Patient Related Outcome Measures

Dovepress
open access to scientific and medical research

Open Access Full Text Article

ORIGINAL RESEARCH

Developing brief fatigue short forms calibrated to a common mathematical metric: is content-balancing important?

This article was published in the following Dove Press journal:
Patient Related Outcome Measures
6 August 2010
Number of times this article has been viewed

Karon F Cook[1]
Seung W Choi[2]
Kurt L Johnson[1]
Dagmar Amtmann[1]

[1]Department of Rehabilitation Medicine, University of Washington, Seattle, WA; [2]Northwestern University Feinberg School of Medicine, Chicago, IL, USA

Abstract: There are clinical and research settings in which concerns about respondent burden make the use of longer self-report measures impractical. Though computer adaptive testing provides an efficient strategy for measuring patient reported outcomes, the requirement of a computer interface makes it impractical for some settings. This study evaluated how well brief short forms, constructed from a longer measure of patient reported fatigue, reproduced scores on the full measure. When the items of an item bank are calibrated using an item response theory model, it is assumed that the items are fungible units. Theoretically, there should be no advantage to balancing the content coverage of the items. We compared short forms developed using a random item selection process to short forms developed with consideration of the items relation to subdomains of fatigue (ie, physical and cognitive fatigue). Scores on short forms developed using content balancing more successfully predicted full item bank scores than did scores on short forms developed by random selection of items.

Keywords: psychometrics, outcomes, quality of life, measurement, fatigue

Introduction

Fatigue is a primary complaint of people with numerous conditions and diseases including multiple sclerosis (MS),[1] stroke,[2] cancer,[3] post-polio,[4] arthritis,[5] and Parkinson's disease.[6] Fatigue and other patient-centered outcomes often are measured using retrospective self-reports. Because clinical time is at a premium and response rates have been found to be significantly lower with longer versus shorter surveys,[7] patient outcomes often are measured using relatively brief scales. Despite their practicality, shorter measures typically yield less reliable scores than do longer measures. An alternative to static scales is computer adaptive testing (CAT), but CAT administrations require a computer interface and may be impractical in some settings. One alternative offered by classical test theory (CTT) methods is to construct parallel tests,[8] but tests that are truly parallel are difficult if not impossible to construct. Item response theory (IRT) methods offer a more promising approach. Once a parent item bank has been calibrated to an IRT model, items can be extracted to comprise one or more short forms and scores can be calibrated to a common mathematical metric. Whereas CTT scoring methods assume that participants respond to the same items, with an IRT model persons' trait-levels can be estimated on a common mathematical metric even when persons respond to different subsets of items. Thus, an IRT calibrated item bank provides the opportunity to develop multiple short forms whose scores are directly comparable.

The purpose of this study was to evaluate how well short forms constructed from a longer measure of patient reported fatigue could reproduce scores on the full

Correspondence: Karon F Cook
801 Cortlandt Street, Houston, TX 77007, USA
Tel +1 713 291 3918
Email karonc2@u.washington.edu

measure. Two methods for developing short forms were compared – selecting items randomly and balancing the content of items based on targeted subdomains. In addition, the impact of number of items in the short forms was explored.

Methods

Sample

In a previous study, a sample of persons with MS (n = 466) responded to the 21-item Modified Fatigue Impact Scale (MFIS).[9] For the current study, the data were reduced to include only MFIS responses of those who completed all 21 items (n = 374; 80%). Participants were recruited through the Multiple Sclerosis Association (MSA) of King County (Washington). MSA members were mailed a survey. Approximately 400 returned completed surveys, about a 55% response rate. Information on nonrespondents is unavailable because the surveys were mailed by the association. Study investigators did not have access to the mailing list.

Measure

The MFIS[9] was developed to assess the impact of fatigue on a variety of daily activities. The item content is included in Table 1. Respondents rate their fatigue over the previous four weeks on a 0–4 scale where 0 = Never, 1 = Rarely, 2 = Sometimes, 3 = Often, and 4 = Almost always. The MFIS can be scored as a general measure of fatigue by summing across all items. Alternatively, subscale scores can be generated to estimate levels of physical (9 items), cognitive (10 items), and psychosocial fatigue (2 items).

Table 1 Modified fatigue impact loadings in a one- and in a two-factor solution (promax rotation)

Item #	orig	new	Item content (item difficulty)	First order model			Bifactor model		
				1 factor	2 factor		General factor	Group factors	
				I	I	II			
1	C	C	I have been less alert	0.780	0.724	0.144	0.883	–	–0.185
5	C	C	I have been forgetful	0.743	0.827	0.000	0.788	–	0.170
11	C	C	I have had difficulty making decisions	0.777	0.828	0.036	0.808	–	0.235
3	C	C	I have been unable to think clearly	0.803	0.844	0.050	0.873	–	0.086
12	C	C	I have been less motivated to do anything that requires thinking	0.797	0.867	0.022	0.831	–	0.337
2	C	C	I have had difficulty paying attention for long periods of time	0.794	0.870	0.015	0.897	–	0.051
15	C	C	I have had trouble finishing tasks that require thinking	0.805	0.881	0.016	0.831	–	0.391
18	C	C	My thinking has been slowed down	0.870	0.917	0.055	0.900	–	0.240
19	C	C	I have had trouble concentrating	0.856	0.942	0.013	0.901	–	0.282
16	C	C	I have had difficulty organizing my thoughts when doing things at home or at work	0.794	0.984	–0.096	0.851	–	0.411
6	P	P	I have had to pace myself in my physical activities	0.691	0.013	0.788	0.479	0.635	–
17	P	P	I have been less able to complete tasks that require physical effort	0.744	0.040	0.911	0.504	0.759	–
10	P	P	I have had trouble maintaining physical effort for long periods	0.759	0.071	0.964	0.505	0.769	–
20	P	P	I have limited my physical activities	0.750	0.072	0.954	0.491	0.787	–
13	P	P	My muscles have felt weak	0.654	0.088	0.854	0.420	0.691	–
14	P	P	I have been physically uncomfortable	0.602	0.092	0.597	0.454	0.498	–
7	P	P	I have been less motivated to do anything that requires physical effort	0.778	0.138	0.755	0.595	0.598	–
4	P	P	I have been clumsy and uncoordinated	0.675	0.185	0.582	0.534	0.456	–
21	P	O	I have needed to rest more often or for longer periods	0.779	0.200	0.689	0.615	0.534	–
8	S	O	I have been less motivated to participate in social activities	0.758	0.250	0.610	0.623	0.504	–
9	S	O	I have been limited in my ability to do things away from home	0.718	0.031	0.869	0.496	0.690	–

Abbreviations: Orig, original categorization of the MFIS; C, cognitive; P, physical; S, psychosocial; New, recategorization of the MFIS; C, cognitive; P, physical; O, other.

Dimensionality assumption

An assumption of IRT models and CTT is unidimensionality; that is, it is assumed that a single latent construct drives the variance in scores. It is well-recognized that the assumption of unidimensionality is never strictly met in the context of health outcomes measurement. A scale of a very narrowly-defined construct could be expected to exhibit good fit to a unidimensional model based on conventional fit criteria,[10] but most health constructs have greater conceptual breadth and require a broader range of indicators.[11] Health outcomes are conceptually complex and never perfectly meet strictly defined unidimensionality assumptions.[12–15]

A number of approaches have been suggested for evaluating model assumptions, and often the findings of several methods are compared.[16–20] Reise and Haviland[14] have recommended comparing first-order unidimensional models with bi-factor models.[19,20] With a bifactor model, in addition to a general factor, there are "group" factors that account for score variation caused by subdomains. For the current study, we considered the results of an exploratory factor analysis (EFA), first-order unidimensional CFA, and a bifactor analysis. The factor analyses were conducted using Mplus software.[21] For the EFA, we used unweighted least squares estimation. For the one-factor and bifactor CFAs, we used weighted least squares with mean and variance adjustment. Because of the categorical nature of the response data, a polychoric correlation matrix was analyzed. Fit was evaluated based on the Comparative Fit Index (CFI),[10] the Tucker–Lewis Index,[22] and the Root Mean Square Error of Approximation (RMSEA).[23,24] To assess local independence, we examined the magnitude of residual correlations.[25,26] The residuals represent the variance not accounted for by the model and, if local independence holds, they should not be substantially correlated.

Development of short forms

The items of the MFIS were divided into 2-, 3-, 4-, and 5-item short forms using two item selection strategies: content balancing and random selection. This resulted in an 8-cell design (4 sizes of short forms × 2 item selection strategies). The items of the MFIS were used to create 10 different short forms within each study cell. Thus, a total of 80 short forms were generated (8 cells × 10 replications).

Within each selection strategy, the 2-item short forms were comprised of unique items; that is, no item appeared in more than one short form. There were not enough items to build wholly unique short forms for the 3- to 5- item conditions, but each short form was comprised of a unique grouping of items. Within each 10 short form study condition, an effort was made to balance the number of short forms for which any given item was selected.

Content balanced short forms

As already noted, the MFIS can be scored as a single scale or as three subscales to measure the impact of cognitive, physical, and psychosocial fatigue. We conducted our own content review of the items and elected, for content balancing purposes, to reclassify one of the physical fatigue items. It was our judgment that item 21, "needed to rest more often or for longer periods", could indicate cognitive as well as physical fatigue. We defined an "other" category that included the two psychosocial items and item 21. Adding this item made content balancing somewhat easier because there were more items from which to choose in populating this subdomain.

Random item short forms

For the second item selection condition, items were randomly assigned to short forms. When this selection resulted in a duplicate short form within a study cell, a new item subset was generated at random so that, within each study cell, there were no short forms with the exact same items.

Calibration of item responses

Responses to the 21 items of the MFIS were calibrated using Master's Partial Credit Model (PCM)[27] and Parscale Software.[28] The PCM is appropriate for calibrating responses to items that offer three or more response options (eg, never/sometimes/always). Scores are obtained based on a derived probability function that models how persons with different levels of the outcome being measured (fatigue in the current study) are likely to respond to items. Fit to the PCM was evaluated using the computer macro, IRTFIT.[29] We report both S-X^2 and S-G^2 fit statistics ($P < 0.01$).[30,31]

A total of 81 scores were generated for each individual in the validation sample. One set of scores was estimated based on responses to all 21 MFIS items (full scale scores). The other 80 were based on responses to each of the 80 short forms.

Analyses

Persons' full scale scores served as a standard by which the short forms were evaluated. Pearson product-moment correlation coefficients were calculated between short form and full scale scores, and within each study cell, the range and average of correlation coefficients were calculated (using Fisher Z transformation).[32] In addition, the root mean squared errors (RMSE) were calculated. The "errors" were defined for the purposes of this study as short form score minus full scale score.

To evaluate the errors associated with short form scores relative to the error associated with full scale scores, we derived confidence intervals for persons full scale scores. These were defined as calibrated full scale scores ± two standard errors. The standard errors were the individual standard errors obtained for each person in the PCM calibration of all 21MFIS items. We calculated the percentage of short form scores in each study condition that were within this range.

Results

A total of 374 persons responded to all items of the MFIS. Data from these respondents were used for the current study. The study population was largely female (79.4%) and overwhelmingly Caucasian (93%), with an average age of 49 years (range of 21–78). Participants reported their course of disease based on a self-report item that displays five figures in which severity of symptoms is plotted against time.[33] Each figure represents a different pattern of symptom severity over time, and respondents are asked to indicate the one that "best describes the course of your MS over time". Of those who responded to this item, 55% selected a plot consistent with "relapsing remitting", 27% with "secondary progressive", and 18% with "primary progressive".

Tests of unidimensionality assumption

The fit statistics for a first-order unidimensional CFA model yielded mixed results. The CFI[10] and TLI[22] were 0.891 and 0.942 respectively, suggesting moderate model fit.[10] The RMSEA, however, was very high (0.331), indicating very poor fit. Half of the residual pairs had correlations >0.10 (n = 106); more than half of these were >0.20 (n = 56).

An EFA was conducted. The ratio of the first and second eigenvalue was 3.95, with the first factor accounting for 60.0% of the variance. The correlation between the first and second factor was 0.57. Item factor loadings obtained in an EFA are provided in Table 1. Loadings from both a one- and two-factor solution are included. The loadings supported a two-factor solution in which cognitive items loaded on the first factor, and all other items loaded on the second factor. The loadings of items categorized as psychosocial in the original categorization and as "other" in our reclassification loaded with the physical items on the second factor.

A bifactor model was fitted in which all items loaded on a general factor, cognitive items loaded on one orthogonal group factor and all other items on a second orthogonal group factor. The fit of this model was substantially better than that of the first-order, one-factor model (CFI = 0.961, TLI = 0.993, RMSEA = 0.105), and only three residual correlations had values greater than absolute value of 0.10. The general factor accounted for 67% of the common variance. The physical and "other" items accounted for 29% of the common variance, and the cognitive factor accounted for only 7%. Because the cognitive specific factor accounted for such a small proportion of variance, we concluded that the data were sufficiently unidimensional for calibration with the partial credit model.[27] Though the items we designated as "other" did not define a separate group factor, for content balancing purposes only, we retained the category.

Fit of items to the PCM

Three of the 21 items (15%) failed to fit the PCM at alpha = 0.01. The items that failed at this criterion were items 16, 18, and 19 (item content reported in Table 1).

Correlations among short form and full scale scores

Pearson product-moment correlations were computed between short form and full scale scores and these were compared across study conditions (short form size and item selection strategy). The results for the ten replications per study cell were summarized by calculating the range of correlation values and the average correlation. Average correlations were calculated by transforming r-values into corresponding z-scores, finding the mean of those scores, and then transforming this value back to an r-value.[32] Figure 1 is a box plot displaying the 10th, 25th, 50th, 75th, and 90th percentiles for each study condition. As the plot indicates results were substantially better for the short forms created based on content-balancing. Even for short forms comprised of only two items, correlations ranged from 0.83 to 0.90 (mean = 0.87). With the 5-item short forms correlations ranging from 0.94 to 0.96 (mean = 0.95). Though the correlations for the short forms based on random selection of items fared relatively well with means of 0.81, 0.89, 0.91, and 0.93 for short forms with 2-, 3-, 4-, and 5-items, respectively, the results were inferior to those obtained with the content-balanced short forms and far more variable.

As expected, short forms with more items performed better than those with fewer items. For the short form sizes evaluated in this study, there was little "leveling-off" of the advantage gained by having more items. The 5-item short forms performed better than the 4-item short forms; 4-item short forms performed better than the 3-item short forms, and so on.

Root mean squared errors (RMSE)

We made the assumption that trait-level estimates based on all 21 items of the MFIS would be superior to estimates based

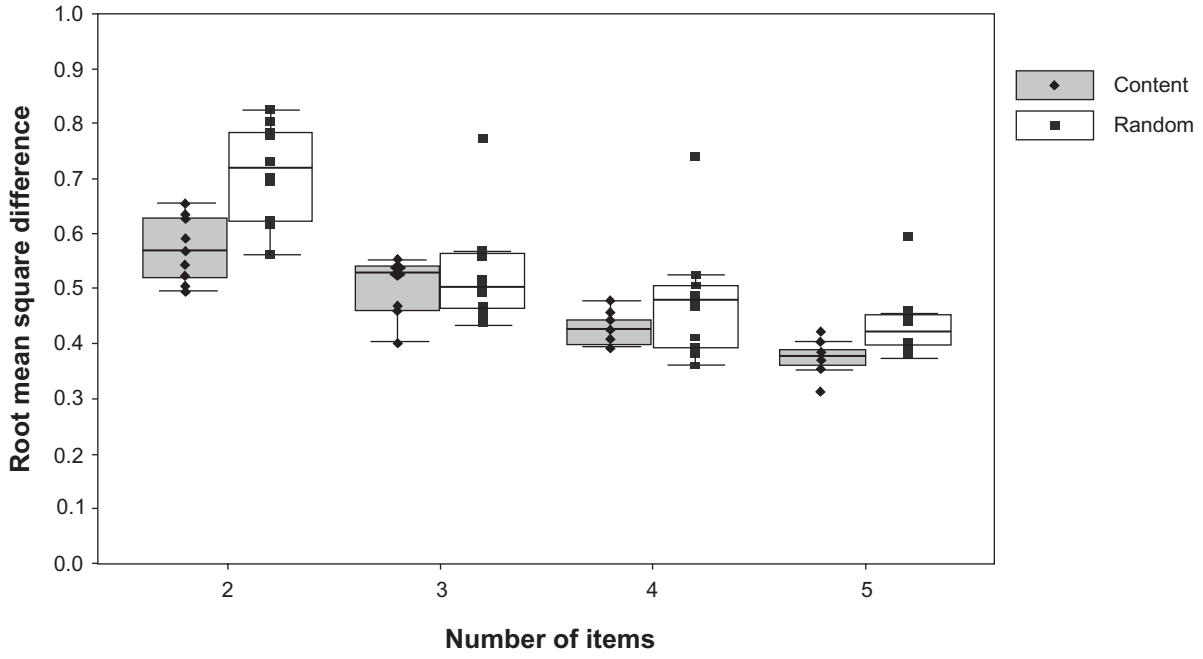

Figure 1 Correlation between short form scores and full scale scores by item selection strategy and number of items.

on fewer items. For this study, therefore, "error" was defined as short form score minus full-scale score. Figure 2 compares the RMSE values calculated based on this definition of error. RMSEs are in the metric of the scale, and their magnitude can be interpreted relative to the range of theta estimates (7.7 logits in the current study). The pattern of RMSE results mirrored the correlation results. Increasing the number of items reduced the observed error as did developing short forms based on content-balancing. The 5-item content-balanced short forms performed particularly well in approximating full-scale scores. The RMSE for these short forms was 0.43, which is 5.6% of total score range.

Though we used the full-scale score as our gold standard, this estimate also has an error associated with it. The IRT

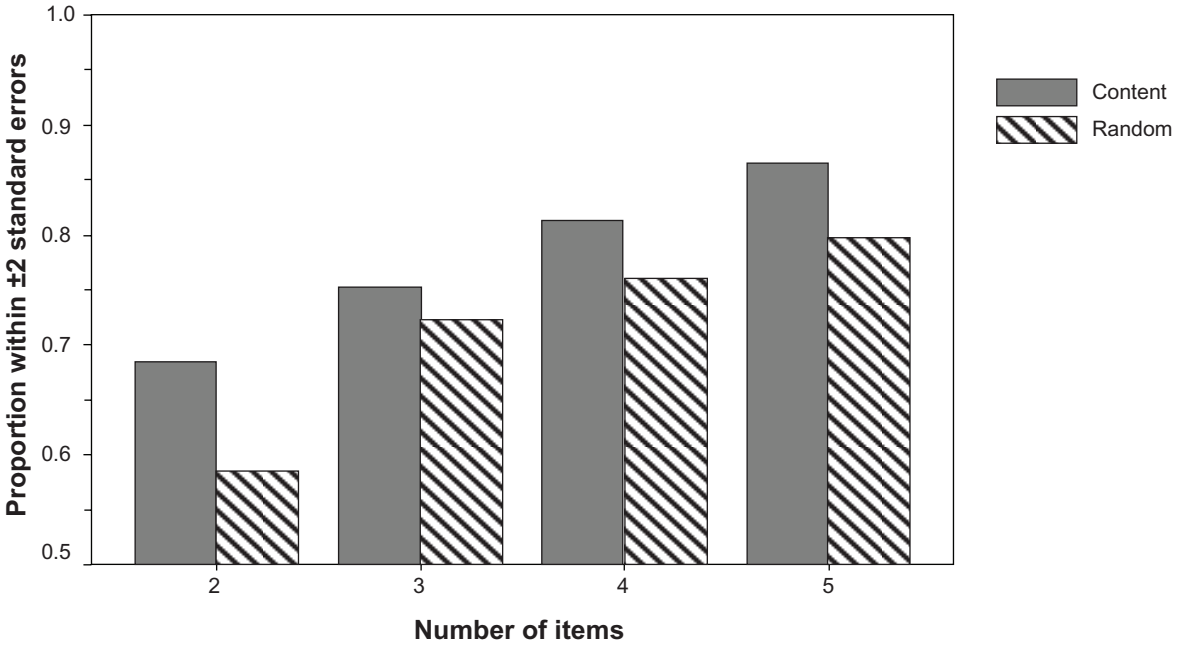

Figure 2 Root mean square errors (short form score minus full scale score) by item selection strategy and number of items.

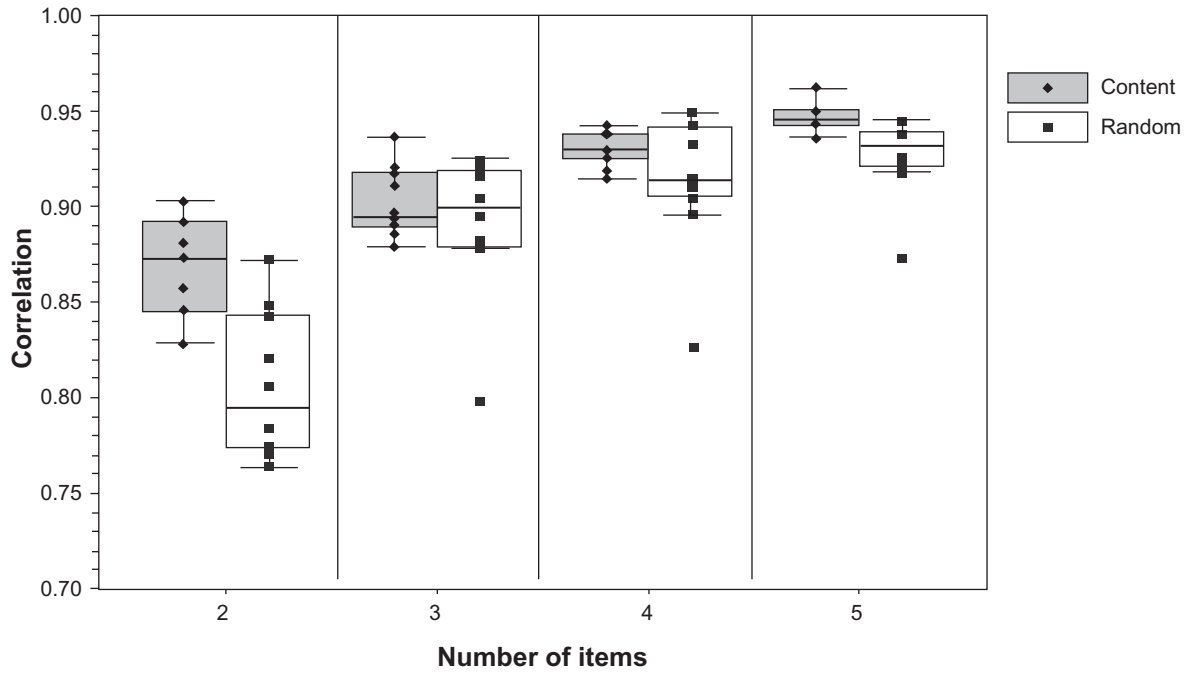

Figure 3 Proportion of scores that differ from full scale score by less than 2 standard errors (based on partial credit model calibration).

calibration outputs a standard error of estimate (SEM) for every person. These vary by trait level. We computed the 95% confidence interval (± 2 SEMs) around each respondent's full-scale trait level estimate and then calculated the proportion of short form scores from each condition that fell within this range. Figure 3 shows the results. Like the previous comparisons, these analyses show the superiority of the content-balanced short forms and the increase in precision gained by adding more items. For example, for the content-balanced short forms the proportion that fell within the ± 2 SEM confidence interval ranged from 0.68 for the 2-item short forms to 0.87 for the 5-item short forms. Of the scores based on short forms comprised of randomly selected items, the proportions falling within the ± 2 SEM confidence interval were 0.58 and 0.80, respectively, for the 2- and 5-item short forms.

Conclusion

We found a clear advantage for using a content-balancing strategy over random selection of items in developing short forms. We did not investigate the impact of difficulty-balancing because of the limits of our item pool. In the current study, short forms developed to be content-balanced proved to be balanced with respect to item difficulty as well. Content- and difficulty-balancing should be compared with a larger item pool to evaluate whether one approach is superior to the other.

Despite the limitations of our study, the results warrant several conclusions. The PCM proved an effective model for developing multiple subscales calibrated to a common mathematical metric, and even very brief subscales produced reasonable approximations of full scale scores, particularly when the subscales were developed to represent the subdomains of the measured construct. Increases in number of items per subscale yielded the expected increases in precision. Future research should further investigate the impact of item content and item parameters in the development of short forms from a parent item bank.

Disclosure

The authors declare no conflicts of interest.

References

1. Higginson IJ, Hart S, Silber E, Burman R, Edmonds P. Symptom prevalence and severity in people severely affected by multiple sclerosis. *J Palliat Care*. 2006;22(3):158–165.
2. van de Port IG, Kwakkel G, Schepers VP, Heinemans CT, Lindeman E. Is Fatigue an Independent Factor Associated with Activities of Daily Living, Instrumental Activities of Daily Living and Health-Related Quality of Life in Chronic Stroke? *Cerebrovasc Dis*. 2006 12;23(1):40–45.
3. Miaskowski C, Cooper B, Paul S, et al. Subgroups of Patients with Cancer with Different Symptom Experiences and Quality-of-Life Outcomes: A Cluster Analysisr. *Oncol Nurs Forum*. 2006;33(5):E79–E89.
4. Yagiz On A, Oncu J, Atamaz F, Durmaz B. Impact of post-polio-related fatigue on quality of life. *J Rehabil Med*. 2006;38(5):329–332.
5. Pollard L, Choy EH, Scott DL. The consequences of rheumatoid arthritis: quality of life measures in the individual patient. *Clin Exp Rheumatol*. 2005;23(5 Suppl 39):S43–S52.
6. Martinez-Martin P, Catalan MJ, Benito-Leon J, et al. Impact of fatigue in Parkinson's disease: the Fatigue Impact Scale for Daily Use (D-FIS). *Qual Life Res*. 2006;15(4):597–606.
7. Edwards P, Roberts I, Clarke M, et al. Methods to increase response rates to postal questionnaires. *Cochrane Database Syst Rev*. 2007(2):MR000008.

8. Crocker L, Algina J. *Introduction to Classical and Modern Test Theory*. Fort Worth: Harcourt Brace Jovanovich College Publishers; 1986.
9. Fisk JD, Ritvo PG, Ross L, Haase DA, Marrie TJ, Schlech WF. Measuring the functional impact of fatigue: initial validation of the fatigue impact scale. *Clin Infect Dis*. 1994;18 Suppl 1:S79–S83.
10. Bentler P. Comparative fit indices in structural models. *Psychol Bull*. 1990; 107:238–246.
11. Reise S, Widaman K, Pugh R. Confirmatory factor analysis and item response theory: two approaches for exploring measurement invariance. *Psychol Bulle*. 1993;114(3):552.
12. Cella D, Yount S, Rothrock N, et al. The Patient-Reported Outcomes Measurement Information System (PROMIS): progress of an NIH Roadmap cooperative group during its first two years. *Med Care*. 2007;45(5 Suppl 1):S3–S11.
13. McDonald R. The dimensionality of test and items. *Br J Math Stat Psychol*. 1981;34:100–117.
14. Reise SP, Haviland MG. Item response theory and the measurement of clinical change. *J Pers Assess*. 2005;84(3):228–238.
15. Reise SP, Waller NG, Comrey AL. Factor analysis and scale revision. *Psychol Assess*. 2000;12:287–297.
16. Bjorner JB, Kosinski M, Ware JE Jr. Calibration of an item pool for assessing the burden of headaches: an application of item response theory to the headache impact test (HIT). *Qual Life Res*. 2003;12(8): 913–933.
17. Bjorner JB, Kosinski M, Ware JE Jr. Using item response theory to calibrate the Headache Impact Test (HIT) to the metric of traditional headache scales. *Qual Life Res*. 2003;12(8):981–1002.
18. Lai JS, Cella D, Chang CH, Bode RK, Heinemann AW. Item banking to improve, shorten and computerize self-reported fatigue: an illustration of steps to create a core item bank from the FACIT-Fatigue Scale. *Qual Life Res*. 2003 Aug;12(5):485–501.
19. Lai JS, Crane PK, Cella D. Factor analysis techniques for assessing sufficient unidimensionality of cancer related fatigue. *Qual Life Res*. 2006;15(7):1179–1190.
20. Stockdale GD, Gridley BE, Balogh DW, Holtgraves T. Confirmatory factor analysis of single- and multiple-factor competing models of the dissociative experiences scale in a nonclinical sample. *Assessment*. 2002;9(1):94–106.
21. *Mplus User's Guide* [computer program]. Version 2. Los Angeles, CA: Muthen and Muthen; 2001.
22. Tucker L, Lewis C. A reliability coefficient for maximum likelihood factor analysis. *Psychometrika*. 1973;38:1–10.
23. Browne MW, Cudeck R. Alternative ways of assessing model fit. In: Bollen KA, Long JS, eds. *Testing Structural Equation Models*. Newbury Park, CA: Sage Publications; 1993;136–172.
24. Steiger MJ, Quinn NP, Marsden CD. The clinical use of apomorphine in Parkinson's disease. *J Neurol*. 1992;239(7):389.
25. Kline RB. *Principles and Practice of Structural Equation Modeling*. New York, NY: The Guilford Press; 1998.
26. McDonald RP. *Test Theory: A Unified Treatment*. Mahway, NJ: Lawrence Earlbaum; 1999.
27. Masters GN. A Rasch model for partial credit scoring. *Psychometrika*. 1982;47(2):149–174.
28. *PARSCALE 3: IRT Based Test Scoring and Item Analysis for Graded Ttems and Rating Scales* [computer program]. Version. Chicago, IL: Scientific Software International, Inc.; 1997.
29. *IRTFIT: A Macro for Item Fit and Local Dependence Tests under IRT Models* [computer program]. Version. Lincoln, RI: QualityMetric Incorporated; 2006.
30. Orlando M, Thissen D. Likelihood-Based Item-Fit Indices for Dichotomous Item Response Theory Models. *Appl Psychol Meas*. 2000;24(1):50–64.
31. Orlando M, Thissen D. Further investigation of the performance of S-X2: An item fit index for use with dichotomous item response theory models. *Appl Psychol Meas*. 2003;27:289–298.
32. Aroian L. A study of Fisher's RA. z distribution and the related F distribution. *Ann Math Statist*. 1941;12:429–448.
33. Bamer AM, Cetin K, Amtmann D, Bowen JD, Johnson KL. Comparing a self report questionnaire with physician assessment for determining multiple sclerosis clinical disease course: a validation study. *Mult Scler*. 2007;13(8):1033–1037.