Differential Item Functioning in the Unified Dyskinesia Rating Scale (UDysRS)

Sheng Luo, PhD,[1]* Yuanyuan Liu, MS,[1]
Jeanne A. Teresi, EdD, PhD,[2,3] Glenn T. Stebbins, PhD,[4]
and Christopher G. Goetz, MD[4]

[1]*Department of Biostatistics, University of Texas Health Science Center, School of Public Health, Houston, Texas, USA* [2]*Columbia University Stroud Center at New York State Psychiatric Institute, New York, NY, USA* [3]*Research Division, Hebrew Home at Riverdale, Bronx, New York, USA* [4]*Department of Neurological Sciences, Rush University Medical Center, Chicago, Illinois, USA*

ABSTRACT

Objective: Test if differential item functioning due to gender, age, race/ethnicity, or education impacts Unified Dyskinesia Rating Scale scores.

Background: Testing rating scales for differential item functioning is a core validation step. If differential item functioning exists, interpretation of item scores must consider secondary influences on dyskinesia ratings.

Methods: Using Unified Dyskinesia Rating Scale translation databases (N = 3,132), we tested uniform and nonuniform differential item functioning. We required confirmation by two independent methods and considered differential item functioning pertinent if McFadden pseudo R^2magnitude statistics exceeded negligible ratings.

Results: No age, race/ethnicity, or education nonuniform differential item functioning was identified. Gender nonuniform differential item functioning occurred for 2 items, both with negligible magnitude. Gender, race, and education uniform differential item functioning was observed for multiple items, all with negligible magnitude.

Conclusions: The Unified Dyskinesia Rating Scale items effectively capture dyskinesia severity without pertinent gender, age, race/ ethnicity, or education influence. © 2017 International Parkinson and Movement Disorder Society

Key Words: Parkinson's disease; dyskinesia; rating scales; clinimetrics; differential item functioning

*Corresponding author: Dr. Sheng Luo, The University of Texas Health Science Center at Houston, 1200 Herman Pressler Drive, Room E815, Houston, TX 77030; sheng.t.luo@uth.tmc.edu

Relevant conflicts of interests/financial disclosures: This research was supported by the International Parkinson and Movement Disorder Society and National Institute of Neurological Disorders and Stroke Grants 5U01NS043127 and R01NS091307. The effort was also supported by the Parkinson's Disease Foundation (PDF) as part of the PDF Rush Research Center of Excellence.

Received: 19 February 2017; **Revised:** 13 April 2017; **Accepted:** 4 May 2017

Published online 00 Month 2017 in Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/mds.27058

The Unified Dyskinesia Rating Scale (UDysRS) was developed as a comprehensive rating tool of dyskinesia in Parkinson's disease (PD).[1] The scale was developed in English with a clinimetric program to provide validated non-English translations.[2,3] Testing a rating scale for differential item functioning (DIF)[4] is a core step to determine if covariates (eg, age, gender) substantially bias any item score. Among people with similar severity levels of dyskinesia and the same probability of responding, DIF occurs for the UDysRS if the probability of an item score differs according to selected covariates. For example, gender-based DIF exists for item 4.1 (communication disability related to dyskinesia) if men and women with the same severity level of dyskinesia responded differently on this item. Two kinds of DIF can occur. In nonuniform DIF (NU-DIF), covariate influences on item scores vary across levels of the dyskinesia trait, whereas in uniform DIF (U-DIF), influences on item scores by the covariate are constant across all trait levels (Figure S1 of Supplementary Material).[5]

We conducted both U-DIF and NU-DIF assessments on UDysRS items on the gender, age, race/ethnicity, and education level.[6] The absence of clinically relevant NU-DIF or U-DIF allows it to be used confidently as a true measure of dyskinesia.

Methods

The UDysRS Dataset

We accessed the cross-sectional combined translation dataset of fully completed UDysRS scores from 13 languages (Chinese, n = 250; English, n = 70; French, n = 250; German, n = 284; Greek, n = 260; Hungarian, n = 256; Italian, n = 252; Japanese, n = 250; Korean, n = 250; Portuguese, n = 256; Slovak, n = 251; Spanish, n = 253; Turkish, n = 250).[3]

Assessing Unidimensionality of the UDysRS

DIF analyses are anchored in the unidimensionality assumption, that is, the items measure a single pertinent trait. To test the unidimensionality of the UDysRS, we conducted confirmatory factor analysis, requiring that the confirmatory fit index was ≥ 0.90 with root means square error of approximation < 0.10.[7]

Sample Sizes for Each Analysis

DIF analyses require that for each item, all possible rating values must have some representation. Because there were no patients scoring in the most severe rating option (4) in many UDysRS items, we combined scores of 3 and 4 as a collapsed designation, termed 3/4. Furthermore, we required at least 5 samples in each of the 0, 1, 2, and 3/4 categories for each UDysRS item.

DIF Determinations

We conducted DIF analysis using 2 independent latent variable models, the iterative hybrid ordinal logistic regression/item response theory (graded response model)[8] approach as realized in the *R* package *lordif*[9] and the multiple indicators multiple causes (MIMIC) model.[10,11] For an item to qualify for DIF designation, we required that both methods independently identify DIF at a significance level corrected for multiple comparisons using a Bonferroni correction.[12]

All items were studied first for NU-DIF, and those without NU-DIF were then analyzed for U-DIF.[12] For items identified with DIF, to determine clinical pertinence (DIF magnitude), we used the McFadden pseudo R^2 magnitude estimate from the *R* package *lordif* and applied the recommended cut-offs of < 0.035 = negligible, 0.035-0.07 = moderate, > 0.07 = large.[13] We considered an item with DIF to be clinically relevant if it exceeded negligible rating. Finally, we examined the combined impact (scale level impact) of multiple identified items with DIF on the UDysRS using the differential test function (DTF) index that compared the test characteristic curves with and without DIF items.[14] The magnitude of the DTF [14] was assessed by a conservative threshold based on Monte Carlo simulations[15,16] (cutoff DTF value = 1.404).

Comparisons

For gender, the analyses compared males and females. For the age-based DIF analyses, we chose 3 age groups (27-51, 52-75, and 76-93) to result in at least 280 cases in each age group. We chose this age divisions to reflect our age ranges (27-93), and they are similar to other reports examining age divisions in PD.[17,18] Based on years of education, we divided the sample into three groups (<7, 7-12, > 12), which resulted in 680 cases in each education group.[19] We chose race/ethnicity categories according to published divisions adopted by the U.S. Office of Management and Budget.[6] Possible categories were White (non-Hispanic), Hispanic, African descent, Asian, Pacific Islander, Native or Endogenous, and other. Whereas the *lordif* model can accommodate multinomial options, MIMIC is restricted to binary comparisons. Therefore, we first conducted comparisons using *lordif*, and, if overall DIF was identified with this strategy, follow-up pairwise comparisons were conducted in *lordif* and MIMIC independently.

Results

Sample Sizes

The full dataset included UDysRS scores for 3,132 patients, but missing data on isolated items or demographic information reduced the samples. In all assessments, however, the sample exceeded 2,500 UDysRS complete scores (Table S1 of Supplementary Material).

Unidimensionality

The confirmatory factor analysis of the full dataset confirmed unidimensionality of the UDysRS. The scale met the criteria of a confirmatory fit index ≥ 0.90 and a root means square error of approximation < 0.10, allowing conduct of the DIF analyses[7] (confirmatory fit index = 0.97, root means square error of approximation = 0.08).

Gender-Based DIF (Upper Part of Table 1)

NU-DIF for gender was identified in 1 historical disability item (speech) and 1 objective disability item (communication). U-DIF for gender was identified in 1 historical disability item (time with dyskinesia). In all cases, the magnitude of the DIF was "negligible." In assessing the combined effects of multiple "negligible" impacts, we did not detect an overall scale-level impact on UDysRS from gender-based DIF using the DTF index score (DTF = 0.0214). (Supplementary Material provides all results for identified DIF.)

Education-Based DIF (Lower Part of Table 1)

None of the items exhibited NU-DIF for education. Education-based U-DIF was found for historical disability ratings for time spent with on-dyskinesia, chewing/swallowing, eating tasks, dressing, and hygiene, although in all cases the magnitude of the DIF was "negligible." We did not detect an overall scale-level impact on UDysRS from education-based DIF using the DTF index score (<7 vs all others = 0.4285; 7-12 vs all others = 0.0144; > 12 vs all others = 1.1297). (Supplementary Material provides all results for identified DIF.)

TABLE 1. Gender- and education-based statistically significant DIF

Item	MIMIC *P* values	*lordif P* values	R^2	Magnitude
Gender-based nonuniform DIF				
Historical disability speech	<.0005	.0001	0.0019	Negligible
Objective disability communication	<.0005	<.00005	0.0025	Negligible
Gender-based uniform DIF				
Historical disability time spent with on-dyskinesia	<.0005	.0005	0.0015	Negligible
Education-based nonuniform DIF				
None	NA	NA	NA	NA
Education-based uniform DIF				
Historical disability time spent with on-dyskinesia				
<7 vs all others	<.0005	<.00005	0.0030	Negligible
Historical disability chewing/swallowing				
>12 vs all others	<.0005	<.00005	0.0035	Negligible
Historical disability eating tasks				
>12 vs all others	<.0005	.00001	0.0021	Negligible
Historical disability dressing				
>12 vs all others	<.0005	<.00005	0.0031	Negligible
Historical disability hygiene				
<7 vs all others	<.0005	<.00005	0.0057	Negligible
>12 vs all others	<.0005	<.00005	0.0079	Negligible

Most of the UDysRS items did not meet the minimal statistical criteria for DIF (see text). The Table lists items with DIF identified by both *lordif* and *MIMIC* as independent approaches (*P* values shown) with McFadden's R^2 (R^2) indicating the impact of the DIF. DIF, differential item functioning; MIMIC, multiple indicators multiple causes; NA, not applicable.

Age-Based DIF

For age-based DIF, none of the Items was identified as having NU-DIF or U-DIF.

Race/Ethnicity-Based DIF (Table 2)

The racial/ethnic groups under consideration were White non-Hispanic, Hispanic, and Asian. We did not have a sufficiently large score representation from other groups. For race/ethnicity-based DIF, none of the items was identified as having NU-DIF. A total of 14 items exhibited race/ethnicity-based U-DIF for White versus other (historical disability ratings for exciting or emotional settings, effects of pain from off-dystonia and dystonia pain; objective impairment ratings for face and right leg/hip; and objective disability ratings for drinking and ambulation), Asian versus other (historical disability ratings for exciting and emotional settings, time off dystonia, effects of off-dystonia separate from pain and effects of pain from off-dystonia and dystonia pain; objective impairment ratings for face, right arm/shoulder, left arm/shoulder, right leg/hip, and left leg/hip; and objective disability ratings of drinking), and Hispanic versus other (historical disability ratings for eating tasks and public/social settings and objective disability ratings for ambulation). In all cases, the impact of U-DIF was negligible. We did not detect an overall scale-level impact on UDysRS using the DTF index score when comparing White versus non-White (DTF = 0.2036) and Hispanic versus non-Hispanic (DTF = 1.0501). The DTF simulation-based threshold was exceeded for Asian versus non-Asian (DTF = 5.0038). (Supplementary Material provides all results for identified DIF.)

Discussion

DIF, often termed *measurement bias*,[12,14-16,20] is essential to test for a full validation of a rating scale and the confident conclusion that the scale is truly measuring the conceptual trait, in this case, dyskinesia severity. The fact that we did not detect DIF of moderate or large magnitude for any item relative to any of the studied demographic elements strongly argues that the UDysRS is effectively capturing dyskinesia severity and is not strongly influenced by gender, age, race/ethnicity, or education. The conclusion is reinforced by our inability to detect a significant combined scale-level impact when multiple "negligible" DIF items occur in the scale. The DTF value above threshold observed for the Asian subsample indicated a small level of impact as evidenced in the graphs of the test characteristic curves for Asians and non-Asians. Although the level of aggregate impact was not sufficient to warrant concern, it is recommended that this finding be investigated further with other datasets.

There are two major differences between this study and our prior MDS-UPDRS DIF analysis.[21] First, because of the unidimensionality of the UDysRS, we could justify performing DIF using the total UDysRS score as the index of dyskinesia severity. In the MDS-UPDRS, because the scale is unidimensional for each part, but not as a total score, our approach necessitated DIF analysis for each part. Second, although the MDS-UPDRS items were not assessed for education-based DIF due to the lack of education information, we can add to our conclusions that the UDysRS scale item performance is not influenced by education level. We acknowledge that educational systems differ by

TABLE 2. Race/ethnicity-based statistically significant DIF

Item	MIMIC *P* values	*lordif P* values	R^2	Magnitude
Race/ethnicity-based nonuniform DIF				
None	NA	NA	NA	NA
Race/ethnicity-based uniform DIF				
Historical disability - eating tasks				
Hispanic vs all others	<.0005	<.00005	0.0023	Negligible
Historical disability public/social settings				
Hispanic vs all others	<.0005	.0003	0.0015	Negligible
Historical disability exciting or emotional settings				
White vs all others	<.0005	<.00005	0.0047	Negligible
Asian vs all others	<.0005	<.00005	0.0072	Negligible
Historical disability time with off-dystonia				
Asian vs. all others	<.0005	<.00005	0.0023	Negligible
Historical disability effects of off-dystonia separate from pain				
Asian vs all others	<.0005	<.00005	0.0028	Negligible
Historical disability effects of pain from off-dystonia				
White vs all others	<.0005	<.00005	0.0018	Negligible
Asian vs all others	<.0005	<.00005	0.0099	Negligible
Historical disability dystonia pain				
White vs all others	<.0005	<.00005	0.0027	Negligible
Asian vs all others	<.0005	<.00005	0.0149	Negligible
Objective impairment face				
White vs all others	<.0005	<.00005	0.0078	Negligible
Asian vs all others	<.0005	<.00005	0.0085	Negligible
Objective Impairment Right Arm/Shoulder				
Asian vs all others	<.0005	<.00005	0.0027	Negligible
Objective impairment left arm/shoulder				
Asian vs all others	<.0005	<.00005	0.0052	Negligible
Objective impairment right leg/hip				
White vs all others	<.0005	<.00005	0.0031	Negligible
Asian vs all others	<.0005	<.00005	0.0119	Negligible
Objective impairment left leg/hip				
Asian vs all others	<.0005	<.00005	0.0091	Negligible
Objective disability drinking				
White vs all others	<.0005	<.00005	0.0044	Negligible
Asian vs all others	<.0005	<.00005	0.0046	Negligible
Objective disability ambulation				
White vs all others	<.0005	.0001	0.0020	Negligible
Hispanic vs all others	<.0005	<.00005	0.0037	Negligible

Most of the UDysRS items did not meet the minimal statistical criteria for DIF (see text). The table lists items with DIF identified by both *lordif* and *MIMIC* as independent approaches (*P* values shown) with McFadden's R^2 (R^2) indicating the impact of the DIF. DIF, differential item functioning.

culture, so the interpretation of DIF absence based on education is limited to conclusions regarding number of years of formal education and not knowledge base.

Although the sample sizes were very large, we were limited by the paucity of item scores in the severe impairment and disability category (4) because all assessments were acquired in outpatient settings where the most severe patients are rarely seen. Hence, we collapsed 3 and 4 categories into a single designation, which may not achieve DIF analysis of the UDysRS as constructed. Moreover, DIF may exist from other covariates such as source of information for parts 1 and 2 (patient, caregiver, or combined patient/caregiver) and rater- or site-based DIF. Our current dataset precluded such additional DIF analysis.

The strengths of our study include the large dataset with worldwide representation across cultures using 1 validated scale. We have been rigorous in our clinimetric approach, requiring that designated items with DIF be identified by 2 independent statistical methods with correction for multiple comparisons. Using the McFadden's R^2 allowed us to interpret the magnitude of identified DIF. The results suggest that the items composing the full UDysRS are highly specific to dyskinesia severity. With the negligible contributions from age, gender, race/ethnicity, and education level, the scale can be viewed as widely applicable and not impacted by these demographic indices. ■

Acknowledgments: The datasets for this study were contributed to the International Parkinson and Movement Disorder Society as part of the international effort to develop validated versions of the UDysRS in multiple languages. We acknowledge the following leaders of these teams who worked with colleagues to examine PD patients using the UDysRS: Chinese, Ruey-Meei Wu; English, Christopher G. Goetz; French, Olivier Rascol; German, Richard Dodel; Greek, Sevesti Bostantjopoulou and Zoe Katsarou; Hungarian, Norbert Kovács; Italian, Angelo Antonini; Japanese, Atsuhi Takeda; Korean, Hee Tae Kim; Slovak, Matej Skorvanek; Spanish, Esther Cubo; Portuguese, Joaquim Ferreira and Francisco Cardoso; Turkish, Cenk Akbostanci.

References

1. Goetz CG, Nutt JG, Stebbins GT. The Unified Dyskinesia Rating Scale: presentation and clinimetric profile. Mov Disord 2008; 23(16):2398-2403.
2. Colosimo C, Martínez-Martín P, Fabbrini G, et al. Task force report on scales to assess dyskinesia in Parkinson's disease: critique and recommendations. Mov Disord 2010;25(9):1131-1142.
3. Goetz CG, Stebbins GT, Wang L, LaPelle NR, Luo S, Tilley BC. IPMDS-sponsored scale translation program: process, format, and clinimetric testing plan for the MDS-UPDRS and UDysRS. Mov Disord Clin Prac 2014;1(2):97-101.
4. Hambleton RK. Good practices for identifying differential item functioning. Med Care 2006;44(11):S182-S188.
5. Mellenbergh GJ. Contingency table models for assessing item bias. J Educ Stat 1982;7(2):105-118.
6. No D. 15: Race and ethnic standards for federal statistics and administrative reporting. Washington, DC: US Office of Management and Budget. 1978 May 4.
7. Brown T. Confirmatory factor analysis for applied research. New York: Guilford Press; 2006.
8. Samejima F. (1968), Estimation of latent ability using a response pattern of graded scores1. ets Research Bulletin Series, 1968:i–169. doi:10.1002/j.2333-8504.1968.tb00153.x
9. Choi SW, Gibbons LE, Crane PK. Lordif: an R package for detecting differential item functioning using iterative hybrid ordinal logistic regression/item response theory and Monte Carlo simulations. J Stat Soft 2011;39(8):1.
10. Muthén B. A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. Psychometrika 1984;49(1):115-132.
11. Jöreskog KG, Goldberger AS. Estimation of a model with multiple indicators and multiple causes of a single latent variable. JASA 1975;70(351a):631-639.
12. Teresi JA. Different approaches to differential item functioning in health applications: advantages, disadvantages and some neglected topics. Med Care 2006;44(11):S152-S170.
13. Jodoin MG, Gierl MJ. Evaluating type I error and power rates using an effect size measure with the logistic regression procedure for DIF detection. Appl Meas Educ 2001;14(4):329-349.
14. Roju NS, Van der Linden WJ, Fleer PF. IRT-based internal measures of differential functioning of items and tests. Appl Psychol Meas 1995;19(4):353-368.
15. Flowers CP, Oshima T, Raju NS. A description and demonstration of the polytomous-DFIT framework. Appl Psychol Meas 1999;23(4):309-326.
16. Kleinman M, Teresi JA. Differential item functioning magnitude and impact measures from item response theory models. Psychol Test Assess Model 2016;58(1):79-98.
17. van Rooden S, Verbaan D, Stijnen T, Marinus J, Van Hilten J. The influence of age and approaching death on the course of non-dopaminergic symptoms in Parkinson's disease. Parkinsonism Relat Disord 2016;24:113-118.
18. Keezer MR, Wolfson C, Postuma RB. Age, gender, comorbidity, and the MDS-UPDRS: results from a population-based study. Neuroepidemiology 2016;46(3):222-227.
19. United Nations Educational, Scientific and Cultural Organization (UNESCO). International Standard Classification of Education. Montreal, Quebec: Unesco Institute for Statistics, 2012.
20. Embretson SE, Reise SP. Item Response Theory. Psychology Press, Mahwah, New Jersey, 2013.
21. Goetz CG, Liu Y, Stebbins GT, et al. Gender-, age-, and race/ethnicity-based differential item functioning analysis of the movement disorder society–sponsored revision of the Unified Parkinson's Disease Rating Scale. Mov Disord 2016;31(12):1865-1873.

Supporting Data

Additional Supporting Information may be found in the online version of this article at the publisher's website.