Published online: 2019-03-13

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 Maritz[1,2], Alan Tennant[1,2], Carolina Saskia Fellinghauer[1], Gerold Stucki[1,2], Birgit Prodinger[1,3], 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

Schlüsselwörter

Alltagsaktivität, Ergebnismessung, Qualitätsmanagement, Rehabilitation, Rasch Modell, Assessmentinstrumente

received 07.09.2018
accepted 10.12.2018

Bibliography

DOI https://doi.org/10.1055/a-0820-4642
Published online: 13.3.2019
Phys Med Rehab Kuror 2019; 29: 224–232

ISSN 0940-6689

Correspondence

Roxanne Maritz, MA
Swiss Paraplegic Research
Guido A. Zäch Straße 4
6207 Nottwil
Switzerland
roxanne.maritz@paraplegie.ch

ABSTRACT

Background The Extended Barthel Index (EBI), consisting of the original Barthel Index plus 6 cognitive items, provides a tool to monitor patients' outcomes in rehabilitation. Whether the EBI provides a unidimensional metric, thus can be reported as a valid sum-score, remains to be examined.

Objective To examine whether the EBI can be reported as unidimensional interval-scaled metric for neurological and musculoskeletal rehabilitation.

Methods Rasch analysis of a calibration sample of 800 cases from neurological or musculoskeletal rehabilitation in 2016 in Switzerland.

Results In the baseline analysis no fit to the Rasch Model was achieved. When accommodating local dependencies with a testlet approach satisfactory fit to the Rasch Model was achieved, and an interval scale transformation table was created.

Conclusion The results support the reporting of adapted EBI total scores for both rehabilitation groups by applying the interval scaled transformation table presented in this study.

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 muskuloskelettale Rehabilitation als eindimensionale intervallskalierte Metrik berichtet werden kann.

Methode Rasch-Analyse einer Stichprobe von 800 neurologischen und muskuloskelettalen Rehapatienten 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.

* NRP74 StARS clinics: cereneo Schweiz – Robinson Kundert; Hôpital du Valais Spital Wallis, Centre Martigny, Sierre, Brig & Saint-Amé – Els De Waele; Klinik Schönberg – Philipp Banz; Kliniken Valens, Rehazentrum Valens, Rehazentrum Walenstadtberg & Rheinburg-Klinik – Stefan Bachmann, Luzerner Höhenklinik Montana – Jean-Marie Schnyder, Reha Rheinfelden – Thierry Ettlin

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 Ethic 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 admission 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 chi^2 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 NEURt2). In Level 2 the rehabilitation group and time point subsamples were aggregated separately (MSKt1&t2, NEURt1&t2, t1MSK&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 Item Conversion on item level original scores (0–64) to the adapted raw score (0–50).

No	Items	EBI 0–64 Categories	EBI 0–50 Categories
1	Feeding	0 2 3 4	0 1 2 3
2	Grooming	0 1 2 3 4	0 1 2 3 4
3	Dressing	0 1 2 4	0 1 2 3
4	Bathing	0 1 2 3 4	0 1 2 3 4
5	Transfer	0 1 2 4	0 1 2 3
6	Mobility	0 1 2 3 4	0 1 2 3 4
7	Stairs	0 1 2 4	0 1 2 3
8	Toilet use	0 1 2 4	0 1 2 3
9	Bowels	0 2 3 4	0 1 2 3
10	Bladder	0 1 3 4	0 1 2 3
11	Expression	0 1 3 4	0 1 2 3
12	Comprehension	0 1 3 4	0 1 2 3
13	Social interaction	0 2 4	0 1 2
14	Problem solving	0 2 4	0 1 2
15	Memory	0 1 2 3 4	0 1 2 3 4
16	Vision/ Neglect	0 1 3 4	0 1 2 3
Min		**0**	**0**
Max		**64**	**50**

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**.

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

Sample	n / CI	Item fit residuals Mean (SD)	Person fit residuals Mean(SD)	chi² p-value	PSI	α	DIF (item No)	Paired t-test (Lower ci %)
MSKt1	200 / 3	− 0.096 (1.814)	− 0.438 (1.130)	0.000	0.861	0.856	gender (2), language (2, 5, 7, 8, 14, 15, 16), insurance (16)	7.5 % (0.0 %)
MSKt2	200 / 3	− 0.675 (2.152)	− 0.169 (0.900)	0.000	0.849	0.902	gender (2, 16), age (2, 3, 13), language (2, 4, 7, 14, 16), nationality (11, 12)	10.0 % (0.0 %)
MSKall	400 / 6	− 0.583 (2.941)	− 0.310 (1.063)	0.000	0.862	0.882	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)	9.3 % (0.0 %)
NEURt1	200 / 3	− 0.764 (2.174)	− 0.310 (1.025)	0.000	0.911	0.941	age (4), language (3, 4, 5, 8, 11, 14)	8.5 % (5.5 %)
NEURt2	200 / 3	− 0.533 (2.576)	− 0.307 (1.175)	0.000	0.918	0.918	language (3, 4, 11, 14), nationality (11), insurance(3, 11)	10.0 % (7.0 %)
NEURall	400 / 6	− 1.009 (3.371)	− 0.328 (1.107)	0.000	0.913	0.941	age (3, 4), language (2, 3, 4, 5, 8, 11, 13, 14), nationality (11), insurance (3)	8.8 % (6.6 %)
t1all	400 / 6	− 0.639 (3.109)	− 0.333 (1.155)	0.000	0.895	0.914	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)	9.3 % (7.1 %)
t2all	400 / 6	− 0.801 (3.101)	− 0.259 (1.084)	0.000	0.896	0.933	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)	8.5 % (6.4 %)
EBIall	800 / 10	− 1.083 (4.451)	− 0.289) (1.108)	0.000	0.896	0.924	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)	7.5 % (6.0 %)
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 rehabilitationt1 = 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

Testlet approaches

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 collapsed into one testlet. In the 5 Testlet version, the cognitive items were divided into the Testlet4 Communication (including items 11-Comprehension and 12-Expression) and 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**.

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

▸ Table 3 Testlet approaches with whole EBI calibration sample (EBIall).

Testlets (item No)	Item fit residuals Mean (SD)	Person fit residuals Mean (SD)	chi2 p-value	PSI	α	DIF (Testlet)	A (PSI)	t-test %	CI-based cond. test of fit
4 Testlets: Self-care (1–4), mobility (5–7), Toileting (8–10), Cognition (11–16)	−0.569 (4.952)	−0.338 (0.881)	0.000	0.780	0.845	age (1, 3), language (1, 2, 3, 4), time-point (1, 2), rehab-group (1, 2, 3, 4)	0.870	2.5 %	only available for two-testlet approach
5 Testlets: Self-care (1–4), mobility (5–7), Toileting (8–10), Communication (11–12), Social cognition (13–16)	−0.891 (3.931)	−0.369 (0.854)	0.000	0.804	0.844	gender (5), language (1, 3, 5), nationality (4), time-point (1, 2, 5),- rehab-group (1, 2, 4, 5)	0.898	2.9 %	only available for two-testlet approach
Two-Testlets: Testlet1 (1, 3, 5, 7, 9, 11, 13, 15), Testlet2 (2, 4, 6, 8, 10, 12, 14, 16)	0.031 (1.676)	−0.476 (0.898)	0.822	0.938	0.950	language (1,2) rehab-group (1,2)	1.047	3.1 %	0.013
Acceptable values	*Not applicable for analyses on testlet level*	*Not applicable for analyses on testlet level*	*>0.01*	*>0.7*	*>0.7*	*No DIF*	*>0.9*	*<5%*	*>0.01*

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

▶ **Table 4** EBI Total Score Transformation Table – adapted 0–50 EBI raw Scores to EBI Interval Scores.

Adapted raw Score	Rasch Estimate	Transformed interval Score
0	−5.358	0.0
1	−4.621	3.4
2	−4.075	6.0
3	−3.670	7.8
4	−3.340	9.4
5	−3.063	10.7
6	−2.823	11.8
7	−2.611	12.8
8	−2.416	13.7
9	−2.235	14.5
10	−2.064	15.3
11	−1.901	16.0
12	−1.745	16.8
13	−1.594	17.5
14	−1.449	18.1
15	−1.309	18.8
16	−1.173	19.4
17	−1.041	20.0
18	−0.912	20.6
19	−0.786	21.2
20	−0.662	21.8
21	−0.539	22.4
22	−0.418	22.9
23	−0.298	23.5
24	−0.178	24.0
25	−0.057	24.6
26	0.064	25.2
27	0.187	25.7
28	0.312	26.3
29	0.439	26.9
30	0.569	27.5
31	0.703	28.1
32	0.841	28.8
33	0.983	29.4
34	1.131	30.1
35	1.284	30.8
36	1.443	31.6
37	1.609	32.3
38	1.783	33.2
39	1.965	34.0
40	2.155	34.9
41	2.354	35.8
42	2.562	36.8
43	2.782	37.8
44	3.013	38.9
45	3.261	40.0
46	3.531	41.3
47	3.836	42.7
48	4.204	44.4
49	4.712	46.7
50	5.412	50.0

musculoskeletal and the neurological group, and for the German and the French speaking region of Switzerland. The invariance of the two language versions supports the quality of the translations from its original in German into French. Notwithstanding this, there remain questions about the relevance of cognitive EBI items in the musculoskeletal population, particularly the meaningfulness of the item 16-Vision/Neglect remains debatable. For more extensive evidence on group invariance, e. g., regarding patient groups and the French language edition of the EBI, further investigation would be needed.

Limitations of the study

The study brings the limitations of secondary data analysis, for example the limited choice of the person factors for the DIF analyses or the lack of information on consistency and accuracy of data entry. The 2-testlet approach is new, and while it was successful in attaining model fit, it loses the granularity of the individual item approach, as no statement about the hierarchy and difficulty of single items or a conceptually related group of items can be made anymore. Of course this does not preclude the latter, but increasingly evidence is emerging that health assessment tools violate the local item independence assumption more often than not, and this has a damaging effect upon traditional scale interpretation [38, 45]. Thus it is difficult to see how some form of testlet solution could be avoided. The 2-testlet approach has the advantage that the total scores of a well-established assessment tool like the EBI can be converted on an interval scale level, without deleting or rescoring items.

In addition, on the levels of the testlet approaches, the analysis for threshold disordering is absent. There is some initial evidence that threshold disordering can be caused by local dependency [36]. If this is the case, it becomes impossible to conclude if the thresholds disordering is a consequence of local dependency or of it is due to item interpretation. As an example, while item thresholds are ordered within their subscales, thresholds can become disordered when subscales are summated together. Further investigations will be needed to confirm the influence of local dependency.

Another general limitation is the potential ceiling effect for the EBI. While the calibration sample used in the current study avoided that problem, by focusing on a broad representation of the EBI score range.

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 which the EBI as frequently used assessment tool in the German 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

[1] Stucki G, Bickenbach J. Functioning: the third health indicator in the health system and the key indicator for rehabilitation. European journal of physical and rehabilitation medicine 2017; 53: 134–138

[2] Nelles G. Neurologische Rehabilitation. Stuttgart: Georg Thieme Verlag, 2004

[3] Stucki G, Prodinger B, Bickenbach J. Four steps to follow when documenting functioning with the International Classification of Functioning, Disability and Health. European journal of physical and rehabilitation medicine 2017; 53: 144–149

[4] Doganay Erdogan B, Leung YY, Pohl C et al. Minimal clinically important difference as applied in rheumatology: An OMERACT rasch working group systematic review and critique. The Journal of rheumatology 2016; 43: 194–202

[5] Gorter R, Fox JP, Twisk JW. Why item response theory should be used for longitudinal questionnaire data analysis in medical research. BMC Med Res Methodol 2015; 15: 55

[6] Grimby G, Tennant A, Tesio L. The use of raw scores from ordinal scales: Time to end malpractice? Journal of rehabilitation medicine 2012; 44: 97–98

[7] Andrich D. Rating scales and Rasch measurement. Expert review of pharmacoeconomics & outcomes research 2011; 11: 571–585

[8] Christensen KB, Kreiner S, Mesbah M. Rasch Models in Health. London / Hoboken: ISTE Ltd / John Wiley & Sons, Inc.; 2013

[9] Prosiegel M, Böttger S, Schenk T et al. The Extended Barthel Index – a new scale for the assessment of diability in neurological patients [German]. Neurologie & Rehabilitation 1996; 7–13

[10] Deutsches Institut für Medizinische Dokumentation und Information DIMDI https://www.dimdi.de/static/de/klassifikationen/icd/icd-10-gm/kode-suche/htmlgm2018/zusatz-06-barthelindex.htm Accessed [2018 Nov 14]

[11] ANQ Nationaler Verein für Qualitätsentwicklung in Spitälern und Kliniken, Bern https://www.anq.ch/wp-content/uploads/2017/12/ANQ_Module_23_Auswertungskonzept.pdf Accessed [2018 June 22]

[12] Federal Statistical Office. [https://www.bfs.admin.ch/bfs/de/home/statistiken/kataloge-datenbanken/publikationen.assetdetail.1940914.html] Accessed [2018 June 22]

[13] Swiss DRG https://www.swissdrg.org/application/files/9315/0997/9991/ST_Reha_Format_und_Inhalt_Daten_2018.pdf Accessed [2018 June 22]

[14] Mahoney FI, Barthel DW. Functional Evaluation: The Barthel Index. Maryland state medical journal 1965; 14: 61–65

[15] Grill E, Stucki G, Scheuringer M et al. Validation of International Classification of Functioning, Disability, and Health (ICF) Core Sets for early postacute rehabilitation facilities: comparisons with three other functional measures. American journal of physical medicine & rehabilitation 2006; 85: 640–649

[16] Haigh R, Tennant A, Biering-Sorensen F et al. The use of outcome measures in physical medicine and rehabilitation within Europe. Journal of rehabilitation medicine 2001; 33: 273–278

[17] Houlden H, Edwards M, McNeil J et al. Use of the Barthel Index and the Functional Independence Measure during early inpatient rehabilitation after single incident brain injury. Clin Rehabil 2006; 20: 153–159

[18] Kwon S, Hartzema AG, Duncan PW et al. Disability measures in stroke: relationship among the Barthel Index, the Functional Independence Measure, and the Modified Rankin Scale. Stroke 2004; 35: 918–923

[19] Marolf MV, Vaney C, Konig N et al. Evaluation of disability in multiple sclerosis patients: a comparative study of the Functional Independence Measure, the Extended Barthel Index and the Expanded Disability Status Scale. Clin Rehabil 1996; 10: 309–313

[20] Jansa J, Pogacnik T, Gompertz P. An evaluation of the Extended Barthel Index with acute ischemic stroke patients. Neurorehabilitation & Neural Repair 2004; 18: 37–41

[21] Bath PM, Iddenden R, Bath FJ et al. Tirilazad for acute ischaemic stroke. The Cochrane database of systematic reviews 2001; CD002087

[22] Jörger M, Beer S, Kesselring J. Impact of neurorehabilitation on disability in patients with acutely and chronically disabling diseases of the nervous system measured by the Extended Barthel Index. Neurorehabilitation & Neural Repair 2001; 15: 15–22

[23] Schuster-Amft C, Henneke A, Hartog-Keisker B et al. Intensive virtual reality-based training for upper limb motor function in chronic stroke: a feasibility study using a single case experimental design and fMRI. Disability & Rehabilitation: Assistive Technology 2015; 10: 385–392

[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

[25] Smith AB, Rush R, Fallowfield LJ et al. Rasch fit statistics and sample size considerations for polytomous data. BMC Medical Research Methodology 2008; 8: 33

[26] R Core Team. R Foundation for Statistical Computing. Vienna: https://www.R-project.org/ Accessed [2018 Nov 14]

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

[28] Hagell P, Westergren A. Sample Size and Statistical Conclusions from Test of Fit to the Rasch Model Accodring to the Rasch Unidimensional Measurement Model (RUMM) Program in Health Outcome Measurement. Journal of Applied Measurement 2016; 17: 416–431

[29] Mallinson T. Rasch Analysis of Repeated Measures. Rasch Measurement Transactions 2011; 251: 1317

[30] Rasch G. Probabilistic models for some intelligence and attainment tests. The University of Chicago Press; Chicago: 1980

[31] Andrich D, Sheridan B, Luo G. Rumm Laboratory Pty Ltd. http://www.rummlab.com.au/ Accessed [2018 Nov 14]

[32] Masters GN. A Rasch model for partial credit scoring. Psychometrika 1982; 47: 149–174

[33] Tennant A, Conaghan PG. The Rasch measurement model in rheumatology: what is it and why use it? When should it be applied, and what should one look for in a Rasch paper? Arthritis and rheumatism 2007; 57: 1358–1362

[34] Andrich D. Components of Variance of Scales With a Bifactor Subscale Structure From Two Calculations of alpha. Educational Measurement: Issues and Practice 2016; 35: 25–30

[35] Wainer H, Kiely G. Item clusters and computer adaptive testing: A case for testlets. Journal of Educational Measurement 1987; 185–202

[36] Wilson M. Detecting and interpreting local item dependence using a family of Rasch models. Applied psychological measurement 1988; 353–364

[37] Tuerlinckx F, De Boeck P. The effect of ignoring item interactions on the estimated discrimination parameters in item response theory. Psychological Methods 2001; 6: 181–195

[38] Lundgren Nilsson A, Tennant A. Past and present issues in Rasch analysis: the functional independence measure (FIM) revisited. Journal of rehabilitation medicine 2011; 43: 884–891

[39] Christensen KB, Makransky G, Horton M. Critical Values for Yen's Q3: Identification of Local Dependence in the Rasch Model Using Residual Correlations. Applied psychological measurement 2017; 41: 178–194

[40] Andrich D. The Polytomous Rasch Model and the Equating of Two Instruments. In: Christensen KB, Kreiner S, Mesbah M. (editors.) Rasch Models in Health. London, UK: ILSTE Ltd; 2013: 164–196

[41] Andrich D, Hagquist C. Real and Artificial Differential Item Functioning in Polytomous Items. Educ Psychol Meas 2015; 75: 185–207

[42] Morris SB, DeShon RP. Combining effect size estimates in meta-analysis with repeated measures and independent-groups designs. Psychological Methods 2002; 7: 105–125

[43] Cohen J. Statistical Power Analysis for the Behavioral Sciences. Hillsdale: Lawrence Erlbaum Associates, 1988

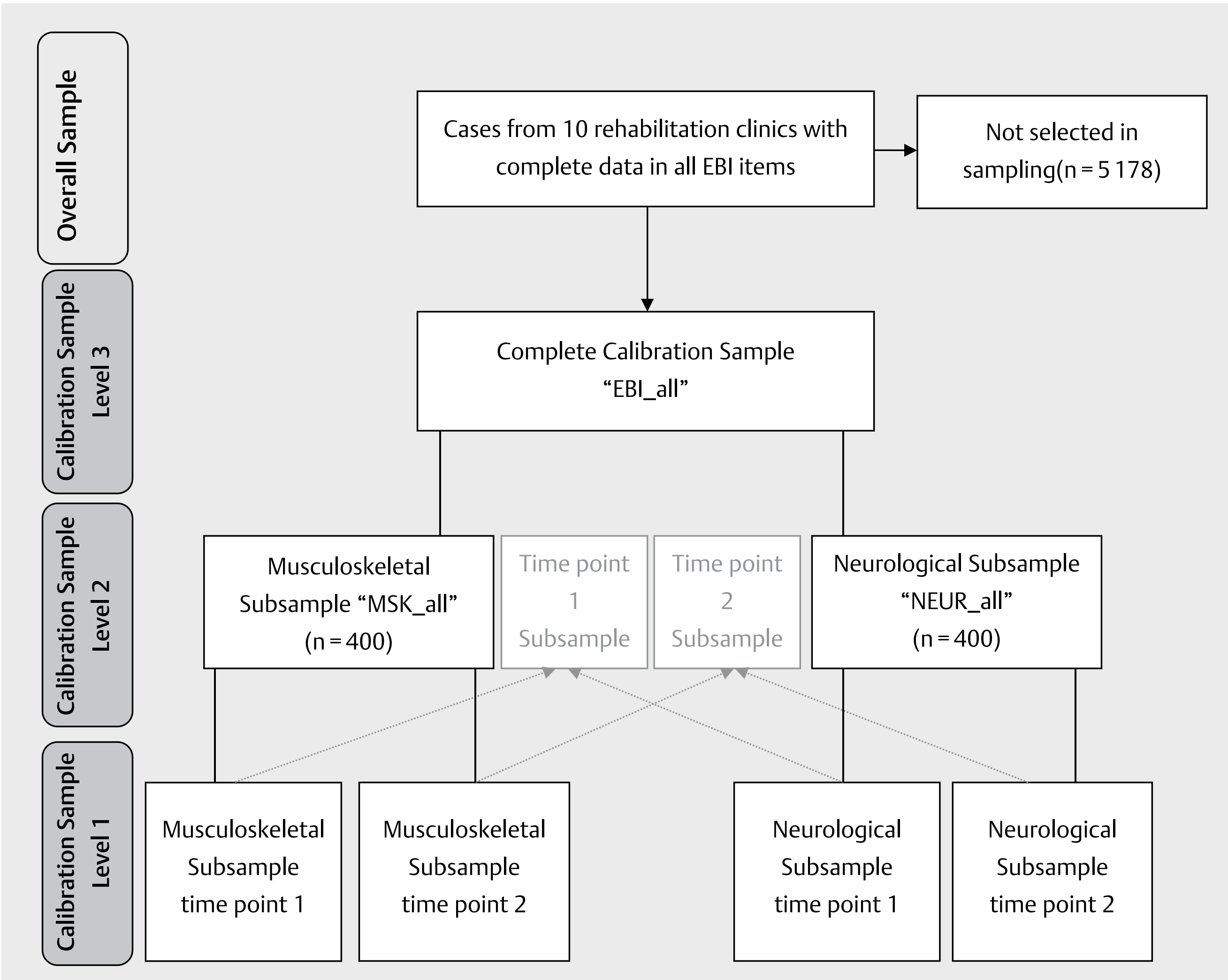
[44] ANQ Nationaler Verein für Qualitätsentwicklung in Spitälern und Kliniken, Bern, Switzerland https://www.anq.ch/wp-content/uploads/2018/02/ANQ_Module_2_EBI.pdf Accessed [2018 Jul 19]

[45] Hammond A, Tennant A, Tyson SF et al. The reliability and validity of the English version of the Evaluation of Daily Activity Questionnaire for people with rheumatoid arthritis. Rheumatology (Oxford, England) 2015; 54: 1605–1615

[46] Mainz J. Defining and classifying clinical indicators for quality improvement. International journal for quality in health care: journal of the International Society for Quality in Health Care 2003; 15: 523–530

[47] Prodinger B, Tennant A, Stucki G. Standardized reporting of functioning information on ICF-based common metrics. European journal of physical and rehabilitation medicine 2018; 54: 110–117

Online Appendix



▶ **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 / CI	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, 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, 5, 8	3, 4, 6, 7, 8	5, 7	5,6	3, 4, 5	10	9	12, 13, 14, 15, 16	11, 13, 14	11, 12, 14, 15	11, 12, 13, 15	11, 13, 14, 16	11, 15
NEUR_t2	200 / 3	1, 2, 5, 6, 7, 8, 9, 10, 12		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, 16	11, 13, 15	11, 12, 14, 15	11, 13, 15, 16	11, 12, 13, 14, 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, 16	11, 13, 15	11, 12, 14, 15	11, 13, 15, 16	11, 12, 13, 14, 16	14, 15
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, 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	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**.**

Sample	n / CI	Item fit residual Mean (SD)	Person fit residual Mean (SD)	chi2 p-value	PSI	α	DIF (Testlet)	T-Test at 5% level (if>5% lower ci)	A	conditional test of fit
MSKt1	200 / 3	−2.182 (1.084)	−0.664 (0.756)	0.847	0.875	0.919	language (T1, T2)	2.0 %	1.017	0.432
MSKt2	200 / 3	−1.265 (1.160)	−0.474 (0.685)	0.167	0.829	0.930	language (T1, T2)	3.0 %	0.976	0.214
MSKall	400 / 6	0.020 (1.210)	−0.465 (0.848)	0.672	0.916	0.929	language (T1, T2)	5.0 % (2.9 % lower ci)	1.063	0.045
NEURt1	200 / 3	0.149 (0.571)	−0.439 (0.821)	0.745	0.954	0.957	age (T1) language (T2)	2.0 %	1.048	0.055
NEURt2	200 / 3	0.170 (0.744)	−0.507 (0.880)	0.858	0.958	0.959	No DIF	4.5 %	1.044	0.457
NEURall	400 / 6	0.114 (0.940)	−0.475 (0.836)	0.665	0.956	0.958	language (T2)	4.5 %	1.047	0.218
t1all	400 / 6	0.066 (1.262)	−0.493 (0.921)	0.919	0.946	0.945	language (T1, T2) rehab-group (T1, T2)	3.0 %	1.057	0.000
t2all	400 / 6	0.079 (1.160)	−0.462 (0.891)	0.738	0.928	0.953	language (T1,T2), rehab-group (T1,T2)	2.8 %	1.036	0.400
EBIall	800 / 10	0.031 (1.676)	−0.476 (0.898)	0.822	0.938	0.950	language (T1,T2), rehab-group (T1,T2)	3.1 %	1.047	0.013
Acceptable values		*Not applicable for analyses on testlet level*	*Not applicable for analyses on testlet level*	*>0.01*	*>0.7*	*>0.7*	*No DIF*	*at least lower ci<5 %*	*>0.9*	*>0.01*

FIM = Functional Independence Measure, MSK = Musculoskeletal rehabilitation, NEUR = Neurological rehabilitationt1 = admission, t2 = discharge, all = combination of time-points or/and rehabilitation-groups, n = sample size, CI = 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 (EBIall).

Analysis Name (Testlets)	Person Factor	Testlet	DIF		
			Uniform Non-uniform	**p-value**	**Interpretation**
Two-testlet (T1, T2)	language	1	Uniform & Non-uniform	0.003420 0.000316	Since both T1 and T2 shows DIF, no testlet is available for anchoring
	language	2	Non-uniform	0.000028	
	rehab-group	1	Uniform & Non-uniform	0.000014 0.000722	
	rehab-group	2	Uniform & Non-uniform	0.000001 0.000022	lowest p-value, basis for Split1
Split1 (T1, T2_MSK, T2_NEUR)	language	T1	Uniform & Non-uniform	0.000164 0.000998	Since T1 shows DIF and T2 is split, not item is available for anchoring.
	language	T2_MSK	Uniform & Non-uniform	0.000000 0.000000	lowest p-value, basis for Split2
	language	T2_NEU	Uniform	0.002907	
Split2 (T1, T2_NEUR, T2_MSK_german, T2_MSK_french)	gender	T2_MSK_ german	Uniform	0.000675	Basis for Split3. T01 from Split2 can be used as an anchor since it has no DIF and is not split.
	language	T2_NEUR	Uniform	0.002603	Basis for Split4, due to results of Split 3.
Split3 (T1, T2_NEUR, T2_MSK_french, T2_MSK_german_fem, T2_MSK_german_mal)	language	T1	Non-uniform	0.001369	Since both T1shows DIF, and T2 is split, no testlet is available for anchoring with the Two-Testlet analysis
	language	T2_NEUR	Uniform	0.004094	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
Split4 – based on Split3 (T1, T2_NEUR_french, T2_NEUR_german, T2_MSK_french, T2_MSK_german_fem, T2_MSK_german_mal)			No DIF present		Split 4b is a second final solution to get rid of all DIF. T01 from Split4b can be used as an anchor since it has no DIF and is not split.

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

Combined Effect Size for repeated measures			
	2-testlet	**Split4**	
Mean person location	1.157	1.19	
SD person location	2.053	2.117	
Correlation of means			0.984
Effect Size			**0.088**

SD = standard deviation