RESEARCH ARTICLE

Item response theory analysis of the Amyotrophic Lateral Sclerosis Functional Rating Scale-Revised in the Pooled Resource Open-Access ALS Clinical Trials Database

ELIZABETH D. BACCI, DOROTA STANIEWSKA, KARIN S. COYNE, STACEY BOYER, LEIGH ANN WHITE, NETA ZACH, JESSE M. CEDARBAUM & THE POOLED RESOURCE OPEN-ACCESS ALS CLINICAL TRIALS CONSORTIUM

Abstract
Our objective was to examine dimensionality and item-level performance of the Amyotrophic Lateral Sclerosis Functional Rating Scale-Revised (ALSFRS-R) across time using classical and modern test theory approaches. Confirmatory factor analysis (CFA) and Item Response Theory (IRT) analyses were conducted using data from patients with amyotrophic lateral sclerosis (ALS) Pooled Resources Open-Access ALS Clinical Trials (PRO-ACT) database with complete ALSFRS-R data (n = 888) at three time-points (Time 0, Time 1 (6-months), Time 2 (1-year)). Results demonstrated that in this population of 888 patients, mean age was 54.6 years, 64.4% were male, and 93.7% were Caucasian. The CFA supported a 4* individual-domain structure (bulbar, gross motor, fine motor, and respiratory domains). IRT analysis within each domain revealed misfitting items and overlapping item response category thresholds at all time-points, particularly in the gross motor and respiratory domain items. Results indicate that many of the items of the ALSFRS-R may sub-optimally distinguish among varying levels of disability assessed by each domain, particularly in patients with less severe disability. Measure performance improved across time as patient disability severity increased. In conclusion, modifications to select ALSFRS-R items may improve the instrument’s specificity to disability level and sensitivity to treatment effects.

Key words: Amyotrophic Lateral Sclerosis Functional Rating Scale, item response theory, disability, psychometric assessment

Introduction
In clinical trials of amyotrophic lateral sclerosis (ALS), commonly used endpoints include assessments of muscle strength (1), respiratory function (2) and mortality (3). However, these assessments can be time-consuming, require patient cooperation and specialized equipment to administer. The Amyotrophic Lateral Sclerosis Functional Rating Scale (ALSFRS) (4) and its revised form (ALSFRS-R) (5) were developed as functional outcome assessments to evaluate clinical aspects of ALS, including bulbar, fine motor, and gross motor function, and respiratory disability. The ALSFRS-R has been and is currently used as a primary or secondary outcome measure in multiple ALS clinical trials (6–11). Psychometric assessments of the ALSFRS-R using classical test theory methods have provided support for the measure’s acceptable internal consistency, reliability, construct validity, and responsiveness (5,12–14). Change in the ALSFRS-R is longitudinally associated with change in other functional measures such as assessments of muscle strength and Schwab and England scores (4,12).

Modern psychometric methods such as Rasch (15) and Graded Response Model (GRM) (16), both forms of Item Response Theory (IRT) (17), can provide valuable information beyond what can be learned using classical test theory methods. For example, these techniques provide information on the scaling metrics and item-level performance of an instrument, as well as specific characteristics of the scale (e.g. response options) that could potentially be modified to provide better measurement.

Using Rasch analysis in cross-sectional data, Franchignoni et al. (18) found substantial misfit in many ALSFRS-R items, including overlapping response options and low person-to-item fit. A subsequent analysis by Franchignoni et al. similarly demonstrated the need for refinement of the
ALSFRS-R (19). However, GRM analysis has not been applied to the ALSFRS-R. The development of the longitudinal Pooled Resources Open-Access ALS Clinical Trials (PRO-ACT) Database (20) provides access to large quantities of ALS patient data from clinical trials, many of which included the ALSFRS-R as a study endpoint. The objective of this study was to conduct a secondary analysis of PRO-ACT Database to investigate the psychometric properties of the ALSFRS-R over time using both Rasch and GRM, including assessments of instrument dimensionality, item fit, item response ordering, and differential item functioning (DIF).

Material and methods

Data source

Data used in the preparation of this article were obtained from the PRO-ACT Database (20). In 2011, Prize4Life, in collaboration with the Northeast ALS Consortium, and with funding from the ALS Therapy Alliance, formed the PRO-ACT Consortium. The data available in the PRO-ACT Database has been volunteered by PRO-ACT Consortium members, and contains more than 8500 fully de-identified records from ALS patients participating in 17 publicly and privately-conducted international ALS clinical trials. Both placebo and treatment arm patients are included in the database.

The ALSFRS-R properties were evaluated using data at three time-points. Time 0 included patients at their first assessment in the trial. Time 1 included all patients who had a follow-up assessment and were administered the ALSFRS-R approximately six months (±45 days) from Time 0. Time 2 included all patients who had a follow-up assessment and were administered the ALSFRS-R approximately six months (±45 days) after Time 1. To capture variability in trial follow-up assessment schedules, a 90-day (±45 days) window was applied at each six-month follow-up. The primary analytic population included all patients in the PRO-ACT database who completed an assessment using the ALSFRS-R at all three time-points, designated as ‘completers’; patients without data at all three time-points were designated as ‘non-completers’.

ALSFRS-R

The ALSFRS-R is composed of 12 items used to assess daily functioning in four domains (with 3 items per domain) – bulbar, gross motor, fine motor, and respiratory function. Each item has five question-specific response options ranging from 0 (complete dependence) to 4 (normal function), resulting in a total sum score ranging from 0 to 48. Although designed as a clinician rating scale, the ALSFRS-R instructions have been adapted to allow patients to respond and caregivers to serve as proxy-respondents (21). All three types of data were included in the database.

Statistical analysis

Descriptive statistics. Descriptive statistics were used to describe the clinical and sociodemographic characteristics of all patients at Time 0, by completer and non-completer status. Independent samples t-tests and χ² tests were performed to determine if complete and non-completers differed systematically on any baseline characteristics. Descriptive statistics were calculated for all items of the ALSFRS-R, including mean (SD), median, range, and percentage of items with responses at the floor and ceiling (defined as 25% or greater responses of 0 or 4, respectively).

Confirmatory factor analysis. Confirmatory factor analysis (CFA) was used to assess fit of the ALSFRS-R as a unidimensional single scale, as well as the fit of the four ALSFRS-R sub-domains individually. All CFAs were conducted using the weighted least squares mean and variance adjusted (WLSMV) estimator in Mplus version 7.11 (22). Model fit was assessed with the comparative fit index (CFI), root mean square error of approximation (RMSEA), and weighted root mean square residual (WRMR). A CFI >0.90, RMSEA <0.07, and WRMR close to 1 were all used to indicate acceptable model fit (23,24).

Rasch analysis and Graded Response Model

The psychometric properties of all ALSFRS-R domains were assessed separately using RUM M2030 (25) for the Rasch (15) modeling and IRTPRO version 2.1 (26) for the GRM (16) at each time-point. Differences between the Rasch model and GRM have been described in detail elsewhere (27) and are briefly described here.

The GRM estimates the probability of correct response (i.e. ability level of the patient accurately corresponding to ability as assessed by the ALSFRS-R) on an item in terms of item difficulty and discrimination. Item difficulty describes the trait level required for a patient to have a 0.50 probability of endorsing a particular response category and higher response categories. Item discrimination indicates the strength of an item score in relation to the latent construct measured by the whole instrument. Thus, items with high discrimination are better able to distinguish between individuals at the response difficulty level.

Through the GRM analysis, item slope values were examined to assess item discrimination. 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 generally indicate an item is too
discriminating, pointing to a high level of dependency between the item score and latent construct or between two item scores (28). In the Rasch model, the discrimination parameter is fixed (does not vary between items) as it is assumed the relationship to the construct is constant for all items. Both approaches are appropriate for ordered categorical item responses as in the ALSFRS-R, but may yield different results in terms of fit and ordering latent scores.

For the Rasch model, the ordering of item thresholds is used to determine if patients with higher levels of the measured attribute (e.g., disability due to ALS) consistently endorse lower scored response options and vice versa. When this does not happen, disordered thresholds occur for a variety of reasons, such as: the item or response phrasing is confusing; the instructions provided to the patient are unclear; or the item responses as stated are indistinguishable from one another in terms of severity. Threshold ordering was assessed by examining category threshold estimates and order, and item characteristic curves (ICCs). In particular, the range of disability for which a particular score (response option) is more probable than the other response options was examined; if that range was narrow, or non-existent, the validity of the response category was questioned.

The fit of each item to model requirements was also assessed using both Rasch and GRM. Item fit using Rasch was assessed using a $\chi^2$ and fit residual (misfit indicated when $\chi^2$ p-value $<0.001$). The fit residual considers the fit of the data in the population (observed data) to the Rasch model. A large negative fit residual value ($<-3.0$) demonstrates 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 detecting differences in severity. For the GRM analysis, item fit was similarly assessed using Pearson’s S-$\chi^2$ fit statistics (misfit indicated when $p<0.001$) (29), which is used to assess the difference between observed and model-based predicted values.

Finally, DIF of each of the ALSFRS-R items, separately by domain, was assessed by sub-groups created using site of symptom onset (bulbar versus limb) using a total $\chi^2$ test. An assessment of DIF is needed to assess the stability of item calibration across different patient populations (28). A significant DIF $\chi^2$ test indicates that the response probabilities differ by type of symptom onset.

**Results**

**Sample characteristics**

A total of 1736 patients had ALSFRS-R data. Of these, 1730 (99.6%) patients had an ALSFRS-R assessment at Time 0, and 888 (51.1%) had an ALSFRS-R assessment at all three time-points. Baseline demographic and clinical characteristics were calculated for all patients with complete ALSFRS-R data at Time 0, and stratified by completer status (Table I). The majority of patients were male (63.4%) and Caucasian (93.2%), with a mean age of 55.2 years. Baseline characteristics were similar between groups. All further analyses report the results of the completer group at all three time-points to enable a comparison of instrument characteristics across time. Parallel analyses were conducted with the non-completer group, with no substantive differences in findings (results not reported).

**ALSFRS-R descriptive analyses**

Descriptive statistics for the ALSFRS-R were calculated at all time-points, with Time 0 and Time 2 scores presented in Tables IIa and IIb. At all time-points, scale scores spanned the entire range. At Time 0, mean scores ranged from 2.7 (moderate difficulty) to 4.0 (performs without difficulty). No items displayed floor effects; however, 10 of 12 items displayed ceiling effects (indicating high percentage of patients with normal function). At Time 2, mean scores on every item were lower, indicating increased disability in patients at approximately 12 months following Time 1. At Time 2, floor effects were observed for one item. Only six items displayed ceiling effects, demonstrating that patients were increasing in disability.

**ALSFRS-R dimensionality**

Confirmatory factor analysis of the total scale score of the ALSFRS-R as one dimension demonstrated poor fit to the data at all time-points (CFI range $= 0.75–0.86$; RMSEA range $= 0.28–0.33$; WRMR $= 0.29–0.31$). However, CFA analyses of each of the four domains individually demonstrated acceptable fit, with little to no variance in each model at all time-points (CFI range $= 1.00$; RMSEA $= 0.00$; WRMR $= 0.00$, for all domains). Standardized coefficients for item loadings within each factor ranged from 0.68 to 0.99.

**Rasch and GRM analysis**

The Rasch model fit within each domain was poor. Items within each of the four domains demonstrated disordered thresholds at all time-points and fit indices indicated item misfit. It was clear that the data did not fit the modeling assumptions of Rasch measurement, thus the results of this analysis are not presented (available upon request). Items modeled with the GRM performed better in terms of fit than when they were modeled with the Rasch. However, select items within each domain still demonstrated poor fit (Tables IIIa, b) and DIF (Table IV). The results presented below are of the GRM analyses only.
Figure 1 shows the ICCs for the bulbar domain. An examination of the ICCs for Item 2 (‘salivation’) indicated that response category 1 (‘marked excess of saliva with some drooling’) did not distinguish well between patients in this and either of the adjacent score categories (‘marked drooling’ and ‘moderately excessive saliva’) at Time 1. Item 3 (‘swallowing’), response category 1 (‘needs supplemental tube feeding’) was similarly problematic at Time 0. Slope values for Item 1 (Time 0) and Items 1 and 3 (Time 2) both exceeded established discrimination cut-off values; however, only Item 1 demonstrated misfit to the model with a large and statistically significant S-$\chi^2$ value. Convergence was

| Item | $n$ | Mean (SD) | Median | Range | n (%) Floor | n (%) Ceiling | n (%) Ceiling |
|------|-----|-----------|--------|-------|-------------|--------------|-------------|
| 1. Speech | 888 | 3.4 (0.90) | 4 | 0.0–4.0 | 15 (1.7%) | 507 (57.1%) | 0 (0.0%) |
| 2. Salivation | 888 | 3.6 (0.73) | 4 | 0.0–4.0 | 5 (0.6%) | 608 (68.5%) | 0 (0.0%) |
| 3. Swallowing | 888 | 3.6 (0.63) | 4 | 0.0–4.0 | 3 (0.3%) | 593 (66.8%) | 0 (0.0%) |
| 4. Handwriting | 888 | 3.1 (0.90) | 3 | 0.0–4.0 | 18 (2.0%) | 348 (39.2%) | 0 (0.0%) |
| 5a. Cutting Food and Handling Utensils (without Gastrostomy) | 876 | 2.9 (1.09) | 3 | 0.0–4.0 | 25 (2.8%) | 306 (34.5%) | 12 (1.4%) |
| 5b. Cutting Food and Handling Utensils (with Gastrostomy) | 11 | 3.0 (1.18) | 3 | 1.0–4.0 | – | 5 (0.6%) | 877 (98.8%) |
| 6. Dressing and Hygiene | 888 | 2.7 (0.97) | 3 | 0.0–4.0 | 23 (2.6%) | 189 (21.3%) | 0 (0.0%) |
| 7. Turning in Bed and Adjusting Bed Clothes | 888 | 3.2 (0.89) | 3 | 0.0–4.0 | 7 (0.8%) | 404 (45.5%) | 0 (0.0%) |
| 8. Walking | 888 | 2.8 (0.92) | 3 | 0.0–4.0 | 13 (1.5%) | 248 (27.9%) | 0 (0.0%) |
| 9. Climbing Stairs | 888 | 2.2 (1.38) | 2 | 0.0–4.0 | 76 (8.6%) | 229 (25.8%) | 0 (0.0%) |
| R1. Dyspnea | 888 | 3.7 (0.60) | 4 | 0.0–4.0 | 1 (0.1%) | 705 (79.4%) | 0 (0.0%) |
| R2. Orthopnea | 888 | 3.9 (0.37) | 4 | 0.0–4.0 | 1 (0.1%) | 803 (90.4%) | 0 (0.0%) |
| R3. Respiratory Insufficiency | 888 | 4.0 (0.22) | 4 | 0.0–4.0 | 1 (0.1%) | 877 (98.8%) | 0 (0.0%) |

Abbreviation. SD: standard deviation.

Bulbar domain. Figure 1 shows the ICCs for the bulbar domain. An examination of the ICCs for Item 2 (‘salivation’) indicated that response category 1 (‘marked excess of saliva with some drooling’) did not distinguish well between patients in this and either of the adjacent score categories (‘marked drooling’ and ‘moderately excessive saliva’) at Time 0. Item 3 (‘swallowing’), response category 1 (‘needs supplemental tube feeding’) was similarly problematic at Time 0. Slope values for Item 1 (Time 0) and Items 1 and 3 (Time 2) both exceeded established discrimination cut-off values; however, only Item 1 demonstrated misfit to the model with a large and statistically significant S-$\chi^2$ value. Convergence was
not reached in the assessment of DIF and Time 0 and Time 1; however, no items displayed DIF at Time 2.

Fine motor domain. Figure 2 shows the ICCs for the fine motor domain. At all time-points the slope value for Item 5 ('cutting food and handling utensils') was large (~10), far exceeding the threshold for acceptable item discrimination. This finding indicates that Item 5 is potentially redundant with the latent construct of disability severity being measured by the ALSFRS-R and may be over-discriminating between individuals with different levels of severity as assessed by this item. All items displayed acceptable fit, with the exception of Item 6 ('dressing and hygiene') at Time 0. Item 5 also displayed DIF at all time-points, while Items 4 and 6 only displayed DIF at Time 0 and Time 1, respectively.

Gross motor domain

Figure 3 shows the ICCs for the gross motor domain. ICCs for this domain show evidence of problematic item responses at all time-points. Specifically, Item 7 ('turning in bed and adjusting bed clothes') response category 0 ('helpless') at Time 0 was not distinct from response category 1 ('can initiate but cannot adjust sheets alone'). Item 9 ('climbing stairs') was the most problematic, as response category 2 ('mild unsteadiness or fatigue') was not distinct from response categories 1 and 3. While all S-χ² values were acceptable for all items at all time-points, the slope value for items 8 and 9 exceeded the recommended discrimination threshold. Items 7 and 8 showed DIF at Time 0, but no items displayed DIF at Time 2.

Respiratory domain

Figure 4 shows the ICCs for the respiratory domain. The fit was very poor for Items R2 and R3 at Time 0, and the lowest threshold could not be estimated because of the lack of responses in that score category (Tables IIIa, b). An examination of the ICCs supported this finding at Time 0, further

| Item | n | Mean (SD) | Median | Range | n (%) Floor | n (%) Ceiling | n (%) Missing |
|------|---|-----------|--------|-------|-------------|--------------|--------------|
| 1. Speech | 888 | 2.6 (1.40) | 3 | 0.0–4.0 | 117 (13.2%) | 321 (36.1%) | 0 (0.0%) |
| 2. Salivation | 888 | 3.1 (1.19) | 3 | 0.0–4.0 | 52 (5.9%) | 437 (49.2%) | 0 (0.0%) |
| 3. Swallowing | 888 | 3.0 (1.21) | 3 | 0.0–4.0 | 62 (7.0%) | 387 (43.6%) | 0 (0.0%) |
| 4. Handwriting | 888 | 2.2 (1.35) | 3 | 0.0–4.0 | 160 (18.0%) | 141 (15.9%) | 0 (0.0%) |
| 5a. Cutting Food and Handling Utensils (without Gastrostomy) | 764 | 1.9 (1.37) | 2 | 0.0–4.0 | 162 (18.2%) | 116 (13.1%) | 124 (14.0%) |
| 5b. Cutting Food and Handling Utensils (with Gastrostomy) | 124 | 1.4 (1.43) | 1 | 0.0–4.0 | 47 (5.3%) | 17 (1.9%) | 764 (86.0%) |
| 6. Dressing and Hygiene | 887 | 1.6 (1.24) | 2 | 0.0–4.0 | 216 (24.3%) | 67 (7.5%) | 1 (0.1%) |
| 7. Turning in Bed and Adjusting Bed Clothes | 887 | 2.1 (1.35) | 2 | 0.0–4.0 | 145 (16.3%) | 163 (18.4%) | 1 (0.1%) |
| 8. Walking | 888 | 2.0 (1.14) | 2 | 0.0–4.0 | 103 (11.6%) | 113 (12.7%) | 0 (0.0%) |
| 9. Climbing Stairs | 888 | 1.2 (1.34) | 1 | 0.0–4.0 | 378 (42.6%) | 88 (9.9%) | 0 (0.0%) |
| R1. Dyspnea | 888 | 3.1 (1.16) | 4 | 0.0–4.0 | 25 (2.8%) | 471 (53.0%) | 0 (0.0%) |
| R2. Orthopnea | 888 | 3.4 (0.93) | 4 | 0.0–4.0 | 13 (1.5%) | 597 (67.2%) | 0 (0.0%) |
| R3. Respiratory Insufficiency | 888 | 3.7 (0.77) | 4 | 0.0–4.0 | 6 (0.7%) | 717 (80.7%) | 0 (0.0%) |

Abbreviation. SD: standard deviation

Table IIIa. Graded response model item parameter estimates by ALSFRS-R domain at Time 0.
demonstrating that the vast majority of patients indicated limited respiratory disability. In addition, for Item R3 (‘respiratory insufficiency’) at all time-points, patients responded with either a response of 4 (‘none’) or 2 (‘continuous use of BiPAP’) rather than 3 (‘intermittent use of BiPAP’), demonstrating that response 3 did not assess a unique level of severity. The \( \chi^2 \) values indicated acceptable fit at all time-points. Only Item R2 at Time 2 had a large slope value. Items R3 at Time 0 and R1 and R2 at Time 1 displayed DIF; however, no items displayed DIF at Time 2.

**Discussion**

Using the large, publically accessible PRO-ACT Database, the current study builds upon previous work (5,12–14,18,19) to provide a more complete understanding of the scaling metrics of the ALSFRS-R. The IRT analyses conducted contribute valuable information to indicate which items and response categories of the ALSFRS-R are performing well, as defined by an IRT model, and which could be revised, in an effort to improve assessment of disability in ALS patients.

Dimensionality assessments in the current study further supported the finding that the ALSFRS-R is better treated as an instrument assessing 4 separate constructs (4,18,30). GRM analyses demonstrating multiple item response categories within the bulbar, gross motor, and respiratory domains did not adequately distinguish between varying levels of disability represented by each category. Items within the respiratory domain were particularly problematic,

| Item | Slope | \( b_1 \) | \( b_2 \) | \( b_3 \) | \( b_4 \) | \( S-\chi^2 \) | df | \( p \)-value |
|------|-------|--------|--------|--------|--------|-----------|----|-------------|
| Bulbar domain | | | | | | | | |
| 1. Speech | 6.50 | -1.19 | -0.78 | -0.32 | 0.37 | 13.62 | 18 | 0.754 |
| 2. Salivation | 3.08 | -1.81 | -1.30 | -0.75 | 0.04 | 40.62 | 26 | 0.034 |
| 3. Swallowing | 4.55 | -1.60 | -1.19 | -0.66 | 0.18 | 31.00 | 21 | 0.073 |
| Fine motor domain | | | | | | | | |
| 4. Writing | 3.73 | -1.02 | -0.50 | -0.03 | 1.09 | 27.28 | 21 | 0.161 |
| 5. Cutting Food and Handling Utensils | 10.82 | -0.72 | -0.04 | 0.42 | 1.03 | 19.17 | 16 | 0.259 |
| 6. Dressing and Hygiene | 3.55 | -0.77 | -0.04 | 0.72 | 1.57 | 47.49 | 21 | 0.001 |
| Gross motor domain | | | | | | | | |
| 7. Turning in Bed and Adjusting Bed Clothes | 2.91 | -1.20 | -0.46 | 0.20 | 1.09 | 42.23 | 25 | 0.017 |
| 8. Walking | 7.13 | -1.26 | -0.61 | 0.55 | 1.14 | 10.64 | 15 | 0.078 |
| 9. Climbing Stairs | 7.61 | -0.16 | 0.57 | 0.82 | 1.29 | 9.82 | 14 | 0.001 |
| Respiratory domain | | | | | | | | |
| R1. Dyspnea | 2.56 | -2.30 | -1.39 | -0.67 | -0.09 | 28.65 | 21 | 0.123 |
| R2. Orthopnea | 4.51 | -2.28 | -1.76 | -1.06 | -0.49 | 34.65 | 19 | 0.015 |
| R3. Respiratory insufficiency | 2.60 | -2.99 | -2.47 | -1.40 | -1.06 | 10.61 | 18 | 0.910 |

**Table IV. Differential Item Functioning by Site of Onset (bulbar versus limb) by ALSFRS-R domain at Time 0, 1, and 2.**

| Item | Total \( \chi^2 \) | \( p \)-value | Total \( \chi^2 \) | \( p \)-value | Total \( \chi^2 \) | \( p \)-value |
|------|------------------|-------------|------------------|-------------|------------------|-------------|
| Bulbar domain | | | | | | |
| 1. Speech | NA | NA | NA | NA | NA | 3.6 | 0.47 |
| 2. Salivation | NA | NA | NA | NA | NA | 4.5 | 0.48 |
| 3. Swallowing | NA | NA | NA | NA | NA | 1.9 | 0.86 |
| Fine motor domain | | | | | | |
| 4. Writing | 78.3 | <0.001 | 5.9 | 0.31 | 2.8 | 0.74 |
| 5. Cutting Food and Handling Utensils | 282.5 | <0.001 | 65.1 | <0.001 | 41.1 | <0.001 |
| 6. Dressing and Hygiene | 6.2 | 0.19 | 31.4 | <0.001 | 4.6 | 0.46 |
| Gross motor domain | | | | | | |
| 7. Turning in Bed and Adjusting Bed Clothes | 71.6 | <0.001 | 7.7 | 0.17 | 4.2 | 0.53 |
| 8. Walking | 24.7 | <0.001 | 8.9 | 0.11 | 1.8 | 0.88 |
| 9. Climbing Stairs | 0.9 | 0.97 | 384.4 | <0.001 | 1.7 | 0.89 |
| Respiratory domain | | | | | | |
| R1. Dyspnea | 0.5 | 0.97 | 37.6 | <0.001 | 3.8 | 0.58 |
| R2. Orthopnea | 9.4 | 0.02 | 69.6 | <0.001 | 2.4 | 0.79 |
| R3. Respiratory Insufficiency | 51.8 | <0.001 | 2.2 | 0.82 | 4.9 | 0.43 |

NA: estimation did not converge; DIF estimates could not be calculated.
A statistically significant total \( \chi^2 \) test indicates the item does not provide uniform measurement across the groups tested.
as the response threshold for category 0 could not be estimated at Time 0 for two items. This finding could be a result of the drastically uneven distribution of responses at Time 0 — only 0.1% of patients were rated 0 (greatest disability) at Time 0 for the three respiratory domain items and over 90% of patients were rated as a 4 (no disability) for two of three respiratory items. This finding indicates either that the current response options for these items do not adequately assess a range of disability in ALS patients with lesser respiratory severity, or patients having their first clinical trial visit are selected or self-selected for having minimal respiratory dysfunction. Similarly, items within each domain exhibited DIF between patients with bulbar versus limb onset that was particularly problematic at Time 0 and Time 1, but resolved by Time 2. This finding could be resultant from uneven score distributions at Time 0, which resolved at Time 2 due disease progression leading to more widespread disease.

The current analysis is unique as ALSFRS-R performance was assessed at multiple time-points. Item fit and parameter performance was better at Time 2 than Time 0, perhaps suggesting the instrument performs better over the course of disease progression. In the respiratory domain, for example, changes were evident in the ICCs for the R1 and R2 respiratory items, indicating response options for these items are more appropriate for assessing respiratory function in a more severe patient population. With the exception of Item 9 (‘climbing stairs’), threshold spacing and individual item fit for all items showed varying levels of improvement over time.

Findings from this analysis can be used to consider revisions to the ALSFRS-R. In its current form, the ALSFRS-R may not be optimized to support sensitive assessment of treatment effects across time in a clinical trial population, as change in some response levels may not accurately correspond to changing severity. Some items (i.e. gross motor

Figure 1. Item Characteristics Curves for Bulbar Domain at Time 0, 1, and 2.
domain) could be modified by clarifying/rewording existing response options. Other items (i.e. respiratory domain) might be improved by presenting additional response options to capture more subtle changes in less disabled patients, allowing for more robust assessment of change given that considerable ceiling effects were present for all but one item at Time 0. These recommendations are supported by previous conclusions that revising the ALSFRS-R may improve its performance both within clinical trial (18) and non-clinical trial populations (19). However, if items of the ALSFRS-R are modified in any way, a full validation study to assess the reliability and validity of the modified measure should be conducted with the modified measure administered following current guidelines.

Potential study limitations should be noted. First, the ALSFRS-R was administered by clinicians, caregivers, or self-administered; however, the PRO-ACT Database did not contain a variable to indicate the reporter (e.g. patient, caregiver, investigator). Instructions used for administering the ALSFRS-R were also not recorded in the database (31). Thus, the impact of mode of administration and instructions for administration could not be analytically examined and/or controlled. Inconsistent administration and rater variability could result in reduced measurement reliability, independent of patient disability severity. Consistent administration could be employed by development and implementation of standard examples to guide individuals administering or responding to the scale. Other limitations of the PRO-ACT database include the omission of a variable to assess disease duration or clinical site. In addition, analyses were restricted to patients who had ALSFRS-R data at all three time-points. Thus, the analysis is subject to attrition bias and exclusion of the most severe patients, as data would be missing for patients who died during the time period analyzed. Thus, these findings may not be

Figure 2. Item Characteristics Curves for Fine Motor Domain at Time 0, 1, and 2.
representative of instrument performance in the most severe patients, nor of patients in a community based population, as all patients within this sample were from a clinical trial sample. In addition, the windows chosen to represent Time 1 and Time 2 were broad (±45 days) – an interval over which change may occur, even within a single patient. The broad visit windows may have accounted for some of the overlapping scaling observed within items. However, a descriptive assessment using a ±15-day window versus the ±45-day window demonstrated that very few patients fell outside the 15-day window, and the mean span of time in which the ALSFRS-R assessment was conducted was almost identical when using either the ±15-day or ±45-day windows. Finally, it should be noted that different analytical techniques (IRT versus CFA, for example) may lead to different conclusions regarding item performance and scaling metrics in the same population (32).

Findings from the current analysis indicate the ALSFRS-R may benefit from a revision for use in a clinical trial population; however, the performance of the measure did improve over time as the study population progressed, suggesting a need to focus on improving response options for patients with milder disability. Modification of response options may increase sensitivity of the instrument across the spectrum of ALS progression, potentially improving the likelihood of detecting true treatment effects. In addition, implementation of standard administration practices should decrease measurement variability arising from poor inter-rater reliability. Such revisions and establishment of standards for training evaluators on administration of the instrument may lead to an increased ability of the measure to accurately detect change at all disability levels. Any revision to the ALSFRS-R or establishment of administration standards must be accomplished by adherence to current guidelines and best practices (33).
Acknowledgement

Funding for this work was provided by Biogen.

Data used in the preparation of this article were obtained from the Pooled Resource Open-Access ALS Clinical Trials (PRO-ACT) Database. As such, the following organizations and individuals within the PRO-ACT Consortium contributed to the design and implementation of the PRO-ACT Database and/or provided data, but did not participate in the analysis of the data or the writing of this report: Neurological Clinical Research Institute, MGH; Northeast ALS Consortium; Novartis; Prize4Life; Regeneron Pharmaceuticals, Inc., Sanofi; Teva Pharmaceuticals Industries, Ltd.

Declaration of interest

KSC, EDB, DS are employed by Evidera, which provides consulting and other research services to pharmaceutical, device, government and non-government organizations. As Evidera employees, they work with a variety of companies and organizations and are expressly prohibited from receiving any payment or honoraria directly from these organizations for services rendered. JC, SB and LAW are employed by Biogen. KSC, EDB, DS were paid consultants to Biogen in connection with the development of this manuscript.

References

1. Council MR. Aid to the investigation of peripheral nerve injuries. In: Medical Research Council. War memorandum. In: Office HMS, editor. 2nd edn. London:1943. pp 11–46.
2. Talakad NS, Pradhan C, Nalini A, Thennarasu K, Raju TR. Assessment of pulmonary function in amyotrophic lateral sclerosis. Indian J Chest Dis Allied Sci. 2009.
3. Gordon PH, Corcia P, Lacomblez L, Pochigaeva K, Abitbol J-L, Cudkowicz M, et al. Defining survival as an outcome measure in amyotrophic lateral sclerosis. Archives of Neurology. 2009;66:758–61.
4. The Amyotrophic Lateral Sclerosis Functional Rating Scale. Assessment of activities of daily living in patients with amyotrophic lateral sclerosis. The ALS CNTF treatment study (ACTS) phase I-II Study Group. Arch Neurol. 1996;53:141–7.
5. Cedarbaum JM, Stambler N, Malta E, Fuller C, Hilt D, Thurmond B, et al. The ALSFRS-R: a revised ALS functional rating scale that incorporates assessments of respiratory function. BDNF ALS Study Group (Phase III). Journal of the Neurological Sciences. 1999;169:13–21.

6. Aggarwal SP, Zinman L, Simpson E, McKinley J, Jackson KE, Pinto H, et al. Safety and efficacy of lithium in combination with riluzole for treatment of amyotrophic lateral sclerosis: a randomized, double-blind, placebo-controlled trial. Lancet Neurology. 2010;9:481–8.

7. Anagnostou E, Rentzos M, Alcakis T, Zouvelou V, Zambelis T, Evdokimidis I. Volume matters: the influence of different botulinum toxin-A dilutions for sialorrhea in amyotrophic lateral sclerosis. Muscle & Nerve. 2013;47:276–8.

8. Beghi E, Chio A, Inghilleri M, Mazzini L, Micheli A, Mora G, et al. A randomized controlled trial of recombinant interferon beta-1a in ALS. Italian Amyotrophic Lateral Sclerosis Study Group. Neurology. 2000;54:469–74.

9. de Carvalho M, Pinto S, Costa J, Evangelista T, Ohana B, Pinto A. A randomized, placebo-controlled trial of memantine for functional disability in amyotrophic lateral sclerosis. Amyotrophic lateral sclerosis: official publication of the World Federation of Neurology Research Group on Motor Neuron Diseases. 2010;11:456–60.

10. Piepers S, Veldink JH, de Jong SW, van der Tweel I, van der Pol WL, Uijtendaal EV, et al. Randomized sequential trial of valproic acid in amyotrophic lateral sclerosis. Annals of Neurology. 2009;66:227–34.

11. Rudnicki SA, Berry JD, Ingersoll E, Archibald D, Cudkowicz ME, Kerr DA, et al. Dexamethasone and efaximoxepin effect on functional decline and survival in subjects with amyotrophic lateral sclerosis in a phase II study: subgroup analysis of demographic and clinical characteristics. Amyotroph Lateral Scler Frontotemporal Degener. 2013;14:44–51.

12. Cedarbaum JM, Stambler N. Performance of the Amyotrophic Lateral Sclerosis Functional Rating Scale (ALSFRS) in multicenter clinical trials. Journal of the Neurological Sciences. 1997;152 (Suppl 1):S1–9.

13. Jenkinson C, Fitzpatrick R, Swash M, Peto V, Group A-HS. The ALS Health Profile Study: quality of life of amyotrophic lateral sclerosis patients and carers in Europe. Journal of Neurology. 2000;247:835–40.

14. Kaufmann P, Levy G, Montes J, Buchsbaum R, Barsdorf AI, Battista V, et al. Excellent inter-rater, intra-rater, and telephone-administered reliability of the ALSFRS-R in a multicenter clinical trial. Amyotrophic lateral sclerosis: official publication of the World Federation of Neurology Research Group on Motor Neuron Diseases. 2007;8:42–6.

15. Andrich D. Rasch models for measurement. Sage Publications; 1988.

16. Samejima F. Estimation of a latent ability using a response pattern of graded scores. Psychometrica Monographs. Richmond, VA: Psychometric Society; 1969. Available from: http://psychometricsociety.org/sites/default/files/pdf/MN17.pdf.

17. Lord FM. A theory of test score (Psychometric Monograph No. 7). Iowa City, IA: Psychometric Society. 1952.

18. Franchignoni F, Mora G, Giordano A, Volanti P, Chio A. Evidence of multidimensionality in the ALSFRS-R Scale: a critical appraisal on its measurement properties using Rasch analysis. Journal of Neurology, Neurosurgery, and Psychiatry. 2013;84:1340–5.

19. Franchignoni F, Mandrioli J, Giordano A, Ferro S. A further Rasch study confirms that ALSFRS-R does not conform to fundamental measurement requirements. Amyotroph Lateral Scler Frontotemporal Degener. 2015 (ahead of print):1–7.

20. Pooled Resource Open-Access ALS Clinical Trials Database. Data Sets. 2013 (December 18, 2013); Available from: https://nctu.partners.org/ProACT/Data.

21. Brooks B. Amyotrophic lateral sclerosis clinimetric scales: guidelines for administration and scoring. Handbook of Clinical Neurological Scales New York: Demos Vermande. 1997:27–9.

22. Muthén L, Muthén B. Mplus User’s Guide. Los Angeles, CA:1998–2004.

23. Steiger JH. Understanding the limitations of global fit assessment in structural equation modeling. Personality and Individual Differences. 2007;42:893–8.

24. Yu CY. Evaluating cut-off criteria of model fit indices for latent variable models with binary and continuous outcomes. Doctoral dissertation; University of California, Los Angeles 2002.

25. RUMM2030. Interpreting RUMM2030 - Part I: dichotomous items. Perth, Australia: RUMM Laboratory Pt Ltd; 2009.

26. Cai L, Thissen D, S dT®. IRTPRO 2.1 for Windows: Item response theory for patient-reported outcomes. Skokie, IL: Scientific Software International, Inc; 2011.

27. Embretson SE, Reise SP. Item response theory for psychologists: Psychology Press; 2000.

28. Reise SP. The Emergence of Item Response Theory Models and the Patient Reported Outcomes Measurement Information Systems. Austrian J Statistics. 2009;38:211–20.

29. Orlando M, Thissen D. Likelihood-based item-fit indices for dichotomous item response theory models. Applied Psychological Measurement. 2000;24:50–64.

30. Campos TS, Rodríguez-Santos F, Esteban J, Vázquez PC, Mora Pardina JS, Carmona AC. Spanish adaptation of the revised amyotrophic lateral sclerosis functional rating scale (ALSFRS-R). Amyotroph Lateral Scler. 2010;11:475–7.

31. Kasarskis EJ, Dempsey-Hall L, Thompson MM, Luu LC, Mendiondo M, Kryscio R. Rating the severity of ALS by caregivers over the telephone using the ALSFRS-R. Amyotroph Lateral Scler. 2005;6:50–4.

32. Hyndon A. Comparing factor analysis and the Rasch model for ordered response categories: an investigation of the scale of gambling choices. J Appl Meas. 2004;5:398–418.

33. Food and Drug Administration (FDA). Guidance for industry on patient-reported outcome measures: use in medical product development to support labeling claims. Federal Register. 2009;74:65132–3.