presented in dotted and solid line respectively. The items with high/low fit statistics were highlighted.).

ISSN (Print) 1225-8962, ISSN (Online) 2287-982X
http://dx.doi.org/10.12674/ptk.2014.21.4.040

sonal care item of the back pain group and sleeping item of the non-LBP group was the easiest task. This logical fashion of the item hierarchy of the Oswestry for the groups was supported by the empirical evidences. The average item calibrations which presented with dotted line for both groups were exactly the same (48.89±4.98 for non-LBP and 48.89±2.26 for LBP group). In addition, the average person ability measures for both groups were differently measured and depicted in a solid line. The average person ability measures were about 70.73±14.57 for the non-LBP and 53.68±5.69 for the LBP group. For the non-LBP group, an obvious ceiling effect (26 of 42 participants) was detected, while all items were able to capture only the participants with low ability level. For the LBP group, the items were well targeted the middle level of participants indicating that the Oswestry is sensitive to measure the LBP group. The item difficulty calibrations well matched person ability measures for the LBP group, while the calibrations of the non-LBP group were poorly targeted the persons.

DIF analysis based on Rasch model

In general, participants with non-LBP had higher scores than participants with LBP except sitting, sleeping, social life, travel and employment items. Figure 2 presents the DIF plots of the Oswestry items for the two groups. Items demonstrated significant DIF were labeled with dotted circles. Four (lifting, personal care, social life and sleeping items) out of ten items were located outside of 95% confidence interval. That is, difficulty ratings between those two groups were different for these items. There were tendencies for non-LBP group to select higher ratings (low ability) on lifting and personal care items and for LBP group to select lower ratings on social life and sleeping items. Similarly the LBP group had a tendency to rate higher ratings for the social life and sleeping items.

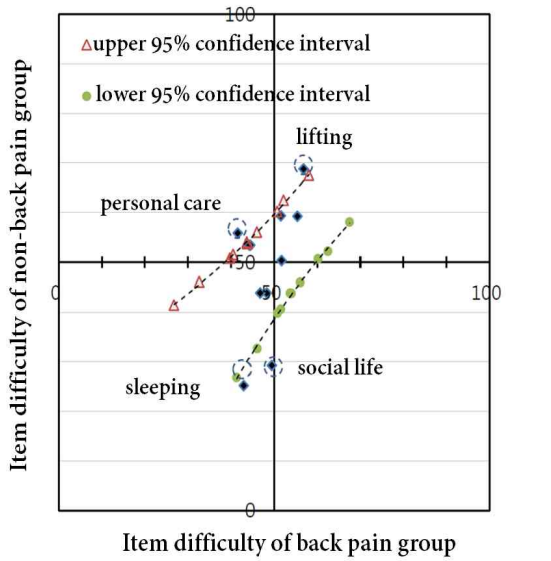

Figure 2. Differential item functioning (DIF) plots for the Oswestry items across two groups (The dashed lines represent the upper and lower 95% confidence intervals. The dashed circled items-lifting, personal care, sleeping, social life items-represent DIF responding differently across the groups. The measures are converted to 0~100 score from logits following the Rasch analysis.).

Discussion

Overall, the Oswestry was found to show construct validity except for the employment item when it applied to participants with LBP, however a few items of the instrument were failed to show the validity when it applied to participants with no LBP. The employment item for the both groups and pain/sleeping items for non-LBP group were problematic fit statistics indicating that the responses to these items were erratic than expected. That is, those items might be measuring another latent trait or in need of further clarification to fit the Rasch model. In this study, the criterion for determining items that did not fit the Rasch model was infit or outfit MnSq≥1.4 or ≤.6 in which several researchers previously presented (Bond and Fox, 2001; Davidson, 2008; Velozo and Peterson, 2001). Although the fit statistics of the employment item for back pain

ISSN (Print) 1225-8962, ISSN (Online) 2287-982X
http://dx.doi.org/10.12674/ptk.2014.21.4.040

group presented a problematic criterion, it was nearly an acceptable range (.58). This item was the one that had replaced with the 'sex' item when Fritz and Irrgang (2001) revised a new version of the Oswestry disability index. This finding would prompt further investigations to clarify if the employment item is appropriate within the Oswestry. As also previously stated, the pain and sleeping item of the non-LBP group was also misfitted. These two items have traditionally been reported as problematic items within the Oswestry (Lu et al, 2013; White and Velozo, 2002). However that was a surprising finding, because the Oswestry was not originally created for either normal populations or other impairment groups.

An empirical evidence of the item difficulty hierarchical order of the Oswestry supported in the logical fashion. That is, participants with LBP had a tendency to rate their disability on the difficult items more severly relative to the easy item. For example, lifting heavy objects is typically one of the leading causes of back related injury compared to the injury from during sleeping or walking. One can expect that the participants would report their disability more severly on the difficult item than sleeping or walking item. However, a surprising finding in the present study was that participants with no LBP appeared to have higher item calibrations (68.79) than participant with LBP (56.78) on lifting item when they were compared simultaneously (Figure 1). Several reasons may be postulated for participants with no LBP having tendency to rate their perception to the lifting function with more severe than participant with LBP. Since participants with no LBP might have not been exposed to LBP conditions and no idea of lifting strategies, these participants may have more of the tendency to view 'lifting' function as being more challenging than viewed by participants with LBP. Or participants with LBP may view the function as being less challenging than the view of participants with no LBP. That is, there may be more strategies aimed at dealing with proper body mechanics in the participants who were recruited from patients currently enrolled for rehabilitation programs. Other items showed similar item hierarchy several studies reported (Davidson, 2008; Lu et al, 2013; White and Velozo, 2002).

As described in Figure 2, four items displayed uniform DIF by the group membership. That is, across the entire ability level, the DIF analysis showed that participants with no LBP perceived their lifting and personal care function to be lower (i.e., more difficulty) than perceived by participants with LBP (68.79 vs. 56.78). In turn, although these participants have no current LBP or history of LBP, their perception to these functions are lower than the other group. This finding was somewhat unexpected because one can believe that participants with no LBP would logically be responded with the easiest response category on those items. This unexpected response was probably due to overly estimated the definition of 'heavy, medium or light weights' from the statement of the 3rd response category on lifting item (i.e., pain prevents me from lifting heavy weights, but I can manage light to medium weights if they are conveniently positioned). In fact, nearly 20% of the participants (8 out of 42) with no LBP selected 2 or higher ratings despite no specific functional disabilities. This may mean that participants with no LBP may be more apprehensive to lifting/personal care functions because they may not have been exposed to such functional limitation or considered their perception to the functions. Similarly, participants with LBP perceived sleeping and social life function to be more difficult than perceived by participants with no LBP (43.01 vs. 25.16 for sleeping and 49.45 vs. 29.25 for social life item respectively). In contrast, Davidson (2008), in a study of the Rasch analysis of three versions of the Oswestry disability index questionnaire, found DIF on the walking item by age. The study showed that persons in 65 year or older group had a tendency to rate their walking ability with higher score (more disability) than expected when the group was com-

ISSN (Print) 1225-8962, ISSN (Online) 2287-982X
http://dx.doi.org/10.12674/ptk.2014.21.4.040

pared to younger group.

A limitation of the current study is the application of the Oswestry instrument to participants with no back pain despite the use of known groups validity. The instrument was not designed to detect any potential disability conditions but to measure disability resulting from LBP. However the primary purpose of this study was to apply the Rasch measurement model to detect DIF items. Further research on the analysis of DIF is needed to test whether the Oswestry items equivalently operates within back related groups across other group membership.

Conclusion

The goals of this study were to demonstrate how to apply Rasch measurement model to generate item difficulty, the hierarchical order and DIF analysis of the Oswestry items with the use of known groups. The Oswestry items showed an acceptable fit statistics for participants with LBP, while three items misfit for participants with no LBP. The hierarchical order of item difficulty was logically supported by the empirical evidences. Four items (lifting, personal care, social life and sleeping items) were detected as uniform DIF indicating the items were differently function across the group membership. Despite the uniform DIF of the Oswestry items across the two groups, all items of the Oswestry were well targeted participants with low back pain.

References

Bond TG, Fox CM. Applying the Rasch Model; Fundamental measurement in the human sciences. 2nd ed. Mahwah, NJ, Lawrence Erlbaum Associates Publishers, 2001:23-28.

Davidson M. Rasch analysis of three versions of the oswestry disability questionnaire. Man Ther. 2008;13(3):222-231.

Fairbank JC, Couper J, Davies JB, et al. The oswestry low back pain disability questionnaire. Physiotherapy. 1980;66(8):271-273.

Fairbank JC, Pynsent PB. The oswestry disability index. Spine (Phila Pa 1976). 2000;25(22):2940-2953.

Finch WH, Hernandez Finch ME. Differential item functioning analysis using a multilevel rasch mixture model: Investigating the impact of disability status and receipt of testing accommodations. J Appl Meas. 2014;15(2):133-151.

Fleishman JA, Lawrence WF. Demographic variation in SF-12 scores: True differences or differential item functioning? Med Care. 2003;41(7 suppl):III75-III86.

Fritz JM, Irrgang JJ. A comparison of a modified oswestry low back pain disability questionnaire and the quebec back pain disability scale. Phys Ther. 2001;81(2):776-788.

Haley SM, Coster WJ, Andres PL, et al. Activity outcome measurement for postacute care. Med Care. 2004;42(1 suppl):I49-I61.

Huang HY. Effects of the common scale setting in the assessment of differential item functioning. Psychol Rep. 2014;114(1):104-125.

Lu YM, Wu YY, Hsieh CL, et al. Measurement precision of the disability for back pain scale-by applying rasch analysis. Health Qual Life Outcomes. 2013;11:119. http://dx.doi.org/10.1186/1477-7525-11-119

Taherbhai HM, Young MJ. Pre-equating: A simulation study based on a large scale assessment model. J Appl Measure. 2004;5(3):301-318.

Teresi JA. Statistical methods for examination of differential item functioning (DIF) with applications to cross-cultural measurement of functional, physical and mental health. J Mental Health Aging. 2001;7(1):31-40.

Velozo CA, Peterson EW. Developing meaningful fear of falling measures for community dwelling elderly. Am J Phys Med Rehabil. 2001;80(9):662-673.

ISSN (Print) 1225-8962, ISSN (Online) 2287-982X
http://dx.doi.org/10.12674/ptk.2014.21.4.040

White LJ, Velozo CA. The use of rasch measurement to improve the oswestry classification scheme. Arch Phys Med Rehabil. 2002;83(6):822-831.

Wright BD, Linacre JM. Reasonable mean-square fit values. Rasch Meas Trans. 1994;8(3):370.

Wright BD, Stone MH. Best Test Design: Rasch measurement. Chicago, MESA press, 1979:93-95.

This article was received September 8, 2014, was reviewed September 8, 2014, and was accepted November 3, 2014.