Modelling item scores of Unified Parkinson's Disease Rating Scale for greater trial efficiency

Yucheng Sheng[1], Xuan Zhou[1], Shuying Yang[1], Peiming Ma[2], and Chao Chen[2]

[1]GlaxoSmithKline
[2]Affiliation not available

July 1, 2020

Abstract

Aim. The multi-part Unified Parkinson's Disease Rating Scale is the standard instrument in clinical trials. A sum of scores for all items in one or more parts of the instrument is usually analyzed. Without accounting for relative importance of individual items, this sum of scores conceivably does not optimize the power of the instrument. The aim was to compare the ability to detect drug effect in slowing down motor function deterioration, as measured by Part III of the Scale - motor examinations - between the item scores and the sum of scores. Methods. We used data from 423 patients in a Parkinson's disease progression trial to estimate the symptom severity by item response modelling; modelled symptom progression using the severity and the sum of scores; and conducted simulations to compare the sensitivity of detecting a broad range of hypothetical drug effects on progression using the severity and the sum of scores. Results. The severity endpoint was far more sensitive than the sum of scores for detecting treatment effects, e.g., requiring 280 versus 570 patients per arm to achieve 60% Probability of Success for detecting a range of potential effects in a 2-year trial. Items related to the left side of the body were most informative; and the domain relevance of tremor items was questionable. Conclusion. This analysis generated clear evidence that longitudinal modelling of item scores can enhance trial efficiency and success. It also prompted the needs for a consensus on the placement of the tremor items in the instrument.

INTRODUCTION

Parkinson's disease (PD) is a chronic and progressive neurodegenerative condition with about 6.2 million patients worldwide[1]. Motor neuron deterioration in the brain is the key characteristic of the disease[2]. Unfortunately, no definitive biomarker for PD has been identified[3]. A Unified Parkinson's Disease Rating Scale (UPDRS) was originally developed as a clinical measure for symptom severity among markedly and severely disabled patients[4]. Later, a Movement Disorder Society version of UPDRS (MDS-UPDRS) was introduced for early diagnosis to measure milder deficit and smaller changes in the early disease stage, focusing on broader and lower ranges in disability than the original UPDRS. The MDS-UPDRS consists of four parts, reflecting different aspects of the clinical manifestation of the disease[5,6]. The outcome of the assessment is a sum of scores (SoS) of multiple items in each part, and a total score (TS) for all parts. Using these composite scores for evaluating disease severity and treatment effects requires large sample sizes to avoid inconclusive drug trials, especially for disease modifying treatments[7]. An alternative analytical approach that can enhance the signal-to-noise ratio would open the path for more efficient and rigorous clinical trials of PD therapies.

Item Response Theory (IRT) modelling describes the relationships between the trait of interest and the items that are used to measure the trait; therefore, it is a promising approach for analyzing itemized scales[8]. Instead of relying on a single composite score of the test, it defines mathematical links for individual items in the instrument to directly estimate a patient's disease severity that the very instrument is designed to measure. For its improved utilization of the data at the item level, IRT has been applied in the research

of several neurological diseases such as Parkinson's disease[12,22,28], Alzheimer's disease[9], multiple sclerosis[10] and schizophrenia[11]. Remarkably, the methodology have shown promise to significantly reduce the size of drug trials[9,22].

Demonstrating the ability to delay motor impairment is essential for a drug aimed to slow down PD progression. Longitudinal IRT models has been developed using MDS-UPDRS to describe the progression of PD[12,22]. The models included the assessments of non-motor domains, as well as interaction terms among items of different domains. The goal of the current analysis was to assess the IRT's ability to enhance the efficiency for detecting drug effect on MDS-UPDRS Part III – motor examinations – which is considered as a more objective endpoint of motor function, hence central to diagnostic and therapeutic assessments. Specifically, the aims were to: i) develop an IRT model for estimating symptom severity using item scores of MDS-UPDRS Part III, ii) use the IRT model to explore relative importance of the items, iii) build longitudinal models to describe symptom progression over time in terms of SoS and symptom severity, and iv) compare the probability of trial success when analyzed using symptom severity or SoS for a potential disease-modifying new treatment with uncertain effect.

METHODS

Data source

We analyzed patient-level data from a *de novo* cohort of the Parkinson's Progression Markers Initiative (PPMI) study - an ongoing multi-cohort observational study to identify biomarkers of Parkinson's disease progression. The study design and its inclusion and exclusion criteria can be found at http://www.ppmi-info.org/wp-content/uploads/2017/02/PPMI-Am11-Protocol.pdf. In brief, patients in this cohort were enrolled within two years of positive diagnosis. They had not taken PD medications for more than 60 days prior to the baseline and were not expected to require PD medications for at least six months from baseline. The MDS-UPDRS observations were collected every 3 months up to 12 months and thereafter every 6 months. We used data that were available as of January 2020.

Symptomatic treatments were allowed at any time during the study. For treated subjects, both Off-Med) and On-Med MDS-UPDRS observations on the same day were recorded. To minimize the symptomatic impact of PD medications and anticipate the intended analysis in the eventual drug trials, this analysis used the Off-Med observations. The MDS-UPDRS Part III assessment included 33 items.

Item response modelling for estimating symptom severity

The concept of the IRT model is shown in Figure 1. The score for each of the 33 items is an ordinal variable in commonly accepted clinical terms: 0 = normal, 1 = slight, 2 = mild, 3 = moderate and 4 = severe. For each item, a graded-response logit model was used for describing the probability of a subject's score for each item[13]:

$$P\left(Y_{\mathrm{ij}} \geq k\right) = \frac{e^{a_j\left(S_i - b_{\mathrm{jk}}\right)}}{1+e^{a_j\left(S_i - b_{\mathrm{jk}}\right)}} \text{Equation 1}$$

$$P\left(Y_{\mathrm{ij}} = k\right) = P\left(Y_{\mathrm{ij}} \geq k\right) - P\left(Y_{\mathrm{ij}} \geq k+1\right) \text{Equation 2}$$

Equation 1 describes $P\ (Y_{ij}\ [?]\ k\ )$ as the probability that the score of subject i for item $j(Y_{ij}\ )$ is at least k, *where* S_i is the severity for subject i ;a_j is called the discrimination parameter for item j , reflecting the ability of the item to differentiate the severity among the patients; and b_{jk} is called difficulty parameter of score k for item j , representing the severity at which there is a 50% probability of obtaining a score $[?]k$ for that item. The probability that the score of subject$i$ for item j $(Y_{ij}\ )$ is k can then be derived in Equation 2.

As such, the 33 item-level graded-probability models described by Equations 1 and 2, one for each item, collectively estimate a severity level for each patient at a given point in time, mirroring the patient's sum of scores (Figure 1). The graphical representation of Equation 1 and Equation 2 are called Item Characteristic Curve (ICC) and Category Characteristic Curve (CCC), respectively; they can be visualized in Figure 1.

The difficulty and discrimination parameters were determined by fitting Equations 1 and 2 to the item scores of the entire dataset; effectively, the severity in the same patient at different visits were estimated independently without correlation. Baseline severity values were assumed to follow a standard normal distribution with a mean of zero and a variance of one. The severity values in subsequent visits were anchored to the baseline, with an estimated shift in their means and variances. This way all the IRT model parameters were identifiable[14]. The distribution of the estimated severity values was plotted over time to explore the disease progression.

Identification of the most informative items

The Fisher information functions in Equation 3 and Equation 4 were used to estimate the information content across the entire severity range[13]:

$I_{\mathrm{jk}}\left(S_{i(t)}\right) = -\frac{\partial^2}{\partial S^2} \log P_{\mathrm{jk}}(S = S_{i(t)})$Equation 3$I_j\left(S_{i(t)}\right) = \sum_{n=0}^{K} I_{\mathrm{jk}}\left(S_{i(t)}\right) P_{\mathrm{jk}}(S_{i(t)})$Equation 4

In these equations,$P_{jk}\ (S_{i\ (t\ )})$ is the probability of responding with score k of item j by subject i with severity$S_{i\ (t\ )}$ at time t . Thus, $I_{jk}(S_{i\ (t\ )})$ is the information for score k of item j from subject i at time$t$, and$I_j\ (S_{i\ (t\ )})$ is the total information for all scores (from the lowest score of0 to the highest score of K) of item j from subject$i$ at time t .

The items were ranked according to their overall informativeness, which was $I_j\ (S\ )$ integrated over time and summed for all subjects.

Longitudinal modelling of symptom progression

Disease progression over the first five years was modelled, separately, in terms of severity and SoS. In the case of the severity, the longitudinal function was estimated from the item scores, while the difficulty and discrimination parameters were fixed to those determined as described above.

Informed by the data pattern shown in the upper panels of Figure 2, a linear function described in Equation 5 was used, where$S_{i\ (t\ )}$, was symptom, in terms of either Severity or SoS, for patient i at time t ;$S_{i,0}$ was baseline; and$Slope_i$ was the progression rate. The $IOV$was the inter-occasion (visit) variability to capture the fluctuation of clinical symptoms.

$S_{i(t)} = S_{i,0} + \mathrm{Slope}_i \bullet time + IOV$Equation 5

Both severity and SoS longitudinal models were fitted to data, including or excluding the tremor items. (See Results section.)

Evaluation of the longitudinal item-response model

The adequacy of the longitudinal item-response model was evaluated in several ways, by comparing the model estimated or simulated data to the actual observations in the trial in terms of item scores and SoS, and over time:

i) To detect any major overall bias in the item-response model, the model estimated CCCs were compared to the distribution of the observations across entire severity range, for each score in each item. ii) To detect score-specific bias at the item level, the proportion of patients having that score was compared between simulated data and the observed data, for each score in each item. iii) To assess the model's longitudinal predictivity, a visual predictive check was conducted to compare the time course of model-simulated and observed SoS values. For additional rigor, the longitudinal visual predictive checks were also conducted for model-simulated proportion of patients having each score for each item.

Assessment of a clinical trial's probability of success

Assuming a drug altered the disease progression rate as shown in Equation 6, where E was the drug effect and all other parameters were the same as in Equation 5, the respective longitudinal models were used to simulate the severity and SoS data for 6000 patients, stratified to either receive the drug treatment or not, according to the assessment schedule in the PPMI trial.

$S_{i(t)} = S_{i,0} + \text{Slope}_i \bullet (1 - E) \bullet time + IOV$ Equation 6

The change from baseline of the simulated data was fitted to a full model which included a drug effect, or a reduced model which did not include any drug effect, for treatment durations of 6, 12, 18 and 24 months. Treatment difference was estimated to generate individual objective function (iOFV) values, which were subject to likelihood ratio test ($p<0.05$) per Monte Carlo Mapped Power (MCMP) method, which has been described in detail elsewhere[29], based on 1000 treatment datasets for a wide range of samples sizes.

A range of potential drug effects – 100 random values from a normal distribution (mean of 0.3 and variance of 0.0169, which generated 5^{th}-95^{th} quantiles of 0.1-0.5) were tested. The collective proportion of trials showing a statistically significant positive drug effect across the entire range was calculated as the Probability of Success (PoS).[15] The PoS was calculated for both severity and SoS endpoints, based on all 33 items or only the non-tremor items (see Results section).

Software

Data modeling and simulation were performed primarily in software NONMEM (ICON, Ellicott City, Maryland, version 7.3) in conjunction with a gfortran (64-bit) compiler using Pirana (version 2.9.7)[16] as an interface. The Laplace integral approximation with -2 times log-likelihood option was used throughout the analyses. The R environment[17] for statistical computing version 3.6.2 was used for simulation and plotting.

RESULTS

The data from the PPMI trial are openly available upon request (https://www.ppmi-info.org/access-data-specimens/). The data used in this analysis were downloaded on 3 Jan 2020; 233607 item level observations from 423 *de novo* PD patients were used in the analysis. The SoS observations are shown in Figure 2 (upper left), and key baseline characteristics are summarized in Table 1.

Item-response model and item importance

An item-response model, including 33 graded-response logit sub-models, one per item, was successfully developed. Figure 2 shows that the pattern of the model-estimated severity data (upper right), including the progression over time, the variability among patients, and the visit-to-visit fluctuation resemble those of the observed SoS (upper left).

The discrimination parameter and four difficulty parameters for all items are shown in Table 2. Score value 4 (severe) was missing from five items (1, 25, 26, 31 and 32). The probability for a patient to score this value, and consequently the corresponding difficulty parameter, could not be estimated for these items.

The information content varied greatly among the items (Figure 3 and Table 2): eight items each held $> 5\%$ of the total information, totaling 65% and with the lowest discrimination parameter being 1.29. All seven items for the left side of the body were among the eight top-informing items.

Conversely, 11 items each held $< 1\%$ information, with the highest discrimination parameter being 0.46. Nine of the ten tremor items were among the 11 least informative ones. Indeed, four of the five items where score 4 (severe) was missing were tremor tests (Table 2). Six items (18, 20, 21, 30, 31 and 32) had mostly score 0 (normal); three were tremor tests (Figure 2, lower right). Several tremor items were even estimated to have a near-zero negative discrimination parameter value, with very wide-ranging difficult parameters. These observations suggested that the parameters were badly estimated for these items and revealed these items' inability to differentiate patients with different levels of symptom severity. Based on these findings, the longitudinal modelling and subsequent estimation of clinical trial PoS were conducted with or without the tremor items.

Longitudinal models for symptom severity and sum of scores

The longitudinal model described in Equation 5 was successfully fitted to both severity data and SoS data, with or without tremor items. The symptom progression rates for both severity and SoS were in turn found to be functions of the baseline: patients with worse symptom at baseline appeared to have slower progression.

When all items were included in the modelling, the progression rate of severity, for the typical patient with a baseline of zero point, was 0.227 points per year. The progression rate of SoS, for the typical patient with a baseline of 19.6 points, was 2.99 points per year. When the tremor items were excluded, the progression rates for severity and SoS were 0.243 and 2.24 points per year, respectively. All model parameters are listed in Table 3.

Adequacy of the longitudinal item-response model

Model-estimated CCCs reflected reasonably well the distribution of observed categories for each item over the range of severity; and the differing steepness and undulation of CCCs among the 33 items suggested these items' varying ability to differentiate severity (Figure 2, lower left). There was good agreement between the observed and model-simulated proportion of each score for each item (Figure 2, lower right). Visual predictive checks further manifested that the final longitudinal IRT model adequately simulated the time course of both Part-III sum of scores (Figure 5) and item scores (Figure 6).

Clinical trial probability of success

The item-response approach consistently demonstrated higher PoS than the SoS approach for all trial durations (Figure 4, upper). For the item-response approach, the PoS was identical whether the tremor items were included or not. For the SoS approach, the PoS estimated based on all items was marginally higher than the one estimated without tremor items. To achieve a 60% PoS in a 2-year trial, the item-response method would require 280 patients per arm and the SoS analysis would require 570 patients per arm. As expected, longer trials produced higher PoS, regardless of the analytical approach.

DISCUSSION

Compared to composite score modelling[20], the IRT approach differentiates the items by their sensitivity level and has shown the potential to reduce trial sample size for detecting drug effects[9,22]. The sample-size saving is an attractive proposition, especially as the field advances towards increasingly personalized medicine, where a certain therapy is expected to be effective only in a small population.

Multi-variable IRT models with item-level interaction across domains have been published; but they were not readily adaptable for application to analysis of Part III alone[12,22]. In this work, we used only items in Part III, aiming to support early development of PD drugs where a Go/No-Go decision hinges on their effect on (the more objective) motor examinations. There is also a differentiating methodological feature of our analysis: the analyses reported by others used the IRT model to simulate the total scores; applied hypothetical drug effects to both the severity endpoint and the simulated total scores; and compared the two endpoints – severity and total score - for the sample size requirement to detect the drug effect. This approach could potentially bias against the total score endpoint, in the event the simulation inflated the noise in total score. In contrast, we applied the drug effect directly to the SoS, as to the severity. In doing so, the two endpoints were treated more fairly.

To compare the sample size requirement between the IRT and the conventional SoS methods, we applied a range of relevant potential reduction in progression rate that a new agent could cause. The normally distributed effects centered at 0.3 and had the 5^{th} – 95^{th} range of 0.1 to 0.5 which has been considered as clinically meaningful effect range for neurodegenerative indications such as Parkinson's disease and Alzheimer's disease[9,22]. While the center of the range represented an effect that's highly relevant and reasonably plausible, the lower and higher tails were respectively less relevant and plausible. As such, the effect levels further away from the center carried less weight in the computation of the overall PoS, which is then effectively the collective power weighted by the distribution of the effect level. We consider this as a useful approach to account for the uncertainty in the eventual effect size that a new agent could produce. Figure 4 lower panel illustrates the (expected) difference between the PoS under this effect distribution and the power under the more extreme effect sizes. For the same sample size, the power for detecting a large treatment effect would be higher than the PoS for detecting a range of potential effects. Under this condition, we found that the IRT method could lead to a tremendous saving of about 50% in sample size compared to the conventional SoS

method. This magnitude of sample size savings is consistent with our recent analysis of a placebo-controlled clinical trial of ropinirole – an established dopaminergic agent.[33]

The tremor tests showed poor discrimination power; they each and collectively held very little information (Table 2). For most of the tremor items, the probability of score 0 (normal) was disproportionally high, regardless of a patient's severity as defined by the overall instrument (Figure 2, lower left and right). Consistent with these observations, the clinical trial PoS was not affected by whether the tremor items were included in the analyses or not (Figure 4 upper). Interestingly, a Rasch measurement theory analysis revealed disordered threshold for several tremor-related items.[34] These observations supported the view that the tremor tests might measure a different construct, hence perhaps should be assessed using a separate and more sensitive scale.[22,31,35]

Interestingly, all seven left-side non-tremor items were among the most informative ones (Table 2). Compared to their right-side counter items, they showed higher discriminatory power (a_j), and generally lower values and narrower ranges of difficulty parameters (b_{j1} to b_{j4}). This was also reflected by the left side's better differentiated ICCs (Figure 2 lower left) and slightly higher proportion of higher scores (Figure 2, lower right). Similarly, Gottipati et al. identified "left hand finger tapping" as the most informative among the sided items[12]. In a previously-reported analysis, we explored the PoS for four different approaches: by IRT and SoS, using all items or only the seven left-side items. For the same sample size, the order of estimated trial PoS was: IRT on all items > IRT on seven items > SoS on seven items > SoS on all items.[34] This order illustrated IRT's ability to enhance signal-noise ratio by item differentiation; indeed, its advantage over SoS was reduced when only the most informative items were included in the analysis. These findings were consistent with earlier analysis of combined Part II and Part III data by Buatois et al.[22]

A recent cross-section analysis also found the discrimination parameters to be higher and difficulty parameters to be lower for the left-side items then for the right-side items.[35] Similar findings were reported from an item-response analysis of multiple latent variables, although that analysis also reported a majority (58%) of the patients having more advanced baseline disability on the right side of the body.[12] The lower difficulty parameters, or worse test performance, for the left side items may be a reflection of most people being right handed, despite neuroimaging and meta-analyses suggesting the dominant side might be affected earlier[25,26,27]. Change of hand preference while the disease progresses has also been reported.[36] This is an area to be investigated further, in different datasets and at different stages of the symptom progression. Another possible reason for the consistent worse performance by the left side was this side being always examined later per UPDRS form. Conceivably, this hypothesis may be tested by randomizing the order of the sided tests.

We introduced an inter-occasion (visit) variability in the longitudinal model to reflect the commonly recognized disease fluctuation; this improved the estimation of the progression rate. The model suggested that patients with lower baseline severity had faster progression, support the report that the progression, when measured by MDS-UPDRS Part III, was slower at the more advanced stage[21]. The effects of other factors such as genotype, comorbidity, age, disease history and diagnostic biomarkers on disease progression remain to be assessed.[23,24,30]

That IRT analysis of MDS-UPDRS Pat III required a smaller sample size is relevant to composite scales used in other indications. Because of the less informative items, composite scores could compromise signal-to-noise ratio. Some instruments are also long, hence physically and mentally exhausting for debilitated patients, and leading to incomplete or poor data. Therefore, a bespoke and shorter instrument is often desired. The development, validation and user training are costly and time consuming; and a new instrument suffers the risk of missing out relevant information when used for assessing a new drug of unestablished profile and lack of comparability with existing data. The IRT approach can enhance signal detection power and reduce sample size through directly accessing and weighting of item-level data of a well-established instrument that's accepted by regulators. When item scores are used directly, incomplete data are still useful. By extension, it may be possible to reduce patient burden by asking each patient to take only a stratified partial test. Other potential applications of this approach include bridging between different versions of an evolving

instrument for meta-analysis or cross-study comparison,[28] and translating clinical trial results to patient outcome expectations. These areas require extensive further research and experience building by the clinical research community.

CONCLUSION

In this work, longitudinal item-response analysis was applied to the data of MDS-UPDRS Part III from the PPMI study. It revealed insight on the relationship between the items of motor examinations and the underlying movement impairment and on the deterioration of the motor function over time. The most useful tests for differentiating symptom severity among patients were those for the left side of the body, and the least useful were the tremor tests. Simulations showed remarkable potential of about 50% sample size reduction by the item-response method, compared to the conventional sum-of-score method, for detecting a range of potential drug effects. We encourage the research community to further explore the full potential of this methodology.

Acknowledgment

This work was conducted using data from Parkinson's Progression Markers Initiatives (PPMI, http://www.ppmi-info.org/) trial, sponsored by The Michael J. Fox Foundation for Parkinson's Research. It would not have been possible without the financial, scientific and personal contributions made by the sponsor, the funding partners, the investigators and the patients of the trial.

Conflict of Interest

The authors conducted this work as salaried employees of GlaxoSmithKline and perceive no conflict of interest.

Funding sources

This work was funded by GlaxoSmithKline employment of all authors.

Data availability

The data from the PPMI trial are openly available upon request (https://www.ppmi-info.org/access-data-specimens/).

References

1. GBD 2015 Disease and Injury Incidence and Prevalence Collaborators. Global, regional, and national incidence, prevalence, and years lived with disability for 310 diseases and injuries, 1990-2015: a systematic analysis for the Global Burden of Disease Study 2015. *Lancet (London, England)* . 2016;388(10053):1545-1602. doi:10.1016/S0140-6736(16)31678-6

2. Villarreal MF, Huerta-Gutierrez R, Fregni F. Parkinson's disease.*Neuromethods* . 2018;138(9):139-181. doi:10.1007/978-1-4939-7880-9_5

3. Miller DB, O'Callaghan JP. Biomarkers of Parkinson's disease: present and future. *Metabolism* . 2015;64(3 Suppl 1):S40-6. doi:10.1016/j.metabol.2014.10.030

4. Movement Disorder Society Task Force on Rating Scales for Parkinson's Disease. The Unified Parkinson's Disease Rating Scale (UPDRS): status and recommendations. *Mov Disord* . 2003;18(7):738-750. doi:10.1002/mds.10473

5. Goetz CG, Fahn S, Martinez-Martin P, et al. 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(1):41-47. doi:10.1002/mds.21198

6. Goetz CG, Tilley BC, Shaftman SR, et al. 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(15):2129-2170. doi:10.1002/mds.22340

7. Kalia L V, Kalia SK, Lang AE. Disease-modifying strategies for Parkinson's disease. *Mov Disord* . 2015;30(11):1442-1450. doi:10.1002/mds.26354

8. Ueckert S. Modeling Composite Assessment Data Using Item Response Theory. *CPT Pharmacometrics Syst Pharmacol* . 2018;7(4):205-218. doi:10.1002/psp4.12280

9. Ueckert S, Plan EL, Ito K, et al. Improved utilization of ADAS-cog assessment data through item response theory based pharmacometric modeling. *Pharm Res* . 2014;31(8):2152-2165. doi:10.1007/s11095-014-1315-5

10. Novakovic AM, Krekels EHJ, Munafo A, Ueckert S, Karlsson MO. Application of Item Response Theory to Modeling of Expanded Disability Status Scale in Multiple Sclerosis. *AAPS J* . 2017;19(1):172-179. doi:10.1208/s12248-016-9977-z

11. Krekels EHJJ, Novakovic AM, Vermeulen AM, Friberg LE, Karlsson MO. Item response theory to quantify longitudinal placebo and paliperidone effects on PANSS scores in schizophrenia. *CPT pharmacometrics Syst Pharmacol* . 2017;(July):543-551. doi:10.1002/psp4.12207

12. Gottipati G, Karlsson MO, Plan EL. Modeling a Composite Score in Parkinson's Disease Using Item Response Theory. *AAPS J* . 2017;(2). doi:10.1208/s12248-017-0058-8

13. Wilson M, Masters GN. *Polytomous Item Response Theory Models* . Vol 58.; 1993. doi:10.1007/BF02294473

14. Lei P-W, Zhao Y. Effects of Vertical Scaling Methods on Linear Growth Estimation. *Appl Psychol Meas* . 2012;36(1):21-39. doi:10.1177/0146621611425171

15. O'Hagan A, Stevens JW, Campbell MJ. Assurance in clinical trial design. *Pharm Stat* . 2005;4(3):187-201. doi:10.1002/pst.175

16. Keizer RJ, Karlsson MO, Hooker A. Modeling and Simulation Workbench for NONMEM: Tutorial on Pirana, PsN, and Xpose. *CPT pharmacometrics Syst Pharmacol* . 2013;2:e50. doi:10.1038/psp.2013.24

17. Team RC. R: A Language and Environment for Statistical ComputingNo Title. 2017.

18. Chalmers RP. mirt : A Multidimensional Item Response Theory Package for the R Environment. *J Stat Softw* . 2012;48(6). doi:10.18637/jss.v048.i06

19. Chalmers RP. Generating Adaptive and Non-Adaptive Test Interfaces for Multidimensional Item Response Theory Applications. *J Stat Softw* . 2016;71(5). doi:10.18637/jss.v071.i05

20. Venuto CS, Potter NB, Ray Dorsey E, Kieburtz K. A review of disease progression models of Parkinson's disease and applications in clinical trials. *Mov Disord* . 2016;31(7):947-956. doi:10.1002/mds.26644

21. Vu TC, Nutt JG, Holford NHG. Disease progress and response to treatment as predictors of survival, disability, cognitive impairment and depression in Parkinson's disease. *Br J Clin Pharmacol* . 2012;74(2):284-295. doi:10.1111/j.1365-2125.2012.04208.x

22. Buatois S, Retout S, Frey N, Ueckert S. Item Response Theory as an Efficient Tool to Describe a Heterogeneous Clinical Rating Scale in De Novo Idiopathic Parkinson's Disease Patients. *Pharm Res* . 2017;34(10):2109-2118. doi:10.1007/s11095-017-2216-1

23. Holden SK, Finseth T, Sillau SH, Berman BD. Progression of MDS-UPDRS Scores Over Five Years in De Novo Parkinson Disease from the Parkinson's Progression Markers Initiative Cohort. *Mov Disord Clin Pract* . 5(1):47-53. doi:10.1002/mdc3.12553

24. Latourelle JC, Beste MT, Hadzi TC, et al. Large-scale identification of clinical and genetic predictors of motor progression in patients with newly diagnosed Parkinson's disease: a longitudinal cohort study and validation. *Lancet Neurol* . 2017;16(11):908-916. doi:10.1016/S1474-4422(17)30328-9

25. Prasad S, Saini J, Yadav R, Pal PK. Motor asymmetry and neuromelanin imaging: concordance in Parkinson's disease. Parkinsonism Relat Disord. 2018;53:28-32

26. Heldmann M, Heeren J, Klein C, et al. Neuroimaging abnormalities in individuals exhibiting Parkinson's disease risk markers. Mov Disord 2018;33(9):1412-1422

27. van der Hoorn A, Burger H, Leenders KL, de Jong BM. Handedness correlates with the dominant Parkinson side: A systematic review and meta-analysis. Mov Disord 2012;27: 206-210

28. Gottipati G, Berges A, Yang S, Chen C, Karlsson M, Plan E. Item response model adaptation for analysing data of different versions of a Parkinson's disease endpoint. Pharm Res 2019; doi.org/10.1007/s11095-019-2668-6

29. Vong C, Bergstrand M, Nyberg J, Karlsson MO. Rapid sample size calculations for a defined likelihood ratio test-based power in mixed-effects models. AAPS J. 2012;14(2):176-186; doi.10.1208/s12248-012-9327-8

30. Ahamadi M, Conrado DJ, Macha S, Sinha V, Stone J, Burton J, Nicholas T, Gallagher J, Dexter D, Bani M, Boroojerdi B, Smit H, Weidemann J, Chen C, Yang M, Maciuca R, Lawson R, Burn D, Marek K, Venuto C, Stafford B, Akalu M, Stephenson D, Romero K; Critical Path for Parkinson's (CPP) Consortium. Development of a disease progression model for leucine-rich repeat kinase 2 in Parkinson's disease to inform clinical trial designs. Clinical Pharmacology and Therapeutics 2020; 107:553-562; doi.org/10.1002/cpt.1634

31. Forjaz MJ, Ayala A, Testa CM, Bain PG, Elble R, Haubenberger D, Rodriguez-Blazquez C, Deuschl G, Martinez-Martin P. Proposing a Parkinson's disease-specific tremor scale from the MDS-UPDRS. Mov Disord. 2015;30(8):1139-43; doi.10.1002/mds.26271

32. Regnault A, Boroojerdi B, Meunier J. et al. Does the MDS-UPDRS provide the precision to assess progression in early Parkinson's disease? Learnings from the Parkinson's progression marker initiative cohort. J Neurol 2019;266:1927–1936

33. Jonsson S, Yang S, Chen C, Plan EL, Karlsson MO. Sample size for detection of drug effect using item level and total score models for Unified Parkinson's Disease Rating Scale data, PAGE 27 (2018) Abstr 8638 [www.page-meeting.org/?abstract=8638]

34 Sheng Y, Yang S, Ma P, Chen C. Item response theory modelling of motor scores to investigate feasibility of reducing proof-of-concept trial for Parkinson's disease. PAGE 27 (2018) Abstr 8545 [www.page-meeting.org/?abstract=8545]

35. de Siqueira Tosin MH, Goetz CG, Luo S, Choi D, Stebbins GT. Item Response Theory Analysis of the MDS-UPDRS Motor Examination: Tremor vs. Nontremor Items [published online ahead of print, 2020 May 29]. Mov Disord. 2020;10.1002/mds.28110

36. Štochl J, Croudace TJ, Brožová H, Klempíř J, Roth J, Růžička E. Changes of hand preference in Parkinson's disease. J Neural Transm (Vienna). 2012;119(6):693-696.

Table 1. Demographics and baseline disease characteristics	**Table 1. Demographics and baseline disease ch**
Patient characteristics	**N=423**
Age (years)	
Mean (SD)	62.1 (9.7)
(Min, Max)	(34.2, 85.2)
Median	62.7
Sex, n (%)	
Male	277 (65.5)
Female	146 (34.5)
Race, n (%)	
White	399 (94.3)

Table 1. Demographics and baseline disease characteristics	Table 1. Demographics and baseline disease ch
Black	7 (1.7)
Asian	10 (2.4)
Other	7 (1.7)
Disease duration (months)	
Mean (SD)	6.7 (6.6)
(Min, Max)	(0, 36.5)
Median	4.1
Dominant hand, n (%)	
Left	38 (9)
Right	375 (88.7)
Mixed	10 (2.4)
MDS-UPDRS Part III score	
Mean (SD)	21.0 (9.0)
(Min, Max)	(4, 51)
Median	20

Table 2. Item-response model parameters and item importance
Item (j)
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30

Table 2. Item-response model parameters and item importance

31
32
33
Parameter definitions: aj: discrimination parameter; bj1 to bj4: difficulty parameters. RUE, right upper extremity; LUE, le

Table 3. Parameters for longitudinal models with or without tremor items	**Table 3. Parameters for longitudinal models with or without tremor items**	**Table 3. Parameters for longitudinal models with or without tremor items**	**Table 3. Parameters for longitudinal models with or without tremor items**	**Table 3. Parameters for longitudinal models with or without tremor items**
Model	Severity (all items)	Severity (non-tremor)	Sum of score (all items)	Sum of score (non-tremor)
Fixed effects (%RSE)	Fixed effects (%RSE)	Fixed effects (%RSE)	Fixed effects (%RSE)	Fixed effects (%RSE)
Baseline	0 (fixed) 0.227 (5)	0 (fixed) 0.243 (5)	19.6 (2) 2.99 (11)	15 (3) 2.24 (11)
Slope ($year^{-1}$)	-0.0545 (20)	-0.0568 (23)	-0.043 (33)	-0.0292 (49)
Effect of baseline on slope				
Inter-individual variability	Inter-individual variability	Inter-individual variability	Inter-individual variability	Inter-individual variability
$\omega^2_{baseline}$ ω^2_{slope}	1 (fixed) [a] 0.0365 [a]	1 (fixed) [a] 0.0436 [a]	0.171 [b] 0.565 [b]	0.252 [b] 0.585 [b]
Inter-occasion variability	Inter-occasion variability	Inter-occasion variability	Inter-occasion variability	Inter-occasion variability
$\sigma_{proportional}$ $\sigma_{additive}$	– 0.181	– 0.197	0.0567 –	0.0665 –
[a] additive variability; [b] exponential variability	[a] additive variability; [b] exponential variability	[a] additive variability; [b] exponential variability	[a] additive variability; [b] exponential variability	[a] additive variability; [b] exponential variability

Figure legends

Figure 1. Item response model for MDS-UPDRS Part III. Left: Item scores relate to underlying severity, which mirrors the sum of score, through Item Characteristic Curves (ICCs). Upper right: the position and steepness of ICCs reflect an item's difficulty and ability to differentiate patient severity, respectively. The blue, pink, green and red curves describe the probabilities of having a score of not lower than 1, 2, 3 and 4, respectively. Lower right: The blue, pink, green, red and yellow Category Characteristic Curves (CCCs) describe the probabilities of having a score of 0, 1, 2, 3 and 4, respectively.

Figure 2. Data pattern and model evaluation. The pattern of the observed Sum-of-Score data (upper left) was reproduced by modeled Symptom Severity (upper right). Model-estimated Category Characteristic Curves (lines) reflected the distribution of observed categories (circles) for each item over the range of symptom severity (lower left). The proportion of the simulated scores were compared with the observed scores (lower right).

Figure 3. Item informativeness. Item information over the whole spectrum of symptom severity shows

some items are far more informative than others. The color-coded areas represent the items, from bottom to top, in the order of decreasing information.

Figure 4. Trial probability of success. Upper: Probability of trial success for detecting a hypothetical drug's ability to slow down disease progression was higher when data were analyzed using Symptom Severity (brown) than using the Sum of Scores (green), where solid and dashed lines reflect analyses including all items and only non-tremor items, respectively. Lower: Comparison of power for detecting drug effect (green: 0.1; blue: 0.5) and overall probability of trial success (brown) for detecting a range of potential drug effects in a one-year trial.

Figure 5. Visual predictive check for the longitudinal item-response model. The time course of the distribution of the observed sum of scores was well reproduced by the longitudinal IRT model (dots: observations; green lines: 5%, 50% and 95% quantiles of the observations; red line: predicted time course of sum of score for a typical patient; bands: 95% confidence intervals of model simulated corresponding quantiles).

Figure 6. The model accurately simulated the time course of the observed proportion of each score for each item. The lines are the proportion of the observed scores of 0 to 5. The bands are the 95% confidence intervals of the model simulation.

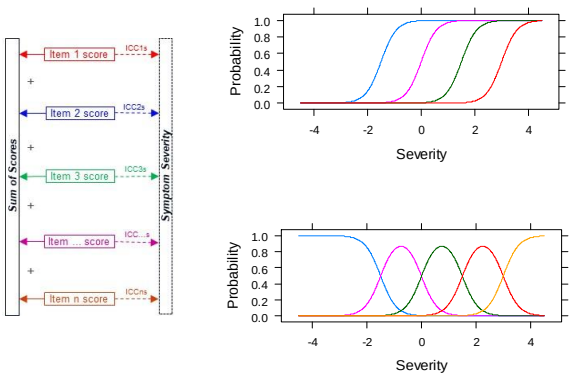

Figure 1. Item response model for MDS-UPDRS Part III. Left: Item scores relate to underlying severity, which mirrors the sum of score, through Item Characteristic Curves (ICCs). Upper right: the position and steepness of ICCs reflect an item's difficulty and ability to differentiate patient severity, respectively. The blue, pink, green and red curves describe the probabilities of having a score of not lower than 1, 2, 3 and 4, respectively. Lower right: The blue, pink, green, red and yellow Category Characteristic Curves (CCCs) describe the probabilities of having a score of 0, 1, 2, 3 and 4, respectively.

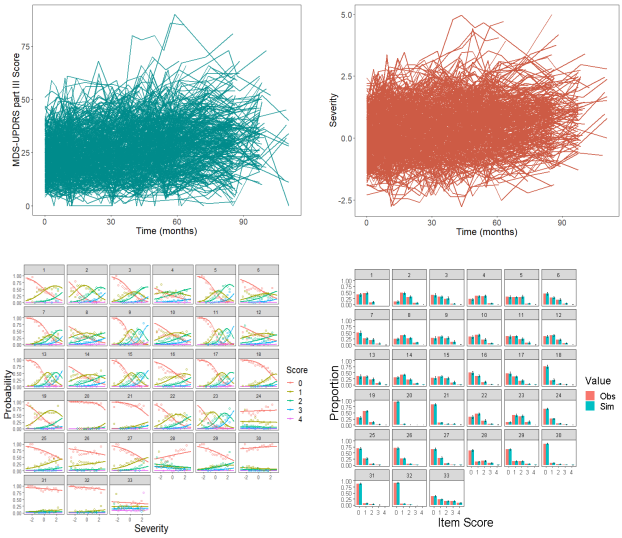

Figure 2. Data pattern and model evaluation. The pattern of the observed Sum-of-Score data (upper left) was reproduced by modeled Symptom Severity (upper right). Model-estimated Category Characteristic Curves (lines) were overlaid on the distribution of observed categories (circles) for each item over the range of symptom severity (lower left). The proportion of the simulated scores were compared to the observed scores (lower right).

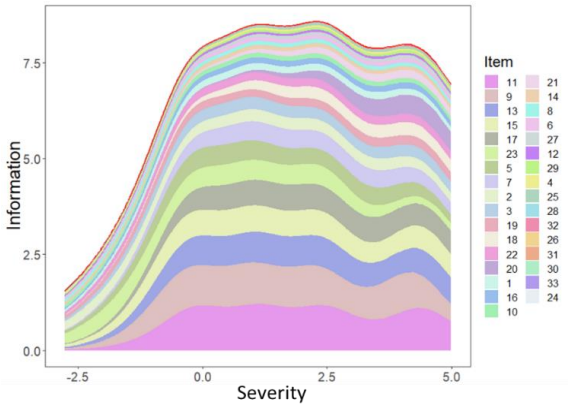

Figure 3. Item informativeness. Item information over the whole spectrum of symptom severity shows some items are far more informative than others. The color-coded areas represent the items, from bottom to top, in the order of decreasing information.

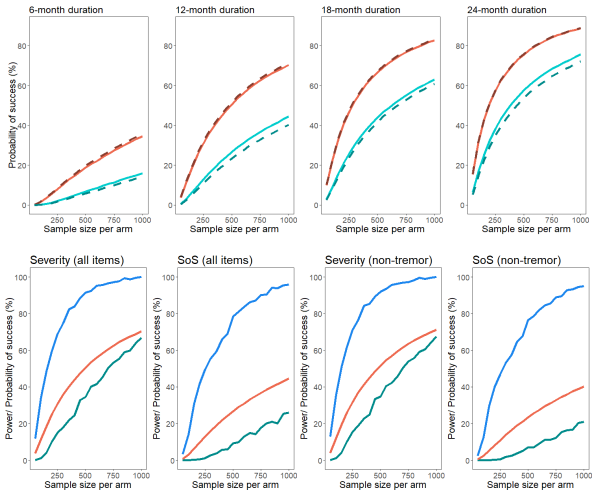

Figure 4. Trial probability of success. Upper: Probability of trial success for detecting a hypothetical drug's ability to slow down disease progression was higher when data were analyzed using Symptom Severity (brown) than using the Sum of Scores (green), where solid and dashed lines reflect analyses including all items and only non-tremor items, respectively. Lower: Comparison of power for detecting drug effect (green: 0.1; blue: 0.5) and overall probability of trial success (brown) for detecting a range of potential drug effects in a one-year trial.

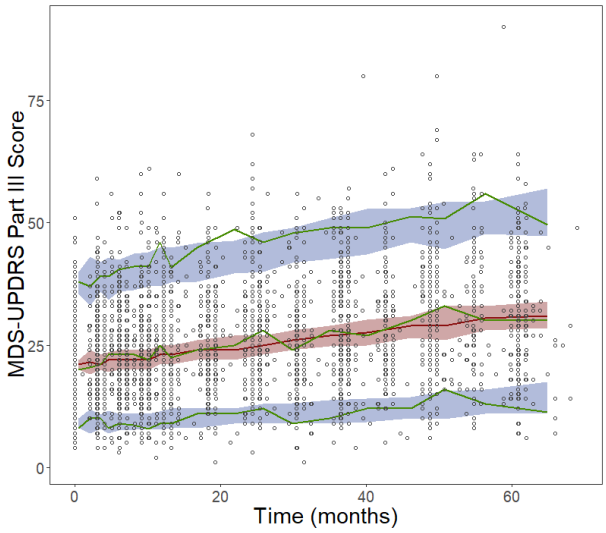

Figure 5. Visual predictive check for the longitudinal item-response model. The time course of the distribution of the observed sum of scores was well reproduced by the longitudinal IRT model (dots: observations; green lines: 5%, 50% and 95% quantiles of the observations; red line: predicted time course of sum of score for a typical patient; bands: 95% confidence intervals of model simulated corresponding quantiles).

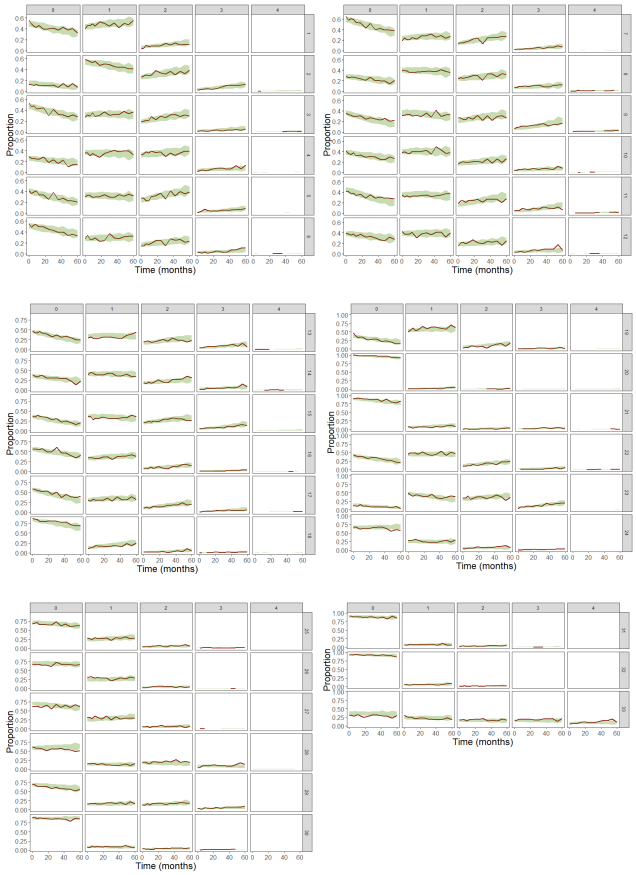

Figure 6. The model accurately simulated the time course of the observed proportion of each score for each item. The lines are the proportion of the observed scores of 0 to 5. The bands are the 95% confidence intervals of the model simulation.