The Extended Barthel Index (EBI) can Be Reported as a Unidimensional Interval-Scaled Metric – A Psychometric Study

Der Erweiterte Barthel Index (EBI) kann als eindimensionale intervallskalierte Metrik berichtet werden – eine psychometrische Studie

Authors
Roxanne Maritz1,2, Alan Tennant1,2, Carolina Saskia Fellinghauer1, Gerold Stucki1,2, Birgit Prodinger1,3, on behalf of the NRP74 StARS clinics*

Affiliations
1 Swiss Paraplegic Research, Nottwil, Switzerland
2 Department of Health Sciences and Health Policy, University of Lucerne, Lucerne, Switzerland
3 Faculty of Applied Health and Social Sciences, Rosenheim, Technische Hochschule Rosenheim, Germany

Key words
activities of daily living, assessment instruments, outcome assessment, rehabilitation, Rasch model, quality management

ZUSAMMENFASSUNG
Hintergrund Der Erweiterte Barthel Index (EBI), der den Barthel Index um 6 kognitive Items ergänzt, ist ein Assessmentinstrument für die Rehabilitation. Ob der EBI eine eindimensionale Metrik liefert und somit als valider Gesamtscore berichtet werden kann, ist unklar.

Ziel Untersuchung ob der EBI für die neurologische und muskulokeskelettale Rehabilitation als eindimensionale intervallskalierte Metrik berichtet werden kann.

Methode Rasch-Analyse einer Stichprobe von 800 neurologischen und muskulokeskettalen Rehabpatienten aus der Schweiz.

Ergebnisse In der Basisanalyse wurde keine Übereinstimmung mit den Annahmen des Rasch-Modells erreicht. Nachdem lokale Item-Abhängigkeiten mit 2 Testlets angepasst wurden, wurde die Übereinstimmung erreicht und eine intervallskalierte Transformationstabelle erstellt.

Konklusion Die Ergebnisse unterstützen die Verwendung eines angepassten EBI Gesamtscores für beide Rehabilitationsgruppen unter Anwendung der intervallskalierten Transformationstabelle.

Introduction

Functioning is the primary outcome in rehabilitation [1]. Global Activities of Daily Living (ADL) assessment tools that aim to assess functioning are essential for the documentation of the rehabilitation progress and its outcome [2, 3]. Sum-scores of such ADL assessment tools are commonly created by simply summing up the scores of individual items, which often deliver only an ordinal scale of a person’s dependency in ADL tasks. There is increasing evidence that treatment decisions based on ordinal level scores can be misinformed [4] as ordinal-level scores can lead to under- or overestimation of the treatment benefit of a person [5]. Therefore, it is essential to transform ordinal measures into interval scales [6]. For this purpose, valid assumptions such as unidimensionality and group invariance need to be established [7].

This issue can be addressed by applying assessment tool data to the Rasch Model. If fit to the Rasch Model can be achieved, and assumptions of local independence and group invariance are supported, an interval-based scoring system can be developed [8].

The Extended Barthel Index (EBI) is such a global ADL tool that is a well-established assessment tool in German speaking countries at the patient, the institutional and the national level [9]. In Germany the EBI is one of the assessment tools used within the ICD-10-GM System as a tool to code restrictions in functioning, that can be relevant for the DRG based payment system [10]. In Switzerland the EBI is one assessment tool used for the national quality monitoring in rehabilitation from the National Association for Quality Development in Hospitals and Clinics (ANQ) [11], part of the CHOP (Swiss classification of treatments for national medical statistics) [12] and will also be part of the DRG based payment system for rehabilitation called ST Reha, that is to be implemented in 2022 [13].

The Extended Barthel Index (EBI) was developed in order to widen the utility of the original Barthel index (BI) [9]. The original BI assesses 10 motor ADL items [14]. The extension of the EBI consists of 6 additional cognitive items, of which 5 are adapted from the FIM™ (Functional Independence Measure), and one – “Vision/Neglect” – is unique to the EBI [9]. Thus, the EBI is a combination of 2 of the most commonly used general outcome measures for rehabilitation, the BI and the FIM™ [15–18]. Due to its simpler rating system and the elimination of some redundant FIM™items the EBI was recommended over the FIM™, as it increases user-friendliness and compliance [19]. While originally intended for patients with multiple sclerosis, the EBI was also validated and is often applied for other neurological patients, e.g., stroke, traumatic brain injury, or Parkinson’s disease [9, 19–23]. Even though the EBI is used for high impact decisions at the patient, institutional and national levels in German speaking countries, no work has been undertaken to-date to explore whether the EBI allows for the calculation of valid sum scores, which would subsequently be eligible for a broad range of statistical analyses. As long as we do not know whether the EBI delivers an ordinal- or interval-scaled unidimensional metric [24] change scores that are based on the EBI can be misleading and have to be interpreted with caution.

Therefore, the objective of the current study was to examine whether the properties of the EBI support its reporting as a unidimensional interval-scaled metric, when administered for national quality monitoring of patients functioning outcomes in neurological and musculoskeletal rehabilitation. This objective resulted in two specific aims: i) To explore the internal construct validity of the EBI and ii) to determine if an interval-scale scoring system of the EBI can be made available.

Methods

Subjects and Setting

We conducted a secondary analysis of data routinely collected for the ANQ for national quality monitoring of rehabilitation clinics in Switzerland. We contacted all 64 Swiss rehabilitation clinics which provided musculoskeletal or neurological rehabilitation data to the ANQ in 2016. Thirty clinics agreed to provide their datasets. As the ANQ data collection permits clinics to choose between different ADL assessment tools, not all datasets contained EBI data. For this study we could include datasets from 10 Swiss rehabilitation clinics containing EBI data with in total 5978 complete cases, representing the German and French Swiss language regions. The datasets included data of the EBI on item level, collected at 2 time points – admission and discharge. Ethical approval of the study was requested from all Swiss Ethnic Commissions, which stated in a declaration of no objection that the project fulfils the general ethical and scientific standards for research with humans and opposes no health hazards.

Measure

The Extended Barthel Index (EBI) is a clinician-administered scale to assess a patient’s need for help with activities of daily living. It consists of 16 items, 10 on physical functioning and 6 on cognitive functioning [9]. The physical functioning items are those from the original Barthel Index [14]: 1-Feeding, 2-Grooming, 3-Dressing, 4-Bathing, 5-Transfer, 6-Mobility, 7-Stairs 8-Toilet use 9-Bowel, and 10-Bladder. The 6 cognitive items are 11-Expression, 12-Comprehension, 13-Social interaction, 14-Problem solving, 15-Memory, and 16-Vision/Neglect. Items 11–15 are adapted from the FIM™. Only item 16 is unique in the EBI. Each item is scored from 0–4, resulting in a total score of 64 [20]. Similar to the BI, not all items represent all categories from 0–4, such as item 1-Feeding that can be scored 0, 2, 3 or 4 (category 1 is missing) or item 13-Social interaction with categories 0, 2, 4 (categories 1 and 3 are missing). The EBI was developed in German [9], the French translation of the EBI used by the participating French speaking clinics, is a non-validated version created by the ANQ.

Sampling

Since a Rasch analysis with a larger sample size is prone to type 1 errors [25], a random stratified calibration sample was obtained using R [26]. The calibration sample contained in total 800 cases, consisting of 4 subsamples containing each 200 cases, each large enough for statistical conclusions and stable item calibration [27, 28]. The 4 subsamples were chosen to equally represent the 2 rehabilitation groups and assessment time points: musculoskeletal cases at admission (MSKt1), musculoskeletal cases at discharge (MSKt2), neurological at admission (NEURt1) and neurological cases at discharge (NEURt2). Cases that were selected for the ad-
mission subsamples were excluded to be selected for the discharge subsamples [29]. Prior to the random selection we deleted all cases with missing values in a variable of interest and all cases with extreme scores (0 or 64) since they cannot be used to estimate item difficulties by the Rasch Measurement Model [30]. In order to be able to give a valuable statement about the whole range of possible total scores of the EBI and the 2 different language regions (German and French) we randomly selected one of each available total scores per subsample and language group. In order to reach 200 cases for each subsample, additional cases were selected by assigning a higher selection probability to rarer total scores in order to best represent the whole range of total scores of the scale. The sampling strategy, with its different subsamples is represented in Online Appendix 1.

Data analysis
We used descriptive statistics to summarize basic sample characteristics and response distributions. In order to reach specific aim I, Rasch analysis was conducted with the RUMM2030 software [31]. The Partial Credit Model was used, as the EBI has polytomous items with varying lengths [32]. The non-continuous nature of the EBI items response categories required recoding into subsequent categories suitable for the Rasch analysis, resulting in a raw adapted total score ranging from 0–50. The conversion of the original scoring (0–64) to the adapted EBI scoring (0–50) on an item basis is presented in Table 1.

Baseline analysis
To test how well the observed EBI data fitted the Rasch Model, we conducted the baseline analysis on all levels of the calibration sample [33]. To do so we ascertained the person and item fit residuals, the reliability indices α and the Person Separation Index (PSI), and the chi² p-value of the item-trait interaction, with the respective acceptable levels represented in the bottom line of the corresponding result table. In addition we investigated local response dependency among items, threshold disordering, and differential item functioning (DIF) for 7 person factors: gender, age (four age groups according to the interquartile ranges), nationality (Swiss or other), insurance status (general, semi-private, private), rehabilitation group (neurological or musculoskeletal rehabilitation), clinic language (German or French) and time point of measurement (admission t1, discharge t2).

Testlet approaches
If the item local independency assumption was not met, testlet approaches combining items into super-items in order to absorb the dependencies in the data were adopted [34–37]. The application of testlets on a related assessment tool, more precisely the FIM™[38] has shown to be an appropriate strategy when dealing with the clustering of items in the underlying subscale structure. In this study we applied 2 different testlet approaches.

Initially, a traditional testlet approach was adopted. This approach emphasises the underlying structure of motor and cognitive items of the EBI. The creation of these testlets was furthermore oriented towards existing local dependencies among items, indicated as standardized residual correlations [39]. Subsequently another testlet approach, referred to as the alternative 2 testlet approach, was used to equally divide items from similar item groups in 2 equally sized testlets, in order to emphasise the ‘sameness’ of the total item set. This alternative testlet approach, which creates 2 super items, has the advantage of gaining access to additional fit and unidimensionality statistics in RUMM2030 such as the conditional test of fit comparing the observed data with the model expectations, while in the same time satisfying the prerequisite that testlets should be equal in length [34]. Both testlet approaches also allow to report the explained common variance associated with the unidimensional latent estimate, obtained within a bi-factor equivalent approach [34]. The acceptable ranges of these additional statistics are as well indicated at the bottom line in the respective result table [40]. We did not report threshold disordering for the testlet approaches, as it does not allow a meaningful interpretation.

To ensure robustness of the analyses, we conducted the baseline analyses and the testlet approach indicating the best fit to the Rasch Model at three aggregation levels of the calibration sample, represented in Online Appendix 1. In Level 1 all four subsamples were analysed separately (MSKt1, MSKt2, NEUR t1 and NEUR t2). In Level 2 the rehabilitation group and time point subsamples were aggregated separately (MSKt1&t2, NEUR1&t2, t1M- SK&NEUR, t2MSK&NEUR). In Level 3 all data were combined, representing the entire calibration sample (EBIall). Likewise, the 3 aggregation levels resulted in nine analytical steps. Throughout, the emphasis of the analyses was upon making the existing EBI work, without the necessity of deleting items or changing its scoring structure other than just making items have consecutive values.

DIF strategy
We analysed DIF in situations in which local dependencies could be accommodated satisfactorily with testlets on the level of the whole calibration sample (EBIall). If lack of invariance between different DIF factors was observed, we split the testlets for the factor with the strongest DIF first and continued, stepwise, until no further DIF was present [41]. We conducted an effect size calculation in order to determine if the splitting makes a substantial difference and should be applied in the final transformation table. The effect size calculation based on the Rasch person estimates from the split and unsplit solutions with estimates from analyses anchored on a DIF free testlet. The effect size calculation was based on the mean of the person estimates, their standard deviations, and the correlation of the split and unsplit version [42]. If the effect size was below 0.2, considered as a small effect size [43], no action was taken to adjust the final transformation table for DIF.

Transformation table
In order to reach specific aim II we sought to create a transformation table in the case that fit to the Rasch Model could be achieved. Based on the solution with the best fit to the Rasch Model, represented by the most satisfactory core values for the whole calibration sample, we constructed an interval-based transformation table of the ordinal adapted EBI total scores (0–50), based on the respective estimates according to the Rasch Model.
### Table 1 (Continued)

<table>
<thead>
<tr>
<th>No</th>
<th>Items</th>
<th>EBI 0–64 Categories</th>
<th>EBI 0–50 Categories</th>
</tr>
</thead>
<tbody>
<tr>
<td>12</td>
<td>Comprehension</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>13</td>
<td>Social interaction</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>14</td>
<td>Problem solving</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>15</td>
<td>Memory</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>16</td>
<td>Vision/ Neglect</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Min</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Max</td>
<td>64</td>
<td>50</td>
<td></td>
</tr>
</tbody>
</table>

### Results

#### Sample characteristics

The calibration sample, containing 800 cases in total, contained 400 cases in each rehabilitation group (MSK, NEUR) and 400 in each time point of assessment (admission t1, discharge t2) as defined in the sampling criteria (Online Appendix 1). EBI sum scores (in the 0–64 scoring) had a mean of 43.7 (SD = 14.6, median = 46). The mean age of the selected cases of the calibration sample was 61 years (min = 18, max = 98). The calibration sample contained 53% (n = 421) male and 47% (n = 379) female cases, 54% (n = 432) were in the German-speaking region of Switzerland and 46% (n = 368) in the French-speaking region, 82% (n = 659) of the sample were Swiss and 18% (n = 141) had another nationality. Insurance status related to 80% (n = 637) general, 11% (n = 88) semi-private, and 9% (n = 75) private.

#### Rasch analysis

**Baseline analyses**

In the 9 baseline analysis steps no fit to the Rasch Model was achieved (Table 2). In all analyses the p-values of the item-trait $\chi^2$ were significant. Furthermore, in all baseline analyses items showed DIF, threshold disordering and local dependency among diverse items. Threshold disordering and local dependency in the baseline analyses are represented in Online Appendix 2.
The traditional testlet approach gave rise to 2 different options – a 4 and a 5 Testlets version of the EBI. For both options, the physical disability items were divided into 3 testlets, with Testlet1 Self-care (including items 1-Feeding, 2-Grooming, 3-Dressing, 4-Washing), Testlet2 Locomotion (including items 5-Transfer, 6-Mobility, 7-Stairs) and Testlet3 Toileting (8-Toilet use, 9-Bowels, 10-Bladder). For the 4 Testlet version all 6 items of the cognitive scale were collected into one testlet. In the 5 Testlet version, the cognitive items were divided into the Testlet4 Communication (including items 1-Feeding, 2-Grooming, 3-Dressing, 4-Washing), Testlet5 (including 13-Social interaction, 14-Problem solving, 15-Memory and 16-Vision/Neglect). For both – the 4 Testlet and the 5 Testlet version, no fit to the Rasch Model was achieved (Table 3).

In the 2-testlet approach, the items were identified as thematic subtopics and then divided equally into the respective 2 testlets: Testlet1 containing items 1-Eating, 3-Dressing, 5-Transfer, 7-Stairs, 9-Bowels, 11-Comprehension, 13-Social interaction, 15-Memory and Testlet2 containing items 2-Grooming, 4-Washing, 6-Mobility, 8-Toilet use, 10-Bladder, 12-Expression, 14-Problem solving, and 16-Vision/Neglect. With the 2-testlet solution, fit to the Rasch Model was achieved across all nine analyses steps. The item-trait chi² statistics were non-significant, the reliability indexes all above 0.85, and the item and person fit estimates showed acceptable values. Furthermore, the conditional test of fit also indicated fit at eight of the nine analysis steps. Most A-values were marginally above 1, indicating some remaining local dependency among the testlets. The core values of the testlet approaches for the whole calibration sample (EBIall) are summarized in Table 3. The core values for the other 8 subsamples of the successful 2-testlet solution can be found in Online Appendix 3.

### Testlet approaches

**Table 2** EBI baseline analyses with different aggregation levels of calibration sample.

<table>
<thead>
<tr>
<th>Sample</th>
<th>n / CI</th>
<th>Item fit residuals Mean (SD)</th>
<th>Person fit residuals Mean (SD)</th>
<th>chi² p-value</th>
<th>PSI</th>
<th>α</th>
<th>DIF (Item No)</th>
<th>Paired t-test (Lower ci %)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MSKt1</td>
<td>200 / 3</td>
<td>-0.096 (1.814)</td>
<td>-0.438 (1.130)</td>
<td>0.000</td>
<td>0.861</td>
<td>0.856</td>
<td>gender (2), language (2, 5, 7, 8, 14, 15, 16), insurance (16)</td>
<td>7.5% (0.0%)</td>
</tr>
<tr>
<td>MSKt2</td>
<td>200 / 3</td>
<td>-0.675 (2.152)</td>
<td>-0.169 (0.900)</td>
<td>0.000</td>
<td>0.849</td>
<td>0.902</td>
<td>gender (2, 16), age (2, 3, 13), language (2, 4, 7, 14, 16), nationality (11, 12)</td>
<td>10.0% (0.0%)</td>
</tr>
<tr>
<td>MSKall</td>
<td>400 / 6</td>
<td>-0.583 (2.941)</td>
<td>-0.310 (1.063)</td>
<td>0.000</td>
<td>0.862</td>
<td>0.882</td>
<td>gender (2, 16), age (2, 3), language (2, 4, 5, 7, 8, 10, 14, 16), nationality (11); insurance (3), time-point (2, 3, 5, 8, 10, 13, 14, 15)</td>
<td>9.3% (0.0%)</td>
</tr>
<tr>
<td>NEURt1</td>
<td>200 / 3</td>
<td>-0.764 (2.174)</td>
<td>-0.310 (1.025)</td>
<td>0.000</td>
<td>0.911</td>
<td>0.941</td>
<td>age (4), language (3, 4, 5, 8, 11, 14)</td>
<td>8.5% (5.5%)</td>
</tr>
<tr>
<td>NEURt2</td>
<td>200 / 3</td>
<td>-0.533 (2.576)</td>
<td>-0.307 (1.175)</td>
<td>0.000</td>
<td>0.918</td>
<td>0.918</td>
<td>language (3, 4, 11, 14), nationality (11), insurance (3, 11)</td>
<td>10.0% (7.0%)</td>
</tr>
<tr>
<td>NEURall</td>
<td>400 / 6</td>
<td>-1.009 (3.371)</td>
<td>-0.328 (1.107)</td>
<td>0.000</td>
<td>0.913</td>
<td>0.941</td>
<td>age (3, 4), language (2, 3, 4, 5, 8, 11, 13, 14), nationality (11), insurance (3)</td>
<td>8.8% (6.6%)</td>
</tr>
<tr>
<td>t1all</td>
<td>400 / 6</td>
<td>-0.639 (3.109)</td>
<td>-0.333 (1.155)</td>
<td>0.000</td>
<td>0.895</td>
<td>0.914</td>
<td>age (4), language (2, 3, 4, 5, 8, 10, 14, 15, 16), rehab-group (1, 2, 3, 4, 5, 8, 11, 13, 14, 15, 16)</td>
<td>9.3% (7.1%)</td>
</tr>
<tr>
<td>t2all</td>
<td>400 / 6</td>
<td>-0.801 (3.101)</td>
<td>-0.259 (1.084)</td>
<td>0.000</td>
<td>0.896</td>
<td>0.933</td>
<td>gender (2, 16), age (2, 3), language (2, 3, 4, 7, 14, 16), insurance (11), nationality (3, 11, 12), rehab-group (1, 2, 3, 4, 14, 15)</td>
<td>8.5% (6.4%)</td>
</tr>
<tr>
<td>EBlall</td>
<td>800 / 10</td>
<td>-1.083 (4.451)</td>
<td>-0.289 (1.108)</td>
<td>0.000</td>
<td>0.896</td>
<td>0.924</td>
<td>gender (2, 16), age (2, 3, 4, 13, 16), language (2, 3, 4, 5, 7, 8, 10, 11, 13, 14, 16), nationality (3, 11), insurance (3, 16), rehab-group (1, 2, 3, 4, 5, 11, 12, 13, 14, 15, 16), time-point (2, 3, 5, 7, 8, 14)</td>
<td>7.5% (6.0%)</td>
</tr>
</tbody>
</table>

Acceptable values

- SD < 1.4
- SD < 1.4
- > 0.01
- > 0.7
- > 0.7
- No DIF
- At least Lower ci < 5%

EBI = Extended Barthel Index, MSK = Musculoskeletal rehabilitation, NEUR = Neurological rehabilitation
1 = admission, t2 = discharge, all = combination of time-points or/and rehabilitation-groups, n = sample size, CI = Class Intervals, SD = standard deviation, PSI = Person Separation Index, α = Cronbach’s alpha, DIF = Differential Item Functioning, ci = Confidence Interval, * only applicable for analyses on the item level

With the 2-testlet solution, fit to the Rasch Model was achieved across all nine analyses steps. The item-trait chi² statistics were non-significant, the reliability indexes all above 0.85, and the item and person fit estimates showed acceptable values. Furthermore, the conditional test of fit also indicated fit at eight of the nine analysis steps. Most A-values were marginally above 1, indicating some remaining local dependency among the testlets. The core values of the testlet approaches for the whole calibration sample (EBIall) are summarized in Table 3. The core values for the other 8 subsamples of the successful 2-testlet solution can be found in Online Appendix 3.

### DIF strategy

The DIF Strategy is presented in more detail in Online Appendix 4. In order to solve the DIF in the fitting 2-testlet solution of the whole calibration sample, Testlet2 was split four times resulting in the following 6 super-items: Testlet1, Testlet2_NEURgerman, Testlet2_NEURfrench, Testlet2_MSKfrench, Testlet2_MSKgerman_female, Testlet2_MSKgerman_male. Testlet1 was the anchor for the comparison of the person estimates of the split and the unsplit version. The resulting effect size amounted 0.09, indicating that...
there is no benefit in splitting the final interval scale transformation into different subgroups. (Online Appendix 5)

Transformation table

Based on the 2-testlet solution an interval scale based transformation table was created for the EBI 0–50 total raw scores, that can be used to transfer the ordinal EBI score into interval EBI scores, when having data on the item level. This transformation is represented in Table 4.

Discussion

Summary of findings

This study examined the psychometric properties of the EBI, providing first evidence of its internal construct validity for neurological and musculoskeletal patients. Even though no fit to the Rasch Model was achieved at the baseline analyses and with the traditional testlet approaches, we could attain model fit by applying an alternative 2-testlet approach. The robustness of the fit was confirmed at all three aggregation levels and subsets of the calibration sample. The evidence of the EBI’s unidimensionality, provides a statement for the internal construct validity of and therefore the reporting of EBI total scores. Furthermore, this study provides an interval scale transformation table of the EBI raw adapted total scores (from 0–50). To avoid bias in reporting change, it is necessary to use the EBI interval scores, as the transformation table shows that changes of a patient at the ends of the score range would be underestimated and changes happening in the middle of the score range would be overestimated if the ordinal EBI raw scores were applied. For example a patient with a EBI raw admission score of 25 and a raw discharge score of 30 would result in a change score of 5 on the raw ordinal basis but only in a change score of 2.9 on the interval level. The transformation table can also be applied for historical analyses when having data on an item level, by applying the conversion table (Table 1). This study therefore further provides evidence for the use of the EBI as an ADL assessment tool, consistent with earlier findings [19].

The application of the 2-testlet approach, that divides similar items equally into 2 clusters, highlighting the sameness of all the items in an assessment tool, was successful in attaining model fit. Noteworthy, this approach puts emphasis on a higher order construct of the EBI, incorporating both motor and cognitive aspects, and is the closest that a 2-testlet approach can get to the actual total score. Still, the EBI can offer different levels of granularity: the level of single items out of which some relate conceptually to each other, e.g., item 6 Mobility and 7 Stairs, the level of sub-scales, e.g., the motor and cognitive subscales, and the level of the overall summary score, that is 16 items indicating the independence of a patient in ADL. Depending on the required use, all 3 levels of granularity are available for reporting. In this study the focus was at the level of the overall summary score – finally represented by 2 super-items – to achieve fit to the Rasch Model.

Furthermore, this study offers first evidence for the EBI’s application for other patients than neurological patients and it is the first investigation of its French translation [44]. The results support that there is no substantial differential item functioning for the
Application in practice

The clinical and practical relevance of this study is twofold: First, this study provides an empirical argument that the EBI items can be summed up to a single total score. This might not appear surprising since the single total score is widely used in practice. However, the unidimensionality of the EBI has not been proven empirically before. This evidence supports the use of the EBI as an assessment tool in practice. Second, the table to transform the raw score into an interval-based score provided in this study for neurological and musculoskeletal patients, allows for the monitoring of patient changes in EBI scores over time in an empirically sound way. Such monitoring is challenged when using the raw ordinal based EBI scores. The transformation table therefore enables a sound comparison of patient or clinic outcomes, which is a key characteristic for learning and improvement processes [46]. In addition, the interval scoring provides an important basis for the application of a standardized reporting system for functioning information [47], in terval scoring provides an important basis for the application of a learning and improvement processes [46]. In addition, the interval scoring provides an important basis for the application of a standardized reporting system for functioning information [47], in terval scoring provides an important basis for the application of a learning and improvement processes [46]. In addition, the interval scoring provides an important basis for the application of a standardized reporting system for functioning information [47], in terval scoring provides an important basis for the application of a learning and improvement processes [46]. In addition, the interval scoring provides an important basis for the application of a standardized reporting system for functioning information [47], in terval scoring provides an important basis for the application of a learning and improvement processes [46]. In addition, the interval scoring provides an important basis for the application of a standardized reporting system for functioning information [47], in terval scoring provides an important basis for the application of a learning and improvement processes [46]. In addition, the interval scoring provides an important basis for the application of a standardized reporting system for functioning information [47], in terval scoring provides an important basis for the application of a learning and improvement processes [46]. In addition, the interval scoring provides an important basis for the application of a standardized reporting system for functioning information [47], in terval scoring provides an important basis for the application of a learning and improvement processes [46]. In addition, the interval scoring provides an important basis for the application of a standardized reporting system for functioning information [47], in terval scoring provides an important basis for the application of a learning and improvement processes [46]. In addition, the interval scoring provides an important basis for the application of a standardized reporting system for functioning information [47], in terval scoring provides an important basis for the application of a learning and improvement processes [46]. In addition, the interval scoring provides an important basis for the application of a standardized reporting system for functioning information [47], in terval scoring provides an important basis for the application of a learning and improvement processes [46]. In addition, the interval scoring provides an important basis for the application of a standardized reporting system for functioning information [47], in terval scoring provides an important basis for the application of a learning and improvement processes [46]. In addition, the interval scoring provides an important basis for the application of a standardized reporting system for functioning information [47], in terval scoring provides an important basis for the application of a learning and improvement processes [46]. In addition, the interval scoring provides an important basis for the application of a standardized reporting system for functioning information [47], in terval scoring provides an important basis for the application of a learning and improvement processes [46]. In addition, the interval scoring provides an important basis for the application of a standardized reporting system for functioning information [47], in terval scoring provides an important basis for the application of a learning and improvement processes [46]. In addition, the interval scoring provides an important basis for the application of a standardized reporting system for functioning information [47], in terval scoring provides an important basis for the application of a learning and improvement processes [46]. In addition, the interval scoring provides an important basis for the application of a standardized reporting system for functioning information [47], in terval scoring provides an important basis for the application of a learning and improvement processes [46]. In addition, the interval scoring provides an important basis for the application of a standardized reporting system for functioning information [47], in terval scoring provides an important basis for the application of a learning and improvement processes [46]. In addition, the interval scoring provides an important basis for the application of a standardized reporting system for functioning information [47], in terval scoring provides an important basis for the application of a learning and improvement processes [46]. In addition, the interval scoring provides an important basis for the application of a standardized reporting system for functioning information [47], in terval scoring provides an important basis for the application of a learning and improvement processes [46]. In addition, the interval scoring provides an important basis for the application of a standardized reporting system for functioning information [47], in terval scoring provides an important basis for the application of a learning and improvement processes [46]. In addition, the interval scoring provides an important basis for the application of a standardized reporting system for functioning information [47], in terval scoring provides an important basis for the application of a learning and improvement processes [46]. In addition, the interval scoring provides an important basis for the application of a standardized reporting system for functioning information [47], in terval scoring provides an important basis for the application of a learning and improvement processes [46]. In addition, the interval scoring provides an important basis for the application of a standardized reporting system for functioning information [47], in terval scoring provides an important basis for the application of a learning and improvement processes [46]. In addition, the interval scoring provides an important basis for the application of a standardized reporting system for functioning information [47], in terval scoring provides an important basis for the application of a learning and improvement processes [46]. In addition, the interval scoring provides an important basis for the application of a standardized reporting system for functioning information [47], in terval scoring provides an important basis for the application of a learning and improvement processes [46]. In addition, the interval scoring provides an important basis for the application of a standardized reporting system for functioning information [47], in terval scoring provides an important basis for the application of a learning and improvement processes [46]. In addition, the interval scoring provides an important basis for the application of a standardized reporting system for functioning information [47], in terval scoring provides an important basis for the application of a learning and improvement processes [46]. In addition, the interval scoring provides an important basis for the application of a standardized reporting system for functioning information [47], in terval scoring provides an important basis for the application of a learning and improvement processes [46]. In addition, the interval scoring provides an important basis for the application of a standardized reporting system for functioning information [47], in terval scoring provides an important basis for the application of a learning and improvement processes [46]. In addition, the interval scoring provides an important basis for the application of a standardized reporting system for functioning information [47], in terval scoring provides an important basis for the application of a learning and improvement processes [46]. In addition, the interval scoring provides an important basis for the application of a standardized reporting system for functioning information [47], in terval scoring provides an important basis for the application of a learning and improvement processes [46]. In addition, the interval scoring provides an important basis for the application of a standardized reporting system for functioning information [47], in terval scoring provides an important basis for the application of a learning and improvement processes [46]. In addition, the interval scoring provides an important basis for the application of a standardized reporting system for functioning information [47], in terval scoring provides an important basis for the application of a learning and improvement processes [46]. In addition, the interval scoring provides an important basis for the application of a standardized reporting system for functioning information [47], in terval scoring provides an important basis for the application of a learning and improvement processes [46]. In addition, the interval scoring provides an important basis for the application of a standardized reporting system for functioning information [47], in terval scoring provides an important basis for the application of a learning and improvement processes [46]. In addition, the interval scoring provides an important basis for the application of a standardized reporting system for functioning information [47], in terval scoring provides an important basis for the application of a learning and improvement processes [46]. In addition, the interval scoring provides an important basis for the application of a standardized reporting system for functioning information [47], in terval scoring provides an important basis for the application of a learning and improvement processes [46]. In addition, the interval scoring provides an important basis for the application of a standardized reporting system for functioning information [47], in terval scoring provides an important basis for the application of a learning and improvement processes [46]. In addition, the interval scoring provides an important basis for the application of a standardized reporting system for functioning information [47], in terval scoring provides an important basis for the application of a learning and improvement processes [46]. In additio
speaking area, could be included. This is beneficial as a standardized reporting of functioning information enables clinicians to continue using assessment tools while still being able to compare and aggregate the information within and across tools or institutions [47].

The 0–50 adapted raw scores proposed in this study can seem confusing when clinicians want to interpret single EBI scores and are used to the original 0–64 scoring system. However, as long as the data is available on the item level in a digital format, this score transformation can be implemented easily in the background of a dataset by a simple look-up table to convert individual item score back to the original, giving a 0–64 range.

Of note, for the EBI, there already exist different scoring systems. The one that is used in Germany in the ICD-10-GM system is different from the one of the original EBI scoring system that was used in this study, having different numbers of categories for certain items and having different item category values ranging from 0–15 [10]. In order to create an interval transformation table for other EBI scoring systems, the Rasch analysis would need to be repeated with data collected with the different scoring systems. The strategy applied in this study would give a good guidance to do so.

Conclusion

The results support the internal construct validity and therefore also the unidimensionality of the EBI for the neurological and the musculoskeletal rehabilitation groups and therefore the reporting of an adapted raw EBI total score. In order to do so the Rasch transformed and interval scaled EBI total scores ranging from 0–50 developed in this study should be used. This interval-based scoring system of the EBI provides the basis to integrate the EBI in a standardized reporting system of functioning information.

Acknowledgements

Thank goes to Dr. L. Menzi Head of Rehabilitation ANQ and K. Schmitt Corporate Development Director of the Swiss Paraplegic Centre for their good advice, the provision of valuable information and the fruitful discussions about the project. This project is part of the cumulative Dissertation of Roxanne Maritz, which is funded by the Swiss National Science Foundation’s National Research Programme “Smarter Health Care” (NRP 74) within the project “Enhancing continuous quality improvement and supported clinical decision making by standardized reporting of functioning”.

Conflict of interest

The authors have no other competing interests to declare.

References


[24] Adroher ND, Prodinger B, Fellinghauer CS et al. All metrics are equal, but some metrics are more equal than others: A systematic search and review on the use of the term ‘metric’. PloS one 2018; 13: e0193861


[27] Linacre JM. Archives of the Rasch Measurement SIG. https://www.rasch.org/rmt/rmt74m.htm Accessed [2018 Nov 14]


Online Appendix

Cases from 10 rehabilitation clinics with complete data in all EBI items

Complete Calibration Sample “EBI_all”

Musculoskeletal Subsample “MSK_all” (n = 400)
Musculoskeletal Subsample time point 1
Musculoskeletal Subsample time point 2

Time point 1 Subsample
Time point 2 Subsample

Neurological Subsample “NEUR_all” (n = 400)
Neurological Subsample time point 1
Neurological Subsample time point 2

Not selected in sampling (n = 5 178)

Calibration Sample
Overall Sample
Level 3
Calibration Sample
Level 2
Calibration Sample
Level 1

Online Appendix 1 Figure Flow chart Calibration Sample, with 3 different aggregation levels.
Online Appendix 2  EBI baseline analysis including threshold disordering and local dependency.

| Sample | n / Cl | Items with threshold disordering | Local Dependency Item No. related items No. | Item1 | Item2 | Item3 | Item4 | Item5 | Item6 | Item7 | Item8 | Item9 | Item10 | Item11 | Item12 | Item13 | Item14 | Item15 | Item16 |
|--------|-------|---------------------------------|------------------------------------------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| MSK_t1 | 200 / 3 | 1, 2, 4, 5, 7, 8, 9, 10, 11, 12, 15, 16 | 8 4, 5, 8 | 3.5, 8 | 3.4, 6, 8 | 5.7, 8 | 6 | 2.3, 4, 5, 6 | 10 | 9 | 12, 14, 15 | 11, 14, 15, 16 | 14, 15 | 11, 12, 13 | 11, 12, 13 | 12 |
| MSK_t2 | 200 / 3 | 1, 2, 4, 5, 7, 8, 9, 10, 11, 12, 15, 16 | 14 | 4 | 4.5, 8 | 2.3, 5, 8 | 3.4, 8 | 8.10 | 3.4, 5, 6 | 10 | 6.9 | 12, 14 | 11, 14 | 14, 15 | 1, 11, 12, 13, 15 | 13, 14 |
| MSK_all | 400 / 6 | 1, 2, 4, 5, 7, 8, 9, 10, 11, 12, 15, 16 | 4 | 4.5, 8 | 2.3, 5, 8 | 3.4, 6, 8 | 5.7, 8 | 6 | 3.4, 5, 6 | 10 | 9 | 12, 13, 14, 15 | 11, 14 | 11, 15 | 11, 12, 13, 15 | 11, 13, 14 |
| NEUR_t1 | 200 / 3 | 1, 2, 3, 4, 5, 7, 8, 9, 10, 12, 16 | 3 | 2.4, 5, 8 | 3.4, 6, 7, 8 | 5.6 | 3.4, 5 | 10 | 9 | 12, 13, 14, 15 | 11, 13, 14, 15 | 11, 12, 13, 15 | 11, 12, 13, 15 | 11, 13, 14, 16 | 11, 15 |
| NEUR_t2 | 200 / 3 | 1, 2, 4, 5, 6, 7, 8, 9, 10, 12, 16 | 3.4 | 2.4, 5, 8 | 2.3, 5, 8 | 3.4, 6, 7, 8 | 5.7, 8 | 5.6, 8 | 3.4, 5, 6, 7 | 10 | 9 | 12, 13, 14, 15 | 11, 13, 15 | 11, 12, 14, 15 | 11, 13, 15, 16 | 11, 14, 15 |
| NEUR_all | 400 / 6 | 1, 2, 4, 5, 6, 7, 8, 9, 10, 12, 16 | 3.4 | 2.4, 5, 8 | 2.3, 5, 8 | 3.4, 6, 7, 8 | 5.7 | 5.6 | 3.4, 5 | 10 | 9 | 12, 13, 14, 15 | 11, 13, 15 | 11, 12, 14, 15 | 11, 13, 15, 16 | 11, 12, 13, 14, 16 |
| t1_all | 400 / 6 | 1, 2, 3, 4, 5, 7, 8, 9, 10, 11, 12, 16 | 2 | 1.3, 4, 8 | 2.4, 5, 8 | 2.3, 5, 8 | 3.4, 6, 7, 8 | 5.7, 8 | 5.6 | 2.3, 4, 5, 6 | 10 | 9 | 12, 13, 14, 15 | 11, 13, 14, 15 | 11, 12, 13, 15 | 11, 12, 13, 14 |
| t2_all | 400 / 6 | 1, 2, 5, 7, 8, 9, 10, 11, 12, 16 | 3.4 | 2.4, 5, 8 | 2.3, 5, 8 | 3.4, 6, 7, 8 | 5.7, 8 | 5.6 | 3.4, 5, 6 | 10 | 9 | 12, 13, 14, 15 | 11, 13, 15 | 11, 12, 14, 15 | 11, 13, 15 | 11, 13, 14 |
| EBI all | 800 / 10 | 1, 2, 5, 7, 8, 9, 10, 11, 12, 16 | 14 | 3.4 | 2.4, 5, 8 | 2.3, 5, 8 | 3.4, 6, 7, 8 | 5.7, 8 | 5.6 | 3.4, 5, 6 | 10 | 9 | 12, 13, 14, 15 | 11, 13, 14, 15 | 11, 12, 14, 15 | 1, 11, 12, 13, 15 | 11, 12, 13, 14 |
Online Appendix 3  EBI two-testlet approach repeated for all aggregation levels of the calibration sample.

<table>
<thead>
<tr>
<th>Sample</th>
<th>n / Cl</th>
<th>Item fit residual Mean (SD)</th>
<th>Person fit residual Mean (SD)</th>
<th>chi² p-value</th>
<th>PSI</th>
<th>α</th>
<th>DIF (Testlet)</th>
<th>T-Test at 5% level (if &gt;5% lower ci)</th>
<th>A</th>
<th>conditional test of fit</th>
</tr>
</thead>
<tbody>
<tr>
<td>MSKt1</td>
<td>200 / 3</td>
<td>−2.182 (1.084)</td>
<td>−0.664 (0.756)</td>
<td>0.847</td>
<td>0.875</td>
<td>0.919</td>
<td>language (T1, T2)</td>
<td>2.0 %</td>
<td>1.017</td>
<td>0.432</td>
</tr>
<tr>
<td>MSKt2</td>
<td>200 / 3</td>
<td>−1.265 (1.160)</td>
<td>−0.474 (0.685)</td>
<td>0.167</td>
<td>0.829</td>
<td>0.930</td>
<td>language (T1, T2)</td>
<td>3.0 %</td>
<td>0.976</td>
<td>0.214</td>
</tr>
<tr>
<td>MSKall</td>
<td>400 / 6</td>
<td>0.020 (1.210)</td>
<td>−0.465 (0.848)</td>
<td>0.672</td>
<td>0.916</td>
<td>0.929</td>
<td>language (T1, T2)</td>
<td>5.0 %</td>
<td>1.063</td>
<td>0.045</td>
</tr>
<tr>
<td>NEURt1</td>
<td>200 / 3</td>
<td>0.149 (0.571)</td>
<td>−0.439 (0.821)</td>
<td>0.745</td>
<td>0.954</td>
<td>0.957</td>
<td>age (T1) language (T2)</td>
<td>2.0 %</td>
<td>1.048</td>
<td>0.055</td>
</tr>
<tr>
<td>NEURt2</td>
<td>200 / 3</td>
<td>0.170 (0.744)</td>
<td>−0.507 (0.880)</td>
<td>0.858</td>
<td>0.958</td>
<td>0.959</td>
<td>language (T2)</td>
<td>4.5 %</td>
<td>1.044</td>
<td>0.457</td>
</tr>
<tr>
<td>NEURall</td>
<td>400 / 6</td>
<td>0.114 (0.940)</td>
<td>−0.475 (0.836)</td>
<td>0.665</td>
<td>0.956</td>
<td>0.958</td>
<td>language (T2)</td>
<td>4.5 %</td>
<td>1.047</td>
<td>0.218</td>
</tr>
<tr>
<td>t1all</td>
<td>400 / 6</td>
<td>0.066 (1.262)</td>
<td>−0.493 (0.921)</td>
<td>0.919</td>
<td>0.946</td>
<td>0.945</td>
<td>language (T1, T2)</td>
<td>3.0 %</td>
<td>1.057</td>
<td>0.000</td>
</tr>
<tr>
<td>t2all</td>
<td>400 / 6</td>
<td>0.079 (1.160)</td>
<td>−0.462 (0.891)</td>
<td>0.738</td>
<td>0.928</td>
<td>0.953</td>
<td>language (T1, T2)</td>
<td>2.8 %</td>
<td>1.036</td>
<td>0.400</td>
</tr>
<tr>
<td>EBlall</td>
<td>800 / 10</td>
<td>0.031 (1.676)</td>
<td>−0.476 (0.898)</td>
<td>0.822</td>
<td>0.938</td>
<td>0.950</td>
<td>language (T1, T2)</td>
<td>3.1 %</td>
<td>1.047</td>
<td>0.013</td>
</tr>
<tr>
<td>Acceptable values</td>
<td>Not applicable for analyses on testlet level</td>
<td>Not applicable for analyses on testlet level</td>
<td>&gt;0.01</td>
<td>&gt;0.7</td>
<td>&gt;0.7</td>
<td>No DIF</td>
<td>at least lower ci &lt;5%</td>
<td>&gt;0.9</td>
<td>&gt;0.01</td>
<td></td>
</tr>
</tbody>
</table>

FIM = Functional Independence Measure, MSK = Musculoskeletal rehabilitation, NEUR = Neurological rehabilitation, t1 = admission, t2 = discharge, all = combination of time-points or/and rehabilitation-groups, n = sample size, Cl = Class Intervals, ci = confidence interval, SD = standard deviation, PSI = Person Separation Index, α = Cronbach’s alpha, DIF = Differential Item Functioning, A = Explained Common Variance
### Online Appendix 4 DIF Strategy for two-testlet approach on the level of the whole calibration sample (EBI_all).

<table>
<thead>
<tr>
<th>Analysis Name (Testlets)</th>
<th>Person Factor</th>
<th>Testlet</th>
<th>DIF</th>
<th>p-value</th>
<th>Interpretation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Two-testlet (T1, T2)</td>
<td>language</td>
<td>1</td>
<td>Uniform &amp; Non-uniform</td>
<td>0.003420 0.000316</td>
<td>Since both T1 and T2 shows DIF, no testlet is available for anchoring</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
<td>Non-uniform</td>
<td>0.000028</td>
<td></td>
</tr>
<tr>
<td></td>
<td>rehab-group</td>
<td>1</td>
<td>Uniform &amp; Non-uniform</td>
<td>0.000014 0.000722</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
<td>Uniform &amp; Non-uniform</td>
<td>0.000001 0.000022</td>
<td>lowest p-value, basis for Split 1</td>
</tr>
<tr>
<td>Split1 (T1, T2_MSK, T2_NEUR)</td>
<td>language</td>
<td>T1</td>
<td>Uniform &amp; Non-uniform</td>
<td>0.000164 0.000998</td>
<td>Since T1 shows DIF and T2 is split, not item is available for anchoring.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>T2_MSK</td>
<td>Uniform &amp; Non-uniform</td>
<td>0.000000 0.000000</td>
<td>lowest p-value, basis for Split2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>T2_NEU</td>
<td>Uniform</td>
<td>0.002907</td>
<td></td>
</tr>
<tr>
<td>Split2 (T1, T2_NEUR, T2_MSK_german, T2_MSK_french)</td>
<td>gender</td>
<td>T2_MSK_german</td>
<td>Uniform</td>
<td>0.000675 0.002603</td>
<td>Basis for Split3. T01 from Split2 can be used as an anchor since it has no DIF and is not split.</td>
</tr>
<tr>
<td>Split3 (T1, T2_NEUR, T2_MSK_french, T2_MSK_german_fem, T2_MSK_german_mal)</td>
<td>language</td>
<td>T1</td>
<td>Non-uniform</td>
<td>0.001369</td>
<td>Since both T1 shows DIF, and T2 is split, no testlet is available for anchoring with the Two-Testlet analysis</td>
</tr>
<tr>
<td></td>
<td></td>
<td>T2_NEUR</td>
<td>Uniform</td>
<td>0.004094</td>
<td>Since T1 cannot be split (as an anchor is needed, and T2 is already split), Split 4 is based on Split2, in which T2_NEUR shows also language DIF</td>
</tr>
<tr>
<td>Split4 – based on Split3 (T1, T2_NEUR_french, T2_NEUR_german, T2_MSK_french, T2_MSK_german_fem, T2_MSK_german_mal)</td>
<td>No DIF present</td>
<td>1.157 1.19</td>
<td>2.053 2.117</td>
<td>0.984 0.088</td>
<td>Mean person location 2.053 2.117 Correlation of means 0.984 Effect Size 0.088 SD = standard deviation</td>
</tr>
</tbody>
</table>

#### Combined Effect Size for repeated measures

<table>
<thead>
<tr>
<th>2-testlet</th>
<th>Split4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean person location</td>
<td>1.157 1.19</td>
</tr>
<tr>
<td>SD person location</td>
<td>2.053 2.117</td>
</tr>
<tr>
<td>Correlation of means</td>
<td>0.984</td>
</tr>
<tr>
<td>Effect Size</td>
<td>0.088</td>
</tr>
</tbody>
</table>

SD = standard deviation

### Online Appendix 5 Cont’d DIF Strategy for two-testlet approach on the level of the whole calibration sample (EBI_all).