HHS Public Access

Author manuscript

Mov Disord. Author manuscript; available in PMC 2017 December 01.

Published in final edited form as:

Mov Disord. 2016 December ; 31(12): 1865–1873. doi:10.1002/mds.26847.

Corresponding author: Christopher G. Goetz, MD, Rush University Medical Center, Suite 755: 1725 W. Harrison Street, Chicago, IL, 60612 USA, Telephone: 312-942-8016; FAX: 312-563-2024; cgoetz@rush.edu.

Supplemental material included as separate file

Author roles
Christopher G. Goetz:
Research project - conception, organization, execution
Statistical analysis - design, review and critique
Manuscript preparation - writing of the first draft, critique and review
Yuanyuan Liu:
Statistical analysis – conduct and review
Manuscript preparation – review and critique
Glenn T. Stebbins:
Research project - conception, organization, execution
Statistical analysis - design, review and critique
Manuscript preparation – writing of first draft, review and critique
Lu Wang:
Statistical analysis –conduct and review
Manuscript preparation – review and critique
Jeanne Teresi :
Statistical analysis – consultation and review
Manuscript preparation - review and critique
Doug Merkitch:
Statistical analysis – review
Manuscript preparation – review and critique
Barbara C. Tilley:
Research project - conception, organization, execution
Statistical analysis – design, supervision at all levels, review and critique
Manuscript preparation – review and critique
Sheng Luo:
Research project - conception, organization, execution
Statistical analysis – design, supervision, conduct and review
Manuscript preparation – review and critique

Financial Disclosers for the past 12 months
Christopher G. Goetz, MD
Consulting or Advisory Board Membership with honoraria: Addex, Avanir, Boston Scientific, CHDI, Clevexel, Kanter Health, Oxford Biomedica, Pfizer, WebMD.
Grants/Research: Funding to Rush University Medical Center from NIH, Michael J. Fox Foundation for research conducted by Dr. Goetz. Dr. Goetz directs the Rush Parkinson's Disease Research Center that receives support from the Parkinson's Disease Foundation and some of these funds support Dr. Goetz's salary as well as his research efforts. He directs the translation program for the MDS-UPDRS and UDysRS and receives funds directed to Rush University Medical Center from the International Parkinson and Movement Disorder Society (IPMDS) for this effort.
Honoraria: American Academy of Neurology, Captain James A Lovell Federal Health Care Center, University of Pennsylvania, University of Rochester
Intellectual Property Rights: none
Ownership interests: none
Royalties: Elsevier Publishers, Oxford University Press, Wolters Kluwer,
Salary: Rush University Medical Center
Yuanyuan Liu, MS
Consulting: none
Honoraria: none
Intellectual Property Rights: none
Ownership interests: none
Royalties: None
Salary: University of Texas School of Public Health (MDS support for project included)
Glenn T. Stebbins, PhD
Consulting and Advisory Board Membership with honoraria: Acadia, Pharmaceuticals, Adamas Pharmaceuticals, Inc., Ceregene, Inc., CHDI Management, Inc., Ingenix Pharmaceutical Services (i3 Research), Neurocrine Biosciences, Inc., Pfizer, Inc..
Grants and Research: National Institutes of Health, Michael J. Fox Foundation for Parkinson's Research, Dystonia Coalition, CHDI, International Parkinson and Movement Disorder Society, CBD Solutions.
Honoraria: International Parkinson and Movement Disorder Society, American Academy of Neurology, Michael J. Fox Foundation for Parkinson's Research, Food and Drug Administration.

Gender-, age- and race/ethnicity-based Differential Item Functioning (DIF) analysis of MDS-UPDRS

Christopher G. Goetz, MD[1], **Yuanyuan Liu, MS**[2], **Glenn T. Stebbins, PhD**[1], **Lu Wang, MS**[2], **Barbara C. Tilley, PhD**[2], **Jeanne A. Teresi, EdD PhD**[3,4], **Douglas Merkitch, BS**[1], and **Sheng Luo, PhD**[2]

[1]Department of Neurological Sciences, Rush University Medical Center, Chicago, IL USA

Intellectual Property Rights: none
Ownership interests: none
Royalties: none
Expert Testimony: none
Salary: Rush University Medical Center
Lu Wang, MS
Consulting: none
Honoraria: none
Intellectual Property Rights: none
Ownership interests: none
Royalties: none
Salary: University of Texas School of Public Health (MDS support for project included)
Barbara C. Tilley, PhD
Consulting and Advisory Boards: Pfizer Data and Safety Monitoring Committee;
Grants/Research: Movement Disorders Society Grant, NIH grants (NINDS, NHLBI, NIMHD, NIGMS), NIH Data and Safety Monitoring Committees, NIA Clinical Trials Advisory Panel, CHDI.
Honoraria: Roche Pharmaceuticals
Intellectual Property Rights: none
Ownership interests: none
Royalties: none
Salary: University of Texas School of Public Health (MDS and CHDI support for project included)
Jeanne A. Teresi, EdD, PhD
Consulting and Advisory Boards: Weill Cornell Medical College, Division of Geriatrics and Palliative Medicine
Grants/Research: NIH grants (NIA, NINR, NINDS, NHLBI, NIMHD, NIAMSD), HSOD, PCORI, NIH Data and Safety Monitoring Committees
Honoraria: none
Intellectual Property Rights: none
Ownership interests: none
Royalties: none
Salary (No support for this effort): salary from New York State Office of Mental Health and Research Division, Hebrew Home at Riverdale
Douglas Merkitch, BS
Consulting: none
Honoraria: none
Intellectual Property Rights: none
Ownership interests: none
Royalties: none
Salary: Rush University Medical Center
Sheng Luo, PhD
Consulting: none
Grants/Research: NIH grants (NHLBI R01NS091307, NCATS 5KL2TR000370), grants from CHDI Foundation, Parkinson Disease Foundation and International Parkinson and Movement Disorder Society
Honoraria: none
Intellectual Property Rights: none
Ownership interests: none
Royalties: none
Salary: University of Texas School of Public Health: (MDS support for project included)

[2]Department of Biostatistics, University of Texas Health Science Center, School of Public Health, Houston, TX, USA

[3]Columbia University Stroud Center at New York State Psychiatric Institute

[4]Research Division, Hebrew Home at Riverdale, RiverSpring Health

Abstract

Objective—Assess MDS-UPDRS items for gender-, age-, and race/ethnicity-based Differential Item Functioning.

Background—Assessing Differential Item Functioning is a core rating scale validation step. For the MDS-UPDRS, Differential Item Functioning occurs if item-score probability among people with similar levels of parkinsonism differ according to selected covariates (gender, age, race/ethnicity). If the magnitude of Differential Item Functioning is clinically relevant, item-score interpretation must consider influences by these covariates. Differential Item Functioning can be Non-uniform (covariate variably influences an item-score across different levels of parkinsonism) or Uniform (covariate influences an item-score consistently over all levels of parkinsonism.

Methods—Using the MDS-UPDRS translation database of over 5,000 PD patients from fourteen languages, we tested gender-, age-, and race/ethnicity-based Differential Item Functioning. To designate an item as having clinically relevant Differential Item Functioning, we required statistical confirmation by two independent methods, along with a McFadden pseudo-R^2 magnitude statistic greater than “negligible.”

Results—Most items showed no gender-, age- or race/ethnicity-based Differential Item Functioning. When Differential Item Functioning was identified, the magnitude statistic was always in the “negligible” range, and the scale level impact was minimal.

Conclusions—The absence of clinically relevant Differential Item Functioning across all items and all Parts of MDS-UPDRS is strong evidence that the scale can be used confidently. As studies of Parkinson's disease increasingly involve multinational efforts and the MDS-UPDRS has several validated non-English translations, the findings support the scale's broad applicability in populations with varying gender, age, and race/ethnicity distributions.

Keywords

Parkinson's disease; MDS-UPDRS; Rating Scales; Clinimetrics; Differential Item Functioning

Introduction

The Movement Disorder Society revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS) was developed to be a comprehensive clinical rating scale covering motor and non-motor elements of Parkinson's Disease (PD).[1,2] The scale has been designated by the NIH Common Data Elements as the recommended scale for the overall assessment of PD.[3] The scale has four Parts, each designed to measure one domain of PD: Part 1 - Non-motor Experiences of Daily Living; Part 2 - Motor Experiences of Daily Living; Part 3 - Motor Examination; and, Part 4 - Complications of Therapy. The scale was developed in English with a clinimetric program to provide validated non-English translations.[4]

Multiple aspects of the clinimetric strengths of the MDS-UPDRS are known, but there has been no examination of the potential for Differential Item Functioning (DIF).[5] Testing a rating scale for DIF is a core step in comprehensive scale validation methodology to determine if covariates, such as age, gender or race/ethnicity, substantially bias any item score. DIF occurs for the MDS-UPDRS when the probability of an item-score differs among people with similar severity levels of a parkinsonism domain or trait (in DIF terminology) embodied by the summary Part score, but who belong to different groups on a covariate such as gender, age, or race/ethnicity. For example, gender-based DIF would be present for item 1.1 (Cognition) if men and women with the same level of Non-motor Experiences of Daily Living (the trait measured by Part 1) responded differently. Depending on the pattern of this gender-based difference, two kinds of DIF can occur. In non-uniform DIF (NU-DIF), covariate influences on item-scores vary across levels of the Parkinsonian trait, having one pattern of influence at lower ranges of the trait measure and a different pattern when the trait measure is higher. In uniform DIF (U-DIF), influences on item-scores by the covariate are constant across all trait levels (Figure).[6] In either case, DIF signals a concern for potential secondary influences on the scale that must be tested further for clinical relevance typically determined by an additional magnitude calculation such as a McFadden R^2 score.[7] Establishing that NU-DIF or U-DIF cannot be identified in MDS-UPDRS items with regards to important covariates allows the scale to be applied across broad populations of subjects with PD without consideration of the covariate. We tested the hypothesis that the MDS-UPDRS items would not demonstrate clinically relevant DIF by conducting both U-DIF and NU-DIF assessments with a focus on the demographic characteristics of gender, age, and officially designated categories for race/ethnicity.[8]

Methods

The MDS-UPDRS dataset

We accessed the cross-sectional international translation dataset for the MDS-UPDRS program that included English (N=877) and 13 non-English validated editions each with a minimum of 350 cases (Chinese [N=350], Estonian [N=352], French [N=350], German [N=450], Greek [N=350], Hebrew [N=383], Hungarian [N=357], Italian [N=378], Japanese [N=365], Korean [N=362], Russian [N=384], Slovakian [N=354], Spanish [N=443]).[4] Most languages were tested in only one country, but some used multiple geographical populations: English (USA, Canada, United Kingdom and Australia); Spanish (Spain, Argentina, Cuba, Mexico and USA); German (Germany and Austria). Each language team translated and back-translated the original validated English MDS-UPDRS, refined the version using Cognitive Pre-testing methodology, and used the translation to examine PD, patients, submitting scores to a central database. Validated versions were designated if pre-specified criteria were met based on Comparative Fit Index methodology.[4] Cases were included in the DIF analysis for a given Part of the MDS-UPDRS if all items were complete in that Part.

Assessing unidimensionality of the MDS-UPDRS parkinsonism domains measured by each Part

DIF analyses are anchored in the assumption that the items being examined measure a single pertinent trait. In the original English MDS-UPDRS validation program, we established

unidimensionality within four clinimetrically sound domains designated as Parts (1: Non-motor Experiences of Daily Living, 2: Motor Experiences of Daily Living; 3: Motor Examination; and 4: Motor Complications).[1,2] Because both *lordif*[9] and Multiple Indicators, Multiple Causes (MIMIC)[10,11] DIF analyses require items to be tested against a unidimensional domain, we began the program by testing unidimensionality of each Part of the MDS-UPDRS in the combined language datasets. To consider the Parts of the MDS-UPDRS as providing four unidimensional domains of parkinsonism, we conducted confirmatory factor analysis for each Part, requiring that the Confirmatory Fit Index (CFI) was > 0.90 with Root Means Square Error of Approximation (RMSEA) < 0.10.[12]

Sample sizes for each analysis

DIF analyses require that for each item, all possible rating values must have some representation. For many MDS-UPDRS items, however, there were no patients scoring in the most severe rating option (4). Therefore, we combined scores of 3 and 4 as a collapsed designation, termed 3/4, allowing the statistical methods to converge mathematically. Further, to conduct our analyses for gender, age, and race/ethnicity, we required data representation of at least 5 subject samples in the 0, 1, 2, and 3/4 categories for each MDS-UPDRS item in a given Part in order to proceed with DIF analysis.

Overall approach

We conducted DIF analysis using two independent latent variable models, the iterative hybrid ordinal logistic regression/item response theory (graded response model)[13] approach as realized in the R package *lordif*[9] and the MIMIC model.[10,11] Following published recommendations, 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 Bonerroni correction. [14] Because Item 1.6 (Features of Dopamine Dysregulation Syndrome) had performed poorly in the original scale assessment,[15] we excluded this item from the Part I analysis.

All items were studied first for NU-DIF and those without NU-DIF were then analyzed for U-DIF.[14] For items identified with DIF, to determine the relevance or magnitude on the overall domain, 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.[7] As a prespecified outcome, we considered an item with DIF to be clinically relevant and of concern for co-variate bias if the McFadden R^2 indicated a moderate or large magnitude. Finally, in each co-variate analysis, for any Part with multiple identified DIF items, we examined their combined impact on the Part, termed Scale Level Impact, using the Differential Test Function (DTF) index that compared the Test Characteristic Curves with and without DIF items.[16] To assess the magnitude of the DTF, we used the recommended chi-square statistic, but in the context of our very large sample size and recognition of possible over-identification of DIF with chi-square,[16] we also calculated more conservative thresholds based on Monte Carlo simulations[17,18] (cutoff DTF value Part 1 = 0.648; Part 2 = 0.702; Part 3 = 2.782; Part 4 = 0.324).

Comparisons

For gender, the analyses compared males and females. For the age-based DIF analyses, we chose three age groups (ages 28-51, ages 52-75, and ages 76-97). This trichotomy of the sample's range resulted in at least 400 cases in each age group. We chose race/ethnicity categories according to published divisions adopted by the US Office of Management and Budget.[8] The prescribed methodology for such determination is one of self-definition by the study subject. These categories were reviewed by each language team before starting each language translation program and adapted for the countries where data would be obtained (i.e., African-American was adapted to African descent). Possible choices were: White (non-Hispanic), Hispanic, African descent or African American, Asian, Pacific Islander, Native or Endogenous, and Other (see Supplemental Material for specific definitions). Whereas the *lordif* model can accommodate multinomal 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

Unidimensionality

The confirmatory factor analysis of the combined translation datasets confirmed unidimensionality within each of the four parts of the MDS-UPDRS. Each Part met our pre-specified criteria for unidimensionality of a CFI ≥ 0.90 and a RMSEA < 0.10, allowing conduct of the DIF analyses[12] (Part 1 CFI = 0.91, RMSEA = 0.08; Part 2 CFI = 0.97, RMSEA = 0.09; Part 3 CFI = 0.94, RMSEA = 0.08; Part 4 CFI = 1.00, RMSEA =0.06).

Sample Sizes

The entire data set included MDS-UPDRS scores for 5,755 subjects, but missing data on isolated items or demographic information reduced the samples. In all assessments however, the sample exceeded 5,000 MDS-UPDRS complete scores for the Part being assessed (Table 1).

Gender-based DIF (Table 2)

All MDS-UPDRS items had sufficient representation of severity scores across all categories (0-3/4) to be analyzed. No item exhibited NU-DIF for gender. Twenty items (2 from Part 1, 6 from Part 2, 10 from Part 3, and 2 from Part 4) met criteria by the two independent methods for gender-based U-DIF, though in all cases the magnitude of the DIF was “negligible” with McFadden R^2 values far below the minimal value to meet a “moderate” magnitude rating. In assessing any combined effects of multiple “negligible” impacts, we did not detect an overall Scale Level Impact on any MDS-UPDRS Part from gender-based DIF using the DTF index score (DTF Part 1 = 0.12; Part 2 = 0.12: Part 3 = 1.11; Part 4 = 0.07; all chi-square p's > 0.995; DTF simulation-based thresholds not exceeded). (Supplemental Material provides all results for identified DIF).

Age-based DIF: Table 3

For NU-DIF, three items in Part 1 met the DIF criteria, but all were negligible in magnitude. No NUDIF was identified in Parts 2, 3, and 4. For U-DIF, 16 Items met the DIF criteria, five showing DIF in all three age-group comparisons, and the other items showing DIF in one or two of the group comparisons. In all cases, the McFadden R^2 values did not meet the pre-specified criteria for moderate or large magnitude impact on the relevant Part of the MDS-UPDRS. Further, based on chi-square statistics, we did not detect an overall Scale Level Impact on any MDS-UPDRS Part from age-based DIF using the DTF index score. Using the simulation-based threshold values, impact was observed for Parts 1 and Parts 3 for the youngest (<52) and oldest (>75) group comparisons (DTF Part 1 < 52 years = 2.01, 52 - 75 years = 0.01; > 75 years = 5.45; Part 2 < 52 years = 0.02, 52-75 years = 0.10, > 75 years = 0.09; Part 3 < 52 years = 2.12, 52 - 75 years = 1.45, > 75 years = 5.10; Part 4 < 52 years = 0.00, 52 - 75 years = 0.11, > 75 years = 0.28). The effect of DIF was small, with the potential for less than a 3-point difference in the total Part 1 score and less that a 4-point difference in the total Part 3 score for both the youngest and oldest groups. (Supplemental Material provides all results for identified DIF).

Race/ethnicity-based DIF (Table 4)

The racial/ethnic groups with sufficient representation for analysis were White non-Hispanic, Hispanic, and Asian. We did not have a sufficiently large score representation to allow inclusion of other groups. For these three groups, Parts 1, 2, and 4 items all had sufficient representation of item scores across categories (0, 1, 2, and combined 3/4) to be studied. For Part 3, in spite of the overall large sample size, our requirement to have at least five subject scores in each of the categories for each Part 3 item was not met, and, therefore, race/ethnicity analyses considered DIF only for Part 1, 2, and 4.

For NU-DIF, only one item met DIF criteria but had negligible magnitude on the overall Part 1 scoring. For U-DIF, eight Items met the criteria for statistical consideration, all of negligible magnitude. We did not detect an overall Scale Level Impact on any MDS-UPDRS Part from race/ethnicity-based DIF using the DTF index score (Part 1 White vs. non-White = 0.34. Asian vs. non-Asian = 0.02, Hispanic vs. non-Hispanic = 0.31; Part 2 White vs. non-White = 0.34. Asian vs. non-Asian = 0.26, Hispanic vs. non-Hispanic = 0.11; Part 4 White vs. non-White = 0.01. Asian vs. non-Asian = 0.03, Hispanic vs. non-Hispanic = 0.04; all chi-square p's > 0.995; DTF simulation-based thresholds not exceeded). (Supplemental Material provides all results for identified DIF).

Discussion

DIF, often termed “measurement bias”,[14,16-19] is essential to test for a full validation of a rating scale and the confident conclusion that the scale is truly measuring the intended condition. Our failure to detect DIF of moderate or large magnitude for any item relative to any of the studied demographic elements strongly argues that the MDS-UPDRS is effectively capturing parkinsonism and is not highly influenced by gender, age, or race/ethnicity. The conclusion is reinforced by our inability to detect a significant combined Scale Level Impact when multiple “negligible” DIF items occur in any Part of the scale. Our

conclusions on Scale Level Impact are anchored in the standard chi-square-based DTF index calculation, but we are interested in the future development and applications of simulation-based cutoffs for this determination.[17,18] Using this method, we identified a small DIF impact on the youngest and oldest age cohorts for Parts I and III, but at this point, we rely on the standard recommended chi-square analysis for our final interpretations.[16]

Although the sample sizes were very large, we were limited by the paucity of item-scores in the severe impairment and disability category (4). For this reason, because DIF statistical programs require representation of all categories, we collapsed 3 and 4 categories into a single designation. We admit that this strategy does not achieve a full DIF analysis of the MDS-UPDRS as constructed, and we have encouraged colleagues to contribute cases with severe PD across the entire program to enrich the current sample. We asked groups to provide us with a representative sample with all Hoehn and Yahr stages represented, but, in an effort to reduce bias, we did not issue administrative directives to submit datasets that covered the entire range of item-scores.

Although we focused on gender, age and race/ethnicity, several other DIF influences could still exist. We were unable to address potential DIF related to source of information for Parts 1 and 2 (patient, caregiver, or combined patient/caregiver). Almost all assessments were from patients.The very low representation of caregiver and combined patient/caregiver files failed to allow those categories to meet our sample size requirements for item score representation needed for analysis. A future study focused on this issue, however, could allow such an analysis. A second issue would be rater- or site-based DIF, but the datasets involved hundreds of sites, each often with multiple raters, also precluding such analyses.

A third issue of potential DIF would be the impact of ON vs. OFF state. This analysis would be particularly interesting, but it is important to emphasize the core premise of DIF so as not to confuse. As we point out in the Introduction, DIF addresses the fundamental issue of whether covariates differentially influence patients' item scores at the same level of the primary trait being studied (parkinsonism). As a group, OFF patients and ON patients differ in this primary trait, because OFF patients are more parkinsonian than ON patients. DIF cannot simply compare these two groups. On the other hand, if a set of patients in the ON state had the same distribution of overall parkinsonism as another set of patients in the OFF state, DIF analysis could be performed. In examining our dataset, only 26% were assessed in the OFF state, and again our sample size did not meet the requirement for a DIF analysis. .

We chose age divisions to reflect our age ranges, and they are similar to other reports examining age divisions in PD.[20,21] The race/ethnicity divisions used in this study were developed by a US panel,[8] but we were careful to review these categories with each language team prior to data collection to ensure that any specific ambiguities would be resolved. Although we anticipated some concerns, in fact, we had only rare questions directed to our administrative team on race/ethnicity designaations. We adapted "African American" to "African Descent" and "Native American" to "Indigenous" for all non-American patients. We did not have a strong representation for subjects of African heritage, but further expansion of our translation program may provide more subjects to allow comprehensive testing. Another underrepresented race/ethnic division was "mixed race",

and we discovered that this term was considered pejorative in many cultures. Specific translated terms used in each program were selected to be as culturally neutral as possible, but the final number of cases with this self-designation was too small to analyze. The method of self-designation for race/ethnicity is standard for the methodology linked to these categories.[8]

The analysis of race/ethnicity DIF is complicated, because divisions blend genetics, culture, environment, education, and potentially health care access.[22]. For example, the Spanish language cohort included individuals from Spain, Argentina, Cuba, Mexico and USA, and thus represented multiple genetic, cultural and environmental factors. Asians represented those of Chinese, Japanese, and Korean ancestry. We admit that an Asian living in the US may be more similar to other US-based PD patients than those actually living in Asia. If groups are to be compared culturally, attempts to examine DIF by geography, either country, or world regions (Northern Europe vs. Southern Europe) could be envisioned. In spite of our large, combined data set, each language, except English, involved approximately 350 subjects, and these subjects often represented several countries where the language is spoken, limiting the sample size available for an individual country. As we acquire more languages and as more groups contribute data to the effort, we can approach these very pertinent questions.

We found very few examples of NU-DIF, and when identified, the magnitude was consistently negligible. In NU-DIF, rather than showing a consistent demographic-based influence, an item is influenced in one way at the lower ranges of the measured trait and changes to another pattern in cohorts of higher overall trait severity levels. This “crossing” of demographic group curves defines NU-DIF, and when it has moderate or large magnitude, it poses higher levels of complexity to item scale interpretation, because demographic influences on a given MDS-UPDRS item response are present, but the precise effect differs as overall disability changes from low to high.[6,14] The absence of pertinent NU-DIF for gender, age and race/ethnicity is particularly important to the validation profile of the MDS-UPDRS and allows scale users to dismiss concerns of shifting influences.

In the original validation studies of the MDS-UPDRS, a classical test theory approach was used.[1,2] The DIF analyses method used here employed an item-response theory model which utilizes a latent variable approach.[16,19] Whereas DIF analysis has not been widely applied to PD or movement disorders rating scales, DIF analyses have been published for scales used neurologically, including the Mini-Mental Status Examination,[23] Mattis Dementia Rating Scale[24] and several depression and quality of life measures.[25] In addition to age, gender, and race/ethnicity, such studies have also focused on educational level. We acknowledge the limitation that our MDS-UPDRS database did not record educational level across the full cohort, so we are not able to examine educational level relative to potential DIF. With new language translations in development, our aim is to include African descent and other groups into the analysis.

With the cited limitations stated, the strengths of our study include the very large dataset with world-wide representation across cultures using one validated scale. We have been rigorous in our clinimetric approach, requiring that designated items with DIF be identified

by two independent statistical methods with embedded correction for multiple comparisons. Using the McFadden's R^2 application provides a rigorous method to assess the statistical importance to each observed DIF finding, allowing us to interpret the magnitude of identified DIF and in this case relegating all identified DIF as negligible. The results allow us to consider the items in the MDS-UPDRS as highly specific to PD impairment and disability. With the negligible contributions from age, gender, and race/ethnicity, the scale can be viewed as widely applicable. Further data collection may allow for additional analyses including the possible effect of other race categories and the potential impact of different languages.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

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 MDS-UPDRS in multiple languages. We acknowledge the leaders of these teams who worked with colleagues to examine PD patients using the MDS-UPDRS: Chinese - Ruey-Meei Wu; English - Christopher G. Goetz; Estonian - Pille Taba; French - Olivier Rascol; German - Richard Dodel; Greek - Sevesti Bostantjopoulou and Zoe Katsarou; Hebrew - Nir Galadi; Hungarian - Norbert Kovács; Italian - Angelo Antonini; Japanese - Yoshi Mizuno; Korean - Hee Tae Kim; Russian - Arseniy Lavrov; Slovakian - Matej Skorvanek; Spanish - Pablo Martinez-Martin.

This research was supported by the International Parkinson and Movement Disorder Society and NINDS grant 5U01NS043127, NCATS grant 5KL2TR000370, and NINDS grant R01NS091307. The effort was also supported by the Parkinson's Disease Foundation (PDF) as part of the PDF Rush Research Center of Excellence.

References

1. Goetz CG, Fahn S, Martinez-Martin P, Poewe W, Sampaio C, Stebbins GT, Stern MB, Tilley BC, Dodel R, Dubois B, Holloway R, Jankovic J, Kulisevsky J, Lang AE, Lees A, Leurgans S, Lewitt PA, Nyenhuis D, Olanow CC, Rascol O, Schrag A, Teresi JA, Van Hilten JJ, Lapelle N. Movement Disorder Society-sponsored revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS): Process, format, and clinimetric testing plan. Mov Disord. 2007; 22:41–47. [PubMed: 17115387]
2. Goetz CG, Tilley BC, Shaftman SR, Stebbins GT, Fahn S, Martinez-Martin P, Poewe W, Sampaio C, Stern MB, Dodel R, Dubois B, Holloway R, Jankovic J, Kulisevsky J, Lang AE, Lees A, Leurgans S, Le Witt PA, Nyenhuis D, Olanow CW, Rascol O, Schrag A, Teresi JA, van Hilten JJ, LaPelle N. Movement Disorder Society UPDRS Revision Task Force. Movement Disorder Society-Sponsored Revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS): Scale Presentation and Clinimetric Testing Results. Mov Disord. 2008; 23:2129–2170. [PubMed: 19025984]
3. www.commondataelements.ninds.nih.gov/pd.aspx
4. 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; 1:97–101.
5. Hambleton RK. Good practices for identifying differential item functioning. Med Care. 2006; 44(Suppl 3):S182–188. [PubMed: 17060826]
6. Mellenbergh GJ. Contingency table models for assessing item bias. J Educ Statis. 1982; 7:105–118.
7. 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; 14:329–349.
8. Office of Management and Budget. DIRECTIVE NO. 15 Race and Ethnic Standards for Federal Statistics and Administrative Reporting. http://wonder.cdc.gov/wonder/help/populations/bridged-race/directive15.html

9. Choi SW, Gibbons LE, Brane PK. lordif: An R package for detecting differential item functioning using iterative hybrid ordinal logistic regression/item response theory and Monte Carlo simulations. J Statistical Software. 2011; 39:1–30.
10. Muthén BO. A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. Psychometrika. 1984; 49:115–132.
11. Jöreskog K, Goldberger A. Estimation of a model of multiple indicators and multiple causes of a single latent variable. J Amer Stat Assoc. 1975; 10:631–639.
12. Brown, TA. Confirmatory Factory Analysis in Applied Research. Guilford Sage Publications, Inc.; New York NY: 2006.
13. Samejima F. Estimation of latent ability using a response pattern of graded scores. Psychometrika. 1969; 34(Monograph Supplement 1):100–114.
14. Teresi JA. Different approaches to differential item functioning in health applications. Advantages, disadvantages and some neglected topics. Med Care. 2006; 44(Suppl 3):S152–S170. [PubMed: 17060822]
15. Goetz CG, Tilley BC, Stebbins GT. Dopamine dysregulation syndrome item from the MDS-UPDRS. Mov Disord. 2012; 27:166.
16. Raju NS, van der Linden WJ, Fleer PF. IRT-based internal measures of differential functioning of items and tests. Appl Psych Meas. 1995; 19:353–368.
17. Flowers CP, Oshima TC, Raju NS. A description and demonstration of the polytomous DFIT framework. Appl Psych Meas. 1999; 23:309–326.
18. Kleinman M, Teresi JA. Differential item functioning magnitude and impact measures from item response theory models. Psych Test Assessment Modeling. 2016; 58:79–98.
19. Embretson, SE., Reise, SP. Item Response Theory for Psychologists. Lawrence Erlbaum; New Jersey: 2000.
20. Keezer MR, Wolfson C, Postuma RB. Age, Gender, Comorbidity, and the MDS-UPDRS: Results from a Population-Based Study. Neuroepidemiology. 2016; 46:222–227. [PubMed: 26967747]
21. van Rooden SM, Verbaan D, Stijnen T, Marinus J, van Hilten JJ. The influence of age and approaching death on the course of nondopaminergic symptoms in Parkinson's disease. Parkinsonism Relat Disord. 2016; 24:113–118. [PubMed: 26774535]
22. Manly JJ. Deconstructing race and ethnicity: implications for measurement of health outcomes. Med Care. 2006; 44(suppl 3):S10–S16. [PubMed: 17060816]
23. Orlando Edelen MO, Thissen D, Teresi JA, Kleinman M, Ocepek-Welikson K. Identification of differential item functioning using item response theory and the likelihood-based model comparison approach. Application to the Mini-Mental State Examination. Med Care. 2006; 44(Suppl 3):S134–S142. [PubMed: 17060820]
24. Teresi JA, Kleinman M, Ocepek-Welikson K. Modern psychometric methods for detection of differential item functioning: application to cognitive assessment measures. Stat Med. 2000; 19:1651–83. [PubMed: 10844726]
25. Teresi JA, Ramirez M, Lai JS, Silver S. Occurrences and sources of Differential Item Functioning (DIF) in patient-reported outcome measures: Description of DIF methods, and review of measures of depression, quality of life and general health. Psychol Sci Q. 2008; 50:538. [PubMed: 20165561]

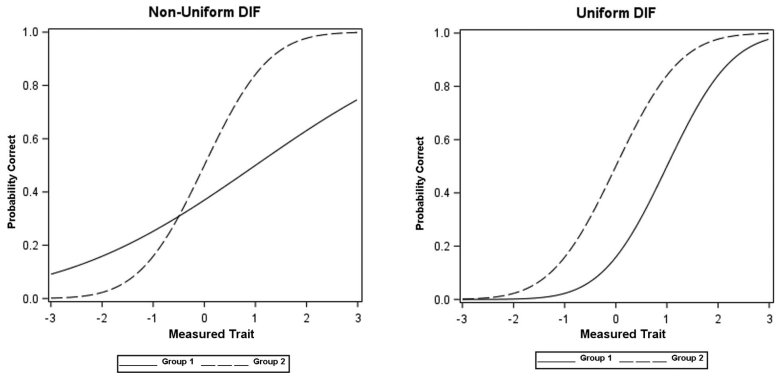

Figure.

The two curves, generated from simulated data, show the differential patterns of non-uniform DIF and uniform DIF for a given item based on two covariates (Group 1, Group 2). In non-uniform DIF (left graph), at low levels of the measured trait, Group 1 scores higher compare to Group 2, but at higher levels of the measured trait, Group 1 scores lower than the other group. In uniform DIF (right graph), Group 1 consistently scores lower than Group 2 across all levels of the measured trait. As an example if men (Group 1) score higher on the Cognitive Impairment item (Item 1.1) than women (Group 2) when both have low Non-motor Experiences of Daily Living trait scores, but score lower on this item when the overall trait score is high, gender-based non-uniform DIF for Item 1.1 would occur; if men consistently have less cognitive impairment and thereby score lower than women on Item 1.1 across all levels of the measured trait of Non-motor Experiences of Daily Living, gender-based uniform DIF exists. (See Supplemental Material for graphs generated from the MDS-UPDRS datasets showing DIF effects for individual items and Parts).

Table 1

Sample size from the master set of English and international translations of the MDS-UPDRS programs

	Part 1	Part 2	Part 3	Part 4
Gender	5547	5546	5326	5562
Age	5381	5375	5159	5397
Race/Ethnicity	5561	5559	5338	5574

Legend: A total of 5755 subjects were included in the master dataset including subjects with missing values. After removing the cases with missing values, the sample sizes listed above were available for DIF analyses. The numbers vary by Part and by age, gender and race/ethnicity.

Table 2

Gender-based DIF

Gender-based U-DIF: Impact magnitude of identified significant DIF

	Item	R^2	Magnitude
Part 1 N=5547	1.4 Anxious Feelings	0.0033	Negligible
	1.9 Pain	0.0019	Negligible
Part 2 N=5546	2.1 Speech	0.0104	Negligible
	2.2 Saliva/drooling	0.0151	Negligible
	2.7 Handwriting	0.0016	Negligible
	2.9 Turning in bed	0.0030	Negligible
	2.10 Tremors	0.0019	Negligible
	2.11 Getting Out of Bed	0.0033	Negligible
Part 3 N=5326	3.1 Speech	0.0153	Negligible
	3.2 Facial expression	0.0120	Negligible
	3.3a Rigidity neck	0.0059	Negligible
	3.3d Rigidity Right LE	0.0029	Negligible
	3.5a Hand movements R	0.0009	Negligible
	3.8a Leg Agility R	0.0032	Negligible
	3.8b Leg Agility L	0.0015	Negligible
	3.9 Arise From Chair	0.0039	Negligible
	3.10 Gait	0.0016	Negligible
	3.12 Postural Stability	0.0079	Negligible
Part 4 N=5562	4.1 Time with Dyskinesias	0.0026	Negligible
	4.2 Functional impact of Dyskinesias	0.0047	Negligible

Legend: For gender, there was no NU-DIF identified. The Table lists items with U-DIF identified by both *lordif* and MIMIC as independent approaches (see Supplemental Material). McFadden's R^2 values and Magnitude of impact are shown in the columns.[7]

Table 3

Age-Based Statistically Significant DIF

Age-Based NU-DIF: Impact magnitude of identified significant DIF				
	Item		**R^2**	**Magnitude**
Part 1 N=5381	1.10 Urinary symptoms	52-75 vs. all others	0.0024	Negligible
	1.12 Lightheadedness	> 75 vs. all others	0.0013	Negligible
	1.13 Fatigue	> 75 vs. all others	0.0018	Negligible
Age-based U-DIF: Impact magnitude of identified significant DIF				
	Item		**R^2**	**Magnitude**
Part 1 N=5381	1.1 Cognition			
		< 52 vs. all others	0.0068	Negligible
		52-75 vs. all others	0.0021	Negligible
		> 75 vs. all others	0.0198	Negligible
	1.2 Hallucinations			
		< 52 vs. all others	0.0069	Negligible
		> 75 vs. all others	0.0147	Negligible
	1.8 Daytime sleepiness	< 52 vs. all others	0.0012	Negligible
	1.11 Constipation			
		< 52 vs. all others	0.0086	Negligible
		> 75 vs. all others	0.0068	Negligible
Part 2 N=5375	2.13 Freezing			
		52-75 vs. all others	0.0015	Negligible
		> 75 vs. all others	0.0012	Negligible
Part 3 N=5159	3.2 Facial expression	< 52 vs. all others	0.0013	Negligible
	3.3b Rigidity right UE	< 52 vs. all others	0.0011	Negligible
	3.3c Rigidity Left UE	< 52 vs. all others	0.0011	Negligible
	3.9 Arising from Chair			
		< 52 vs. all others	0.0036	Negligible
		52-75 vs. all others	0.0090	Negligible
		> 75 vs. all others	0.0207	Negligible
	3.10 Gait			
		< 52 vs. all others	0.0034	Negligible
		52-75 vs. all others	0.0039	Negligible
		> 75 vs. all others	0.0131	Negligible
	3.11 Freezing			
		< 52 vs. all others	0.0031	Negligible
		> 75 vs. all others	0.0016	Negligible
	3.12 Postural Stability			
		< 52 vs. all others	0.0046	Negligible

Age-Based NU-DIF: Impact magnitude of identified significant DIF

	Item		R^2	Magnitude
	3. 13 Posture	52-75 vs. all others	0.0036	Negligible
		> 75 vs. all others	0.0126	Negligible
	3.14 Global Spontaneity	< 52 vs. all others	0.0092	Negligible
		52-75 vs. all others	0.0030	Negligible
		> 75 vs. all others	0.0170	Negligible
	3.16a Tremor right UE	52-75 vs. all others	0.0019	Negligible
		> 75 vs. all others	0.0041	Negligible
		52-75 vs. all others	0.0021	Negligible
		> 75 vs. all others	0.0036	Negligible
Part 4 N=5397	4.6 Painful Off dystonia	>75 vs. all others	0.0044	Negligible

Legend: Most of the MDS-UPDRS 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 in Supplemental Materials). McFadden's R^2 and impact Magnitude are shown in the columns.[7]

Table 4

Race/ethnicity-based statistically significant DIF

Race/ethnicity-Based NU-DIF: Impact magnitude of identified significant DIF				
	Item		**R^2**	**Magnitude**
Part 1 N=5561	1.12 Lightheadedness	White vs. all others	0.0010	Negligible
Race/ethnicity-Based U-DIF: Impact magnitude of identified significant DIF				
	Item		**R^2**	**Magnitude**
Part 1 N=5561	1.2 Hallucinations			
		White vs. all others	0.0020	Negligible
		Asian vs. all others	0.0071	Negligible
	1.7 Sleep problems	Asian vs. all others	0.0034	Negligible
	1.11 Constipation			
		White vs. all others	0.0076	Negligible
		Asian vs. all others	0.0070	Negligible
		Hispanic vs. all others	0.0024	Negligible
	1.13 Fatigue	Asian vs. all others	0.0041	Negligible
Part 2 N=5559	2.2 Saliva/drooling			
		White vs. all others	0.0012	Negligible
		Hispanic vs. all others	0.0014	Negligible
	2.3 Swallowing	Hispanic vs. all others	0.0010	Negligible
	2.10 Tremors			
		White vs. all others	0.0039	Negligible
		Asian vs. all others	0.0058	Negligible
	2.13 Fatigue	Asian vs. all others	0.0011	Negligible

Legend: Three groups had sufficient item representation to be compared: Whites (non-Hispanic), Hispanics, and Asians. Each group was compared against the combined comparator groups. Most of the MDS-UPDRS items did 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 in Supplemental Materials. McFadden's R^2 and impact Magnitude shown in columns.[7]