MSJ Experimental Translational Clinical

Original Article

Analysis of the psychometric properties of the Multiple Sclerosis Impact Scale-29 (MSIS-29) in relapsing–remitting multiple sclerosis using classical and modern test theory

Multiple Sclerosis Journal – Experimental, Translational and Clinical

2: 1–13

DOI: 10.1177/2055217316673235

ED Bacci, KW Wyrwich, GA Phillips, T Vollmer and S Guo

Abstract

Background: Investigations using classical test theory support the psychometric properties of the original version of the Multiple Sclerosis Impact Scale (MSIS-29v1), a disease-specific measure of multiple sclerosis (MS) impact (physical and psychological subscales). Later, assessments of the MSIS-29v1 in an MS community-based sample using Rasch analysis led to revisions of the instrument's response options (MSIS-29v2).

Objective: The objective of this paper is to evaluate the psychometric properties of the MSIS-29v1 in a clinical trial cohort of relapsing–remitting MS patients (RRMS).

Methods: Data from 600 patients with RRMS enrolled in the SELECT clinical trial were used. Assessments were performed at baseline and at Weeks 12, 24, and 52. In addition to traditional psychometric analyses, Item Response Theory (IRT) and Rasch analysis were used to evaluate the measurement properties of the MSIS-29v1.

Results: Both MSIS-29v1 subscales demonstrated strong reliability, construct validity, and responsiveness. The IRT and Rasch analysis showed overall support for response category threshold ordering, person-item fit, and item fit for both subscales.

Conclusions: Both MSIS-29v1 subscales demonstrated robust measurement properties using classical, IRT, and Rasch techniques. Unlike previous research using a community-based sample, the MSIS-29v1 was found to be psychometrically sound to assess physical and psychological impairments in a clinical trial sample of patients with RRMS.

***Keywords*:** Multiple Sclerosis Impact Scale, reliability, validity, responsiveness, Rasch model, item response theory, graded response model, relapsing–remitting multiple sclerosis

Date received: 26 April 2016; accepted: 7 September 2016

Introduction

Patient-reported outcomes (PROs) are increasingly being used in clinical trials to evaluate how a disease affects health and well-being from the patient's perspective.[1] The importance of incorporating the patient view in clinical research is reflected in the development of organizations such as the Patient-Centered Outcomes Research Institute (PCORI), established to support research designed to improve patient care through a patient-centeredness approach (www.pcori.org). In patients with multiple sclerosis (MS), various PROs have been developed, such as assessments of functional ability (Functional Assessment of Multiple Sclerosis[2]), health-related quality of life (e.g. Hamburg Quality of Life Questionnaire in Multiple Sclerosis[3]), and symptoms (Patient-Reported Indices for Multiple Sclerosis[4]).

One instrument increasingly incorporated into clinical trials of MS[5–9] is the Multiple Sclerosis Impact Scale (MSIS-29),[10] a disease-specific PRO developed to examine the physical and psychological

Correspondence to:
Elizabeth Dansie Bacci
Evidera Inc, 1417 4th Ave., Suite 510, Seattle, WA 98101, USA
elizabeth.bacci@evidera.com

ED Bacci, KW Wyrwich
Evidera Inc, Outcomes Research, USA

GA Phillips,
Biogen, Value-Based Medicine, USA

T Vollmer,
University of Colorado School of Medicine, Department of Neurology, USA

S Guo,
Evidera Inc, Modeling and Simulation, USA

impact of MS. The measure consists of two subscales, a 20-item scale measuring physical impact and a nine-item scale measuring psychological impact. All items have a Likert-type response format (''*Not at all,'' ''A little,'' ''Moderately,'' ''Quite a lot,''* and *''Extremely''*). Multiple investigations using traditional psychometric analyses based on classical test theory have been conducted to assess the psychometric properties of the instrument, providing evidence of the instrument's reliability, validity, and responsiveness.[11–20]

As with all scales, additional validation assessments are required in a range of populations, using a variety of methods. These methods include modern psychometric techniques like Rasch analysis and Item Response Theory (IRT) used to evaluate item-level performance of a scale. Some of the benefits of using both of these newer psychometric approaches include an ability to: examine latent trait estimates that do not vary with the characteristics of the population, estimate item difficulty and discrimination, assess person fit to a measure, and determine if response categories are ordered properly and function as intended.[21]

To this end, the MSIS-29 has been evaluated using Rasch measurement.[22] Hobart and Cano[22] examined the properties of the MSIS-29 using Rasch measurement in a community-based sample of 1725 individuals in the United Kingdom (UK), finding that the five-category item scoring did not function as intended for nine items in the physical impact subscale and one item in the psychological subscale. There were either too many or overlapping response options, thus the MSIS-29 was revised from its original five-category item scoring (MSIS-29 version 1 (MSIS-29v1)) to a four-category scoring (MSIS-29v2),[22] including categories of ''*Not at all,*'' ''*A little,*'' ''*Moderately,*'' and ''*Extremely.''* In a subsequent Rasch investigation of the MSIS-29v1 in an Australian community-based sample, Ramp and colleagues[23] found 11 of 20 MSIS-29 physical impact items demonstrated some threshold disordering, concluding response options categories for this subscale should be reduced from five to three (i.e. ''*A little bit*,'' ''*Moderately*,'' and ''*Quite a bit''* could be replaced by ''*Moderately*'') to improve item performance.[23]

However, the performance of the MSIS-29v1 has not been evaluated: (1) using a clinical trial-based sample (versus community-based populations), or (2) under a less restrictive IRT model. The philosophical difference between the application of the Rasch model and a less restrictive IRT model is important to recognize. In the Rasch paradigm previously used among the community samples,[22,23] the emphasis is on identifying and studying measurement anomalies in the data disclosed by the Rasch model. However, other IRT models introducing additional fit parameters (e.g. slopes) emphasize the opportunity for finding a model that best characterizes the given data for an instrument that has demonstrated strong measurement properties, with any challenges to that fit assisting the research team to better understanding specific measurement problems. Therefore, objectives of the current analyses were to: (1) confirm the psychometric properties of the MSIS-29v1 using classical test theory to assess for scale reliability, construct validity, and ability to detect change in patients with relapsing–remitting multiple sclerosis (RRMS) enrolled in a 52-week clinical trial; (2) assess item performance of the MSIS-29v1 using a Graded Response Model (GRM) IRT analysis; and (3) evaluate the MSIS-29v1 using Rasch analysis in this clinical trial sample.

Methods

Study design and data source

Data used for this analysis were from the SELECT (NCT00390221) study,[24] a 52-week randomized, double-blind, placebo-controlled multicenter study conducted to assess the efficacy and safety of daclizumab high-yield process (DAC HYP) in patients with RRMS, where reducing the annualized relapse rate was the primary endpoint. Patients were randomized into one of three groups and received 150 mg DAC HYP, 300 mg DAC HYP, or placebo, administered subcutaneously every four weeks for 52 weeks. Institutional review board approval was obtained prior to patient enrollment.

Eligible patients for SELECT were men and women between 18 and 55 years, diagnosed with RRMS according to McDonald criteria,[25] had an Expanded Disability Status Scale (EDSS) score between 0.0 and 5.0,[26] and had experienced ≥ 1 confirmed MS relapse in the 12 months before randomization or ≥ 1 new gadolinium-enhancing lesion on the brain as confirmed by magnetic resonance imaging ≤ 6 weeks prior to randomization. A total of 621 patients were enrolled in SELECT; the current study population consisted of the modified intention-to-treat (ITT) population, defined as all ITT patients who received ≥ 1 dose of DAC HYP or placebo and completed ≥ 1 post-baseline (Week 12, 24 or 52) MSIS-29v1 assessment.

Statistical methods

Classical test theory. Three psychometric properties of the MSIS-29v1 physical and psychological subscales were examined using classical test theory, including reliability (internal consistency and test-retest), convergent validity, and responsiveness. A description of methods is provided in the online supplement.

Modern test theory

GRM. As the primary analysis, the psychometric scaling of the MSIS-29v1 physical and psychological subscales was examined separately using Samejima's[27] GRM at baseline and Week 52. The GRM of IRT is appropriate for ordered categorical item responses. The two sets of items were assessed for ordering of item characteristic curves (ICCs), slope and item fit, and person-item fit. An inspection of ICCs is used to determine if patients with high levels of the measured attribute (e.g. physical impact of MS) consistently endorse high-scoring response options indicating greater severity across all items, while patients with low levels should endorse low-scoring responses. The items of the MSIS-29v1 were developed to have ordered categorical response thresholds, where threshold parameters represent the trait level needed to have a 50% probability of responding in category k or higher. Disordered thresholds occur when respondents inconsistently endorse response categories (e.g. someone with greater physical impact endorses a response option indicating lower physical impact).

The slope, or discrimination parameter, represents the strength of the association between the item and the underlying construct. Higher values are associated with items that are better able to discriminate between adjacent trait levels, and provide greater information about a patient than less discriminating items. However, slope parameters > 4.0 were used to indicate that an item is possibly redundant with the latent variable.[28] Item fit was also assessed using the likelihood ratio S-G^2 and Pearson's S-X^2 fit statistics,[29] used to assess the difference between observed values and model-based predicted values. A value of $p < 0.001$ was used to indicate misfit.

Finally, distributions of item threshold location and person location estimates were reviewed to determine if the thresholds of the item set cover the range of severity demonstrated by the patient population. The axis for such displays is on a logit scale and represents the assumed unidimensional measure of the latent variable, in this case severity of MS impact. Ideally, items in a scale should be able to successfully measure the range of severity as demonstrated by the individuals completing the scale. MULTILOG IRT software was used to fit the GRM.[30]

Rasch analysis. As an additional analysis, model fit of the MSIS-29v1 in this clinical trial sample was assessed using a Rasch measurement approach.[31] Similar to the GRM, the greater a patient's physical and psychological impact relative to the degree of impact assessed by an item, the higher the probability of a positive response to that item. However, the Rasch model assumes that all items have uniform discrimination power between high and low severity, thus the slope is fixed and the modeling is more restrictive. Like the GRM, the properties of both subscales were assessed using Rasch measurement through an examination of ICC ordering, item fit statistics, and person-item threshold distributions, in addition to response threshold ordering. Using Rasch measurement, an item was marked as misfitting using a chi-square and fit residual. The chi-square value is a measure of the interaction between each item and the trait (i.e. impact of MS) being measured by those items; misfit was considered when the chi-square p value of an item was less than the alpha value ($p = 0.05$) with a Bonferroni correction. The fit residual considers the fit of the data in the population (observed data) to the Rasch model; a large negative fit residual value demonstrates an over-discriminating item (< -3.0); that is, the information provided by this item does not add additional value to the measurement. A high positive residual value (> 3.0) demonstrates that the item is under-fitting, indicating that the item is not discriminating differences in severity. The software RUMM2030[32] was used for the Rasch analyses.

Results

Baseline demographics, clinical characteristics, and PRO scores for the ITT efficacy population from SELECT ($N = 600$) are shown in Table 1. Across groups, most patients were female (63%–68%), with 1.3–1.4 relapses in the past year and a mean EDSS score of 2.6–2.8. All baseline characteristics and PRO scores were similar across groups, thus all further analyses collapsed across treatment and placebo groups. Relatively few (<5%) patients were missing any PRO items.

Table 1. Baseline demographics and characteristics from SELECT.

Characteristic	DAC HYP 150 mg ($n = 201$)	DAC HYP 300 mg ($n = 203$)	Placebo ($n = 196$)
Age, y	35.2 (9.1)	35.4 (8.6)	36.9 (9.0)
Female, *n* (%)	136 (67.7)	132 (65.0)	123 (62.8)
Disease duration, y	4.5 (5.0)	3.8 (4.0)	4.1 (5.3)
Number of relapses in past year	1.4 (0.7)	1.3 (0.7)	1.3 (0.6)
EDSS score	2.8 (1.1)	2.6 (1.2)	2.7 (1.2)
MSIS-29v1			
Physical Impact Subscale	24.7 (20.2)	24.0 (19.5)	26.3 (22.0)
Psychological Impact Subscale	28.6 (21.5)	29.6 (20.7)	29.5 (22.5)
SF-12			
PCS	42.9 (9.9)	43.1 (9.0)	42.5 (10.0)
MCS	46.1 (11.5)	45.5 (11.0)	46.4 (10.2)
EQ-5D			
VAS	72.0 (17.4)	72.1 (18.1)	71.2 (18.3)
Summary Health Index	0.7 (0.2)	0.7 (0.2)	0.7 (0.2)

y: years; DAC HYP: daclizumab high-yield process; EDSS: Expanded Disability Status Scale; EQ-5D: EuroQol 5-Dimensions; MCS: mental component summary; MSIS-29v1: Multiple Sclerosis Impact Scale; PCS: physical component summary; SF-12: Short-Form Health Survey-12; VAS: visual analog scale.
Values are reported as mean (standard deviation), except where noted.

Classical test theory

The results of the assessment of the reliability, validity, and responsiveness of the MSIS-29v1 physical and psychological impact scales are described and presented in the online supplement.

Modern test theory

GRM analysis. A visual examination of the ICCs displayed no disordering and only one item with a response option that overlapped with an adjacent response (Figures 1(a) and (b)). Specifically, Figure 1(a) (b) shows that the response option ''*moderately''* for MSIS-29v1 psychological impact item Q2 at baseline overlapped with the response options ''*A little''* and ''*Quite a lot*.''

Table 2(a) and (b) present the item slopes and fit statistics, which indicated that item discriminations were moderate to high for all items at baseline and Week 52. For two items (Q12-physical impact subscale, Q5-psychological impact subscale), the slopes exceeded the 4.0 threshold at Week 52. However, the item fit statistics S-G^2 and S-X^2 demonstrated every item fit the predicted GRM model at both time points for both subscales; no p value was less than 0.001.

Figures 2(a) and (b) provide baseline and Week 52 person-item threshold maps for the MSIS-29v1 physical and psychological impact scales. For the physical impact domain at baseline, the thresholds are well distributed; however, there is evidence for a floor effect as the sample is concentrated in the lower half of the item threshold location range. This indicates the scale is assessing more severe impact than present in the current sample. However, this floor effect is less pronounced at Week 52. The psychological impact scale at baseline and Week 52, in contrast, displays a distribution of item threshold locations more appropriate for the current population as more item thresholds are found in the lower region that better match the person location distribution.

Rasch analysis. In the Rasch analysis of this clinical trial data, all category thresholds for all of the MSIS-29v1 items were ordered properly at both time points, with all five response options assessing an independent range on the scale (online supplement Tables S3(a) and (b)). This finding was supported by a visual examination of the ICC plots (not shown), which displayed no disordering. However, the ICC plots demonstrated that the response option *''A little''* in Q11 of the physical impact subscale and Q2 of the psychological subscale was not completely distinct from responses options *''Not at all''* and ''*Moderately*,'' reflecting the findings of the GRM for Q2 of the psychological subscale. The item fit statistics from the Rasch analysis (Tables S3(a) and

Table 2. (a) Graded response model item parameters and fit statistics for MSIS-29v1 Physical and Psychological Impact Subscales—baseline.

			Category threshold[a]				Item fit statistics[b]			
Number	Descriptor	Slope[a]	β_1	β_2	β_3	β_4	S-G^2	S-G^2 p value	S-X^2	S-X^2 p value
Physical Subscale										
Q1	Do physically demanding tasks	3.11	−0.90	0.07	0.94	2.28	50.02	0.4724	50.65	0.4477
Q2	Grip things tightly	2.07	0.15	0.97	1.98	3.31	26.60	0.9868	26.85	0.9855
Q3	Carry things	2.27	−0.46	0.42	1.35	2.23	48.36	0.4988	42.43	0.7349
Q4	Problems with balance	2.47	−0.84	0.15	0.96	2.11	41.42	0.8533	41.62	0.8480
Q5	Difficulties moving about indoors	2.97	−0.03	0.84	1.82	2.85	25.69	0.9875	24.90	0.9910
Q6	Being clumsy	2.79	−0.52	0.51	1.46	2.28	35.81	0.8344	34.47	0.8726
Q7	Stiffness	2.36	−0.26	0.62	1.62	2.87	52.59	0.3369	49.40	0.4570
Q8	Heavy arms and/or legs	2.44	−0.70	0.30	1.23	2.39	43.30	0.7698	40.97	0.8413
Q9	Tremor of arms/legs	1.84	−0.20	0.84	1.77	2.97	41.15	0.7473	38.77	0.8267
Q10	Spasms in limbs	1.73	−0.07	0.90	1.81	3.21	43.52	0.6940	44.30	0.6639
Q11	Body not doing what you want it to do	2.48	−0.08	0.89	1.53	2.59	43.09	0.5530	42.05	0.5977
Q12	Having to depend on others to do things for you	3.38	0.16	0.85	1.44	2.14	16.95	1.0000	16.60	1.0000
Q13	Limitations in social/leisure activities at home	2.78	−0.06	0.94	1.74	2.94	28.67	0.9839	26.95	0.9917
Q14	Being stuck at home more than would like	2.33	−0.08	0.72	1.36	2.25	35.49	0.8688	34.02	0.9044
Q15	Difficulties using hands in everyday tasks	2.52	0.13	0.91	1.77	2.60	49.29	0.3055	47.04	0.3891
Q16	Having to cut down time spent on work/daily activities	2.75	−0.37	0.50	1.47	2.45	45.95	0.5976	45.46	0.6173
Q17	Problems using transport	2.75	0.12	0.85	1.55	2.33	41.60	0.6949	39.22	0.7829
Q18	Taking longer to do things	3.44	−0.54	0.42	1.21	2.37	25.91	0.9900	25.42	0.9918
Q19	Difficulty doing things spontaneously	2.54	−0.03	0.68	1.35	2.17	24.42	0.9963	23.17	0.9980
Q20	Needing to go to the toilet urgently	1.69	−0.11	0.77	1.51	2.50	55.23	0.2838	51.46	0.4162
Psychological subscale										
Q1	Feeling unwell	2.25	−0.66	0.54	1.57	2.74	34.16	0.7298	33.16	0.7695
Q2	Problems sleeping	1.44	−0.15	1.09	1.87	2.86	47.59	0.2914	51.88	0.1663
Q3	Mentally fatigued	2.60	−0.45	0.54	1.34	2.37	51.65	0.1026	49.66	0.1407
Q4	Worries about our MS	1.89	−1.18	0.25	1.14	2.19	49.98	0.1588	47.65	0.2204
Q5	Anxious or tense	3.39	−0.64	0.44	1.25	2.18	47.64	0.0929	47.24	0.0994
Q6	Irritable, impatient, or short-tempered	2.62	−0.71	0.25	1.21	2.22	28.61	0.8649	29.37	0.8412
Q7	Problems concentrating	2.75	−0.54	0.50	1.35	2.44	39.68	0.3516	37.91	0.4277
Q8	Lack of confidence	3.09	−0.38	0.63	1.43	2.19	30.13	0.7024	29.00	0.7522
Q9	Feeling depressed	3.20	−0.15	0.83	1.59	2.40	28.26	0.8485	26.58	0.8978

[a]Using IRT software MULTILOG. [b]Using SAS macros IRTFIT. MSIS-29v1: Multiple Sclerosis Impact Scale.

Table 2. (b) Graded response model item parameters and fit statistics for MSIS-29v1 Physical and Psychological Impact Subscales—Week 52.

			Category threshold[a]				Item fit statistics[b]			
Number	Descriptor	Slope[a]	β_1	β_2	β_3	β_4	S-G^2	S-G^2 p value	S-X^2	S-X^2 p value
Physical Subscale										
Q1	Do physically demanding tasks	3.25	−0.95	0.01	0.95	2.20	35.15	0.8272	32.64	0.8964
Q2	Grip things tightly	2.42	−0.14	0.69	1.62	2.68	27.20	0.9834	25.79	0.9905
Q3	Carry things	2.71	−0.56	0.30	1.09	2.30	51.33	0.3080	46.87	0.4778
Q4	Problems with balance	2.65	−0.90	0.03	1.05	1.96	34.85	0.9365	34.26	0.9455
Q5	Difficulties moving about indoors	3.48	−0.19	0.69	1.57	2.79	34.29	0.7953	33.07	0.8362
Q6	Being clumsy	3.18	−0.68	0.37	1.27	2.28	35.39	0.7886	33.93	0.8372
Q7	Stiffness	2.87	−0.47	0.43	1.35	2.45	44.47	0.4095	42.72	0.4834
Q8	Heavy arms and/or legs	2.55	−0.69	0.25	1.08	2.42	51.43	0.2367	48.39	0.3378
Q9	Tremor of arms/legs	1.99	−0.27	0.74	1.68	2.68	39.56	0.7012	38.80	0.7309
Q10	Spasms in limbs	1.97	−0.37	0.68	1.67	2.75	29.38	0.9843	29.02	0.9862
Q11	Body not doing what you want it to do	3.22	−0.28	0.69	1.47	2.30	31.11	0.9114	29.24	0.9459
Q12	Having to depend on others to do things for you	4.07	−0.10	0.68	1.41	2.19	24.35	0.9679	23.33	0.9779
Q13	Limitations in social/leisure activities at home	3.31	−0.28	0.73	1.61	2.41	22.34	0.9923	21.12	0.9957
Q14	Being stuck at home more than would like	3.18	−0.34	0.51	1.19	1.99	28.77	0.9401	27.85	0.9542
Q15	Difficulties using hands in everyday tasks	2.94	−0.10	0.82	1.63	2.63	37.08	0.6865	36.55	0.7085
Q16	Having to cut down time spent on work/daily activities	3.08	−0.64	0.35	1.41	2.59	25.91	0.9900	25.64	0.9910
Q17	Problems using transport	3.01	−0.10	0.68	1.45	2.20	32.07	0.9092	31.97	0.9114
Q18	Taking longer to do things	3.81	−0.63	0.31	1.13	2.33	32.18	0.8363	30.96	0.8729
Q19	Difficulty doing things spontaneously	3.23	−0.27	0.55	1.14	1.94	20.26	0.9973	18.56	0.9990
Q20	Needing to go to the toilet urgently	2.05	−0.38	0.50	1.23	2.22	38.93	0.7604	37.60	0.8064
Mental subscale										
Q1	Feeling unwell	2.46	−0.73	0.49	1.58	2.56	37.95	0.3805	36.68	0.4371
Q2	Problems sleeping	1.78	−0.19	0.77	1.65	2.66	48.23	0.1235	50.05	0.0912
Q3	Mentally fatigued	3.13	−0.56	0.54	1.42	2.17	22.48	0.9347	21.80	0.9476
Q4	Worries about our MS	2.11	−0.81	0.43	1.44	2.46	35.38	0.5912	34.24	0.6442
Q5	Anxious or tense	4.77	−0.50	0.47	1.35	2.34	21.45	0.8734	20.81	0.8937
Q6	Irritable, impatient, or short-tempered	2.81	−0.83	0.30	1.15	2.49	35.99	0.4221	35.59	0.4407
Q7	Problems concentrating	3.07	−0.54	0.46	1.40	2.44	24.48	0.9081	23.18	0.9372
Q8	Lack of confidence	3.32	−0.42	0.63	1.39	2.21	27.81	0.8339	24.70	0.9226
Q9	Feeling depressed	3.67	−0.16	0.75	1.46	2.20	29.88	0.6700	29.83	0.6722

[a]Using IRT software MULTILOG. [b]Using SAS macros IRTFIT. MSIS-29v1: Multiple Sclerosis Impact Scale.

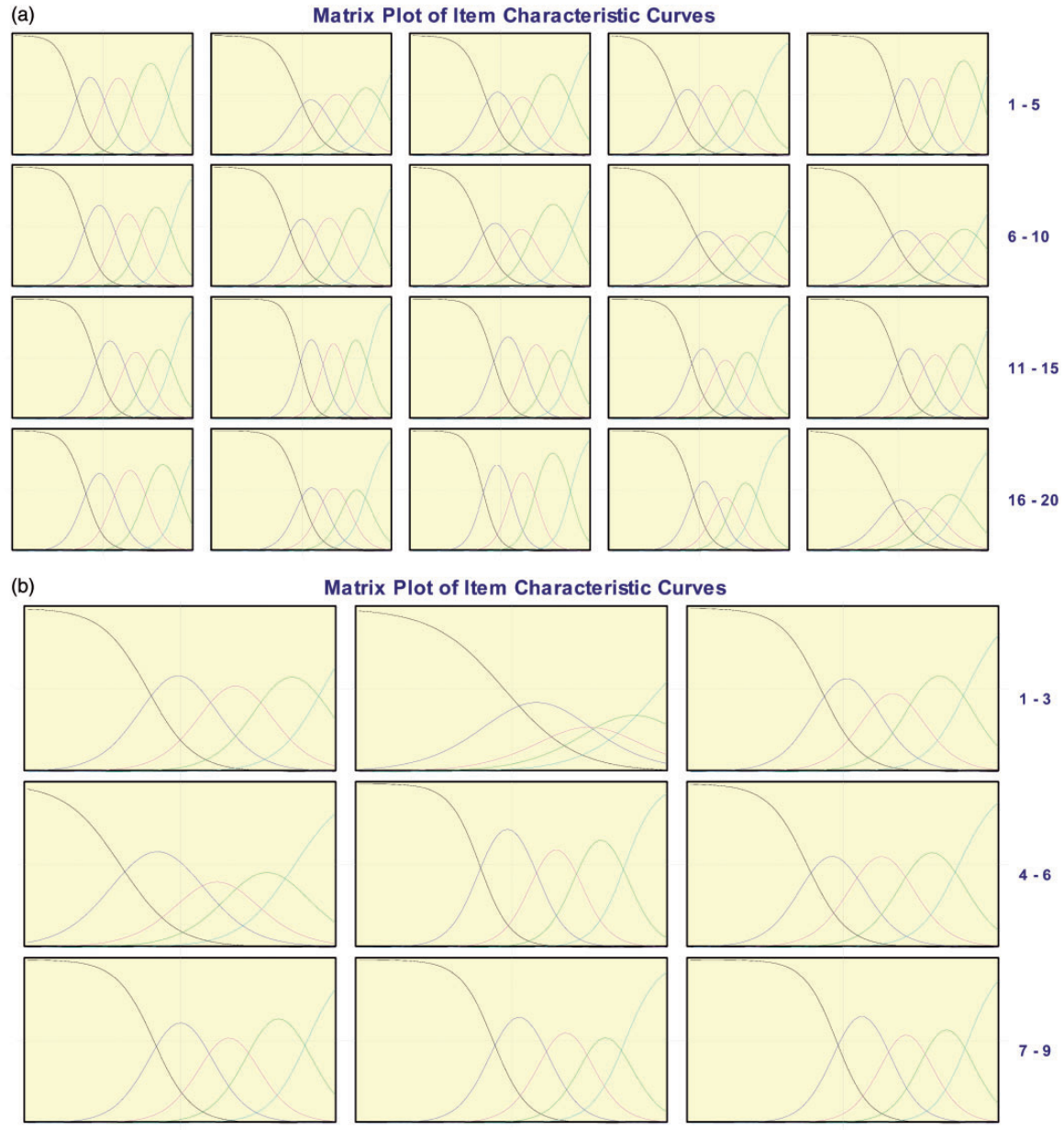

Figure 1. (a) Graded response model item characteristic curves for Multiple Sclerosis Impact Scale (MSIS)-29v1. (a) Physical impact subscale and (b) psychological impact subscale—baseline..

(b)) indicated item fit was acceptable for 75% and 70% of the physical impact items at baseline and Week 52, respectively, and 67% and 77% of the psychological impact items. For example, large fit residuals and statistically significant chi-square values were found for items Q18 and Q20 of the physical impact subscale at baseline.

Finally, the person-item threshold maps for both subscales of the MSIS-29v1 at baseline and Week 52 indicated both scales generally assess the entire range of patient responses (Figures S2(a) and (b)). However, for both subscales at baseline and Week 52, the lower end of the person severity distribution (least severe patients) was not assessed well by the MSIS-29v1 items when modeled using Rasch analysis. Specifically, the logit range for item responses did not match the logit range for the person responses at the lower end of the scale.

Discussion

The aim of the study was to use classical and modern test theory methods to assess the

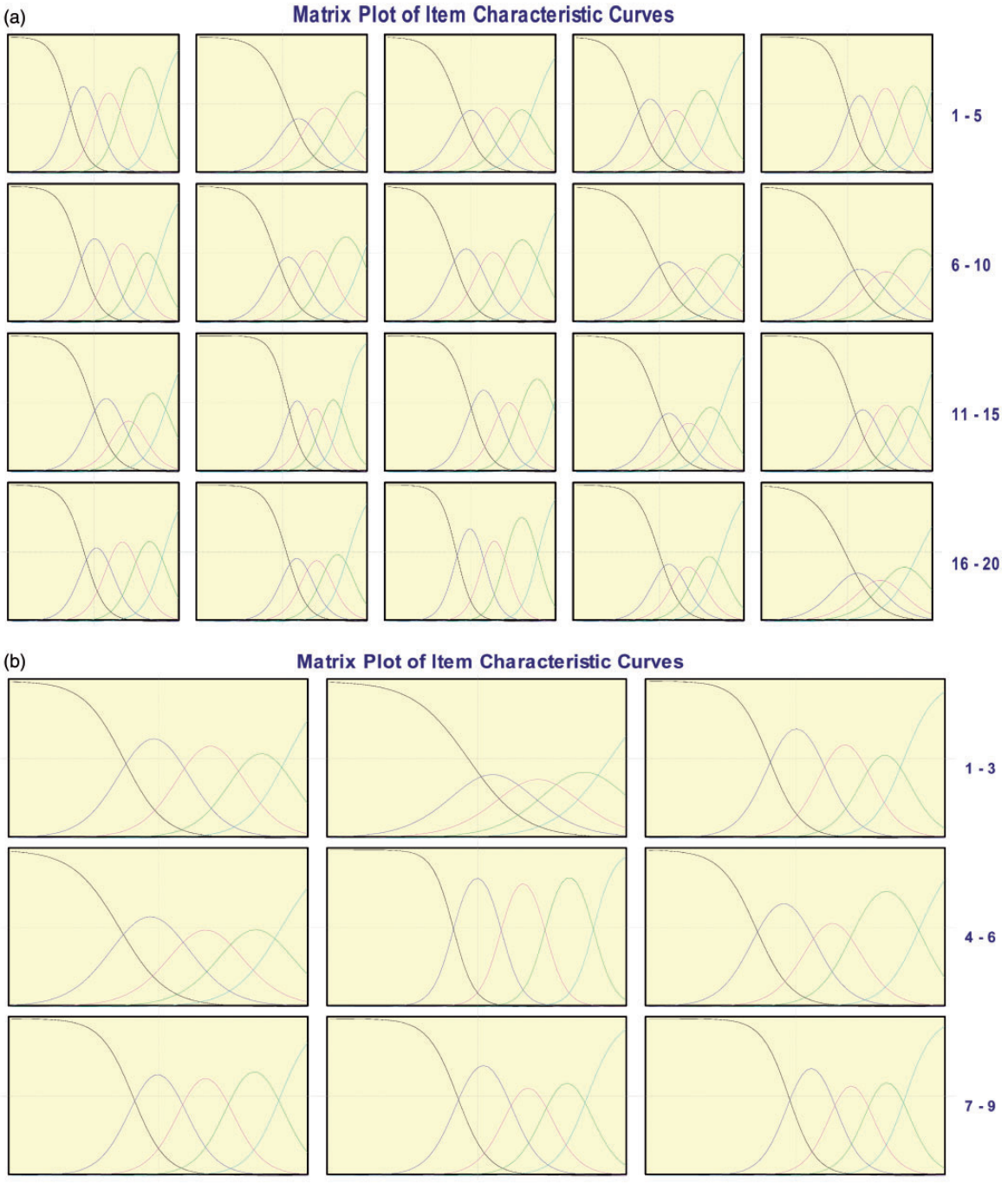

Figure 1. (b) Graded response model item characteristic curves for Multiple Sclerosis Impact Scale (MSIS)-29v1. (a) Physical impact subscale and (b) psychological impact subscale—Week 52.

psychometric properties of the MSIS-29v1 in a clinical trial population. Multiple analytic techniques were used to assess the properties of the MSIS-29v1 at various time points in a sample of patients with RRMS enrolled in a 52-week clinical trial. Through these analyses, evidence was generated to indicate that the MSIS-29v1 functions well in a clinical trial population across time.

Much like multiple previous studies using classical test theory methods in community-based populations,[11–20] the current study using the SELECT clinical trial population found support for the internal consistency and test-retest reliability, construct validity, and responsiveness of the MSIS-29v1 longitudinally over 52 weeks in patients with RRMS.

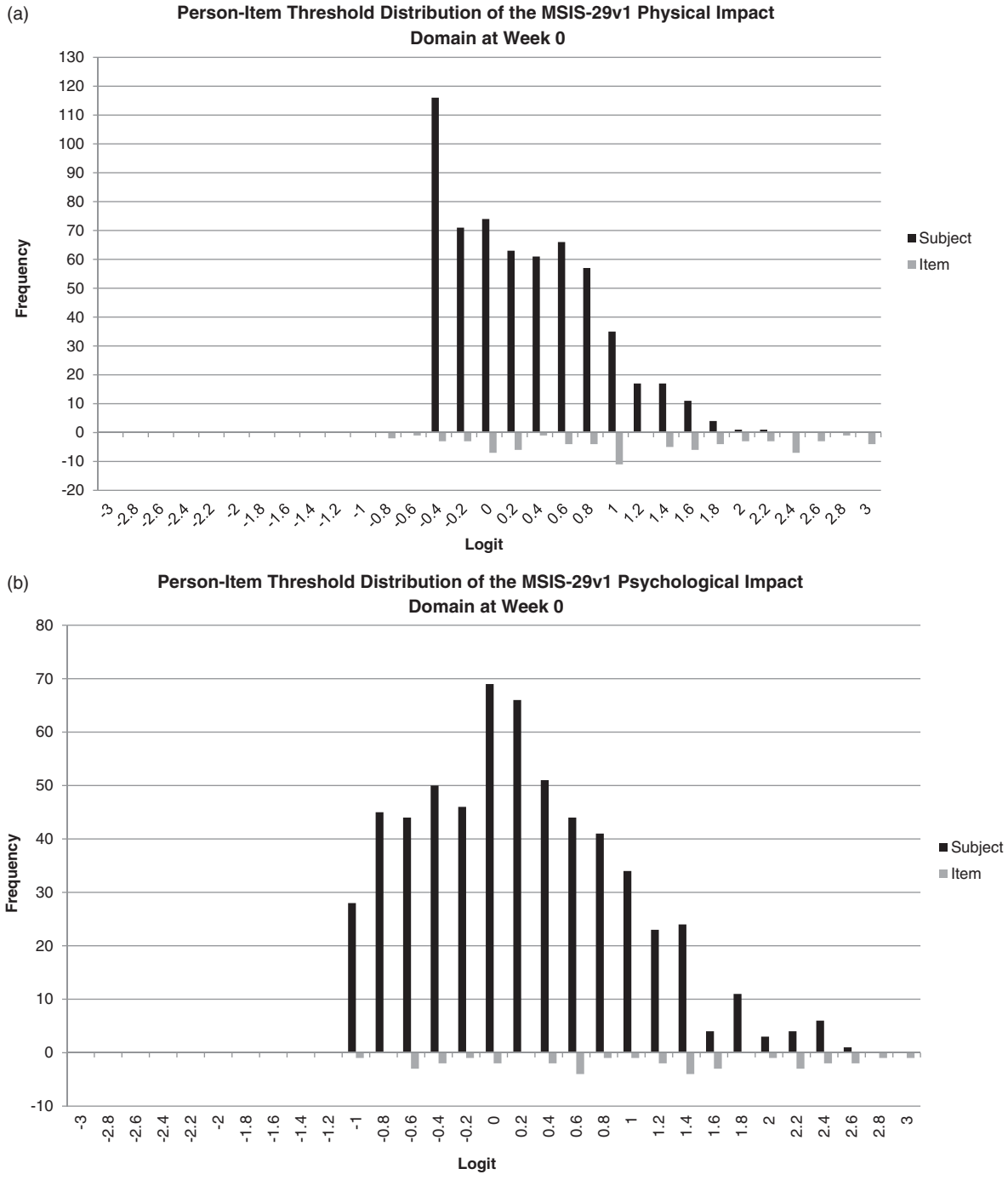

Figure 2. (a) Graded response model person-item threshold distribution for Multiple Sclerosis Impact Scale (MSIS)-29v1. (a) Physical impact subscale and (b) psychological impact subscale—baseline.

Complementing these findings, overall support was established for the psychometric properties of both subscales of the MSIS-29v1 using the modern psychometric method of GRM. These findings included evidence of ordered item-response categories through an inspection of ICCs, acceptable indicators of item fit, and a high degree of person fit to the measure. In an extensive re-evaluation of the MSIS-29v1 using Rasch analysis, Hobart and Cano[22] provided evidence that the five-category scoring system did not function properly in their community-based population, as disordered thresholds were found in nine items in the physical impact subscale and one item in the psychological

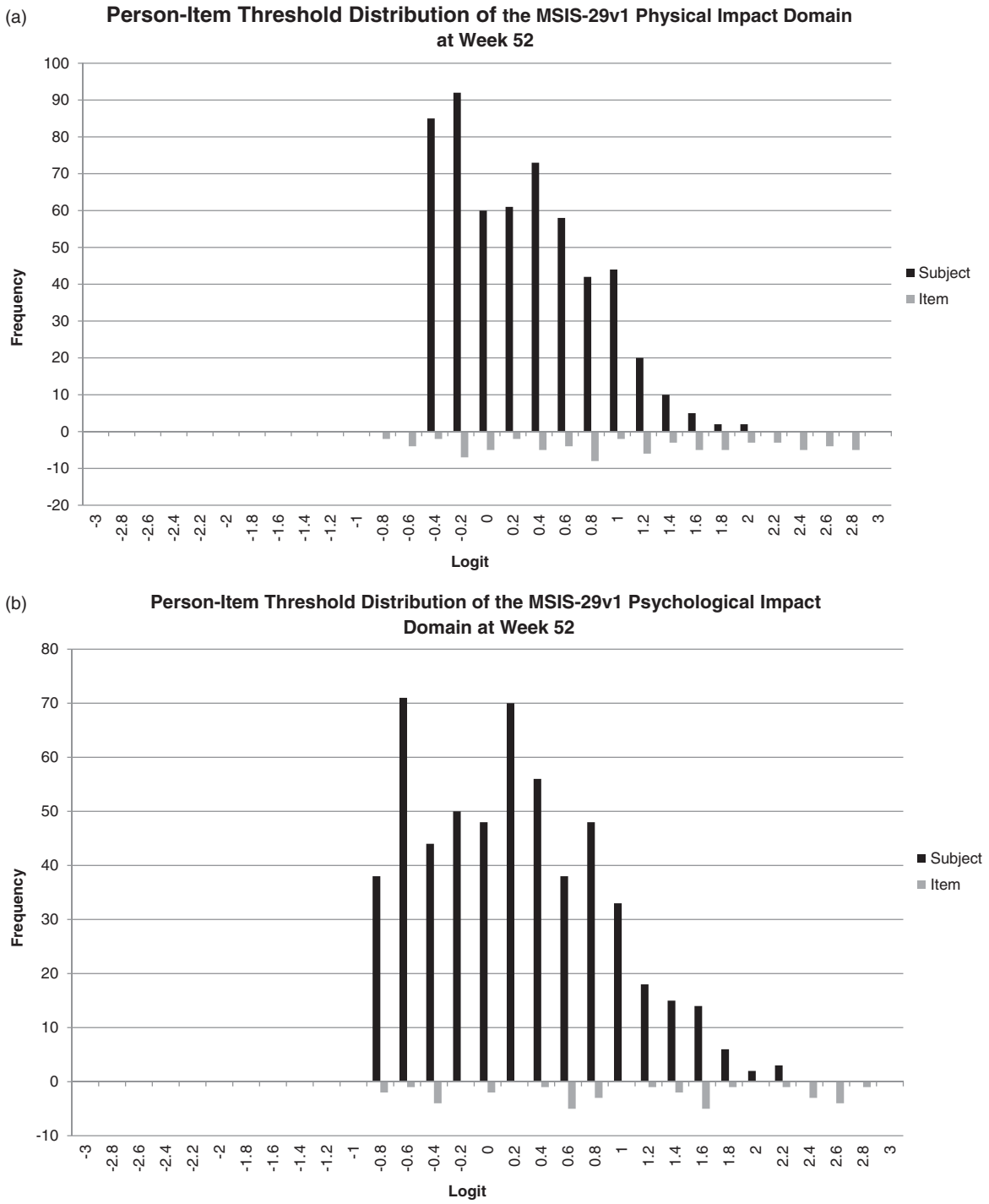

Figure 2. (b) Graded response model person-item threshold distribution for Multiple Sclerosis Impact Scale (MSIS)-29v1. (a) Physical impact subscale and (b) psychological impact subscale—Week 52.

impact subscale. In addition to problematic response options, Hobart and Cano[22] indicated the person-item fit was also poor in both subscales. Large fit residuals and significant chi-squares supported their conclusions that many items of the MSIS-29v1 did not fit, prompting a revision and creation of the MSIS-29v2. A subsequent investigation of the MSIS-29v1 by Ramp and colleagues[23] using Rasch measurement similarly concluded that there was a need to revise the scale response options; however, other indicators of fit were acceptable.

In the current investigation using GRM, all items contained response categories that were ordered properly; however, ICCs indicated one item had mild overlapping of thresholds. These findings indicated response options were informative and uniquely distinguishable from the RRMS patient's underlying physical and psychological impact, providing evidence that the MSIS-29v1 response options are acceptable for clinical trial use. The person-item fit in the current sample was also acceptable, with evidence of a small floor effect in the MSIS-29v1 physical domain, implying that the severity of impact from MS measured by the scale is generally in correspondence to the population severity. These conclusions were supported by acceptable statistical indicators of individual item fit.

The differences between the current findings and those of previous investigations[22,23] could possibly be due to the fit of the measure to different study populations with different disease characteristics (i.e. community versus clinical trial) or the appropriateness of the mathematical model underlying the statistical methods used to assess the properties of the MSIS-29v1 (i.e. Rasch versus GRM). Thus, we replicated our analysis using Rasch measurement. The Rasch item threshold estimates provided no evidence of threshold disordering and the ICCs indicated that all but two items contained item-response categories that all assessed an independent range on the scale. One item (psychological subscale Q2, ''*problems sleeping''*) detected as potentially problematic using GRM was also problematic using Rasch. Person-item fit was also similar using Rasch for all but the least severe (healthiest) patients, where no items matched their severity. Finally, indicators of item fit were less supportive using Rasch than GRM; however, nearly all items still displayed acceptable fit under the Rasch model at both time points. Thus, the differences in study findings could be due to differences in the severity of the patient population, with the instrument functioning less well in more severe/progressive patients with great disease duration and higher EDSS scores who were present in the community samples.[22,23] However, further research using both analytic methods in a clinical trial population is needed.

Strengths of the current analysis include the use of multiple analytic techniques longitudinally in a sample of patients with few missing data, while the inclusion of only RRMS patients on the lower end of the disability scale is a limitation. In addition, while all items demonstrated acceptable fit, two items had slopes that were more discriminating than model expectations (>4.0). These findings indicate a need to further investigate the performance of these items in a more severe population.

Moreover, the person-item maps indicate that the MSIS-29v1 does not measure as well among the least impaired SELECT trial patients compared to trial patients with the greatest limitations because the instrument does not include items difficult enough to tap this top range of abilities. A key implication of this finding in the clinical trial setting is that MSIS-29v1 improvements over time in physical or psychological functioning among the highest performing patients may not be well captured, and the resulting mean change scores comparing treatments and/or placebo groups may be biased toward the null for effective treatments among RRMS patients.

In conclusion, the MSIS-29v1 is a generally psychometrically sound instrument for measuring the physical and psychological impact of MS. Overall, this comparison of the psychometric properties of the MSIS-29v1 using GRM and Rasch analyses support the hypothesis that the MSIS-29v1 functions well in a clinical trial sample of patients with RRMS and may be an important PRO to include in future clinical trials.

Conflicts of interest

The author(s) declared the following potential conflicts of interest with respect to the research, authorship, and/or publication of this article: Elizabeth D Bacci, Kathleen W Wyrwich and Shien Guo are full-time employees of Evidera Inc. Glenn Phillips is a full-time employee of Biogen. Timothy Vollmer has provided consulting services to Acorda, Biogen, Consortium of MS Centers, DeltaQuest, Genentech, Novartis, Novartis Canada, Novartis Japan, Teva, Teva Canada, Xenoport, Mylan, and Medscape, and provided clinical research services to Accelerated Cure Project, Acorda, Avanir, Biogen, EMD Serono, Genzyme, Jensen Research, MedImmune, National Institutes of Health (NIH), Novartis, Ono Pharmaceuticals, Rocky Mountain MS Center, Teva Neuroscience, Vaccinex, and Roche.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by Biogen and AbbVie Biotherapeutics Inc.

Notes

All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki Declaration and its later amendments or comparable ethical standards.

Informed consent was obtained from all individual participants included in the study.

References

1. Food and Drug Administration. Guidance for industry on patient-reported outcome measures: Use in medical product development to support labeling claims. *Federal Register* 2009; 74: 65132–65133.
2. Cella DF, Dineen K, Arnason B, et al. Validation of the functional assessment of multiple sclerosis quality of life instrument. *Neurology* 1996; 47: 129–139.
3. Gold SM, Heesen C, Schulz H, et al. Disease specific quality of life instruments in multiple sclerosis: Validation of the Hamburg Quality of Life Questionnaire in Multiple Sclerosis (HAQUAMS). *Mult Scler* 2001; 7: 119–130.
4. McKenna SP, Doward LC, Twiss J, et al. International development of the patient-reported outcome indices for multiple sclerosis (PRIMUS). *Value Health* 2010; 13: 946–951.
5. Gnanapavan S, Grant D, Morant S, et al. Biomarker report from the phase II lamotrigine trial in secondary progressive MS—neurofilament as a surrogate of disease progression. *PloS One* 2013; 8: e70019.
6. Kostecki J, Zaniewski M, Ziaja K, et al. An endovascular treatment of Chronic Cerebro-Spinal Venous Insufficiency in multiple sclerosis patients—6 month follow-up results. *Neuro Endocrinol Lett* 2011; 32: 557–562.
7. Rice CM, Mallam EA, Whone AL, et al. Safety and feasibility of autologous bone marrow cellular therapy in relapsing-progressive multiple sclerosis. *Clin Pharmacol Ther* 2010; 87: 679–685.
8. Thomas S, Thomas PW, Kersten P, et al. A pragmatic parallel arm multi-centre randomised controlled trial to assess the effectiveness and cost-effectiveness of a group-based fatigue management programme (FACETS) for people with multiple sclerosis. *J Neurol Neurosurg Psychiatry* 2013; 84: 1092–1099.
9. Zajicek J, Ball S, Wright D, et al. Effect of dronabinol on progression in progressive multiple sclerosis (CUPID): A randomised, placebo-controlled trial. *Lancet Neurol* 2013; 12: 857–865.
10. Hobart J, Lamping D, Fitzpatrick R, et al. The Multiple Sclerosis Impact Scale (MSIS-29): A new patient-based outcome measure. *Brain* 2001; 124: 962–973.
11. Bosma L, Sonder J, Kragt J, et al. Detecting clinically-relevant changes in progressive multiple sclerosis. *Mult Scler* 2015; 21: 171–179.
12. Costelloe L, O'Rourke K, Kearney H, et al. The patient knows best: Significant change in the physical component of the Multiple Sclerosis Impact Scale (MSIS-29 physical). *J Neurol Neurosurg Psychiatry* 2007; 78: 841–844.
13. Costelloe L, O'Rourke K, McGuigan C, et al. The longitudinal relationship between the patient-reported Multiple Sclerosis Impact Scale and the clinician-assessed Multiple Sclerosis Functional Composite. *Mult Scler* 2008; 14: 255–258.
14. Gray O, McDonnell G and Hawkins S. Tried and tested: The psychometric properties of the multiple sclerosis impact scale (MSIS-29) in a population-based study. *Mult Scler* 2009; 15: 75–80.
15. Hoogervorst EL, Zwemmer JN, Jelles B, et al. Multiple Sclerosis Impact Scale (MSIS-29): Relation to established measures of impairment and disability. *Mult Scler* 2004; 10: 569–574.
16. Learmonth YC, Hubbard EA, McAuley E, et al. Psychometric properties of quality of life and health-related quality of life assessments in people with multiple sclerosis. *Qual Life Res* 2014; 23: 2015–2023.
17. McGuigan C and Hutchinson M. The multiple sclerosis impact scale (MSIS-29) is a reliable and sensitive measure. *J Neurol Neurosurg Psychiatry* 2004; 75: 266–269.
18. Riazi A, Hobart JC, Lamping DL, et al. Multiple Sclerosis Impact Scale (MSIS-29): Reliability and validity in hospital based samples. *J Neurol Neurosurg Psychiatry* 2002; 73: 701–704.
19. Riazi A, Hobart JC, Lamping DL, et al. Evidence-based measurement in multiple sclerosis: The psychometric properties of the physical and psychological dimensions of three quality of life rating scales. *Mult Scler* 2003; 9: 411–419.
20. Schäffler N, Schönberg P, Stephan J, et al. Comparison of patient-reported outcome measures in multiple sclerosis. *Acta Neurol Scand* 2013; 128: 114–121.
21. Hays RD, Morales LS and Reise SP. Item response theory and health outcomes measurement in the 21st century. *Med Care* 2000; 38(9 Suppl): II28–II42.
22. Hobart J and Cano S. Improving the evaluation of therapeutic interventions in multiple sclerosis: The role of new psychometric methods. *Health Technol Assess* 2009; 13: iii, ix–x, 1–177.
23. Ramp M, Khan F, Misajon RA, et al. Rasch analysis of the Multiple Sclerosis Impact Scale MSIS-29. *Health Quality Life Outcomes* 2009; 7: 58.
24. Gold R, Giovannoni G, Selmaj K, et al. Daclizumab high-yield process in relapsing–remitting multiple

sclerosis (SELECT): A randomised, double-blind, placebo-controlled trial. *Lancet* 2013; 381: 2167–2175.

25. Polman CH, Reingold SC, Edan G, et al. Diagnostic criteria for multiple sclerosis: 2005 revisions to the ''McDonald Criteria''. *Ann Neurol* 2005; 58: 840–846.
26. Kurtzke JF. Rating neurologic impairment in multiple sclerosis: An Expanded Disability Status Scale (EDSS). *Neurology* 1983; 33: 1444–1452.
27. Samejima F. Estimation *of Latent Ability Using a Response Pattern of Graded Scores (Psychometric Monograph No. 17)*. Richmond, VA: Psychometric Society. Retrieved from http://www.psychometrika.org/journal/online/MN17.pdf.
28. Reise SP. The emergence of item response theory models and the patient reported outcomes measurement information systems. *Austrian Journal of Statistics* 2009; 38: 211–220.
29. Orlando M and Thissen D. Likelihood-based item-fit indices for dichotomous item response theory models. *Appl Psychol Meas* 2000; 24: 50–64.
30. Thissen D, Chen WH and Bock RD. *MULTILOG* (version 7). Lincolnwood, IL: Scientific Software International, 2003.
31. Andrich D. *Rasch models for measurement*. Newbury Park, CA: Sage, 1988.
32. RUMM2030. (2009) *Interpreting RUMM2030—Part I: Dichotomous items*. Perth, Australia: RUMM Laboratory Pt Ltd, 2009.