

Testing measurement invariance of the patient-reported outcomes measurement information system pain behaviors score between the US general population sample and a sample of individuals with chronic pain

Hyewon Chung · Jiseon Kim · Karon F. Cook · Robert L. Askew · Dennis A. Revicki · Dagmar Amtmann

Accepted: 20 June 2013 / Published online: 4 July 2013
© Springer Science+Business Media Dordrecht 2013

Abstract

Purpose In order to test the difference between group means, the construct measured must have the same meaning for all groups under investigation. This study examined the measurement invariance of responses to the patient-reported outcomes measurement information system (PROMIS) pain behavior (PB) item bank in two samples: the PROMIS calibration sample (Wave 1, $N = 426$) and a sample recruited from the American Chronic Pain Association (ACPA, $N = 750$). The ACPA data were collected to increase the number of participants with higher levels of pain.

Methods Multi-group confirmatory factor analysis (MG-CFA) and two item response theory (IRT)-based differential item functioning (DIF) approaches were employed to evaluate the existence of measurement invariance.

Results MG-CFA results supported metric invariance of the PROMIS–PB, indicating unstandardized factor loadings with equal across samples. DIF analyses revealed that impact of 6 DIF items was negligible.

Conclusions Based on the results of both MG-CFA and IRT-based DIF approaches, we recommend retaining the original parameter estimates obtained from the combined samples based on the results of MG-CFA.

Keywords Multi-group confirmatory factor analysis · Differential item functioning · Item response theory · Patient outcome measures · Pain measurement · Psychometrics

H. Chung (✉)
Department of Education, Chungnam National University, 99
Daehak-ro, Yuseon-gu, Daejeon 305-764, Korea
e-mail: hyewonchung7@gmail.com

J. Kim
Department of Rehabilitation Medicine, University of
Washington, Box 359612, Seattle, WA 98195, USA

K. F. Cook
Department of Medical Social Sciences, Northwestern
University Feinberg School of Medicine, 625 N. Michigan Ave.,
Suite 2700, Chicago, IL 60611, USA

R. L. Askew
Department of Rehabilitation Medicine, University of
Washington, Box 356590, Seattle, WA 98195, USA

D. A. Revicki
Outcomes Research, United BioSource Corporation, 7101
Wisconsin Ave, Suite 600, Bethesda, MD 20814, USA

D. Amtmann
Department of Rehabilitation Medicine, University of
Washington, Box 354237, Seattle, WA 98195, USA

Abbreviations

ACPA American Chronic Pain Association
CFA Confirmatory factor analysis
DIF Differential item functioning
IRT Item response theory
MG-CFA Multi-group confirmatory factor analysis
PB Pain behavior
PROMIS Patient-reported outcomes measurement information system

Introduction

Pain behaviors (PBs) are behaviors that communicate to others that a person is experiencing pain [1–3]. PB is an important outcome in studies of persons living with chronic

pain [4, 5] because PBs may predict development of disability [6].

The National Institutes of Health's (NIH) patient-reported outcomes measurement information system (PROMIS) included PBs among its targeted outcomes [7]. All PROMIS measures were developed as item banks, and candidate items were administered to a large sample of predominantly healthy community participants (Wave 1). Few individuals from Wave 1 reported higher levels of pain, requiring additional data collection. Participants were recruited from the American Chronic Pain Association (ACPA) with higher levels of pain and completed an online survey that included the PROMIS–PB items. The data from Wave 1 and from the ACPA were combined for the purpose of calibrating the items. However, calibration of the PB items was conducted in the combined sample without an investigation of measurement invariance.

Measurement invariance means that the same construct is measured similarly across groups. For instance, cancer pain typically has unique emotional components not necessarily found in other types of chronic pain, and this emotional component of the pain might influence several dimensions of PBs. Researchers may be concerned that test score differences observed in various subgroups are due to measurement instrument problems rather than true differences in the trait being measured. Lack of measurement invariance has been mainly investigated using two methods: multi-group confirmatory factor analysis (MG-CFA) and item response theory (IRT).

MG-CFA procedures are commonly employed to test for measurement equivalence [8–12]. The main question underlying tests of measurement equivalence across groups is whether certain factor analytic parameters such as loadings, intercepts, error variances, factor variances, factor covariances, and factor means can be assumed equivalent across groups [10, 12, 13].

In the IRT framework, when a lack of measurement invariance occurs at item level, it is referred to as differential item functioning (DIF). DIF is defined as “a difference in the probability of endorsing an item across comparison groups when the scores are on a common metric” [14]. Several researchers have investigated similarities and differences of the two models in detecting a lack of measurement invariance [14–16]. Stark et al. [14] reported both CFA and IRT methods showed similar results in detecting DIF across a majority of simulated conditions. The authors found that the CFA approach performed slightly worse than the IRT approach in dichotomous data; however, it performed better under condition of polytomous data with a small sample size. The authors also pointed out that testing measurement invariance via the IRT approach seemed more complicated than the CFA approach. In the current study, we explored measurement invariance across Wave 1 and ACPA samples

with both MG-CFA and IRT-based DIF approaches. Evidence of measurement invariance provides support for using the PROMIS–PB score to compare observed differences in group means for both healthy and clinical samples. The data for the study were collected in the process of instrument development, and the study design is described in detail in Cella et al. [7]. The purpose of the current study was to investigate the level of measurement invariance of the PROMIS–PB across a sample of individuals from the general population who are generally healthy and a sample of individuals with chronic pain.

Methods

Participants

The PROMIS Wave 1 data included 21,133 research participants. Of these, 19,601 were recruited from an internet panel (YouGovPolimetrix; www.polimetrix.com), and 1,532 were recruited from primary research sites associated with the PROMIS network. A detailed description on Wave 1 data collection is available at <http://www.nihpromis.org/science/> calibration testing. For purposes of this study, only the data from participants who responded to the full bank and had no missing data were used.

As described above, the sample size for Wave 1 was quite large; however, few individuals reported higher levels of pain. With IRT models, a sufficient number of responses in every response category are essential for precise estimates of item parameters [17]. Thus, research participants with chronic pain were recruited through the ACPA. Eligibility requirements included being 21 years of age or older and having one or more chronic pain conditions for at least 3 months prior to the survey.

Analyses

Three levels of measurement invariance were tested using the MG-CFA approach. The first and weakest level, *configural invariance* [18], assumes that the same pattern of item-factor loadings exists across groups being compared; the same items must have nonzero loadings on the same factors. *Metric invariance* [19] requires, additionally, that unstandardized factor loadings be invariant across the comparison groups. *Scalar invariance* is the strongest level of invariance [18, 20] and requires that all the assumptions of configural and metric invariance be met. In addition, the scale's item intercepts be invariant across groups.

Mplus software 6.1 [21] was used with weighted least-squares mean and variance adjusted (WLSMV) estimation. Several fit indices were used in the current study: χ^2 , comparative fit index (CFI) [22], Tucker–Lewis index

(TLI) [23], and root mean square error of approximation (RMSEA) [24, 25]. CFI and TLI values above 0.90 are considered acceptable [13, 26], and RMSEA values of <0.08 are considered to indicate adequate fit [27].

In the MG-CFA approach, fit of a baseline model is compared to the fit of increasingly constrained models. The χ^2 difference test is utilized to compare the fit of two nested models [28–30]. A nonsignificant χ^2 difference supports the less parameterized model (i.e., the addition of the extra parameters does not significantly improve model fit). To account for the sensitivity of the χ^2 difference test to sample size, α -level of 0.05 for χ^2 difference test was used. Additionally, a difference of <0.01 in the Δ CFI index was used to support the less parameterized model [9, 10]. Note that the model fit was compared only when both models of interest individually fit the data.

Additionally, DIF was analyzed with the R software package Lordif [31]. The Lordif utilizes an ordinal logistic regression framework, and the graded response (GR) model is used for IRT trait estimation [32]. Two criteria were considered to detect meaningful DIF in the current study: (1) <0.13 pseudo R^2 statistic [33] and (2) 10 % changes in beta [31, 34, 35].

Following Cook et al. [36] approach, the impact of DIF on the scores was assessed; a Pearson correlation between DIF-adjusted person scores and the original person scores was calculated to examine the existence of meaningful impact of DIF on the scores. A strong magnitude of correlation would suggest that adjusting for DIF would make a negligible difference in the person scores. This indicates that item parameters calculated when combining all groups together could be used without concern for substantial impact of DIF on person's scoring.

Items

The PROMIS–PB item bank provided good coverage of the PB construct [37]. A census-weighted subsample of the PROMIS Wave 1 data was used to anchor the PROMIS scores on a T -score metric ($M = 50$; $SD = 10$) [38]. The PROMIS–PB items have a seven-day time frame and are rated on a six-point scale that ranges from 1 = *had no pain* to 6 = *always*. Because of low frequencies of responses, categories 1 and 2 (never) were subsequently combined.

Results

Initial analyses

Initial analyses were conducted using data from all 36 items administered to combined PROMIS and ACPA

samples. The initial model, however, had poor fit: χ^2 (594, $N = 1,176$) = 8,397.010, $p < .01$, CFI = 0.894, TLI = 0.888, RMSEA = 0.106 (from 0.104 to 0.108). We investigated potential local dependency among items because it can cause biased parameter estimates. To identify the potential local dependency and to modify model specifications, residual correlations and modification indices were inspected. Any items with absolute values of residual correlations >0.20 indicate local dependency [39]. Based on the results, nine items were eliminated due to the potential local dependency: PB2, “When I was in pain I became irritable”; PB9, “When I was in pain I became angry”; PB16, “When I was in pain I appeared upset or sad”; PB23, “When I was in pain I asked one or more people to leave me alone”; PB24, “When I was in pain I moved stiffly”; PB29, “When I was in pain I used a cane or something else for support”; PB31, “I limped because of pain”; PB43, “When I was in pain I walked carefully”; and PB53, “When I was in pain I moved my arms or legs stiffly.” A schematic flow of the item analysis used in the present study is illustrated in Fig. 1.

Descriptive analysis

A total of 426 PROMIS Wave 1 (Male = 192, Female = 234) and 750 ACPA participants (Male = 136, Female = 610, missing = 4) participants were included in the current study. Table 1 describes demographic and clinical details of the samples. The PROMIS Wave I and ACPA samples were statistically different on age, t (1,172) = 4.990, $p < .001$, gender, χ^2 (1, $N = 1,172$) = 96.922, $p < .001$, ethnicity, χ^2 (1, $N = 1,170$) = 50.485, $p < .001$, marriage status, χ^2 (2, $N = 1,119$) = 7.137, $p < .001$, and education χ^2 (4, $N = 1,174$) = 30.957, $p < .001$.

MG-CFA approach

Configural invariance

A configural invariance model (i.e., the same pattern of item-factor loadings across groups) was tested across the comparison groups. The findings supported configural invariance between the PROMIS and ACPA samples: χ^2 (648, $N = 1,176$) = 3,453.968, $p < .01$, CFI = 0.904, TLI = 0.896, RMSEA = 0.086 (from 0.083 to 0.089) (Table 2).

Metric invariance

A metric invariance model (i.e., equal constraints on unstandardized item-factor loadings across groups) also supported good fit: χ^2 (675, $N = 1,176$) = 3,486.512, $p < .01$, CFI = 0.904, TLI = 0.900, RMSEA = 0.084

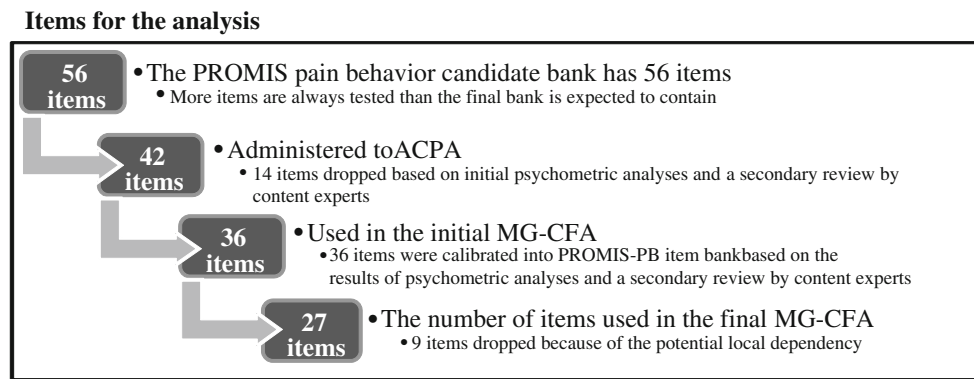


Fig. 1 A schematic flow of item analysis

Table 1 Demographics between the PROMIS Wave 1 sample and the ACPA sample for pain behavior

	PROMIS Wave 1 (N, %)		ACPA (N, %)	
Gender				
Male	192	45.07	136	18.13
Female	234	54.93	610	81.33
Missing	–	–	4	0.54
Ethnicity				
White	339	79.58	695	92.67
Nonwhite	87	20.42	49	6.53
Missing	–	–	6	0.80
Marriage status				
Never-married	58	13.62	65	8.67
Married/living with partner in committed relationship	296	69.48	479	63.87
Separated/divorced/widowed	72	16.90	149	19.86
Missing	–	–	57	7.60
Education				
Less than high school grad	8	1.88	13	1.73
High school grad/GED	63	14.79	118	15.73
Some college/technical degree/AA	144	33.80	362	48.27
College degree (BA/BS)	128	30.05	166	22.13
Advanced degree (MA, PHD, MD)	83	19.48	89	11.87
Missing	–	–	2	0.27

PROMIS patient-reported outcomes measurement information system, ACPA American Chronic Pain Association

(from 0.081 to 0.087). Next, the model fit of the configural and metric invariance models was compared. The Chi-square difference test results were statistically significant: $\Delta\chi^2 (\Delta df = 27) = 428.170$, $p < .01$, indicating that some unstandardized factor loading values were statistically different between PROMIS Wave I and ACPA samples. Since the χ^2 difference is sensitive to relatively larger sample sizes, CFI difference test (ΔCFI) is frequently used in testing measurement invariance [9, 10]. Dissimilar to the

χ^2 difference, a decrease of <0.01 in the CFI value ($\Delta CFI = 0.00$) was found in the nested model comparison, supporting the same unstandardized factor loading values between PROMIS and ACPA samples.

Scalar invariance

After finding support for both configural and metric invariance, the authors examined the PROMIS–PB for scalar invariance (i.e., invariance of the unstandardized item thresholds across groups). The results did not support scalar invariance: $\chi^2 (771, N = 1,176) = 9,085.440$, $p < .01$, CFI = 0.716, TLI = 0.742, RMSEA = 0.135 (from 0.133 to 0.138).

IRT-based DIF approach

The criterion of pseudo R^2 (i.e., classifying pseudo $R^2 < 0.13$ as negligible DIF) resulted in no items being detected as DIF. Using the DIF criterion of 10 % beta change, seven items were identified as having meaningful DIF. The correlation between the original and adjusted scores was 0.98, indicating no concern for substantial impact of DIF on person's scoring when combining all groups together.

Discussion

The current study examined the measurement invariance of PB items using MG-CFA across two samples to evaluate whether the construct of PBs is the same in healthy people and those with chronic pain. The PROMIS Wave 1 community sample was comprised predominantly of healthy participants, and the ACPA sample was comprised exclusively of individuals living with chronic pain. There is still little consensus in the literature in regard to the level of equivalence necessary for inferring measurement invariance across groups. Horn and McArdle required metric invariance to sure that the same constructs are measures

Table 2 Results of testing measurement invariance of the pain behavior items across PROMIS Wave 1 and ACPA using MG-CFA

Measurement invariance models	Overall fit indices						Comparative fit indices				
	χ^2	<i>df</i>	CFI	TLI	RMSEA	RMSEA 90 % CI	Model comparison	$\Delta \chi^2$	Δdf	<i>p</i>	Δ ACFI
1. Configural	3,453.968	648	0.904	0.896	0.086	0.083–0.089	1 versus 2	428.170	27	<.01	Δ 0.00
2. Metric	3,486.512	675	0.904	0.900	0.084	0.081–0.087	n.a.	n.a.	n.a.	n.a.	n.a.
3. Scalar	9,085.440	771	0.716	0.742	0.135	0.133–0.138	n.a.	n.a.	n.a.	n.a.	n.a.

Measurement invariance tests included 27 items after deleting PB2, PB9, PB16, PB23, PB24, PB29, PB31, PB43, and PB53 items from the initial 36 items
CFI comparative fit index, *TLI* Tucker–Lewis index, *RMSEA* root mean square error of approximation, *CI* confidence interval, *n.a.* not applicable

across groups [19]. Chen, Sousa, and West argued that comparing means across groups could be meaningful after confirming the existence of scalar invariance [40]. Reise, Widaman, and Pugh, however, claimed that a form of partial loading invariance is actually required to permit across-group comparisons [16]. The findings of the current study supported measurement invariance at the level of metric invariance, but not at the level of scalar invariance.

Conclusions and recommendations

Had the PROMIS-PI failed to support either configural or metric invariance, we might need to consider a remedy such as re-calibrating the item bank or removing items that function differently in the two compared groups. The results from this study found that a subset of 27 PROMIS–PB items met all but the strictest form of measurement invariance. Based on IRT-based DIF analysis results, it was concluded that although statistically significant DIF was identified using 10 % beta change, the adjustments for DIF would result in negligible changes in person scores since correlations between adjusted and nonadjusted scores were approximately 0.98. For this reason, it was concluded that any DIF in this item set among the MS and APCA groups could be disregarded. This implies that the instrument measures the same construct in both healthy and clinical including those with chronic pain. Based on the findings of the current study, we conclude that using the originally obtained parameter estimates from the combined sample of PROMIS Wave I and ACPA participants are acceptable, and the instrument can be scored and used as originally published.

The current study could use only 27 of the 36 items in the PROMIS–PB item bank mainly due to local dependence. Local dependence may cause biased parameter estimates [41, 42], and thus, we recommend that the PROMIS–PB address the local dependence in the item bank or utilize testlets to handle local dependence among the items [42]. In summary, the results of the current study support the use of PROMIS–PB item parameters obtained from the combined general population and chronic pain

sample. The construct of PBs appears to function in the same way in a community sample as well as in people living with chronic pain. As a result, the PROMIS–PB score can be used to compare mean differences between groups.

Acknowledgments The project described was supported by Award Number 3U01AR052177-06S1 from the National Institute of Arthritis and Musculoskeletal and Skin Diseases. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute of Arthritis and Musculoskeletal and Skin Diseases or the National Institutes of Health.

References

1. Fordyce, W. E. (1976). *Behavioral methods for chronic pain and illness*. St. Louis, MO: C. V. Mosby.
2. Keefe, F. J., Williams, D. A., & Smith, S. J. (2001). Assessment of pain behaviors. In D. C. Turk & R. Melzack (Eds.), *Handbook of pain assessment* (pp. 170–187). New York, NY: Guilford Press.
3. Waters, S. J., Dixon, K. E., Keefe, F. J., Ayers, S., Baum, A., McManus, C., et al. (2007). *Cambridge handbook of psychology, health and medicine* (2nd ed., pp. 300–303). Cambridge UK: Cambridge University Press.
4. Hadjistavropoulos, T., Herr, K., Turk, D. C., Fine, P. G., Dworkin, R., Helme, R., et al. (2007). An interdisciplinary expert consensus statement on assessment of pain in older persons. *The Clinical Journal of Pain*, 23(1 Suppl), S1–S43.
5. Turk, D. C., Dworkin, R. H., Revicki, D. A., Harding, G., Burke, L. B., Cella, D., et al. (2008). Identifying important outcome domains for chronic pain clinical trials: An IMMPACT survey of people with pain. *Pain*, 137(2), 276–285.
6. Jensen, M. P. (1997). Validity of self-report and observational measures. In T. S. Jensen & J. A. Turner (Eds.), *Proceedings of the 8th world congress on pain: Progress in pain research and management*. Seattle, WA: IASP Press.
7. Cella, D., Riley, W., Stone, A., Rothrock, N., Reeve, B., Yount, S., et al. (2010). Initial item banks and first wave testing of the patient-reported outcomes measurement information system (PROMIS) network: 2005–2008. *Journal of Clinical Epidemiology*, 63(11), 1179–1194.
8. Cheung, G. W., & Rensvold, R. B. (1999). Testing factorial invariance across groups: A reconceptualization and proposed new method. *Journal of Management*, 25(1), 1–27.
9. Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indices for testing measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, 9(2), 233–255.

10. French, B. F., & Finch, W. H. (2006). Confirmatory factor analytic procedures for the determination of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, 13(3), 378–402.
11. King, W. C., & Miles, E. W. (1995). A quasi-experimental assessment of the effect of computerizing noncognitive paper-and-pencil measurements: A test of measurement equivalence. *Journal of Applied Psychology*, 80(6), 643–651.
12. Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, 3(1), 4–70.
13. Kline, R. B. (2010). *Principles and practice of structural equation modeling* (3rd ed.). New York, NY: The Guilford Press.
14. Stark, S., Chernshenko, O. S., & Drasgow, F. (2006). Detecting differential item functioning with confirmatory factor analysis and item response theory: Toward a unified strategy. *Journal of Applied Psychology*, 91(6), 1292–1306.
15. Meade, A. W., & Lautenschlager, G. J. (2004). A comparison of item response theory and confirmatory factor analytic methodologies for establishing measurement equivalence/invariance. *Organizational Research Methods*, 7(4), 361–388.
16. Reise, S. P., Widaman, K. F., & Pugh, R. H. (1993). Confirmatory factor analysis and item response theory: Two approaches for exploring measurement. *Psychological Bulletin*, 114(3), 552–566.
17. Choi, S. W., Cook, K. F., & Dodd, B. G. (1997). Parameter recovery for the partial credit model using MULTILOG. *Journal of Outcome Measurement*, 1(2), 114–142.
18. Meredith, W. (1993). Measurement invariance, factor analysis, and factorial invariance. *Psychometrika*, 58(4), 525–543.
19. Horn, J. L., & McArdle, J. J. (1992). A practical and theoretical guide to measurement invariance in aging research. *Experimental Aging Research*, 18(3), 117–144.
20. Steenkamp, E. M. J., & Baumgartner, H. (1998). Assessing measurement invariance in cross-national consumer research. *Journal of Consumer Research*, 25(1), 78–90.
21. Muthén, L. K., & Muthén, B. O. (1998–2010). *Mplus user's guide* (6th ed.). Los Angeles, CA: Muthén & Muthén.
22. Bentler, P. M. (1980). Multivariate analysis with latent variables: Causal modeling. *Annual Review of Psychology*, 31(1), 419–456.
23. Tucker, L. R., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika*, 38(1), 1–10.
24. Byrne, B. M. (1998). *Structural equation modeling with LISREL, PRELIS, and SIMPLIS*. Hillsdale, NJ: Lawrence Erlbaum.
25. Steiger, J. H., & Lind, J. C. (1980). Statistically-based tests for the number of common factors. In *Paper presented at the annual spring meeting of the Psychometric Society*, Iowa City, IA.
26. Woehr, D. J., Arciniega, L. M., & Lim, D. H. (2007). Examining work ethic across populations. A comparison of the multidimensional work ethic profile across three diverse cultures. *Educational and Psychological Measurement*, 67(1), 154–168.
27. Browne, M., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. Bollen & J. Long (Eds.), *Testing structural equation models* (pp. 136–162). London, England: Sage.
28. Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, 6(1), 1–55.
29. Marsh, H. W., Hau, K., & Wen, Z. (2004). In search of golden rules: Comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing Hu and Bentler's (1999) findings. *Structural Equation Modeling: A Multidisciplinary Journal*, 11(3), 320–341.
30. Sivo, S. A., Fan, X., Witta, E. L., & Willse, J. T. (2006). The search for “optimal” cutoff properties: Fit index criteria in structural equation modeling. *The Journal of Experimental Education*, 74(3), 267–288.
31. Choi, S. W., Gibbons, L., & Crane, P. K. (2011). Lordif: An R package for detecting differential item functioning using iterative hybrid ordinal logistic regression/item response theory and Monte Carlo simulations. *Journal of Statistical Software*, 39(8), 1–30.
32. Crane, P. K., Gibbons, L. E., Jolley, L., & van Belle, G. (2006). Differential item functioning analysis with ordinal logistic regression techniques: DIFdetect and difwithpar. *Medical Care*, 44(Suppl 3), S115–S123.
33. Zumbo, B. D. (1999). *A handbook on the theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary framework for binary and Likert type (ordinal) item scores*. Ottawa, ON: Directorate of Human Resources Research and Evaluation Department of National Defense.
34. Crane, P. K., Van Belle, G., & Larson, E. B. (2004). Test bias in a cognitive test: Differential item functioning in the CASI. *Statistics in Medicine*, 23(2), 241–256.
35. Crane, P. K., Gibbons, L., Ocepek-Weiklson, K., Cook, K., Cella, D., Narasimhalu, K., et al. (2007). A comparison of three sets of criteria for determining the presence of differential item functioning using ordinal logistic regression. *Quality of Life Research*, 16(Suppl 1), 69–84.
36. Cook, K. F., Bombardier, C. H., Bamer, A. M., Choi, S. W., Kroenke, K., & Fann, J. R. (2011). Do somatic and cognitive symptoms of traumatic brain injury confound depression screening?. *Archives of Physical Medicine and Rehabilitation*, 92, 818–823.
37. Revicki, D. A., Chen, W., Harnam, N., Cook, K. F., Amtmann, D., Callahan, L. F., et al. (2009). Development and psychometric analysis of the PROMIS pain behavior item bank. *Pain*, 146(1–2), 158–169.
38. Liu, H. H., Cella, D., Gershon, R., Shen, J., Morales, L. S., Riley, W., et al. (2010). Representativeness of the PROMIS internet panel. *Journal of Clinical Epidemiology*, 63(11), 1169–1178.
39. Cook, K. F., Teal, C. R., Bjorner, J. B., Cella, D., Chang, C. H., Crane, P. K., et al. (2007). IRT health outcomes data analysis project: An overview and summary. *Quality of Life Research*, 16(Suppl 1), 121–132.
40. Chen, F. F., Sousa, K. H., & West, S. G. (2005). Testing measurement invariance of second-order factor models. *Structural Equation Modeling: A Multidisciplinary Journal*, 12(3), 471–492.
41. Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement*, 30(3), 187–213.
42. Steinberg, L., & Thissen, D. (1996). Uses of item response theory and the testlet concept in the measurement of psychopathology. *Psychological Methods*, 1(1), 81–97.