Short communication

Successful use of the Unified Dyskinesia Rating Scale regardless of PD- or dyskinesia-duration

Xuehan Ren[a,1], Jeffrey Lin[a,1], Sheng Luo[b,*], Christopher G. Goetz[c], Glenn T. Stebbins[c], Esther Cubo[d]

[a] Department of Biostatistics and Data Science, School of Public Health, University of Texas Health Science Center at Houston, Houston, TX, USA
[b] Department of Biostatistics and Bioinformatics, Duke University Medical Center, Durham, NC, USA
[c] Department of Neurological Sciences, Rush University Medical Center, Chicago, IL, USA
[d] Neurology Department, Hospital Universitario Burgos, Avda Islas Baleares 3, 09006, Burgos, Spain

ARTICLE INFO

ABSTRACT

Objective: We assessed differential item functioning (DIF) in the Unified Dyskinesia Rating Scale (UDysRS) to evaluate bias risk from the duration of Parkinson's Disease (PD) and duration of dyskinesia.
Background: Assessing DIF is a core validation step for rating scales. If DIF is present for an item, interpretation must consider influences from the tested covariates. DIF can be uniform or non-uniform, depending on the consistency of influence from the given covariate across all levels of dyskinesia.
Methods: Using a large UDysRS database (N = 2313), uniform and non-uniform DIF related to the duration of PD and/duration of dyskinesia were tested. Unidimensionality of UDysRS was first confirmed using confirmatory factor analysis. DIF analysis was conducted using two independent latent models. DIF in an item was confirmed if both methods independently identified DIF at a significance level using Bonferroni correction. McFadden pseudo R^2 measured clinical relevancy of DIF magnitude (negligible, moderate, and large) for items identified with DIF, and items with DIF were considered clinically relevant if they exceeded a negligible designation.
Results: Most items did not show uniform or non-uniform DIF based on PD and dyskinesia duration in isolation or in combination. For all items where DIF was identified, the magnitude statistic was in the negligible range (McFadden pseudo $R^2 < 0.035$) and the combined impact of multiple identified DIF items on UDysRS likewise did not exceed the negligible designation.
Conclusion: The absence of clinically relevant DIF suggests that the UDysRS can be applied across all patients regardless of their PD- or dyskinesia-duration.

1. Introduction

The Unified Dyskinesia Rating Scale (UDysRS) is a comprehensive rating scale for dyskinesia in Parkinson's disease (PD) [1]. The scale was developed in English and translated to other languages through a standardized program [2]. Assessing a rating scale for differential item functioning (DIF) is a core validation step in determining if any item scores are substantially biased by covariates [3]. Bias introduced by DIF occurs when the probability of a score is affected by a covariate, such as age or gender, in individual scores for respondents with similar severity. Previous work has been done to examine DIF in the UDysRS due to the specific covariates of age, race/ethnicity, education, and gender [4]. In this analysis, we continue to examine DIF to see if the duration of

PD, the duration of dyskinesia, or their interaction contribute to any bias in item scores. The importance of these covariates in relation to DIF is due to the nonlinear course of many motor and non-motor features [5,6] across duration of disease. By examining the independent and interactive effects of these covariates on DIF in the UDysRS, the differential probability of an item score can be examined among patients with similar levels of dyskinesia. Two types of DIF can occur: uniform or non-uniform. Uniform DIF (U-DIF) occurs when covariates have consistent levels of influence on item-score across all levels of dyskinesia. Non-uniform DIF (NU-DIF) occurs when covariates have varying levels of influence on item-score across different levels of dyskinesia. Both U-DIF and NU-DIF assessments were conducted on all 26 items in four parts of the UDysRS. Absence of duration based DIF in UDysRS

---

* Corresponding author.
E-mail address: sheng.luo@duke.edu (S. Luo).
[1] Denotes equal contributions to first authorship.

items would allow the scale to be applied across patients without consideration of how long patients have had PD or dyskinesia.

2. Methods

2.1. The UDysRS dataset

The cross-sectional combined International Parkinson and Movement Disorder Society's translation UDysRS dataset consisting of completed scores from 12 languages (N = 3057) was used. After removing subjects with missing PD and dyskinesia durations the sample size was N = 2313 (Traditional Chinese [N = 34], French [N = 243], German [N = 126], Greek [N = 257], Hungarian [N = 256], Italian [N = 0], Japanese [N = 181], Korean [N = 229], Russian [N = 247], Slovak [N = 246], Spanish [N = 244], Turkish [N = 250]).

2.2. Test for unidimensionality

DIF analysis assumes unidimensionality for the test, that is, the items being examined measure a single pertinent trait. Unidimensionality of the UDysRS was tested by conducting confirmatory factor analysis and requiring the Confirmatory Fit Index (CFI) > 0.90 with Root Mean Square Error Approximation (RMSEA) < 0.10 [7].

2.3. Sample sizes for each analysis

DIF analysis also requires that for each item, all possible rating values must have some representation. However, for many items, there were no patients who scored in the most severe category (option 4). Therefore, options 3 and option 4 were combined into a single category termed 3/4. In addition, each category (0, 1, 2, 3/4) was required to have at least five samples for each UDysRS item.

2.4. DIF determinations

The DIF analysis was conducted using two independent latent variable models: the iterative hybrid ordinal logistic regression/item response theory (graded response model) approach implemented in the R package *lordif* [8] and the multiple indicators multiple causes (MIMIC) model [9]. An item qualifies for DIF designation if both methods identify DIF at a significance level determined by a Bonferroni correction for multiple comparisons [10]. All items were first tested for NU-DIF. If NU-DIF was not present in the item, then it was analyzed for U-DIF [10]. For items identified to have DIF, the McFadden Pseudo R^2 statistic from the *lordif* R package was used to determine the clinical relevancy (DIF magnitude). The recommended cutoffs were set as < 0.035 = negligible, 0.035–0.07 = moderate, > 0.07 = large [11], with an item considered to be clinically relevant if it exceeds the negligible rating. Lastly, the combined impact (scale level impact) of multiple items identified with DIF was examined using the differential test function (DTF) index to compare the test characteristic curves of items with and without DIF [12]. The magnitude of the DTF was assessed using a conservative threshold derived from Monte Carlo simulations (DTF cutoff value = 1.404) [13]. Interaction terms between the durations of PD and dyskinesia were then tested on items with significant DIF using the same procedure.

2.5. Comparisons

For both the duration of PD and the duration of dyskinesia, the sample was divided into 5 groups (< 5 years, 5–9, 10–14, 15–19, ≥20). This resulted in at least 200 cases in each group for the duration of PD and at least 40 cases in each group for the duration of dyskinesia. The *lordif* model is able to accommodate multinomial options. However, because the MIMIC model is restricted to binary comparisons, the comparisons were first conducted using *lordif* and if overall DIF was identified, follow-up pairwise comparisons were conducted in both *lordif* and MIMIC independently.

3. Results

3.1. Unidimensionality

The confirmatory factor analysis confirmed the unidimensionality assumption in the UDysRS. The scale has a confirmatory fit index CFI = 0.97 > 0.90 and a root mean square approximation RMSEA = 0.09 < 0.10 [7].

3.2. DIF for PD duration (Table 1)

NU-DIF for PD duration was identified in items 1 and 24. U-DIF was identified in items 9 and 11. In all cases, the DIF was of negligible magnitude. An overall scale-level impact due to the combined effects of multiple "negligible" impacts was not discovered using the DTF index score (< 5 vs all others = 1.2246; 5–9 vs all others = 0.0165; 10–14 vs all others = 0.0976; 15–19 vs all others = 0.0883; ≥20 vs all others = 0.0985; DTF cutoff value = 1.404).

3.3. DIF for dyskinesia duration (Table 1)

No items were detected to have NU-DIF for dyskinesia duration. Item 20 was detected to have U-DIF, however, the effects were negligible. DTF did not detect an overall scale impact for any of the groups (< 5 vs all others = 0.0892; 5–9 vs all others = 0.3057; 10–14 vs all others = 0.4385; 15–19 vs all others = 0.8752; ≥20 vs all others – item 15 had negative slope parameters; DTF cutoff value = 1.404).

3.4. DIF for interactions between PD and dyskinesia duration (Table 2)

Item 1 was detected to have NU-DIF with negligible effects for subjects with PD duration < 5 years and dyskinesia duration < 5 years, as well as for subjects with PD duration 10–14 years and dyskinesia duration 5–9 years. Item 24 was detected to have NU-DIF with negligible effects for subjects with PD duration < 5 years and dyskinesia duration < 5 years. Item 11 was detected to have uniform DIF with negligible effects for subjects with PD duration < 5 years and dyskinesia duration < 5 years.

Table 1
PD and Dyskinesia Duration – based DIF.

Item	MIMIC P-value	Lordif P value	R-square	Magnitude
PD duration-based nonuniform DIF				
Historical disability time spent with on-dyskinesia (Item 1)				
< 5 vs all others	< 0.00005	< 0.00005	0.0035	Negligible
Objective disability drinking (Item 24)				
< 5 vs all others	< 0.00005	0.0001	0.0031	Negligible
PD duration-based uniform DIF				
Historical disability walking/balance (Item 9)				
≥20 vs all others	< 0.00005	< 0.00005	0.0029	Negligible
Objective disability drinking (Item 11)				
< 5 vs all others	< 0.00005	0.0002	0.0021	Negligible
Dyskinesia duration-based nonuniform DIF				
None				
Dyskinesia duration-based uniform DIF				
Objective disability trunk (Item 20)				
10-14 vs all others	< 0.00005	0.0002	0.0023	Negligible

Legend: Most items did not meet the criteria for establishing DIF, while items with significant DIF had negligible magnitude. P-values for the two independent methods of establishing DIF (*lordif*, MIMIC) are shown as well as the McFadden's R^2 value to show the impact of the DIF.

Table 2
DIF for interaction terms.

Item	MIMIC P-value	Lordif P value	R-square	Magnitude
PD duration X Dyskinesia duration-based nonuniform DIF				
Historical disability time spent with on-dyskinesia (Item 1)				
PD duration (< 5) X Dyskinesia duration (< 5) vs all others	< 0.00005	< 0.00005	0.0032	Negligible
PD duration (10–14) X Dyskinesia duration (5–9) vs all others	< 0.00005	< 0.00005	0.0029	Negligible
Objective disability drinking (Item 24)				
PD duration (< 5) X Dyskinesia duration (< 5) vs all others	< 0.00005	< 0.00005	0.0029	Negligible
PD duration X Dyskinesia duration-based uniform DIF				
Objective disability drinking (Item 11)				
PD duration (< 5) X Dyskinesia duration (< 5) vs all others	0.0001	< 0.00005	0.0024	Negligible

Legend: Interactions between significant items were explored. Most interaction terms did not have significant DIF. P-values for the two independent methods of establishing DIF (*lordif*, MIMIC) are shown as well as the McFadden's R^2 value to show the impact of the DIF.

4. Discussion

DIF, often known as measurement bias [10,12,13], is essential to test in order to fully validate a rating scale. A lack of DIF allows for the confident conclusion that the rating scale is indeed measuring the desired trait, which in the case of UDysRS is dyskinesia severity. Failure to detect DIF of moderate or large magnitude in our analysis suggests that the UDysRS is not strongly influenced by the duration of PD or the duration of dyskinesia. As such, in conjunction with the prior documentation of DIF independence, we consider the scale to be an effective rating scale for measuring dyskinesia severity. This conclusion is reinforced by the fact that we were unable to detect a scale level impact even when considering the combined effects of multiple "negligible" DIF items.

Although the sample size is very large, not many item scores from the tested sample fell into the severe impairment and disability rating categories (option 4). This distribution of case severities may have been primarily related to the assessment methodology of the original datasets that drew almost exclusively from outpatient and ambulatory clinics, where patients with severe conditions were likely underrepresented. Therefore, categories 3 and 4 had to be merged into a single category. However, we note that this strategy may not achieve a full DIF analysis as it deviates from the original structure of the UDysRS. Future analyses may also look at DIF from covariates such as the source of information (patient, caregiver, combination of patient/caregiver). As more and more languages are assembled from the translation program, other possible analyses could focus on DIF related to culture or geography, for example, separating European-based and Pan-American based Spanish or linguistic differences where Latin-based languages or Germanic based languages could be considered as covariates. Moreover, missing data across languages were excluded in the study. We have identified 10 items (items 1, 9, 14, 16, 18–22, 26) with statistically significant different response patterns from the subjects included vs. excluded in the study, due to the large study sample size. This can be a potential of bias.

UDysRS scores are associated with the durations of PD and dyskinesia [14]. In this study, however, we determined that UDysRS scores do not differ significantly based on the durations of PD or dyskinesia when conditioned on patients with similar levels of dyskinesia. Furthermore, we did not detect DIF in the interactions between PD duration and dyskinesia duration. Therefore, the UDysRS is highly specific to measuring dyskinesia severity with negligible influence from the PD and dyskinesia durations despite their association.

Author roles

X. Ren: Statistical analysis – conduct and review. **J. Lin:** Manuscript preparation – writing of first draft, review and critique. **S. Luo:** Statistical analysis – conduct and review. **CG. Goetz:** Research project – conception, organization, execution. **GT. Stebbins:** Research project – conception, organization, execution. **E. Cubo**: Research project – conception, organization, execution. All authors have approved the final article.

Ethical compliance statement

This study has been approved by Duke University Medical Center (DUMC) Institutional Review Board (IRB). Informed patient consent was not necessary for this work. We confirm that we have read the Journal's position on issues involved in ethical publication and affirm that this work is consistent with those guidelines.

Author disclosures related to this research

The authors received compensation from the Movement Disorder Society for the management of this program.

Conflicts of interest

The authors declared that there is no conflicts of interest in preparing this manuscript.

Acknowledgements

References

[1] C.G. Goetz, J.G. Nutt, G.T. Stebbins, The unified dyskinesia rating scale: presentation and clinimetric profile, Mov. Disord. 23 (16) (2008) 2398–2403.
[2] C.G. Goetz, G.T. Stebbins, L. Wang, N.R. LaPelle, S. Luo, B.C. Tilley, IPMDS-sponsored scale translation program: process, format, and clinimetric testing plan for the MDS-UPDRS and UDysRS, Mov. Disord. Clin. Pract. 1 (2) (2014) 97–101.
[3] R.K. Hambleton, Good practices for identifying differential item functioning, Med. Care 44 (11) (2006) S182–S188.
[4] S. Luo, Y. Liu, J.A. Teresi, G.T. Stebbins, C.G. Goetz, Differential item functioning in the unified dyskinesia rating scale (UDysRS), Mov. Disord. 32 (8) (2017) 1244–1249.
[5] C.V. Thuy, J.G. Nutt, N.H.G. Holford, Progression of motor and nonmotor features of Parkinson's disease and their response to treatment, Br. J. Clin. Pharmacol. 74 (2012) 267–283.
[6] C.G. Goetz, G.T. Stebbins, B. Ouyang, Visual plus nonvisual hallucinations in Parkinson's disease: development and evolution over 10 years, Mov. Disord. 26 (2011) 2196–2200.
[7] T. Brown, Confirmatory Factor Analysis for Applied Research, Guilford Press, New York, 2006.
[8] S.W. Choi, L.E. Gibbons, P.K. Crane, Lordif: an R package for detecting differential item functioning using iterative hybrid ordinal logistic regression/item response theory and Monte Carlo simulations, J. Stat. Softw. 39 (8) (2011) 1.
[9] K.G. Joreskog, A.S. Goldberger, Estimation of a model with multiple indicators and multiple causes of a single latent variable, JASA 70 (351a) (1975) 631–639.
[10] J.A. Teresi, Different approaches to differential item functioning in health applications: advantages, disadvantages and some neglected topics, Med. Care 44 (11) (2006) S152–S170.

[11] M.G. Jodoin, M.J. Gierl, Evaluating type I error and power rates using an effect size measure with the logistic regression procedure for DIF detection, Appl. Meas. Educ. 14 (4) (2001) 329–349.

[12] N.S. Roju, W.J. Van der Linden, P.F. Fleer, IRT-based internal measures of differential functioning of items and tests, Appl. Psychol. Meas. 19 (4) (1995) 353–368.

[13] M. Kleinman, J.A. Teresi, Differential item functioning magnitude and impact measures from item response theory models, Psychol. Test Assess. Model. 58 (1) (2016) 79–98.

[14] E. Cubo, S. Luo, S. Calvo, P. Martinez-Martín, Differences in the unified dyskinesia rating scale based on age, Parkinson's disease and dyskinesias duration, Mov. Disord. 33 (2018).