한국전문물리치료학회지 2016년 23권 4호 71-80 ISSN (Print) 1225-8962, ISSN (Online) 2287-982X
Phys Ther Korea 2016;23(4):71-80 https://doi.org/10.12674/ptk.2016.23.4.071

Concordance Rate Between the Ratings of Clinician and Self Ratings of Worker on a Functional Capacity Evaluation

Bong-sam Choi, PhD, MPH, PT
Dept. of Physical Therapy, College of Health and Welfare, Woosong University
Advanced Institute of Convergence Sport Rehabilitation, Woosong University

Abstract

Background: Functional capacity evaluations (FCEs) are designed to systematically assess the capacity to perform work-related tasks and to determine worker's ability to return to the previous job following work-related injuries. These evaluations may be rated either by clinician or worker. There has been a lack of consensus between the two scoring methods.

Objects: This study aimed: 1) to confirm if the data are fit to the Rasch rating scale model and 2) to investigate the item-level concordance rate between the ratings of clinician and injured worker of the FCE.

Methods: A cross-sectional study was conducted with a sample (n=124) of a rehabilitation program with the Occupational Rehabilitation Data Base for workers with low back pain. The functional capacity evaluation at admission and discharge was administered to clinicians and workers. The data were analyzed using both classical test theory-based Pearson's r and intra-class coefficient followed by item-level analysis with Rasch rating scale model.

Results: All items of the FCE, except sitting items rated by clinician at admission and handling items rated by both clinician and worker throughout admission and discharge, were acceptable fit statistics with minor out of ranges for a misfit criterion. This may indicate that the items of the FCE overall fit to the Rasch rating scale model. Few problematic items responding differently to clinician and worker both at admission and discharge were detected with the differential item functioning analysis despite the excellent concordance rate using the two conventional statistics-sitting and handling items at admission and handling item at discharge.

Conclusion: The item-level speculations using Rasch analysis of the FCE demonstrate that the ratings of clinician and self ratings of worker were psychometrically acceptable though there was an apparent discrepancy between the raters both at admission and discharge.

Key Words: Functional capacity evaluation; Low back pain; Measurement; Rasch analysis; Rehabilitation.

Introduction

Sustained pain or disability due to work-related injuries commonly connect to poor physical performance capacity of injured workers. An extensive research paper on approximating the functional capacity to accomplish work-related tasks has been published by applying indirect measure such as questionnaire-based evaluations (Cutler et al, 2003; Fishbain et al, 1994; Gross, 2004; Gross and Battie, 2004; James et al, 2016; Ratzon et al, 2015; Soer et al, 2008; Trippolini et al, 2015). Despite the limitations of indirect measure, the functional capacity evaluations (FCE) have played an essential role in what rehabilitation clinicians may encounter in the care of the workers with work-related functional deficits (Kuijer et al, 2012; Soer et al, 2008). Several FCEs have been developed for the last few decades such as the Isernhagen work system and physical work performance evaluation, only a few functional

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

capacity evaluation systems are currently in use at many clinical settings (Chen, 2007). Of these evaluations, an instrument measuring residual functional capacity based on the dictionary of occupational title (DOT) was firstly developed by Fishbain and colleagues and widely accepted in clinical settings (Fishbain et al, 1999). The evaluation includes 17 DOT items measuring standing, walking, sitting, lifting, carrying, pushing, pulling, climbing, balancing, stooping, kneeling, crouching, crawling, reaching, handling, fingering, and feeling function (United States. Department of Labor, 1986). The evaluation was found to be psychometrically optimal (Fishbain et al, 1999; Velozo et al, 2006). There was another instrument classifying injured workers into a returning-to-work model similar to that of Fishbain's, called the Occupational Rehabilitation Data Base (ORDB) FCE (Velozo and Santopoalo, 1994). The FCE testing the capability of workers' returning to work was designed for rehabilitation clinicians to evaluate injured workers as well as for injured workers to rate themselves in terms of functional deficits.

However, there has been little consensus on a subtle discrepancy between clinician-generated vesus client-generated measures or proxy versus self-reported measures (Bovend'Eerdt et al, 2011; Davis et al, 2007; Magaziner et al, 1997; van der Linden et al, 2006; Velozo et al, 2006). Several studies provide some supports for the use of proxy-report versus self-reported measures, while both measures are proved some discrepancies (Davis et al, 2007; Lim et al, 2014). Likewise, for the discrepancy between professional clinician and client view of the evaluations, investigators presented convincing arguments of which measures are more valid in representing low correlations of the two measures (Bovend'Eerdt et al, 2011; Velozo et al, 2006). In general, clinician-generated measures are often considered more objective and valid than the client generated measures, especially on pain-related measures for functioning. That is, although clients have a tendency to rate their functional deficits more severe than what they were actually able to perform functional tasks, in a review of FCE for the clinical use, studies encouraged using client-reported measures due to conserving administrative time and stable psychometric property (Choi and Park, 2012; Velozo et al, 2006). To date, only two studies on the issue of discrepancy between clinician and client have been published. Consequently, less is known about how these two measures correspond with each other, particularly on the FCE (Choi and Park, 2012; Velozo et al, 2006).

Concordance or agreement rate represents the extent to which measures are assigned to the same score on particular variables. There are several ways to examine the concordance rate of two measures such as Pearson's r, intra-class correlation coefficient (ICC), kappa statistic, and percent agreement (Beninato et al, 2009; De Civita et al, 2005; Lim et al, 2014). Additionally, some studies on a conceptual variable such as quality of life suggest the use of Student's t-test combining Pearson's r to test the concordance rate between raters (Bray et al, 2010; Eiser and Morse, 2001; White-Koning et al, 2007). Of these statistics, the kappa has generally been thought to be the most robust statistic when taking into account the concordance rate, in other words inter-rater reliability (McHugh, 2012; Viera and Garrett, 2005). However, all these statistics based on score-level with raw data do not provide meaningful information at the item-level (i.e., scale dependent), but only overall information on the magnitude in which those two measures are coincided with each other. In addition, the score-level statistics with raw data vary sample to sample (i.e., sample dependent) (Velozo et al, 1999).

Rasch analysis focusing on the psychometric properties of item-level within assessment can examine the item-level stability of measures (Linacre, 2002). Rasch analysis can be applied to assessments to estimate item difficulty and person ability based on the probability of getting a particular rating in response to an individual item within assessment. These two estimates are calibrated by converting ordinal rating scores into linear measures (Wright and Linacre, 1989) and represented as being invariant measures on the traits over time (i.e., scale independent) (Haley et al,

2004; Velozo et al, 1999). The property of invariant item difficulty measures allows estimating person ability measures in view of items within assessment and remaining unchanged item statistics from sample to sample (i.e., sample independent) (Rouquette et al, 2016; Taherbhai and Young, 2004, Velozo et al, 1999). These invariant item difficulty can be compared by differential item functioning (DIF) method identifying items that appear to be shown how group membership is differently respond to particular items (Huang, 2014; Teresi, 2001). The concordance rate between two rater perspectives to their disability levels are commonly be compared by scatter plotting those item statistics. An inquiry has become a major focus on the level of agreement between those two perspectives, clinician and injured worker ratings, on FCE.

The purposes of this study are: 1) to confirm if the data are fit to the Rasch rating scale model and 2) to investigate the item-level concordance rate between the ratings of clinician and injured worker of the FCE.

Methods

Instruments and Participants

The reported data of this study was reinstated from a larger project to develop a FCE based on the DOT (Velozo and Santopoalo, 1994). The project was aimed at unraveling how well items of the FCE capture the functional capacity for injured workers who were incorporated return-to-work issues into their rehabilitation program at 28 rehabilitation institutes between August 22 and November 11, 1994 in Chicago, IL. The FCE incorporated in the project comprises 10 DOT job factors that are typically included in measuring functions for injured workers: standing, walking/running, sitting, lifting, carrying, pushing/pulling, climbing, stooping/crouching/kneeling, reaching, handling/fingering. The items were scored on a four-point scale: 1-severely impaired, 2-moderately impaired, 3-mildly impaired, 4-not impaired. The FCE consists of two rater versions (i.e., clinician and injured worker) evaluating the functional status of the 10 job factors.

A total of 230 workers who were incorporated in the ORDB project were selected and only data from workers with low back pain (n=124) was analyzed in this study. The worker with low back pain and clinician were administered the DOT-based FCE at the time of admission to the work rehabilitation program and at the time of discharge from the program. Written informed consent was obtained from the participants and clinicians before administering the FCE. Eighty-two of the 124 participants (66%) were male and forty two (34%) were female. The average age was 38.1 ranged from 21 to 65 years of age.

Data Analysis

Data were analyzed with Winsteps ver. 3.57.2 (Linacre, Chicago, IL, USA) using Rasch rating scale model in order to confirm if the data were fit to the Rasch model for item-level comparisons. Based on goodness of fit test statistic, mean square (MnSq) value of fit statistics can be calculated by the program. The fit statistic is analogous to the chi-square statistic divided by its degree of freedom. The statistic is commonly expected to obtain close to one and used to identify items that did not fit the Rasch model. Model misfit was determined with range of <.6 and >1.4 for infit/outfit and >2.0 for Z-score standardized statistic (Bond and Fox, 2001; Linacre, 2002). High MnSq values may indicate a non-modeled noise or a source of variance in the data. That is, particular items with high value tend to be measuring a different construct rather than what they were intended to measure. Similarly, Low MnSq value may indicate overly predicting model. That is, the model would be too good to be true for predicting the data (Linacre, 2004). Additionary, Rasch analysis provide estimated item difficulty and person ability on the same linear continuum based on the individual's response on each item. The item difficulty are presented as log-odds units or logits.

In the context of a conventional approach to determine the extent to which ratings of two raters

agree with each other, both Pearson's r and ICC at the time of admission and discharge was analyzed using SPSS program version 23. The DIF method was followed by the presentation to determine if there are differences in response between the ratings of clinician and worker of the FCE. The DIF method applied in the present study demonstrated the level of concordance rate between clinician and injured worker for the 10 DOT items of the FCE. The scatter plots set clinician ratings as the X-axis and injured worker ratings as the Y-axis with a 95% confidence interval (Wright and Stone, 1979).

Results

1. Dimensionality of the FCE at admission and discharge

As an evidence of confirming the extent to which the FCE items represent a unidimensional construct, the fit statistics from the Rasch analysis were inspected. Table 1 through Table 4 represent fit statistics for the FCE items in order of item difficulty rated by clinicians and injured workers at the time of admission and discharge from the rehabilitation programs. By applying the previous standards (Bond and Fox, 2001), two items are detected as erratic response items either with low or high MnSq values (i.e., sitting item rated by clinician at admission and handling item rated both by clinician and worker throughout the whole evaluations). In addition, 5 items represented as darker shade show very high MnSq values (i.e., >1.4) carrying item rated by clinician at admission, 2) stooping, sitting and reaching items rated by clinician at discharge, and 3) sitting and pushing/pulling items rated by worker at discharge). Of these noisy items with high/low fit statistics, it was determined that all items were in acceptable ranges of infit and outfit statistics due to slightly out of misfit criterion. As presented as lighter shade, these two items with erratic patterns (i.e., sitting item rated by clinician at admission and handling items rated both by clinician and worker both at admission and discharge) show very high MnSq values, indicating that these items are poorly measuring the FCE not only on the targeted persons but also on outlier at extreme ends on ability continuum. Overall, all items except the sitting and handling items fit to the Rasch rating scale model, indicating measuring a unidimensional construct on FCE.

2. Concordance rate between clinician and worker ratings

The concordance rate in the Pearson's r and ICC between clinician and worker ratings on the FCE items are excellent both at admission and discharge

Table 1. Fit statistics for clinician ratings at admission

Items (measure order)	Difficulty (logits)	Infit MnSq[a]	ZSTD[b]	Outfit MnSq	ZSTD
Lifting	1.46	.67	-2.9	.65	-2.7
Carrying	1.29	.58	-3.9	.59	-3.5
Stooping	1.24	.72	-2.5	.79	-1.6
Pushing/Pulling	.80	.85	-1.2	.84	-1.3
Standing	.41	1.29	2.2	1.25	1.9
Walking	.26	.72	-2.5	.75	-2.1
Climbing	-.10	1.05	.5	1.02	.2
Sitting	-.35	1.84	5.5	2.01	5.9
Reaching	-1.58	.99	.0	.88	-.6
Handling	-3.42	1.44	2.1	1.71	1.5

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

Table 2. Fit statistics for worker ratings at admission

Items (measure order)	Difficulty (logits)	Infit MnSq[a]	ZSTD[b]	Outfit MnSq	ZSTD
Stooping	1.01	.82	-1.5	.89	-.7
Lifting	.72	.82	-1.5	.80	-1.5
Carrying	.55	.64	-3.3	.66	-2.8
Pushing/Pulling	.47	.67	-2.9	.64	-3.0
Standing	.43	1.08	.7	1.04	.3
Sitting	.13	1.22	1.7	1.17	1.3
Walking	.13	1.04	.4	1.20	1.5
Climbing	.05	.99	.0	.99	.0
Reaching	-1.31	.89	-.9	.92	-.6
Handling	-2.18	1.89	5.4	1.95	4.8

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

Table 3. Fit statistics for clinician ratings at discharge

Items (measure order)	Difficulty (logits)	Infit MnSq[a]	ZSTD[b]	Outfit MnSq	ZSTD
Stooping	1.45	.57	-3.1	.52	-3.3
Lifting	1.35	.72	-2.0	.65	-2.3
Carrying	1.20	.71	-2.0	.66	-2.2
Standing	.66	.96	-.2	.95	-.2
Pushing/Pulling	.61	.79	-1.3	.74	-1.5
Climbing	.29	1.25	1.4	1.10	.6
Walking	.12	.90	-.6	.79	-1.1
Sitting	-.13	1.36	1.8	1.56	2.3
Reaching	-1.91	1.54	2.4	1.29	.8
Handling	-3.63	1.62	2.4	9.90	3.9

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

Table 4. Fit statistics for worker ratings at discharge

Items (measure order)	Difficulty (logits)	Infit MnSq[a]	ZSTD[b]	Outfit MnSq	ZSTD
Stooping	1.07	.81	-1.3	.76	-1.3
Lifting	.93	.85	-1.1	.81	-1.1
Standing	.77	.89	-.7	.89	-.6
Climbing	.45	1.11	.7	1.08	.5
Carrying	.38	.74	-1.9	.75	-1.6
Sitting	.26	1.48	2.9	1.49	2.6
Pushing/Pulling	.09	.45	-4.7	.45	-4.2
Walking	-.04	.97	-.2	.97	-.2
Reaching	-1.20	1.10	.7	.97	-.1
Handling	-2.27	1.77	3.7	1.50	1.3

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

Table 5. FCE mean scores and the concordance rate between clinician and worker ratings

	Clinician rating		Worker rating		Pearson's r	ICC[b]
	Mean (logits)	SD[a]	Mean (logits)	SD		
Admission	.001	1.510	.000	.986	.982*	.947*
Discharge	.001	1.602	.044	1.032	.999*	.953*

[a]standard deviation, [b]intra-class correlation coefficient, *p<.001.

(Table 5). At admission, the mean values of the both ratings are identical, while the ratings of worker at discharge are slightly higher than that of the ratings of clinician, indicating worker have a tendency of rating more difficult on the FCE.

3. DIF analysis using Rasch analysis

For the item-level analysis for the FCE items, the DIF analysis between clinician and worker at the time of admission and discharge are conducted and presented in Figure 1. At admission point, four out of 10 items were located outside of the 95% confidence interval (CI) lines. This indicates that the ratings between clinician and worker were differently responded for those items. That is, the clinician rated two items (i.e., lifting and carrying) more difficult and two items (i.e., sitting and handling) easier than the worker did. Since sitting item rated by clinician and handling item rated by both clinician and worker was misfit to the Rasch model, these items were no longer providing important information on the DIF analysis. Likewise at discharge point, three out of 10 items were located outside of the 95% CI. The clinician rated carrying item more difficult and reaching/handling item easier than the worker did. From the previous fit statistics analysis (Table 3 and 4), all items except handling item were showing acceptable fit statistics despite slightly out of range. Therefore the handling item rated by clinician and worker at discharge was determined as an unimportant item in the DIF analysis.

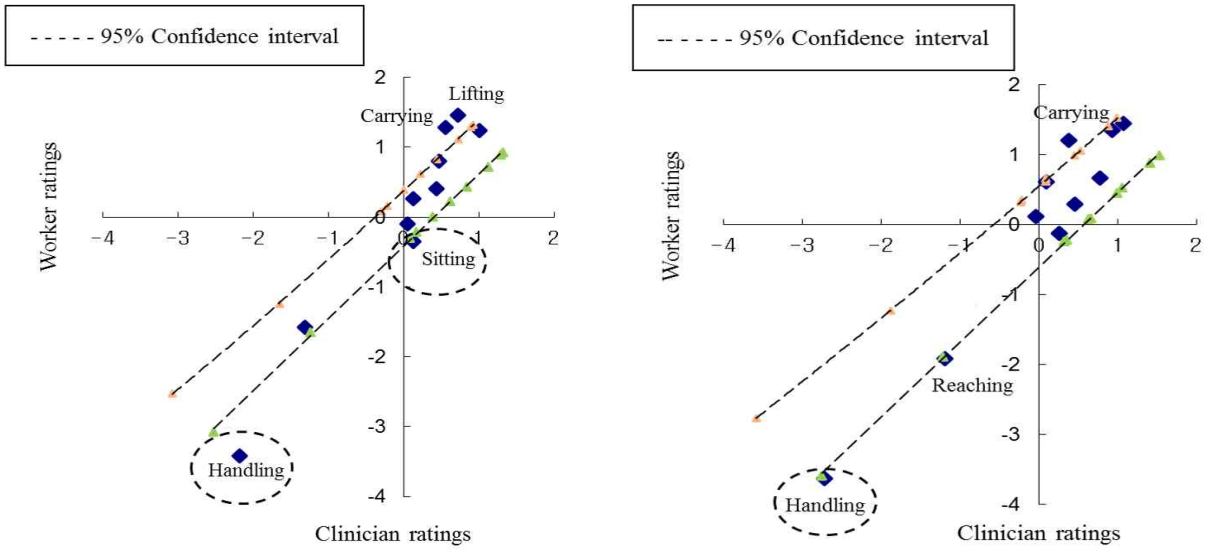

Figure 1. Differential item functioning (DIF) plots for the functional capacity evaluation (FCE) items across clinician and worker ratings at admission and discharge (The dashed lines represent the upper and lower 95% confidence intervals. The carrying and lifting item at admission and carrying and reaching items represent DIF responding differently across the ratings of clinician and worker. The dashed circled items, handling on both at admission/discharge and sitting at admission, represent unacceptable items for the DIF analysis. The measures are converted to logits following the Rasch analysis.).

Discussion

This study explored the item-level differences in concordance rates of the ratings between clinician and injured worker on the FCE. Traditionally, several classical test theory (CTT)-based statistical procedures, such as Pearson's r, kappa, or ICC, have been suggested by many studies to determine the concordance rate between two measures (Beninato et al, 2009; Bray et al, 2010; De Civita et al, 2005; Eiser and Morse, 2001; Lim et al, 2014; White-Koning et al, 2007). Those CTT-based statistics are often failed to provide the insight on item-level information. That is, the statistics offer overall agreements between measures at the score level. However Rasch analysis can provide for promising means to compare the concordance at the item-level of the relationship between two measures. In present study, although the ratings between clinician and worker showed excellent concordance rates, few problematic items were identified by the use of the DIF method following the Rasch analysis. In DIF analysis, lifting and carrying item at admission and carrying and reaching item at discharge were detected as items differently responding to the clinician and worker. The hierarchically ordered item difficulties of the FCE showed a slightly larger discordances on the items of carrying and lifting at admission and the items of carrying and reaching at discharge. The clinician had a tendency to rate lifting and carrying items at admission and carrying item at discharge point more difficult than worker did. Conversely, the worker had a tendency to rate reaching item more difficult than the clinician did at discharge point. The DIF analysis as well as the fit statistics detected handling and sitting items at admission and handling item at discharge as problematic items.

With some varied fit statistics, all items except handling and sitting of the FCE was found to show a unidimensional construct. In this study, the criterion for determining unacceptable items for the unidimensional construct was infit or outfit MnSq$\geq$1.4 or $\leq$.6 in which many authors previously suggested and presented (Bond and Fox, 2001). Although the fit statistics of few items presented a problematic criterion, those were nearly acceptable ranges (i.e., shaded items in Table 1 through Table 4). Handling item was misfit throughout all ratings both at admission and discharge, while sitting item was misfit only on clinician rating at admission. An critique whether or not functional capacity batteries should contain the handling item has been an issue (Fishbain et al, 1994; Fishbain et al, 1999, Velozo et al, 2006). Fishbain and colleagues (1994; 1999) proposed including the handling item as an item of DOT-based job factors because the item was a part of manual dexterity. However some authors were strongly opposed to use the item in any FCE in their factor analysis study identifying the handling item always loaded on a different factor or construct (Velozo et al, 2006). In the present study, item-level analysis using Rasch analysis consistently showed the handling as a problematic item with high fit statistic. Additionally, not only confirmed the handling item problematic, sitting item showed such high infit and outfit statistics (1.84 and 2.01, respectively). These high fit statistic suggest that there may be a non-modeled noise or variance, indicating the sitting item may show 84-101% more variance than expected. Since the sitting task is easy for workers with low back pain, persons with high ability may be rated with either high or low ratings on the sitting item. It is interesting to note that the erratic pattern of the ratings of sitting item at admission point are rated by clinicians. That is, the clinicians may have more of a tendency to view sitting activity as being more challenging than when viewed by workers experiencing low back pain.

The present study that explores the concordance rate between clinician and worker's ratings has produced a surprising finding that could serve as a caution to assess how much both the raters agree with each other on the FCE. However, few problematic items with some discrepancies in raters, items responding differently to clinicians and workers, were detected with the Rasch-based item-level analyses detected. That is, few items of the FCE were differ-

ently responded to clinicians and workers, although the traditional statistics (i.e., Pearson's r and ICC) were excellent, indicating that the two measures were consistent in the concordance rates (ranging from .947 to .999 at p=.001). A reason may be postulated for the traditional statistics being different from the item-level DIF analysis following the Rasch analysis in the concordance rate. In general, CTT-based statistics do not provide information about the extent to which measures are assigned to the same score over time at the item-level. However, the Rasch analysis (i.e., one-parameter item response theory model) can overcome the limitation by applying item-level analysis such as the DIF method (Lim et al, 2014). Rasch item difficulty estimates are mathematically transformed logit scales, which allow invariant estimates being never changed from sample to sample or from a test to another test (Linacre, 2004). Therefore the invariant property as well as the item-level statistics appeared to be a useful method to tap all the underlying discrepancy on which the score-based CTT statistics would hardly discriminate.

While the concordance rate based on both Pearson's r and ICC showed excellent concordance rate between the ratings of clinicians and workers both at admission and discharge, there were apparent differences in the two raters. From the DIF analysis at admission, four items (lifting, carrying, sitting and handling items) were differently responded to the two raters. Of these items, both handling and sitting item were determined to be unacceptable for the DIF analysis due to their high fit statistics, indicating no item-level information being provided by the items. At admission point, there were tendencies for the ratings of clinician to view functional limitations as being more difficult on lifting and carrying items relative to the ratings of workers. At discharge, the tendency of response pattern for the carrying item remained unchanged, while reaching item conversely responded. That is, the concordance rate of lifting and carrying items at admission and carrying and reaching items at discharge were problematic to assess the functional capacity of workers with low back pain. A question then arises as to why this discrepancy between the admission and discharge points happened. The discrepancy is of concern. The clinicians at admission is more likely than workers to be serious on lifting and carrying functions, while the clinicians at discharge is less likely than workers to be serious on reaching function. One may hypothesize that clinicians may be faced with many challenges to educate the injured workers with low back pain for some safety skills of lifting and carrying function throughout the whole stay in rehabilitation program as well as carrying function at discharge. From the workers' perspective, workers may be faced with more challenging on reaching function relative to clinicians.

The present study carries some inherent limitations because the analysis did not include the standard error of those ratings as well as the point measure correlations for the 10 items. It is unknown whether the determination in selecting acceptable items for the concordance rate was optimal. In addition, the problematic items were included in actual analyses due to their conceptual importance to back-related functional deficits. It may be a source of latent errors that this study did not include. Future researches on the concordance rate considering those factors are required for precisely measuring function for workers with low back pain.

Conclusion

The Rasch analysis may provide some findings that challenge those obtained from the CTT-based statistical procedures. Item-level statistics using the Rasch analysis detected few differences between the ratings of clinician and worker for viewing worker's functional deficits resulting from low back pain. Overall, Rasch analysis of the FCE demonstrates that those two ratings were psychometrically acceptable except the sitting and handling items. These would provide clinicians with some supports of worker's perspectives on all items except handling and sitting items.

References

Beninato M, Portney LG, Sullivan PE. Using the International Classification of Functioning, Disability and Health as a framework to examine the association between falls and clinical assessment tools in people with stroke. Phys Ther. 2009;89(8):816-825. https://doi.org/10.2522/ptj.20080160

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.

Bovend'Eerdt TJ, Dawes H, Izadi H et al. Agreement between two different scoring procedures for goal attainment scaling is low. J Rehabil Med. 2011;43(1):46-49. https://doi.org/10.2340/16501977-0624

Bray P, Bundy AC, Ryan MM, et al. Health-related quality of life in boys with Duchenne muscular dystrophy: Agreement between parents and their sons. J Child Neurol. 2010;25(10):1188-1194. https://doi.org/10.1177/0883073809357624

Chen JJ. Functional capacity evaluation & disability. Iowa Orthop J. 2007;27:121-127.

Choi BS, Park SY. Responsiveness comparisons of self-report versus therapist-scored functional capacity for workers with low back pain. Phys Ther Korea. 2012;19(3):91-97. https://doi.org/10.12674/ptk.2012.19.3.091

Cutler RB, Fishbain DA, Steele-Rosomoff R, et al. Relationships between functional capacity measures and baseline psychological measures in chronic pain patients. J Occup Rehabil. 2003;13(4):249-258.

Davis E, Nicolas C, Waters E, et al. Parent-proxy and child self-reported health-related quality of life: Using qualitative methods to explain the discordance. Qual Life Res. 2007;16(5):863-871.

De Civita M, Regier D, Alamgir AH, et al. Evaluating health-related quality-of-life studies in paediatric populations: Some conceptual, methodological and developmental considerations and recent applications. Pharmacoeconomics. 2005;23(7):659-685.

Eiser C, Morse R. Can parents rate their child's health-related quality of life? Results of a systematic review. Qual Life Res. 2001;10(4):347-357.

Fishbain DA, Abdel-Moty E, Cutler R, et al. Measuring residual functional capacity in chronic low back pain patients based on the Dictionary of Occupational Titles. Spine (Phila Pa 1976). 1994;19(8):872-880.

Fishbain DA, Cutler RB, Rosomoff H, et al. Validity of the dictionary of occupational titles residual functional capacity battery. Clin J Pain. 1999;15(2):102-110.

Gross DP. Measurement properties of performance-based assessment of functional capacity. J Occup Rehabil. 2004;14(3):165-174.

Gross DP, Battié MC. The prognostic value of functional capacity evaluation in patients with chronic low back pain: Part 2: Sustained recovery. Spine (Phila Pa 1976). 2004;29(8):920-924.

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.

James CL, Reneman MF, Gross DP. Functional capacity evaluation research: Report from the second international functional capacity evaluation research meeting. J Occup Rehabil. 2016;26(1):80-83. https://doi.org/10.1007/s10926-015-9589-y

Kuijer PP, Gouttebarge V, Brouwer S, et al. Are performance-based measures predictive of work participation in patients with musculoskeletal disorders? A systematic review. Int Arch Occup Environ Health. 2012;85(2):109-123. https://doi.org/10.1007/s00420-011-0659-y

Lim Y, Velozo CA, Bendixen RM. The level of agreement between child self-reports and parent proxy-reports of health-related quality of life in boys with Duchenne muscular dystrophy. Qual Life Res. 2014;23(7):1945-1952. https://doi.org/10.1007/s11136-014-0642-7

Linacre JM. Optimizing rating scale category effectiveness.

J Appl Meas 2002;3(1):85-106.

Linacre JM. Rasch model estimation: Further topics. J Appl Meas. 2004;5(1):95-110.

Magaziner J, Zimmerman SI, Gruber-Baldini AL, et al. Proxy reporting in five areas of functional status. Comparison with self-reports and observations of performance. Am J Epidemiol. 1997;146(5):418-428.

McHugh ML. Interrater reliability: The kappa statistic. Biochem Med (Zagreb). 2012;22(3):276-282.

Ratzon NZ, Amit Y, Friedman S, et al. Functional capacity evaluation: Does it change the determination of the degree of work disability? Disabil Health J. 2015;8(1):80-85. https://doi.org/10.1016/j.dhjo.2014.08.004

Rouquette A, Côté SM, Hardouin JB, et al. Rasch modelling to deal with changes in the questionnaires used during long-term follow-up of cohort studies: A simulation study. BMC Med Res Methodol. 2016;16(1):105. https://doi.org/10.1186/s12874-016-0211-6

Soer R, Groothoff JW, Geertzen JH, et al. Pain response of healthy workers following a functional capacity evaluation and implications for clinical interpretation. J Occup Rehabil. 2008;18(3):290-298. https://doi.org/10.1007/s10926-008-9132-5

Taherbhai HM, Young MJ. Pre-equating: A simulation study based on a large scale assessment model. J Appl Meas. 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 Ment Health Aging. 2001;7(1):31-40.

Trippolini MA, Dijkstra PU, Geertzen JH, et al. Construct validity of functional capacity evaluation in patients with whiplash-associated disorders. J Occup Rehabil. 2015;25(3):481-492. https://doi.org/10.1007/s10926-014-9555-0

United States Department of Labor. Dictionary of Occupational Titles. 4th eds. Supplement. Washington, DC, United States Government Printing, 1986:189-192.

van der Linden FA, Kragt JJ, Hobart JC, et al. Proxy measurements in multiple sclerosis: Agreement between patients and their partners on the impact of multiple sclerosis in daily life. J Neurol Neurosurg Psychiatry. 2006;77(10):1157-1162.

Velozo CA, Choi B, Zylstra SE, et al. Measurement qualities of a self-report and therapist-scored functional capacity instrument based on the Dictionary of Occupational Titles. J Occup Rehabil. 2006;16(1):109-122.

Velozo CA, Kielhofner G, Lai JS. The use of Rasch analysis to produce scale-free measurement of functional ability. Am J Occup Ther. 1999;53(1):83-90.

Velozo CA, Santopoalo R. Training Manual: Occupational rehabilitation data base manual. Chicago, University of Illinois at Chicago, 1994:12-25.

Viera AJ, Garrett JM. Understanding interobserver agreement: The kappa statistic. Fam Med. 2005;37(5):360-363.

White-Koning M, Arnaud C, Dickinson HO, et al. Determinants of child-parent agreement in quality-of-life reports: A European study of children with cerebral palsy. Pediatrics. 2007;120(4):e804-e814.

Wright BD, Linacre JM. Observations are always ordinal; measurements, however, must be interval. Arch Phys Med Rehabil. 1989;70(12):857-860.

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

This article was received October 7, 2016, was reviewed October 7, 2016, and was accepted November 7, 2016.